14.1 Introduction

Video databases are widely available either in broadcasters or in production/post-production houses or, increasingly so, in social media sites,
like YouTube. The video information stored in such databases is immense in size. It is typically (manually) annotated by text and tags/keywords. Nowadays, video search is performed mainly using such keywords.

Decades ago, the search for just a few words in a large number of text files posed significant difficulties. Similarly, today, the search for specific content in digital video files is difficult. While text search engines are widely available today, content-based video retrieval is still an open issue. Many techniques have been proposed to lead the way to the new and exciting world of content-based video retrieval. These studies concentrate mainly in video analysis and (hopefully automatic) description and also on video indexing, search and retrieval from video databases. In this Chapter, low-level and semantic video analysis and description techniques are presented. Furthermore, the MPEG-7 standard on audiovisual data description is overviewed. MPEG-7 profiles that are suitable for audiovisual content description in XML files (e.g., AVDP) are detailed. Finally, the current form and functionalities of the audiovisual archives are described. New topics, like label propagation for video annotation, relevance feedback and user profiling are presented as well.

14.2 Hierarchical video structure

The hierarchical video decomposition presented in Figure 14.2.1 is a common practice serving two purposes. It exposes the semantic video characteristics and can provide a framework for the spatiotemporal analysis and description of the video content.

First a temporal video decomposition can be performed. A video (e.g., a movie) consists of a sequence of scenes. A scene is a sequence of shots. A shot is a sequence of consecutive video frames, which are captured without interruption by a stationary or a continuously moving camera. Thus, a movie scene, which contains alternating views of two persons consists of multiple shots. A scene is defined as a collection of one or more continuous shots, focusing on an object or objects of interest, or on physical scene, or on a semantically meaningful part of the story (narrative). For example, shots showing a person walking in a corridor and entering a room could constitute a scene, if the person has been captured by different cameras. Three camera shots, showing three different persons walking down a corridor may constitute a scene, if the object of interest is the corridor and not the persons. Video scene must not be confused with physical scene (also called set), where action takes place, thought they sometimes coincide semantically. In the following sections, we shall present methods for segmenting a video stream into shots and scenes.
14.2.1 Shot/scene cut and transition detection

There are various types of shot transitions. A shot cut is an abrupt shot change. A fade-in/fade-out is a slow change in shot luminance which usually leads to, or starts with, a black frame. A dissolve takes place when there is a spatial blend of the frames of two successive shots for the duration of the transition, so that the luminance of the images of the first shot decreases and that of the second shot increases. A wipe occurs when the pixels of the second shot gradually replace those of the first shot with a local motion, e.g., from left to right. Of course, many other gradual shot transition types are possible. All these shot transitions are very popular tools for movie editing. The extent of their use though depends on the editor’s style. For example, once popular fancy shot transitions go out of style and the classical simple transitions (e.g., the abrupt cuts) reemerge.

Abrupt changes are easier to detect, compared to gradual ones. The techniques for abrupt cut detection can form the basis for more recent techniques for the detection of gradual shot transitions. Most abrupt shot cut detection methods examine two video frames $f$ and $f'$ and define a dissimilarity metric $D(f, f')$. If its value exceeds a threshold, a transition between two shots is detected. Various similarity/dissimilarity measures
can be used to this end, meaning, e.g., color/texture/motion/object shape similarity within a shot (and dissimilarity between two successive shots). The simplest such metric between two consecutive video frames is the Mean Square Difference (MSD) of their pixel luminances. Also, the simple color, or luminance, or motion histogram difference between the frames \( f \) and \( f' \) can also be used to this end. More complicated methods based on the same principles can be used for gradual shot transition detection.

Shot detection can also be considered as a video frame clustering problem over time. The video frames within each shot should be similar. They should be dissimilar across consecutive shots. Once video frames are clustered into consecutive shots, shot transitions are easy to detect at or close to the shot start/end frames. The same color/texture/motion/object shape similarity measures can be used for video frame clustering, as the ones used for shot transition detection.

Another approach to shot cut detection utilizes human interaction to improve performance, while, at the same time, preserving content integrity. A human annotator selects a few key frames in various representative shots. Then, a label propagation method can propagate the shot labels to other frames automatically, to finally compile a list of video shots. Content (e.g., color/texture/motion) similarity between successive labeled and unlabeled video frames can be used for label propagation. This procedure requires limited time and effort, in comparison to manual shot annotation.

Shots can be classified into several classes, most notably in long shots, medium shots and close-ups, as already described in Chapter 6. This can be typically done by finding the region of the main actor body and comparing its size with the video frame size. Long shots contain a lot of background (and the actor body). Medium shots contain part of the human body. Close-ups contain the actor face. Therefore, human body and face localization is performed first, in order to decide on the shot type.

Video scenes have a highly semantic nature. Therefore, scene detection is much more abstract and difficult than shot detection. As audio (music) is used by movie directors to underlie scene continuity across shots, audio analysis (e.g., detection of same musical motif or background noise) can be used to group shots into scenes or to detect scene transitions. Alternatively, shots can be detected by clustering consecutive shots, e.g., based on their background similarity denoting action in the same physical scene.

### 14.2.2 Key frame selection and video summarization

Given a long video sequence, a user may want to extract a number of key frames for video description and fast browsing. Their number may vary from 5% to 10% of the total frame number in the original video. Some-
times they are also called representative frames. The key frames should be able to summarize well the video content. Therefore, this task is called video summarization. Unfortunately, video content description is highly subjective and application-dependent. Therefore, there is no mathematical model, which defines the exact requirements for key frame selection. Many techniques for key frame extraction are based on shot cut detection, while other approaches employ the visual content and motion analysis. An easy way to perform video summarization is to select few representative video frames per shot, e.g., the first ones or the central ones (in terms of shot duration) or the central ones, in terms of video frame similarity. In the last choice, few video frames are selected per shot, which are the most similar to the rest of the shot video frames. Frequently, shot key frames are considered to be the ones which correspond to local minima of motion velocity, i.e., the key frames are recognized by the stillness of their visual content. In other cases, outlying video frames are selected to complement the central ones for shot summarization. This philosophy is based on the mathematical (and journalistic) notion of information: rare events carry more information than frequent events, e.g., when a person bites a dog, this event is interesting, the opposite event is not.

When video snippets, rather than video frames, are used for video description, then we have video skimming. It is performed along the same principles with video summarization, the difference being that we use short video snippets, rather than individual video frames.

Video summarization can be used to create an image gallery to describe a video visually. Video skimming is closer to trailer production. However, trailer production is a more sophisticated task, typically created by experts, since it is designed to capture audience attention (and interest to watch the full movie). Video skimming is more difficult in 3DTV, since the human eye can get very tired by fast depth transitions between the various 3DTV video snippets. Therefore, skimming must be performed with great care in 3DTV, in order to produce gradual depth changes and an acceptable overall result.

14.3 Spatiotemporal video description

In many cases, video metadata are created (usually manually) up to shot or scene level. This is understandable, since manual video annotation is a time consuming and unmerciful task. However, the current (semi)-automatic video analysis tools can provide detailed annotation granularity at frame or object level, as already described in Chapter 5. Therefore, we can perform spatial decomposition of a video frame into objects and
Objects, including persons and faces, can be detected by object/person/face detection algorithms. They are typically described by object *Regions Of Interest* (ROI), as shown in Figure 14.3.2. Objects can be tracked over time within a video shot, thus creating moving ROIs (called VideoObjects in video coding). This results in a spatiotemporal video decomposition into moving object ROIs that have a certain motion trajectory, as shown in Figure 14.3.3.

Shot background can be found as well, typically by moving object detection and subtraction. It can be described by a full frame image, or by a panoramic image, or by a 3D graphics model.

Humans are, very frequently, the primary actors in a movie. Therefore, the description of their presence, identity, status and activities is a very important task in video analysis leading to *human-centered* (also called *anthropocentric*) video descriptions. For example, facial image clustering allows us to automatically cluster facial images of the same actor into one cluster. Face recognition allows to tag facial image clusters with the corresponding actor names. Facial expression recognition allows to recognize
and tag various expressions in facial images, e.g., smile, anger. Human activity recognition allows to recognize human actions, e.g., walk, run, bend. Such anthropocentric video analysis tools are at various degrees of maturity and adoption. For example, face and even smile detection are currently used in various cameras to aid inexperienced users. Although they are error-prone, they provide rich metadata at object, shot and scene level that are very useful for a variety of applications, e.g., in video surveillance, human-centered computer interfaces and in semantic video search. When such video descriptions are stored in an XML file, fast searches can be performed of the form 'find a video shot where actor X smiles', 'find a video scene where actors X and Y converse'.

14.4 Semiautomatic approaches for video description and search

Semiautomatic approaches have also been proposed for content-based video retrieval. Most videos are annotated manually. Different people may have different semantic interpretations for the same video scene. Thus, to eliminate this confusion, adaptive and flexible approaches can be used. One approach is to use automatic tools for video analysis and description, as presented in the previous section. Then the human annotator can edit and amend the automatically extracted video descriptions. Another approach is to let an annotator annotate one video or some of it scenes/shots/frames/objects and then use label propagation tools to propagate annotations from one video to another one or from one shot to another one and so on, at varying granularity levels.

Another approach is to continue annotating the video during video search. Each time a user searches for a video with text-based query, text annotation tables are searched. When the query is satisfied, the query terms are added to the annotation text. Using such relevance feedback mechanisms that combine retrieval and annotation, the system can include different descriptions for the same video clip, while preserving, at the same time, semantic soundness.
14.5 Multimodal description of audiovisual content

Both audio and video features can be employed for audiovisual content description and retrieval. The general structure of audio and image feature fusion is shown in Figure 14.5.1. The mean audio intensity is used as a measure of shot significance. It has been shown that shots, which have a higher audio intensity, are more significant than those that have a lower than average audio intensity. For audio streams, the audio features are extracted from low-level audio characteristics. For video streams, the visual features are extracted using motion estimation with luminance histograms and pixel differences. Then, each feature sequence extracted from either audio or video channels is used for the identification of audiovisual content semantics. This way, e.g., in case of scenes containing humans, four different shot types can be identified: dialogue, monologue, action and generic video.

Video description techniques, using both the audio and video channels, have been proposed and have been further improved with the addition of text (subtitle) processing. Figure 14.5.2 shows the block diagram of such a method. Initially, the input audiovisual stream is split in an audio and a video stream. Subsequently, the audio stream can be classified in at least four classes: speech, music, environment noise and silence. Different music genres, environmental noise types or audio events (e.g., explosions) can be detected/recognized. In case of speech, the audio stream is further segmented in different temporal segments, depending on the current speaker. Therefore, speaker turns should be detected. Speaker turn detection can be assisted by subtitle information, in case it is available. Simultaneously, video shot detection and key-frame extraction are performed in the video stream. Following this, color/motion correlation between shots is calculated and an extended shot clustering is performed, so that shots, whose
objects or the background color are closely correlated (for example shots which were acquired in the same 3D physical scene), are grouped together. In the next step, the results of audio and video analysis are merged to improve shot clustering into video scenes. At this point, the results of speaker change detection can be combined with the results of color-based shot grouping, to find the scene transitions. The fusion rule is that shots within audio segments corresponding to one speaker, which are chromatically correlated, are grouped and marked as correlated. In other words, a shot sequence will be grouped in a scene, only when the correlation analysis of the visual content and the audio segmentation detect a common scene transition. In order to detect key persons, as far as audio is concerned, the audio segments, marked by the speaker change points, are further grouped. In video, the shots can be grouped based on key-frame or facial region clusters. Then, the following heuristic rules are used for the detection of potential key persons in audio and video data:

1) The speech/image duration ratio of human actors is usually higher than the corresponding speech/image duration ratio of other persons. This is the case, e.g., for anchorpersons in news broadcasts.
2) The main actor ROIs are more centrally located in the video frame.

3) The shot type is defined according to the main actor ROI in relation to the video frame size.

Semantic particularities can be observed in some video genres. E.g., in news broadcasts, the temporal dispersion of the speech/image duration of the anchorpersons is higher than the corresponding speech/image duration ratio of other persons. That is, the anchorperson will appear from the start until the end of a news broadcast. In the news video case, the visual and audio analysis results can be recombined, resulting to a valid anchorperson shot detection. The fusion rule used here is the logical conjunction between the anchorperson video shots and their speech segments.

Another method involving both visual and audio content for video indexing is shown in Figure 14.5.3. Audio processing starts with speech-silence detection, which is performed using the mean volume of the audio signal and the zero crossings rate, leading to signs of potential scene changes. The audio features are extracted only from voiced frames. For this reason, the video frames must be divided in voiced and unvoiced ones. This division is based on the low- to high-frequency power spectrum ratio of the audio channel, which is calculated using the Short Term Fourier Transform (STFT). In the voice segments, voice analysis can lead to speaker recognition. The extracted audio features can model well the speaker. Training audio data from various speakers can be used to this end. Furthermore, Text To Speech (TTS), also called speech recognition, methods can be used to transcribe the speech signal into text. Subtitle information can assist voice detection, when available. In this case, TTS may be optional. Video processing starts with the detection of shot transitions, using color differences. Subsequently, face detection is performed, assuming that faces can be characterized by skin-like color and have elliptical shape. Facial analysis follows, to estimate first the eyes and subsequently the mouth position. Mouth tracking is used to determine if a person speaks or not (visual speech). Cross-modal audio and video analysis facilitates the detection of shot transitions, using the detection of speech-silence periods. Additionally, face analysis facilitates speaker identification, especially in case when strong background noise corrupts the audio channel, e.g., in street scenes.
The MPEG-7 content description standard

In October 1996, the MPEG standardization committee, initiated an effort to create a digital media content description standard, for various media data types. The audiovisual data which can be described in MPEG-7 may include static images, 2D graphics, three-dimensional object models, audio, speech, video and composition information on how all these elements are combined in a multimedia scenario. The MPEG-7 standard does not aim at a specific application. The MPEG-7 representation tools do not depend on digital media content encoding and storage modes. However, the MPEG-7 standard can exploit the advantages existing in the encoded digital content (e.g., in MPEG-4 format). MPEG-7 aims to standardize:

a) A basic core of Descriptors (D) which can be used for the description of the various features of the multimedia content.

b) Predefined descriptor structures and their relations, which are called Description Schemes (DS).

Figure 14.5.3: Audiovisual content analysis.
c) A language for the determination of the descriptors and the description schemes, which is called Description Definition Language (DDL).

d) The encoded representation of descriptors for efficient storage and fast access to multimedia data.

14.6.1 MPEG-7 parts

The MPEG-7 standard is organized in parts, as shown in Table 14.6.1. The basic functional parts which comprise the MPEG-7 standard are described in the following sections.

Table 14.6.1: MPEG-7 standard parts.

<table>
<thead>
<tr>
<th>Part number</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Systems</td>
</tr>
<tr>
<td>2</td>
<td>Description Definition Language</td>
</tr>
<tr>
<td>3</td>
<td>Visual Description Tools</td>
</tr>
<tr>
<td>4</td>
<td>Audio Description Tools</td>
</tr>
<tr>
<td>5</td>
<td>Multimedia Description Schemes</td>
</tr>
<tr>
<td>6</td>
<td>Reference Software</td>
</tr>
<tr>
<td>7</td>
<td>Conformance</td>
</tr>
<tr>
<td>8</td>
<td>Extraction/use of descriptors</td>
</tr>
<tr>
<td>9</td>
<td>Profiles and levels</td>
</tr>
<tr>
<td>10</td>
<td>Schema definitions</td>
</tr>
</tbody>
</table>

The MPEG-7 Systems part includes the tools required for the preparation of the MPEG-7 descriptions for the more efficient transfer and storage of digital media, the terminal architecture and the regulative interfaces.

The MPEG-7 Description Definition Language part allows for the creation of new description schemes and new descriptors, if needed, and the extension/modification of existing description schemes. It is based on the XML language. As XML has not been specifically designed for multimedia content description, it was extended specifically for the MPEG-7 standard. As a result, this language can be separated in the following logical components:

- structural components of the XML scheme language,
- data type components of the XML scheme language,
• specific MPEG-7 extensions.

The MPEG-7 visual description tools offer basic structures and descriptors for the following basic visual characteristics: color, texture, shape, motion, locality and faces. Each category consists of elementary and advanced descriptors. In the next section, the visual content description technologies of MPEG-7 will be summarized.

The MPEG-7 audio description tools provide structures for audio content description. A set of low-level audio descriptors uses these structures for audio features that are encountered in many applications (e.g., spectral, parametric and temporal characteristics of audio signals). MPEG-7 also provides a suit of high-level audio description tools, which are more application-dependent. These high-level tools include general audio recognition tools and description indexing tools, tools for the description of the instrument timbre, speech content description tools, audio signal description schemes. Melody description tools enable queries by humming (a song).

The MPEG-7 Multimedia Description Schemes (MDS) constitute the set of description tools (descriptors and description schemes) for both basic and multimedia entities. The basic entities are general characteristics used in all digital media, e.g., vector, time and spatial coordinates. Except for this set of general description tools, more complex description tools have been standardized for multimodal media description (e.g., video, 3D graphics and audio). These description tools can be grouped in different classes, depending on their functionality:

1) content description: it represents structural aspects of the multimedia content, e.g., decomposition into spatial/temporal segments;

2) content management: it contains information on the creation, coding, and use of the multimedia content;

3) content organization: it analyzes and clusters multimedia content, e.g., into collections;

4) navigation and access: it specifies summaries and other ways to access, browse and navigate in the multimedia content;

5) user interaction: it describes user interface issues, user preferences and usage record, related to the audiovisual material consumption.

The MPEG-7 reference software, also called eXperimentation Model (XM) simulation platform, provides simulation for MPEG-7 Ds, DSs and DDLs. Beyond normative components, several non-normative components may be needed during simulation. The applications consisting of the corresponding data structures and programs are divided in two types: server applications (retrieval) and client applications (search, filtering).
The part of MPEG-7 conformance contains instructions and compatibility check procedures of each MPEG-7 implementation.

MPEG-7 extraction and use of descriptions presents, in an informative way, how some descriptions can be extracted and used. Since MPEG-7 is a vast collection or more than 1500 multimedia descriptors, Profiles and MPEG-7 Levels have been proposed, so that limited sets of MPEG-7 descriptors and tools are tuned to particular application needs. Such standard profiles and levels are collected in the respective MPEG-7 part. Finally, all MPEG-7 schemes are collected in the MPEG-7 Schema Definition part.

14.6.2 The MPEG-7 visual standard

A very important aim of MPEG-7 is to provide standardized descriptions of stored image or video streams and standardized headers (low-level visual descriptors), which help the user, or the applications, to identify, classify or filter images or video. These low-level descriptors can be used to compare, filter or display images or video, based only on visual content descriptions, or, if necessary, in combination with plain text-based searches.

MPEG-7 descriptors can be grouped in two categories: generic visual descriptors and specialized visual descriptors. The former ones include color, shape and motion characteristics, while the latter ones include, e.g., detection, localization and identification of human faces. This section focuses on generic descriptors, which can be used in most applications.

A particular effort has been made in the design of efficient color descriptors to be used in the detection of chromatically similar images. There is no generic color descriptor that can be used for all applications. Therefore many generic descriptors were standardized, each one being appropriate for a special visual similarity function. For example, several color spaces are included in the standard, notably RGB, $YC_bC_r$, HMMD and HSV. Descriptors for dominant color and color layout (spatial color distribution) are also supported.

MPEG-7 has defined appropriate texture descriptors, which can be used in various applications. The homogeneous texture descriptors describe texture directionality, roughness and regularity and are mostly suitable for the quantitative characterization of homogeneous texture. They use frequency characteristics to describe the image texture, its energy and energy variability. They are based on filter banks, which are sensitive to scaling and orientation (e.g., Gabor filters). Other suitable descriptors are acquired in the frequency domain, by calculating the mean value and standard deviation of transform (e.g., DFT) coefficients. Non-homogeneous texture
descriptors (e.g., edge histograms) capture the spatial edge distribution, towards describing directional textures existing, e.g., in images of wood, sea waves and cloth.

Finally, MPEG-7 has defined a number of shape descriptors, which can be used to describe 2D and 3D shapes. Objects, whose shape features are best expressed by contour information, as the one shown in Figure 14.6.1a, can be described using the MPEG-7 contour shape descriptors. Objects, whose shape is best described by the region they occupy, like the one shown in Figure 14.6.1b, can be described by the MPEG-7 region shape descriptor. 2D/3D shape descriptors can be used when the shape of a three-dimensional object can be approximately described by a limited number of 2-dimensional projections, which are taken as its two-dimensional snapshots from different view angles. Thus, the two-dimensional MPEG-7 shape descriptors can be used to describe each one of the two-dimensional shapes, which are projections of the three-dimensional object. Therefore, the detection of similarities between three-dimensional objects can be performed by matching multiple pairs of two-dimensional shots, one for each object. 3D object shapes can be described by polygonal (e.g., triangular) meshes, as done in VRML files.

Motion description in a video sequence, by using motion fields, can be particularly heavy in terms of bits per video frame, even if the motion fields are sparse. MPEG-7 has developed descriptors which capture basic motion characteristics from the motion field in concise and effective descriptions.

Motion activity descriptors. In a video segment (e.g., a scene, a shot or a given set of consecutive frames), the overall motion activity level or speed is captured by the motion activity descriptor. It describes if a scene is slow, fast or very fast. Examples of high activity are sport or action scenes. Furthermore, scenes from news broadcasts can be considered as low-activity ones, especially the anchorperson shots. The motion activity intensity descriptor measures the motion intensity. Optionally, the motion direction, the spatial distribution of the motion activity and the temporal
distribution of the motion activity can also be extracted as motion activity descriptors and be used for motion similarity detection.

Camera motion descriptor. The motion of the physical or virtual camera can be described by the camera motion descriptor. This descriptor yields information on total camera motion parameters, which exist at a given time instance in a shot. These parameters can, in certain applications, be supplied directly by the camera. It is also possible that these camera parameters are estimated from the video sequence itself, using an appropriate camera model and camera motion estimation. The descriptor can be used for searching video sequences based on certain total motion parameters, such as shots with a high zooming activity or shots with mainly camera pan motion.

Motion trajectory descriptor. The object motion can also be described separately for each independently moving object, using the motion trajectory descriptor. The object trajectory is typically found by object tracking, as described in Chapter 5. It describes object displacement as a function of time and allows object trajectory matching for object motion-based video search. One possible application is visual surveillance, e.g., searching for objects which move close to a specific region, or objects (e.g., cars) which move faster than a speed limit. Another application is human activity or gesture recognition, where hand/limb/foot trajectories can be employed to this end.

14.7 MPEG-7 based audiovisual content description

A considerable amount of research effort has been invested over the last years to improve MPEG-7 ability to deal with semantic video content description. For example, an anthropocentric description scheme for video content analysis, called ANTHROPOS-7 has been proposed. Its main assumption is that humans (actors) are the most important entity in most movies. The basic idea is to observe humans and their environment in video shots and organize the video content description according to our perception about humans (and their context/background). Therefore, this description introduces an MPEG-7 structure for semantic video content description regarding humans, their status/activities and the relevant context (scene background, props, objects). The so-called Detailed Audio Visual Profile (DAVP) MPEG-7 profile has also been proposed, which is based mainly on the MPEG-7 part 5 (MDS) and provides a way to describe
single multimedia content entities, based on a comprehensive structural
description of the content. Finally, an industrial application of the Core
Description Profile (CDP), namely the Metadata Production Framework
(MPF) has been proposed.

The MPEG-7 Audio Visual Description Profile (AVDP) provides a stan-
dard way to store high- and low-level video descriptions that are extracted
either by (semi)automatic video content analysis or by manual annotation.
AVDP is based on the previous DAVP and MPF description schemes and
was designed to benefit both broadcasters and media industry, by creat-
ing a normative layer for video content description. Such descriptions are
useful in video archival, search and retrieval, either in production archives
or in broadcasting archives or even in user-fed social media video content
search, e.g., in YouTube. For example, humans and their activities can be
detected and tracked in a video, using automatic face detection, tracking
and recognition and the appropriate descriptions can be stored in XML
files. Then, we can search these files instead of searching the raw video
content, e.g., for finding videos depicting person X, while walking or smil-
ing. Needless to say that text or keyword search is impossible in many
cases, since manual annotations are either completely missing or inade-
quate. Even if such manual annotations do exist, e.g., in production or
broadcasting archives, their granularity is very gross. For example, we can
retrieve videos where person X walks, but we cannot spot the exact video
segment where this event happens indeed.

AVDP was introduced so as to solve two main problems arising form
the MPEG-7 architecture, namely complexity and interoperability. It is
well known that MPEG-7 is a vast collection of (more than 1500) mul-
timedia descriptors and description schemes. Therefore, its use (as is)
in video content description is questionable. Moreover, interoperability
issues arise, since full interoperability between two different MPEG-7 doc-
uments is only assured by having knowledge on how the standard has been
used. Profiles are a rather old mechanism, which restricts standards to
a specific set of applications. This is done (e.g., in the case of MPEG-
7) by selecting appropriate subsets of Ds and DSs from MPEG-7 and by
defining new constraints on their use. For instance, interoperability can
be considered in a profile by restricting the use of several DSs to follow
a single and unambiguous operational way. AVDP provides all Ds and
DSs, as well as corresponding constraints, needed for audiovisual content
description. The most important AVDP aspects are the following ones.
TemporalDecomposition DS (TD) decomposes a video stream in scenes,
shots, or video segments. SpatialDecomposition DS (SD) describes regions
of interest (ROIs) corresponding to objects within frames. SpatioTemporal-
Decomposition DS (STD) describes moving regions of interest over
time, corresponding to object trajectories. In this case, objects can be
human bodies, faces, or scene props, or any other physical object. *MediaSourceDecomposition* DS (MSD) decomposes an audiovisual segment in its constituent audio and video channels. Video events and concepts can be described by using the MPEG-7 *StructuredAnnotation* (SA). It contains several fields that can be used for VideoSegment description, e.g., Who, WhatObject, How, WhatAction.

In Table 14.7.1, some of the most important MPEG-7 descriptors are illustrated that are used in AVDP for the above-mentioned decompositions. In the following, we shall concentrate on video content description, because audio content description is beyond the scope of this book. An illustrative example of the schematic representation of an AVDP descriptor for shot description is shown in Figure 14.7.1. In scene/shot boundary detection, we try to detect boundaries of scenes/shots, aiming at a temporal decomposition of the audiovisual content into different scenes/shots. Its results can be stored in an AudioVisual Temporal Decomposition (TD). Each resulting AVS will include only some temporal information, e.g., time codes for shot start/end.

### 14.8 Manual and semiautomatic video annotation

Video content description can typically have a header providing general information, such as:

![Schematic representation of scene/shot and key frame/key segment description in AVDP.](image-url)
Table 14.7.1: MPEG-7 Description Schemes used in the AVDP MPEG-7 profile.

<table>
<thead>
<tr>
<th>Description scheme</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>TemporalDecomposition (TD)</td>
<td>This description scheme (DS) acts as a tool that performs temporal decomposition of the video in multiple temporal segments, such as Video Segments (VS) or Audio-Visual Segments (AVS).</td>
</tr>
<tr>
<td>MediaSourceDecomposition (MSD)</td>
<td>It is used in the AVDP context to decompose an audiovisual segment (or an entire audiovisual content) into its audio and video channels.</td>
</tr>
<tr>
<td>VideoSegment (VS)</td>
<td>It provides a way to describe a video segment in the visual content. Each segment is defined by the starting time point of the VS and its time span.</td>
</tr>
<tr>
<td>SpatioTemporalDecomposition (STD)</td>
<td>It enables a video segment to be decomposed into spatiotemporal parts, namely MovingRegions, in order to store information related to moving objects.</td>
</tr>
<tr>
<td>MovingRegion (MR)</td>
<td>It is used to describe, for example, a moving object by storing the spatiotemporal object behavior, e.g., the spatial coordinates of the bounding box of the object and their change over time.</td>
</tr>
<tr>
<td>SpatialDecomposition (SD)</td>
<td>It is used for the spatial decomposition of a frame. The result is one or more regions of interest (StillRegion) within a video frame that may depict an object, face etc.</td>
</tr>
<tr>
<td>StillRegion (SR)</td>
<td>It is used to describe a still object, by defining its spatial region within a video frame. A StillRegion can also denote an entire video frame.</td>
</tr>
<tr>
<td>StructuredAnnotation</td>
<td>It is used to annotate concepts, events, human actions etc.</td>
</tr>
<tr>
<td>FreeTextAnnotation</td>
<td>It can be used to annotate a video segment with free text.</td>
</tr>
</tbody>
</table>
• Movie information
  – Title
  – Genre (comedy, drama, action, documentary, sport, etc.)
  – Date of release
  – Director
  – Cast (actors/actresses, athletes, etc.);
• Date of ingestion
• Movie rating (e.g., appropriate for adults only)
• Target audience (men/women, adults, elderly etc.)
• Technical data (bit-rate, video resolution, type of compression, etc.)

The above information is typically manually fed or updated during content ingestion and is contained in the header of the XML file describing the video.

Most videos are annotated manually. Different people may have different semantic interpretations for the same video scene. Thus, to eliminate this confusion, adaptive and flexible approaches can be used in a semi-automatic way for video annotation. One approach is to use automatic tools for video analysis and description, as presented in previous sections. Then the human annotator can edit and amend the automatically extracted video descriptions. Another approach is to let an annotator annotate one video or some of it scenes/shots/frames/objects and then use label propagation tools to propagate annotations from one video to another one or from one shot to another one and so on, at varying granularity levels.

Another approach is to continue annotating the video during video search. Each time a user searches for a video with text-based query, text annotation tables are searched. When the query is satisfied, the query terms are added to the annotation text. Using such relevance feedback mechanisms that combine retrieval and annotation, the system can include different descriptions for the same video clip, preserving at the same time semantic soundness.

Information concerning the use of the audiovisual content can be archived as well, such as:

• Number of views
• Views frequency
• Views per user group
• Relevance feedback information, if available
• Quality of experience, e.g., likes/dislikes
• Engagement information
• Geographic provenance of audience.

The last three entries are very frequently used in social media databases, such as YouTube. In this case, users provide also other forms of free-text annotations, such as tags and comments that can be valuable in content description, despite the inaccuracies existing in user-fed data.

14.9 Audiovisual content archives

Currently, the audiovisual archives in broadcasters all over the world are migrated from analogue to digital storage technologies. In parallel, the audio and video channels are enriched with different kinds of metadata, ranging from basic technical information up to sophisticated levels of automatic feature extraction and content description. There are various workflows for archive data storage, search and retrieval. The most common workflows are based on a two-step approach, where a first search is based on a highly compressed mirror image (proxy or low-res) of the original file, providing faster search results and less load for the storage devices and networks. In a second step, the original file (high-res) is retrieved from the archive. Such workflows utilize the different quality levels of hierarchical compression schemes, like JPEG2000 used in digital cinema. For content search, lower quality levels of the JPEG2000 file can be used.

In many broadcasters, audiovisual content archiving is still tape-based. However, most of them are in the process of migrating to file-based audiovisual content handling. In tape-based systems, content search is realized with the help of proprietary, sometimes "home-grown", systems, which link assets to the tape archive. As part of the above-mentioned migration to file-based content handling, they are partially replaced/amended by digital Media Asset Management (MAM) systems. Such systems facilitate the work of editors, schedulers, material administrators, etc., within all processes of digital movie scheduling. They are modular, flexible and extensible software systems. They handle movie documentation, license management, material management, trailer production, retrieval, slot scheduling, programme scheduling, trailer scheduling and, possibly, content editing. The scheduling carried out within programme and trailer scheduling is based on the movie documentation data saved in the MAM system. For movies, data such as reference numbers, broadcasting and scheduling information, cast lists, texts, awards, etc., are maintained. Material management involves the recording of and dealing with data connected with movie
material (material sets, time code locations etc.). Material managers request material for scheduled movies via an interface and assign it to the scheduled film version. Material information comprises, among others, the exact film length, time code locations, video and audio formats (4:3 or 16:9, Dolby Digital or stereo, etc.). There are different possibilities for audiovisual data retrieval, e.g., by starting queries for a detailed search, full-text search or combined search. For example, scheduling data, broadcasting data, license information, etc., can be retrieved. For a combined search, a filter can be used to narrow down the search area. MAMs can support either SDTV or HDTV content. Sometimes, content archiving can be performed via two MAM systems, one for the production/post-production domain and one specific for the program domain (ingestion, check, transmission). The metadata archived for describing the audiovisual content in the production/post-production domain can be: a) technical metadata (e.g., time code, audio/video quality, duration) documented by technicians before ingestion and during post-production and b) content-related metadata (mainly used when searching for news programs) documented by specialized personnel, after ingestion and post-production. The metadata archived for the audiovisual content in the program domain are mainly technical ones, whereas content-related data are very basic ones (name of the director, etc.). In general, so far, metadata creation/annotation is mostly manual, error-prone and very time consuming. To make things more complicated, various broadcasters use rather proprietary or very customized MAM solutions. Even language thesauri are not standardized. Therefore, metadata interoperability and transfer from one MAM system to another one is generally difficult. Furthermore, the current metadata granularity is a far cry from the one that can provided, e.g., by AVDP in conjunction with automated video analysis tools.

In terms of systems that can allow TV viewers to browse and search for broadcasted conventional TV content, there has certainly been some progress in the interaction between users and content, starting from the inclusion of Electronic Program Guides (EPG) services, as part of DVB or other DTV systems. EPG allows the users to browse through all the different services being broadcasted along a timeline, which extends from now to a time in the (near) future. Certainly, more advanced features can be added to the EPG, like programming of future recordings or alerts, or more information regarding the movie scenario, cast and awards.

Finally, 3DTV content archiving provides additional requirements and open questions, which are not yet addressed by the workflows and procedures that are currently in place for SDTV and HDTV. While archival of 3DTV material is pretty much manageable now, since only limited 3DTV content is available and produced, it is foreseen that proper procedures for archiving, indexing and retrieval of 3DTV content that takes into ac-
count the particularities of this medium, the rich information that it caries (depth being the most obvious one) and the needs of end users (broadcasters, production and post-production houses) and of the general public (in terms of 3D quality) should be set in place. 3D video content analysis and description is also very useful in the production of classical (single video channel) movie and SDTV/HDTV content, since in many cases, multiview camera setups are used in production and 3D video metadata are essential during post-production. The EU funded 3DTVS project aimed at providing solutions for 3DTV content description, archival and search. To this end, it extended the AVDP profile to cover 3DTV content.

14.9.1 Visual content indexing techniques

In the previous sections, various techniques were presented for analyzing video streams and extracting semantic information. The next step is to present techniques for creating the appropriate indexing structure of the acquired information to facilitate video retrieval. Hashing is a widely known technique for data indexing. Indexing techniques are different if we use low-level video descriptors (e.g., color, motion, or shape histograms) than high-level (semantic) video descriptors (e.g., based on face/person detection and tracking). In the following, we shall present the principles for low-level video indexing. Indexing can be done either at video segment or video frame level.

Low-level video indexing uses color, motion, or shape/texture video descriptors, e.g., in the form of histograms. Grayscale or color histograms can be used to this end, having a number of bins. Local edge directions in predefined image regions can be used for texture description. For each region, the histograms of four edge directions (horizontal, vertical and two diagonal) can be calculated using an edge detector, e.g., a Sobel or Canny one. Motion vector histograms can be used as well. By performing histogram quantization, we expect that visually similar video frames correspond to the same hash table entries. Each entry may have multiple indices to the video frames, which are stored in the database. When the database is queried using an image as the query example, the query image is subjected to analysis, using the same procedure and video frames, which have the same index in the hash table, are retrieved.

14.9.2 Relevance feedback and user profiling

Using a relevance feedback mechanism in an audiovisual database, a user can assign the relevance factor to a video returned to him after a
query. This mechanism is useful for the efficiency of the audiovisual content search system, since it facilitates the refinement of the search and retrieval routines. The information that could be stored in the database, in order to be used by this relevance feedback mechanism should generally contain the following information:

- information about all queries submitted by the users (e.g., type of query, textual data or fields included in the query);
- relations between a user and his queries (a user may be related to multiple queries);
- relations between a query and the resulting content (multiple videos may be the result of a query);
- the relevance factor that a user assigns to a resulting video of one of his/her queries.

Given this structure in the database, the relevance feedback mechanism is described as follows. Firstly, a user submits a query to the search system and a list of videos is returned to him/her. Then, the user rates each video with a factor that indicates the relevance of his/her expected output with the results provided by the system. Then, the query and its relevance to a resulting video (provided by the user) can be stored. This information can be used to assess the efficiency of the retrieval mechanism and to refine it accordingly.

*User profiling* and *personalization* can be another feature of an audiovisual content search system. It describes the ability of the search and retrieval system to act in a user-specific manner, i.e., according to user preferences. This means that the results of a query of a specific user depend on the information about his/her preferences and his role. To this end, for every user, two types of information, namely user personal details (e.g., identity, role) and user preferences must be stored in the database. This is accomplished through user profiling. Personal details rarely change. However, in contrast, user preferences change over time and should be always updated, according to the user’s activity. An example of the first type of information is the role of the user in terms of system usage; for instance, he/she can be an archivist, journalist, etc. Thus, depending on his/her role, the provided system functionalities and access rights can vary with the user. As for the second type of information, it can indicate how much a user is interested in specific video genres (e.g., documentaries or sports video). Also, a history of the queries he/she has submitted must be stored as well. To facilitate the initial profiling procedure, example user profiles can be provisioned, so that a profile can be assigned to each new user, automatically or manually. Later on, user profiles can be automatically
or manually altered, depending on the overall users’ activity. Preference information can be extracted by analyzing the user queries, search and retrieval results and viewing history. The query history provides raw information about the user preferences. The search and retrieval results, on the other hand, provide a direct way to infer user preferences by examining, for example, his/her relevance feedback on videos returned to a user during his search. Lastly, the viewing history of a user is also a direct way to infer preference information.

14.10 Conclusions

Today, thanks to the abundance of audiovisual content in the Internet, the actual use and value of digital video depends on the development of efficient description, storage, indexing and retrieval techniques. Content-based retrieval is an active research and development field. Many research prototypes and innovative techniques have been developed during the last decade, notably MPEG-7 and AVDP. Some of these have been incorporated in commercial products. Some of the audiovisual content description tools have affected the MPEG-7 standardization activities and were incorporated in its current version.

MAM systems used in production and in broadcasting offer audiovisual metadata storage. Users are provided such metadata, e.g., in the form of EPGs during broadcasting. However, the capabilities of existing systems are only a far cry of the potential in this area. The semantic gap between the users’ needs (mainly semantic video search/description) and the currently available technology (mainly low- to middle-level video characteristics) continues to exist. Strong research and development effort is needed, in order to develop fully operational tools for video description and retrieval.