

Jens-Rainer Ohm

Multimedia Content Analysis

Table of Contents

1	Introduction.....	1
1.1	Context.....	1
1.2	Applications.....	3
2	Preprocessing	9
2.1	Nonlinear filters.....	11
2.1.1	Median filters and rank-order filters.....	11
2.1.2	Morphological filters.....	15
2.1.3	Polynomial filters.....	19
2.2	Amplitude-value transformations.....	20
2.2.1	Amplitude mapping characteristics.....	21
2.2.2	Probability distribution modification and equalization.....	22
2.3	Interpolation.....	24
2.3.1	Zero and first order interpolation basis functions.....	25
2.3.2	LTI systems as interpolators.....	27
2.3.3	Spline, Lagrangian and polynomial interpolation.....	28
2.3.4	Interpolation on 2D grids.....	33
2.4	Multi-resolution representation.....	37
2.5	Locally adaptive filters.....	43
2.5.1	Steerable smoothing filters.....	43
2.5.2	Iterative smoothing (diffusion filters).....	45
2.6	Problems.....	48
3	Signal and Parameter Estimation.....	51
3.1	Expected values and probability description.....	51
3.2	Observation and degradation models.....	56
3.3	Estimation based on linear filters.....	57
3.3.1	Inverse filters.....	57
3.3.2	Wiener filters.....	58
3.4	Least-squares estimation.....	60
3.5	Singular value decomposition.....	65
3.6	ML and MAP estimation.....	67
3.7	Parameter estimation and fitting.....	69
3.8	Outlier rejection.....	71
3.9	Correspondence analysis.....	74

3.10	State modeling and estimation	77
3.10.1	Markov processes and random fields	77
3.10.2	Hidden Markov models	80
3.10.3	Kalman filters	81
3.10.4	Particle filters	84
3.11	Problems	84
4	Features of Multimedia Signals	87
4.1	Color	87
4.1.1	Color space transformations	88
4.1.2	Representation of color features	97
4.2	Texture	102
4.2.1	Texture analysis based on occurrence counts	104
4.2.2	Texture analysis based on statistical models	107
4.2.3	Spectral features of texture	110
4.2.4	Inhomogeneous texture analysis	114
4.3	Edge analysis	115
4.3.1	Edge detection by gradient operators	115
4.3.2	Edge characterization by second derivative	119
4.3.3	Edge finding and consistency analysis	121
4.3.4	Edge model fitting	124
4.3.5	Description and analysis of edge properties	125
4.4	Salient feature detection	127
4.5	Contour and shape analysis	132
4.5.1	Contour fitting	132
4.5.2	Contour description by orientation and curvature	136
4.5.3	Geometric features and binary shape features	140
4.5.4	Projection and geometric mapping	144
4.5.5	Moment analysis of region shapes	154
4.5.6	Region shape analysis by basis functions	158
4.6	Motion analysis	159
4.6.1	Projection of 3D motion into the image plane	159
4.6.2	Motion estimation by the optical flow principle	163
4.6.3	Motion estimation by matching	168
4.6.4	Estimation of non-translational motion parameters	178
4.6.5	Estimation of motion vector fields at object boundaries	180
4.7	Disparity and depth analysis	183
4.7.1	Coplanar stereoscopy	183
4.7.2	Epipolar geometry	186
4.7.3	Camera calibration	189
4.8	Audio signal features	193
4.8.1	Audio feature extraction on the timeline	194
4.8.2	Time domain features	196
4.8.3	Spectral domain features	202
4.8.4	Cepstral domain features	206

4.8.5	Harmonic features.....	207
4.8.6	Multi-channel features.....	212
4.8.7	Perceptual features.....	213
4.8.8	Semantic features.....	215
4.9	Problems	217
5	Feature Transforms and Classification.....	223
5.1	Feature value normalization and transforms	223
5.1.1	Normalization of feature values.....	225
5.1.2	Eigenvector analysis of feature value sets	226
5.1.3	Independent component analysis	228
5.1.4	Non-negative matrix factorization	229
5.1.5	Generalized Hough transform.....	231
5.1.6	Derivation of statistical representations	232
5.2	Distance metrics.....	238
5.2.1	Vector distance metrics.....	238
5.2.2	Distance metrics related to comparison of sets	241
5.2.3	Similarity of probability distributions.....	243
5.2.4	Distance metrics based on prior knowledge about classes	249
5.3	Compressed representation of feature data	251
5.4	Feature-based comparison.....	253
5.5	Reliability	257
5.5.1	Reliability criteria	257
5.5.2	Quality of classification	260
5.6	Classification methods	264
5.6.1	Linear classification of two classes.....	265
5.6.2	Generalization of linear classification	270
5.6.3	Nearest-neighbor classification.....	273
5.6.4	Classification without prior knowledge	274
5.6.5	Maximum a posteriori ('naïve Bayes') classification	281
5.6.6	Artificial neural networks	284
5.7	Belief, plausibility and evidence.....	289
5.8	Problems	292
6	Signal Decomposition.....	295
6.1	Spatial segmentation of pictures	296
6.1.1	Segmentation based on sample classification	297
6.1.2	Region-based methods	302
6.1.3	Contour-based methods	304
6.1.4	Segmentation based on 'energy minimization'	305
6.2	Segmentation of video signals	311
6.2.1	Key picture and shot transition detection.....	312
6.2.2	Segmentation by background differencing	313
6.2.3	Object tracking and spatio-temporal segmentation.....	314
6.2.4	Combined segmentation and motion estimation	320

6.3	3D surface and volume reconstruction.....	321
6.3.1	3D point cloud generation.....	322
6.3.2	3D surface reconstruction	323
6.3.3	3D volume reconstruction.....	325
6.3.4	Projection based description of 3D shapes.....	326
6.4	Decomposition of audio signals.....	329
6.4.1	Temporal segmentation of audio	329
6.4.2	Audio source separation.....	329
6.5	Problems	331
7	Signal Composition, Rendering and Presentation.....	333
7.1	Composition and mixing of multimedia signals.....	333
7.2	Mosaicking and stitching	338
7.3	Synthesis of picture content.....	341
7.4	Warping and morphing	345
7.5	Virtual view synthesis.....	347
7.6	Frame rate conversion.....	352
7.7	View-adaptive and stereoscopic rendering of image and video signals	356
7.8	Composition and rendering of audio signals.....	359
7.8.1	Sound effects	361
7.8.2	Spatial (room) features.....	364
A	Fundamentals and definitions.....	367
A.1	Fundamentals of signal processing and signal analysis	367
A.2	Fundamentals of stochastic analysis and description	376
A.3	Vector and matrix algebra.....	385
B	Symbols and Variables	391
C	Glossary and Acronyms.....	397
D	References.....	399
E	Index.....	413

tent-based aspects that are relevant in each of those. Basically, the annotation by content-describing metadata could be done manually, semi-automatic or automatic, where Fig. 1.1b shows a typical procedure for automatic generation, which is most relevant for ease of use. After acquisition and digitization, *preprocessing* is often applied, which shall improve the signal for the purpose of improving subsequent analysis steps. In this context, *linear or nonlinear filters* are employed; methods to increase the resolution by *interpolation* give densely sampled, quasi continuous signals. *Multi-resolution processing* is often applied for improved stability and scale invariance of the subsequent *feature extraction*. Examples of multimedia signal features are color, texture, shape, geometry, motion, depth and 3D structure for images and video; spatial, temporal, spectral, and cepstral characteristics, pitch, tempo, melody, phonemes etc. for audio and speech. Features that are invariant under different capturing conditions are highly desirable. If multiple features are used, a *feature transform* is useful, which shall provide a more compact feature representation in a different feature space or a sub-space which is more suitable for the subsequent *classification*. The last step is the classification itself, which consists of a suitable concatenation, weighting and comparison of the extracted features, which are usually matched against feature constellations known a priori. By this, mapping into semantic classes can be performed, which then allows describing the signal at a higher level of abstraction.

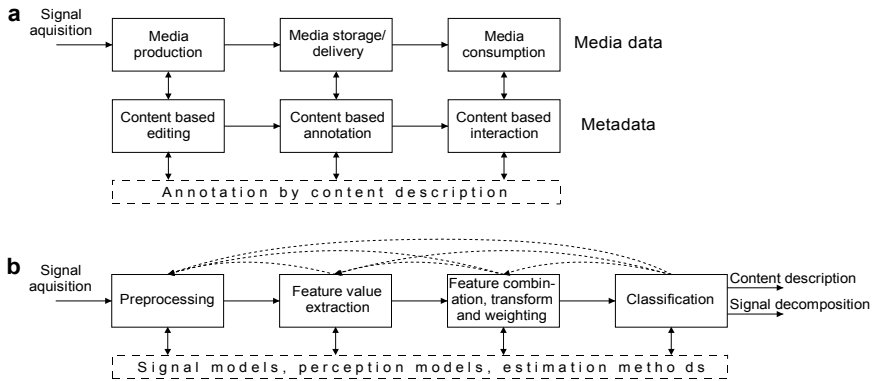


Fig. 1.1. **a** Media production, delivery and consumption chain **b** Processing chain for automatic multimedia content analysis and recognition

Most of the processing steps described here are based on signal models or statistical models, and need to involve estimation methods. On the other hand, when features are extracted and grouped into classes or categories, this knowledge also helps for better adaptation of the models and estimators. Hence, the four basic building blocks shown in the figure should not simply be understood as being forward connected. A recursive and iterative view is more appropriate as indicated by the dotted lines. For example, a classification hypothesis can be used to perform a feature extraction once again on the basis of an enhanced model, which is

Web cams	Persons, objects, events	Feature constellations characterizing persons, objects or events
Audiovisual communication	Appearance and actions of persons (talk/no talk); tracking of persons	Motion, silence, face features, localization features
Sports event augmenting	Automatic analysis of distances, comparison of time behavior of different runners, scene augmentation	Motion, time-line and spatial localization features, feature constellations characterizing persons, objects or events
Automation, inspection, service	Identification of objects or events; unexpected events	Feature constellations characterizing states, objects or events
Signal identification for copyright protection	Similarity with reference items, which must be stable under modifications	Signal footprints, fingerprints, watermarks
Smart cameras and microphones	Optimum scene properties for capture, focus on preferred objects; trigger acquisition in case of pre-defined events	Adjustment of illumination, color, tracking of camera motion or ego-motion of objects; localization of objects or persons; characterization of objects or events

In communication systems, the content description by itself can also significantly contribute to reduce the necessary bandwidth in transmission and storage: Signals which can clearly be identified as undesirable by a compact feature description do not need to be transmitted or stored at all. In some cases, *real-time analysis* of feature data is necessary for this (e.g. in surveillance, smart cameras, real-time communication); in other cases, more complex analysis methods may be performed off-line. In particular for retrieval from archives, databases, or scheduled program streams, the index data relating to the items can be pre-computed and stored either along with the media or in a separate database.

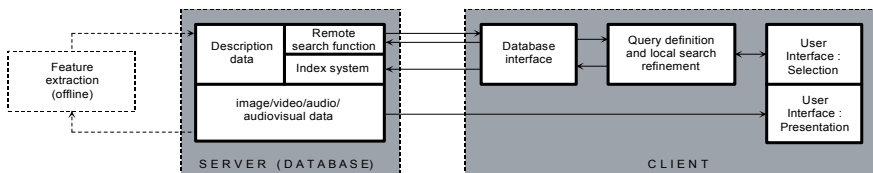


Fig. 1.2. Block diagram of a distributed retrieval application

Fig. 1.2 shows an example block diagram of a distributed retrieval application, where media data shall be found in a remote database. As database indexing systems usually provide powerful and efficient search functions, it is necessary to map the feature-based distance criteria (as resulting from a specific query defini-

tion) into the respective remote search function of the database, which could be defined as an application program interface (API). If not all desired comparison methodologies are supported by the given database system, it is also possible to perform a search refinement locally at the client, after a remote pre-selection at the server (database) side. It is indeed not useful to perform exhaustive search at the client side, as even transmission of a huge amount of compact metadata could be undesirable, if the number of items in the database is large. After the indices of a limited number of most similar items have been determined, the associated media items themselves would be retrieved from the database system and be presented to the user.

By using standardized metadata description formats it is possible to build *interoperable* and *distributed systems* for content-aware applications. For example, a multimedia signal retrieval task could simultaneously look up multiple databases, media or web servers, each of which would preferably accommodate the same schema of feature description. If this is not the case, transcoding of the metadata format is necessary, which typically is costly and time-consuming, such that fast responses of the retrieval system are impossible; furthermore, precision of the description may be lost. Examples of multimedia related metadata description standards are the *Resource Description Framework* (RDF) of the World Wide Web Council (W3C), the *Metadata Dictionary* of the Society of Motion Picture and Television Engineers (SMPTE), the *Dublin Core Metadata Initiative* (DCMI), and MPEG-7 (ISO/IEC 15938: *Multimedia Content Description Interface*).

Regarding the focus of this book in multimedia signal processing, the MPEG-7 standard is interesting, as it directly includes methods to describe low-level features of audiovisual signals. By this, MPEG-7 could also be interpreted as an ‘audiovisual language’, which fills a gap that exists due to the fact that signal features can be described by numbers expressing feature states, rather than by text. In case of major parts of the MPEG-7 standard, normative specification only covers the representation of the content description, while generation of descriptions and consumption of descriptions are regarded as application specific aspects (see Fig. 1.3)¹. The normative representation needs however to relate both to the syntactic structure of a description and to the semantic meaning of description elements, such that the generation of these data should not be random.

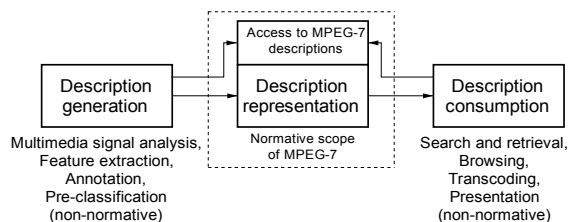


Fig. 1.3. Normative and non-normative elements in an MPEG-7 application

¹ Some elements of MPEG-7, e.g. *compact descriptors for visual search* defined in part 13, also define parts of feature extraction as normative.

is generally not possible for the case of a nonlinear system². When the combining function does not change depending on the coordinate position where it is applied, the nonlinear system is shift invariant.

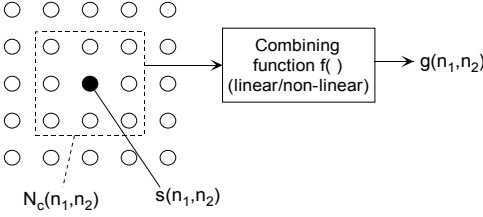


Fig. 2.1. Principle of linear or nonlinear 2D filtering using a finite symmetric neighborhood system

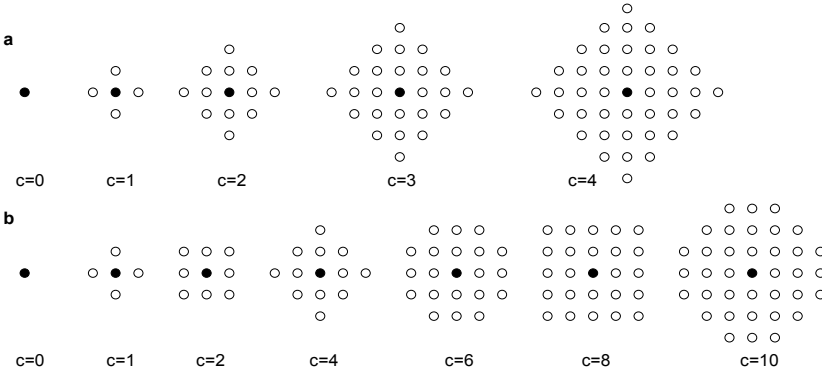


Fig. 2.2. Homogeneous 2D neighborhood systems $\mathcal{N}_c^{(P)}(n_1, n_2)$ with $P=1$ (a) and $P=2$ (b) for various values of c .

Symmetric neighborhood systems are often used in context of linear and non-linear filtering of images, avoiding shifting and degenerative effects of local structures in the output. This property is fulfilled by a *homogeneous neighborhood system*, where samples at positions \mathbf{m} establish the neighborhood of a sample at position \mathbf{n} according to a maximum distance norm of order P ³:

$$\mathcal{N}_c^{(P)}(\mathbf{n}) = \left\{ \mathbf{m} = [m_1 \ \dots \ m_\kappa]^T : 0 < \sum_{i=1}^{\kappa} |m_i - n_i|^P \leq c \right\}, (P \vee c) \geq 0. \quad (2.1)$$

The parameter c influences the size, while P influences the shape of the neighbor-

² The system transfer function of polynomial filters (Sec. 2.1.3) could be mapped into a higher-order spectral transfer function by applying multi-dimensional Fourier transform.

³ For images, the number of dimensions is $\kappa=2$. In case of symmetric neighborhood systems, the current sample at position \mathbf{n} is also a member of the corresponding neighborhood systems of any of its neighbors, $\mathbf{m} \in \mathcal{N}(\mathbf{n}) \Leftrightarrow \mathbf{n} \in \mathcal{N}(\mathbf{m})$.