Jens-Rainer Ohm

Multimedia Content Analysis

Springer

Table of Contents

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.

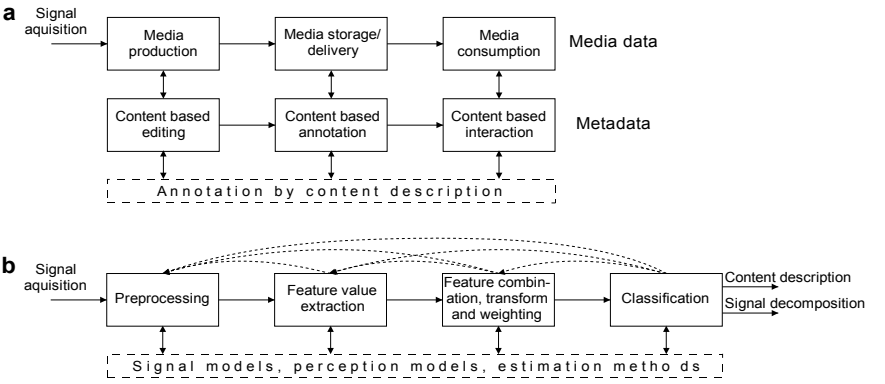

Fig. 1.1. **a** Media production, delivery and consumption chain **b** Processing chain for automatic multimedia content analysis and recognition

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.

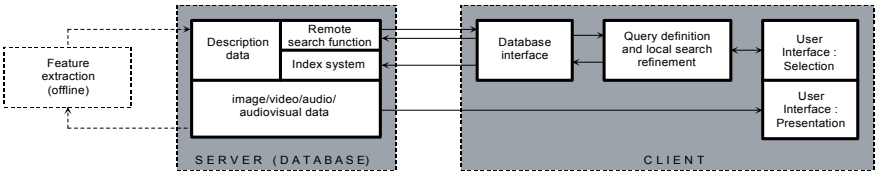

Fig. 1.2. Block diagram of a distributed retrieval application

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.

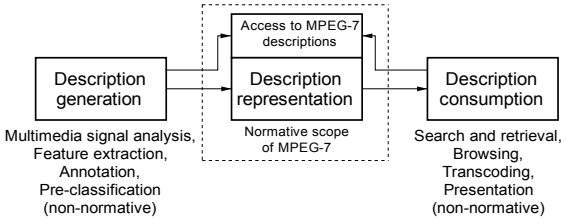

Fig. 1.3. Normative and non-normative elements in an MPEG-7 application

[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 system[2]. When the combining function does not change depending on the coordinate position where it is applied, the nonlinear system is shift invariant.

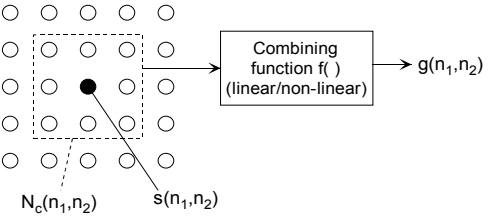

Fig. 2.1. Principle of linear or nonlinear 2D filtering using a finite symmetric neighborhood system

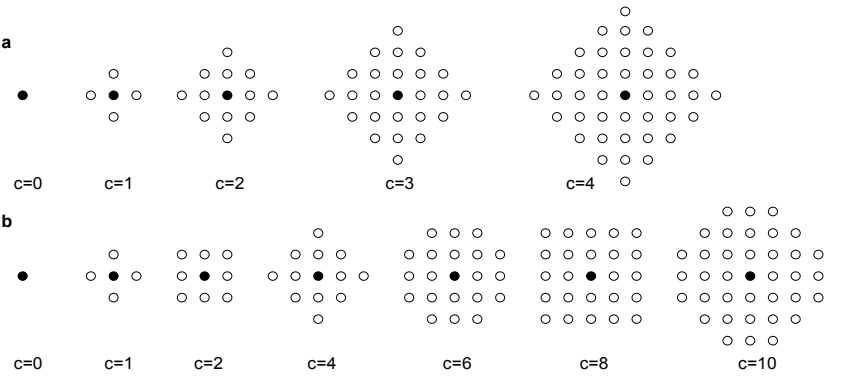

Fig. 2.2. Homogeneous 2D neighborhood systems $\mathcal{N}_c^{(P)}(n_1,n_2)$ with $P=1$ (**a**) and $P=2$ (**b**) for various values of c.

Symmetric neighborhood systems are often used in context of linear and nonlinear 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 **m** establish the neighborhood of a sample at position **n** according to a maximum distance norm of order P[3]:

$$\mathcal{N}_c^{(P)}(\mathbf{n}) = \left\{ \mathbf{m} = \begin{bmatrix} m_1 & \dots & m_\kappa \end{bmatrix}^{\mathrm{T}} \;:\; 0 < \sum_{i=1}^{\kappa} \left| m_i - n_i \right|^P \le c \right\}, \left(P \vee c \right) \ge 0 \,. \qquad (2.1)$$

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 **n** is also a member of the corresponding neighborhood systems of any of its neighbors, $\mathbf{m} \in \mathcal{N}(\mathbf{n}) \Leftrightarrow \mathbf{n} \in \mathcal{N}(\mathbf{m})$.