Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 19 (2004) 437–456
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Using MPEG standards for multimedia customization

João Magalhães[a,*], Fernando Pereira[b]

[a] *Instituto Superior de Engenharia de Lisboa, Portugal*
[b] *Instituto Superior Técnico, Instituto de Telecomunicações, Lisboa, Portugal*

Received 12 August 2003; received in revised form 24 December 2003; accepted 20 February 2004

Abstract

The multimedia content delivery chain poses today many challenges. The increasing terminal diversity, network heterogeneity and the pressure to satisfy the user preferences are raising the need for content to be customized in order to provide the user the best possible experience. This paper addresses the problem of multimedia customization by (1) presenting the MPEG-7 multimedia content description standard and the MPEG-21 multimedia framework; (2) classifying multimedia customization processing algorithms; (3) discussing multimedia customization systems; and (4) presenting some customization experiments.

Keywords: Universal Multimedia Access; Multimedia customization; Multimedia adaptation; Content description; Usage environment description; MPEG-7; MPEG-21

1. Introduction

The exploding variety of multimedia information is nowadays a reality since everyone has a camera, a scanner or another device that almost instantly generates multimedia content. Most often the content author wishes to share its masterpiece with everyone, but the variety of terminals and networks may be a problem if he/she wants everyone to see his/her work with the best possible quality. Typically, when a terminal accesses content to which it was not designed for, the user experience is rather poor. The scenario described misses a bridging element between all the components involved, which should take into account their characteristics and assure an efficient and consistent inter-working: in other words, an efficient way to "access any information from any terminal", eventually after some content adaptation.

The access to multimedia information by any terminal through any network is a new concept referred in the literature as Universal Multimedia Access (UMA) [30,31]. The objective of UMA technology is to make available different presentations of the same information, more or less complex, e.g., in terms of media types or bandwidth, suiting different terminals, networks and user preferences. In UMA scenarios, and in order to more easily and efficiently customize the desired content, it is essential to have available descriptions of the parts that have to be matched/bridged—the content and the usage environment:

- *Content description*: Information on the content features—e.g., resolution, bit-rate, motion,

*Corresponding author.
E-mail address: jmag@deetc.isel.ipl.pt (J. Magalhães).

0923-5965/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2004.02.004

color, pitch, temporal structure, genre—which may be instrumental to perform an efficient customization of that content; if an adequate content description is available, there is no need to extract on-the-fly content features to perform an adequate adaptation.

- *Usage environment description*: Information on the usage conditions—e.g., terminal, network, user preferences, natural environment—which determine the quality of the experience to be provided through an adequate customized variation of the pretended content. If no usage environment description is available, it is difficult to provide adapted content adequately fitting the consumption conditions.

These descriptions have to be matched by some module in the content delivery chain that will produce and implement a decision to perform a set of content customization operations to provide the user with the content for the best possible experience.

MPEG has dedicated large efforts to the standardization of tools for such application scenarios in terms of three major dimensions: content coding (MPEG-1, MPEG-2 and MPEG-4), content description (MPEG-7 [17]) and usage environment description (MPEG-21 [20]). The major objective of this paper is to discuss the role of the various MPEG standards in the context of multimedia customization scenarios and to contribute for a better organization and understanding of the multimedia customization problem.

This paper is organized as follows: Section 2 presents the MPEG-21 vision of a multimedia framework aiming to enable the transparent and augmented use of multimedia resources across a wide range of networks, devices, and communities; a brief description of MPEG-7 capabilities which may fit in the MPEG-21 framework is also made in this section. Section 3 organizes the several types of multimedia processing algorithms which may be used for the content matching process to the usage environment characteristics. Section 4 overviews multimedia customization systems available in the literature and proposes a rather generic system architecture [14] fitting in the MPEG-21 multimedia framework. Finally, Sections 5 and 6 present some experiments and the final remarks regarding the presented work.

2. The MPEG-21 multimedia framework

Considering all the issues inherent to a content delivery chain (and not only in the UMA perspective), the MPEG-21 standard [17] proposes to define a multimedia framework for the transparent multimedia delivery and consumption by all players in the delivery and consumption chain. In this framework, many standard technologies are needed to provide the various functionalities required such as coding, multiplexing, synchronization, description, rights expression, rights management, etc. The MPEG-21 framework will make use of the relevant available standards providing efficient solutions for some of these functionalities, e.g. MPEG-4 for coding and MPEG-7 for description, and will develop new standards whenever required.

In the context of the MPEG-21 multimedia framework, the main entities are Users and Digital Items, see Fig. 1:

- An *MPEG-21 User* is any entity that interacts within the multimedia framework: it can be a content creator, a content distributor, or a content consumer (end user). Users include individuals, consumers, communities, organisations, corporations, consortia, and governments. Users are identified specifically by their relationships to others Users for a certain

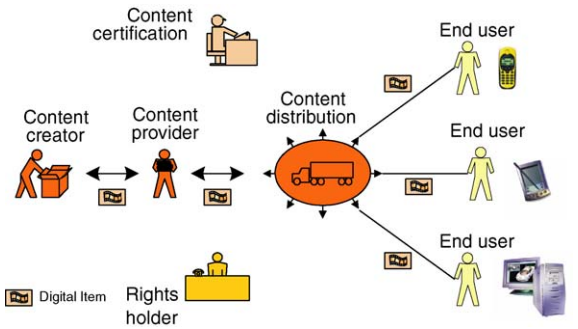

Fig. 1. A representation of the MPEG-21 multimedia framework.

interaction. From a purely technical perspective, MPEG-21 makes no distinction between a "content provider" and a "consumer"—both are Users in the MPEG-21 multimedia framework.

- The *Digital Item* is the fundamental unit of distribution and transaction among Users in the MPEG-21 multimedia framework. It is a structured digital object with resources (the content), unique identification and corresponding metadata (e.g. MPEG-7 description). The structure relates to the relationships among the parts of the Digital Item, both resources and metadata.

Once the content (in the form of Digital Items resources) will be exchanged in the defined framework, there will be entities that will offer content customization functionalities to achieve an optimal end user experience. Therefore, MPEG-21 sets the trail to create a complete UMA system, where such entities will play the role of the "bridging element between the parts that have to be matched/bridged"—the multimedia content and the usage environment.

2.1. Multimedia content description

To facilitate the development of powerful applications for multimedia information retrieval, customization, distribution, and manipulation, some knowledge about the multimedia content characteristics is essential in any content delivery chain [4]. The best the content is known, the more efficient it may be processed, whatever the type of processing to be applied. Content description typically considers two types of features: features about the content but that cannot be extracted directly from the content, such as titles and names, and features conveying information that is present in the content, such as colors, melodies, or events. The second type of features may be low-level or high-level depending on their abstraction or semantic level [21]. Both low-level and high-level (semantic) features may be useful to decide on the best customizations to be performed, e.g. content segment semantics is useful to create a summary based on related user preferences while motion activity may be important for the filtering of violent action segments.

MPEG-7 addresses the multimedia content description problem at different levels: it offers a wide range of description tools that consider both low-level features such as color and pitch as well as high-level features such as the name of the characters in a scene or the title of a movie. MPEG-7 provides a set of description tools intended to characterize the audiovisual content in terms of the type of features listed above. The standard separates the descriptions from the content but provides linking mechanisms between the content and the descriptions. Also more than one description may exist for the same content depending on the needs the descriptions intend to address. The types of description tools specified by the MPEG-7 standard are [17]:

- *Descriptors (D)*: Represent a feature, and define the syntax and semantics of the feature representation.
- *Description Schemes (DS)*: Specify the structure and semantics of the relationships between their components, which may be both Descriptors, and Description Schemes.
- *Description Definition Language (DDL)*: Allows the creation of new Description Schemes, as well as the extension of existing Description Schemes.
- *Systems Tools*: Support the multiplexing of descriptions, synchronization of descriptions with the associated content, binary representation for efficient storage and transmission, management and protection of intellectual property, etc.

MPEG has invested a great effort in the standardization of MPEG-7 description tools for UMA applications [33]. This makes MPEG-7 the most powerful content description solution available for UMA environments and thus has been chosen for describing content in the context of the UMA system later described in this paper. The MPEG-7 UMA related description tools are grouped into three categories:

- *Media description tools*: Describe the media; typical features include the storage support, the

coding format, coding parameters, and the identification of the media. One of the most powerful tools regards transcoding hints, which improves the quality and reduces the complexity of transcoding applications (e.g. regions of interest, motion vectors).

- *Content structure and semantic description tools*: Describe the audiovisual content from the viewpoint of its structure and semantics; the content is organized in segments that represent its spatial, temporal or spatial–temporal structure.
- *Content navigation and access*: Facilitate the browsing and retrieval of audiovisual content and the management of different versions of the same content, notably summaries, views and variations. With MPEG-7 summarization tools, multiple summaries of the same content can be specified at different levels of detail, without the need for generating or storing multiple variations of the content.

These description tools are built upon the MPEG-7 Multimedia Description Schemes basic elements (schema tools, basic data types, links and media localization) which provide the foundation for the development of MPEG-7 description tools [17].

To create MPEG-7 descriptions, some amount of analysis is typically required to extract the features values. Fig. 2 illustrates a rather generic diagram of the multimedia features extraction process, where knowledge is learned from previous analysis, to improve future feature extraction and classification (this analysis ranges from low-level to high-level). There are many papers in the literature describing the extraction of both low-level features and high-level features. Wang et al. [36], present an overview of multimedia content low-level analysis techniques: while the extraction of low-level characteristics typically poses no significant problem but the choice of what to extract to achieve the desired goals, the same is not true when semantic characteristics have to be extracted by means of automatic analysis. To create semantic descriptions, semi-automatic analysis tools allowing humans to complement and/or train the analysis algorithms may be essential. Snoek and Worring present in [29] a review of multimodal indexing techniques for high-level information extraction.

Several developments have been presented in terms of high-level analysis, the most notable using machine learning/pattern recognition algorithms to identify objects, and detect concepts and conceptual relations from multimedia content. Vasconcelos et al. [34] and Naphade et al. [24] used Bayesian networks to characterize multimedia semantic features, e.g. indoor/outdoor, forest, sky, water, explosions, rocket launches; similar work has been done by Adams et al. [1] where features from audio, video and text modalities are individually classified in an earlier stage, to be later combined for the detection of semantic concepts; Benitez et al. [3] proposed a way to discover and measure statistical relationships among concepts, from images and corresponding text annotations; finally, Tansley et al. [32] proposed a four layer semantic representation framework, where each layer encodes information at an increasingly symbolic level.

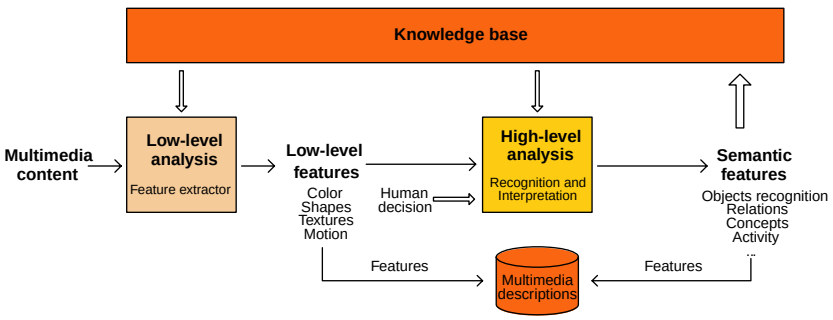

Fig. 2. Low-level and high-level features extraction.

All these works achieve valuable results by employing learning-based approaches to extract high-level information from multimedia content. Semantic content analysis is nowadays a very active research field, with promising advances in upcoming years—the reader is referred to [36,29,28], where the content analysis problem is further discussed.

2.2. Multimedia usage environment description

The previous section presented MPEG-7 as the standard solution to describe the multimedia content to be accessed. The other side of the UMA problem, the usage environment, may also be very heterogeneous due to the different terminals, networks and so forth that may be present. With a standard solution providing information on the usage environment key dimensions, it would be much easier for an application to customize its content and services to the usage environment conditions. Moreover it would improve applications portability since an application could be developed so that it checks the (standard) conditions in which it would be running and behave accordingly.

Part 7 of the MPEG-21 standard—called Digital Item Adaptation (DIA) [19]—has been created mainly to address the usage environment description problem. A diversity of dimensions characterizing the usage environment may enter in the equation that rules the content adaptation/delivery strategy. MPEG-21 DIA considers four major dimensions for usage environment characterization, which have been proposed to MPEG by the authors of this paper [15]:

- *Terminal characteristics*: The most commonly used software with which a user accesses multimedia content is a Web browser and its plug-ins. The browser is dependent on the hardware and software characteristics on the top of which it is running: multimedia decoding software, display resolution, display size, number of colors, audio capabilities, input capabilities (e.g. keyboard type), etc.
- *Network characteristics*: The access network can be the cause of very annoying effects for the user: delay, bandwidth shortage, channel errors, etc. The access network should be described as completely as possible to prevent as much as possible delays or pauses in the content rendering.
- *Natural environment characteristics*: This usage environment dimension includes the natural features regarding the surrounding usage environment that may influence the content adaptation: location, illumination, altitude, temperature, etc.
- *User preferences*: The last element in the content chain: the human user. This dimension holds information regarding his/her preferences, such as genre, and advertisement tastes, but also about disabilities. More general information on preferences can also be used such as food and accommodation preferences for advertising or retrieval.

In a broader sense, the term "usage environment" concerns all user related information that can be described and can be used also for other purposes than multimedia content customization. For example, credit card numbers are part of the user environment but this information is not that important from the point of view of content customization (although it is essential for billing purposes).

2.3. The MPEG wrapping

Starting with the more traditional coding tools such as MPEG-1, and MPEG-2, and the recent scalable video coding tools such as MPEG-4 Fine-Grain-Scalability (FGS), and passing through MPEG-7 content description, MPEG standardization culminates with the MPEG-21 multimedia framework which offers a wrapper to allow all the pieces in a multimedia customization chain to integrate and interact with each other. Fig. 3 depicts a possible configuration of a multimedia customization chain using all MPEG standards. The Digital Item (DI) and the DIA Usage Environment Description are concrete representations of the two sides of this chain.

At the server side, there is the DI with its resources (the content variations) and the corresponding

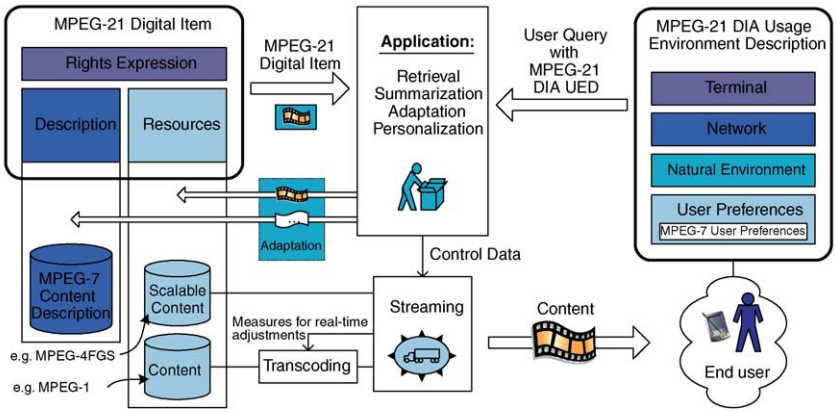

Fig. 3. MPEG wrapping for multimedia customization (in the form of retrieval, summarization, adaptation, or personalization).

descriptions. The content resources may exist in several resolutions and formats; for example, the same Digital Item may have a non-scalable resource variation (e.g. MPEG-1) and a scalable resource variation (e.g. MPEG-4 FGS).

At the end user side, the MPEG-21 DIA UED describes the environment (terminal, network, natural environment, and user preferences) where the content is to be consumed. When the user performs a query or defines a choice, his request is accompanied by the DIA UED, thus enabling a customizing application to explore this information to create the right content variation to provide the best possible user experience.

Finally, the application at the center of Fig. 3 is responsible for matching the user query (and the associated DIA UED) and the Digital Item, either by selecting the most adequate available variation, or by performing some adaptation. When processing a user query, the customizing application creates an adapted variation of the Digital Item to be sent to the user—the new variation and its corresponding description may also be added to the DI resources available at the server.

The user query response may be delivered through a network, eventually using a real-time connection. In this case, the streaming module will stream the scalable or non-scalable content to the user; in the case real-time transcoding is been performed, it may happen that real time adjustments to the transcoding process are implemented using measures which characterize, for example, the network fluctuations.

3. Multimedia customization processing

Several content customization techniques may have to be combined to achieve the optimal result in terms of final user experience. Fig. 4 presents a possible categorization and a non-exhaustive list of adaptation operations that a content customization engine may perform to the basic media types: text, image, audio, speech, video and synthetic content. The content customization operations are divided in two major categories:

- *Selection*: Supposing that several variations of the same content or even several alternative content pieces addressing different kinds of user constraints are available, content selection corresponds to the identification of the most adequate content asset from those available to be sent to the user. The selected variation may be already adequate enough or may need further adaptation as explained below. Existing content variations may include the same or different data types (e.g. video replaced by an image or text converted to speech using a cross-modal transformation). Content selection may involve sending different information to different users not only based on their technical

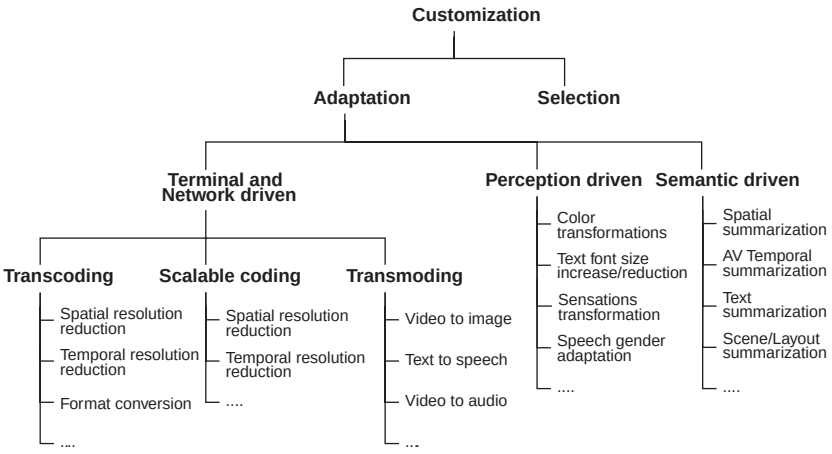

Fig. 4. Organization of multimedia content customization solutions.

capabilities, e.g. bit-rate and video decoder, but also on their location or preferences: for example, a different advertising banner depending on their local temperature for a higher impact (e.g. ice cream shop versus tea house). The several content variations may be organized as an MPEG-21 Digital Item, categorized as choices according to some description features, e.g., bit-rate or genre.

- *Adaptation*: Involves the transformation of the content asset above selected (one of the resources in the context of an MPEG-21 Digital Item) according to some criteria, if the available variation is not already adequate enough. This process typically requires a lot more of computational effort compared with the simple customization by selection, since for some content transformations heavy signal processing algorithms may have to be performed, many times with a low delay. It is here proposed to cluster the various customization by adaptation solutions by using three major classes: terminal and network driven, perception driven, and semantic driven.

It must be pointed out that the adopted adaptation solution may be a combination of several adaptation techniques, e.g. a summary with the goals in a football match at a lower bit-rate which combines semantic abstraction with transcoding.

3.1. Terminal and network driven adaptation

The consumption device and the network capabilities may pose a serious barrier to the content distribution and consumption. This category of adaptations aims to reduce the content consumption requirements by adjusting the source content to the device and network capabilities. Major examples of terminal and network driven adaptation solutions are transcoding, transmoding and scalable coding.

This type of adaptation may be achieved through signal processing operations (e.g. in time, space or frequency) applied in the compressed domain (e.g. DCT domain) or, at least, decoding as less as possible the bitstream to decrease the complexity of the process—typically called transcoding. When the terminal does not have the capability to consume certain media types, transmoding or modality conversion may be the solution; for example, it may imply the conversion of video to text or video to images to match the terminal decoding capabilities. Scalable coding techniques organize the content bitstream into consumption layers, which are truncated depending on the terminal/network resources, thus

avoiding any expensive real-time signal processing operations.

3.1.1. Transcoding

Transcoding is intended to decrease the required content resources and thus matching the available network/terminal consumption capabilities, keeping the same content modality. Examples are format conversion, temporal/spatial resolution reduction, higher frequency DCT coefficients removal, or quantization step reduction. The relevant signal-processing algorithms will require different resources depending on how the transcoding technique is implemented:

- *Uncompressed domain adaptation*: This type of adaptation techniques typically requires large resources in terms of memory and CPU (assuming the content available is coded). Even though this approach entails a lower implementation cost, it becomes quite inefficient since the content must be passed to the uncompressed domain to be adapted and then recompressed. Besides the high computational costs (notably for real-time implementations), this solution also implies a quality reduction, e.g. due to the accumulation of the coding (quantization) errors. Format conversion using this technique entails a certain loss of quality, due to the encode–decode–encode processing chain.
- *Compressed domain adaptation*: Compressed domain adaptation techniques offer faster processing, consume fewer resources, and in principle provide better quality; however, this type of processing may be more complex in terms of implementation since adequate transcoding algorithms have to be developed [35,11,7]. Format conversions in the compressed domain can also be more easily achieved when using similar coding formats, e.g. H.263 to MPEG-4 Simple profile.

Transcoding adaptation requires signal-processing techniques that may consume large resources, e.g. in terms of computational power and memory. Thus it may become quite expensive to achieve a large-scale adaptation deployment when many on-the-fly adaptations are required.

3.1.2. Transmoding

When the usage environment conditions do not allow to consume the content with its original media types, a modality transformation can be used to fit the consumption capabilities in terms of media types able to be consumed; for example, to transform text into audio, to extract key-frames from a video, or to use only the audio part of a movie. Fig. 5 shows a case where only one key-frame would be used for each segment, or the speech track would be converted into text. While the conversion of video to images may be rather simple, e.g., using a key-frame selection algorithm, other conversions such as voice to text may be much more complex to provide acceptable experiences.

3.1.3. Scalable coding

Scalable coding intrinsically assumes that the content shall be distributed through heterogeneous networks and consumed by heterogeneous terminals. Thus, the content format already provides several consumption layers, more or less depending on the granularity adopted, to satisfy different consumption resources, decoupling the coding and adaptation processes. Content customization can

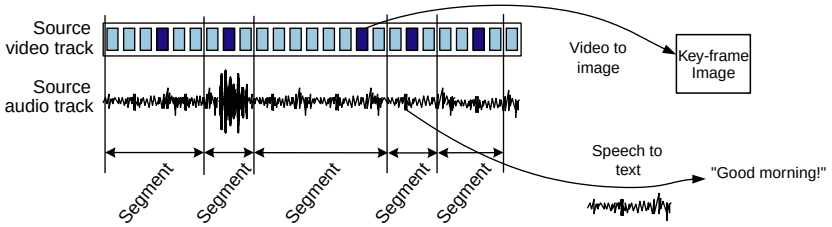

Fig. 5. Modality conversion: video to images and speech to text.

use two major techniques to process scalable content:

- *Scalable content truncation*: A scalable coding technique compresses the data in question into multiple consumption layers with more or less granularity. Typically, one of the compressed layers is the base layer, which can be independently decoded and provide a basic quality. The other layers are enhancement layers, which can only be decoded together with the base layer, therefore successively providing better quality. The complete bitstream (i.e., the combination of all the layers) provides the highest quality. Therefore scalable content does not require much adaptation processing when the access conditions vary, it is just a question of truncating the total bitstream depending on the constraints to be imposed, e.g. a bit-rate limitation. In Fig. 6(a), the various blue layers on the right can be dynamically used according to the quality that the user resources allow for. The types of scalability (e.g., spatial, temporal, SNR), its granularity as well as the efficiency of scalable coding techniques are today a hot research topic [37,2]. Considering the importance of the scalability concept in the context of the MPEG-21 framework, MPEG is currently involved in the development of a new scalable video coding standard which should provide fine grain scalability at almost no cost in terms of coding efficiency regarding the most advanced non-scalable solutions, e.g. H.264/AVC.
- *Scalable content with a bitstream syntax description*: This type of technique is based on the use of an XML based description of the (scalable) content bitstream syntax. The bitstream XML description may follow a generic bitstream syntax description language [26], which supplies structures that can describe the bitstream syntax of the content to be processed, see Fig. 6(b). Whenever an adaptation has to be performed, the XML bitstream description is adapted instead of adapting directly the bitstream. Next, the transformed XML description is used to generate the adapted version of the bitstream by simply parsing the bitstream. If the bitstream syntax description language is generic enough, the adaptation engine can process the content without knowing its coding format. The MPEG-21 DIA specification [19] already defines the so-called Bitstream Syntax Description Language (BSDL) and the generic Bitstream Syntax Schema (gBS Schema). BSDL is a normative language based on XML Schema making possible to design specific Bitstream Syntax Schemas (BSs) describing the syntax of particular scalable media resource formats. While BSDL makes it possible to design specific Bitstream Syntax Schemas describing the

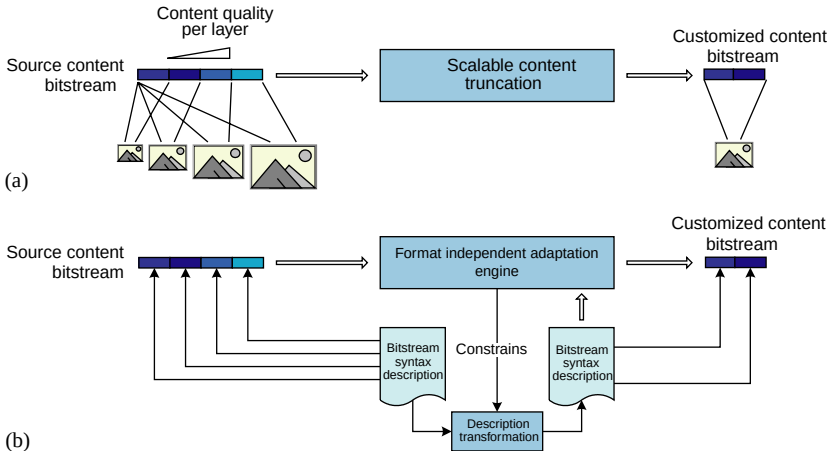

Fig. 6. Scalable content adaptation: (a) scalable content truncation; (b) scalable content adaptation based on its binary syntax description.

syntax of particular media resource formats, gBS Schema enables resource format independent bitstream syntax descriptions (gBSDs) to be constructed. As a consequence, it becomes possible for resource format agnostic adaptation engines to transform bitstreams and associated syntax descriptions.

Nowadays, video distribution in heterogeneous networks mostly uses multiple stream transmission (simulcast) or scalable content techniques, either temporal or spatial (e.g., the PacketVideo streaming server [25]), since real-time processing is really minimal in these cases.

3.2. Perception driven adaptation

Perception driven adaptation regards transformations according to some user preferred sensation, or assisting a user with a certain perception limitation or condition. The user perception of the content will be different from its original version, for example, to address the needs of users with visual handicaps, e.g. color blind deficiencies, or specific preferences, e.g. in terms of visual temperature sensations. This class of customization operations considers all types of adaptations related to some human perceptual preference, characteristic or handicap:

- *Sensation based*: Audio and visual sensations conveyed by the content can be modified according to user preferences. For example: consider a color temperature adaptation providing different sensations in terms of warm or cold color temperatures [16] or a spoken track for which the gender or emotion can be modified like in story telling applications.
- *Handicap assistance*: Certain user handicaps may be compensated or minimized in terms of content consumption by content transformations such as image color transformations, text color changes, and text size increasing. For example, a color blind person may need a transformation of the colors in the content into different levels of luminance.
- *Natural environment*: The natural (physical) environment may limit the user perception capabilities and thus the user may need specific adaptations to maximize his/her access conditions. For instance, if a user is too far from the screen, then the text font size may have to be increased; if a user is driving, his/her attention (and his/her eyes) is focused on the road, and thus an adaptation to produce audio content only may have to be provided.

Most the human perception related adaptations must be performed in the compressed or uncompressed domain using adequate signal-processing techniques. As such, this type of adaptations is rather difficult to implement in large scale (in a server) in real-time due to the significant amount of computational resources that would be needed.

3.3. Semantic driven adaptation

Semantic adaptations involve the temporal and/or spatial reduction of a certain multimedia asset, e.g. temporal duration or number of regions of interest, implying a certain degree of semantic knowledge. For example, the adaptation may create a smaller duration variation of the structured content, usually called summary, by selecting the temporal (and may be also spatial) segments that are more relevant according to some criteria (e.g. user preferences).

To accomplish such customizations, the (spatial and temporal) content structure and the content semantics is essential data (available in a description) to perform any type of semantic adaptation. Semantic customization may be achieved in several ways, notably:

- *Temporal Summarization*: The content is reduced to a smaller duration variation of the source by selecting only some of the content segments according to some (semantic) criteria [24]; sometimes also explicit duration constraints are imposed, e.g. a summary with the most violent moments of a movie but shorter than 2 min. Considering segmented content as shown in Fig. 7, where the temporal segmentation may be described using MPEG-7 tools, the customization engine will select the relevant segments, matching the content description

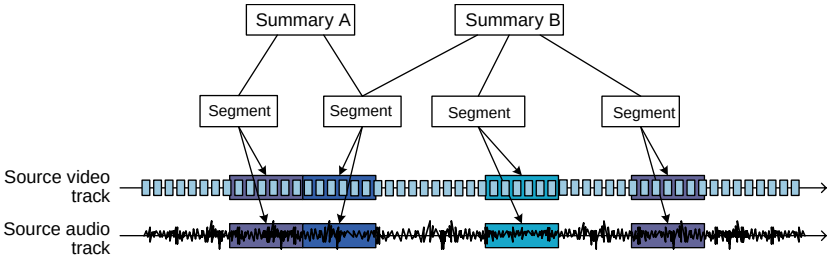

Fig. 7. Summarization of temporally segmented content.

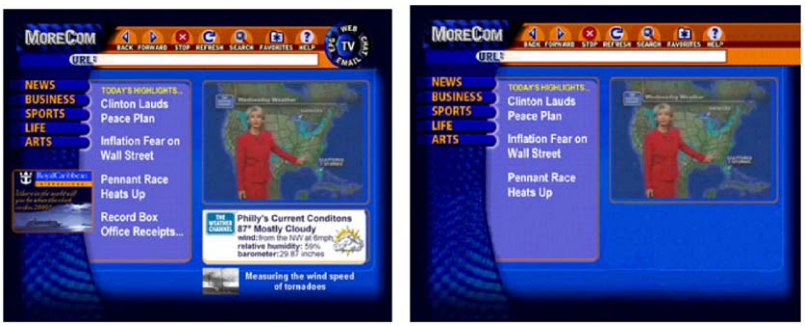

Fig. 8. Spatial/scene summarization example.

with the user preferences, to create the most adequate audiovisual summary of the source material.

- *Spatial/scene summarization*: Semantic adaptation may also be applied to a spatially segmented content (a scene or layout) [8]. A scene is composed by several individual (spatial) content segments, each of them with a semantic description [23]. The summarization of a scene may involve the evaluation of each individual spatial segment in terms of semantic relevance for the concerned user, and the consequent customization by removing some of the spatial segments or by attributing them a lower quality or resolution if spatial summarization is combined with some kind of transcoding. Fig. 8 illustrates a scene (spatial) summarization example where several spatial segments were removed from the scene, thus summarizing it to a lower complexity scene more adjusted to the user interests.

Both Figs. 7 and 8 examples illustrates summarization by segment selection, but some adaptation of each segment individually may also be considered; for example, it is rather common to combine summarization with adaptation techniques such as segment spatial resolution reduction, segment quality reduction, or segment temporal resolution reduction.

The creation of summaries depends very much on the richness and precision of the MPEG-7 descriptions in terms of content structure and content semantics. This is not such a major problem for other types of adaptations such as transcoding because they typically depend on features that are very objective and almost always available in the content description such as the coding format or the video spatial resolution.

4. MPEG-based Universal Multimedia Access system

In principle, content customization can be performed in three different places along the multimedia chain: (1) at the content server, (2) at a proxy server, and (3) at the end user terminal.

The customization of multimedia data at the terminal is typically not feasible due to the limited (terminal and network) resources normally available on most terminals, notably mobile terminals. However there are cases where some degree of customization is performed at the end user terminal, for example, due to the associated advantages in terms of user preferences privacy.

Multimedia customization systems proliferate in the literature in the form of retrieval, summarization, adaptation or personalization systems; they all use one or a combination of the processing techniques described in the previous section. Also, there are systems which use previously extracted features (descriptions or annotations), while others extract them on real-time. Several systems for UMA have already been described in the literature: [1,10,5,6,8,23,22,18,13,12] provide examples of such systems, many of them also using standard tools for some of their modules.

The systems presented in [1,10,5,6] target semantic video retrieval/temporal summarization applications. For that purpose, Adams et al. [1] view the multimedia semantic analysis as a pattern recognition/machine learning problem, where the semantic concept to be detected is learned by a probabilistic network which detects the concept in certain shots. Features from individual modalities (audio, video and text) are individually classified in an earlier stage, to be later combined for the detection of semantic concepts defined in a restricted lexicon. The system uses MPEG-7 descriptions created by tools that the authors made publicly available. A query by keyword retrieval application is used to access the video database. Joyce et al. [10] present an architecture for content retrieval and navigation using a four layers data representation, where trainable agents create links between the layers, linking the content to semantic concepts. A query allows finding similar objects, the concepts related to that object, and the objects matching a certain concept. Chang et al. [5] describe a real-time semantic summarization system where the produced bit-rate is content-based and dynamically varying according to the event importance. The system discriminates the video segments by streaming video for the important segments, and just audio and still pictures for the less important segments. Finally, the approach adopted by Ekin et al. [6] is based on a deterministic approach where a semantic concept is detected when a set of predetermined cinematic features meet together. The detected concepts include goals (finding a sequence of shots with certain characteristics), the referee (based on the referee corresponding size invariant shape on a close-up shot) and the penalty-box (finding some box lines). The user can access three types of summaries: slow-motion shots, goals, and an extension of the two with object-based features (using the penalty-box and referee detection results).

The spatial/scene summarization or content structure customization is the objective of the systems described in [8,23,22,18]. Hwang et al. [8] present a content structure/scene aware system which employs heuristics to adapt the content to mobile terminals (terminal display driven adaptations). This system does not use previously extracted or annotated descriptions; user preferences are also ignored. The authors propose two new heuristics: the ''generalized outlining transform'' to detect repeated scene patterns in a multimedia document and the ''selective elision'' transform to selectively eliminate parts of the repeated scene patterns. In practice, these detected and eliminated scene patterns correspond to menus, tables, lists, etc., in a Web page. Nagao et al. [23,22] present a system based on an adaptation server and a description server which stores the descriptions by URL. The authors have three types of descriptions: linguistic (to describe text, e.g. noun, noun phrase, verb, adnoun or adverb), annotations (to comment non-textual elements) and multimedia (to describe video content). HTML gives the scene structure to the document, where each element (text, voice, image, and video) is identified and described by an external XML description. Mohan et al. [18] present a system with a single content server which concentrates all functions (adaptation and description server). This work proposes a content organization scheme, the InfoPyramid, which structures content variations in terms of its modality and quality. MPEG-7 provides a similar description tool to manage content variations. Both Nagao and Mohan works perform text

summarization, image transcoding, and video summarization (based on manual annotations) and perform scene summarization, by converting tables to lists, evaluating the importance of each content segment in the overall document, and consequently removing or relatively adapting each content segment.

Some multimedia customization systems [13,12], dedicate a strong emphasis to the usage environment description. Ma et al. [13] present a system where there are no descriptions available and consequently the server (origin or proxy) must extract both content and usage environment characteristics on-the-fly (e.g. bandwidth measurement, terminal characteristics discovery through the HTTP query). The system is composed by an adaptation module, an adaptation decision engine (also responsible for the content analysis), and a user/client/network characteristics discovery module, which uses several heuristics, e.g., looking for patterns in the HTTP headers, and monitoring user behavior (e.g. if the time between page requests is too small, it may be an indication of a link bottleneck, which can be compensated by adapting the content to reduce the data size). Lum et al. [12] propose a content adaptation system, which is quite centered on the user context. A proxy-based architecture similar to the one in [13] is used to implement a decision engine. The authors devise a method to quantitatively measure the QoS of a content piece as an n-dimensional value. In the preprocessing stage, the decision engine creates a search space consisting of all the possible adaptation decisions (each decision corresponds to a node in the search space). Then each search space node is scored according to the user's preferred QoS. During real-time operation, the optimal node is located by a negotiation algorithm that examines the search space nodes based on the user terminal capabilities, network parameters, and content metadata.

Inspired in the systems presented above, Fig. 9 shows a rather generic architecture of a UMA system. This system, implemented in the context of this paper, deploys the customization engine at the content server and at a proxy server; also it includes a multimedia customization server and other elements more associated to the access to content. The overall system follows the MPEG-21 vision, where multiple Users interact with the purpose to provide the end user the best possible multimedia experience. Each system element performs precise functions; no function overlap is present, meaning that other architectures can be derived from this one by accumulating functions into a single element or by further splitting any of the modules.

The function of each module in the generic UMA system shown in Fig. 9 is:

- *Content server*: Acts as the content source.
- *MPEG-7 description tool*: Analyzes the multimedia content available at the content server and generates MPEG-7 descriptions that will be stored into an MPEG-7 description server (local or remote to this tool).
- *MPEG-7 description server*: Stores the MPEG-7 descriptions received from the MPEG-7 description tool; in this database, at least one MPEG-7 description exists for each piece of content. The MPEG-7 description tool consists in a content analyzer that scans the multimedia

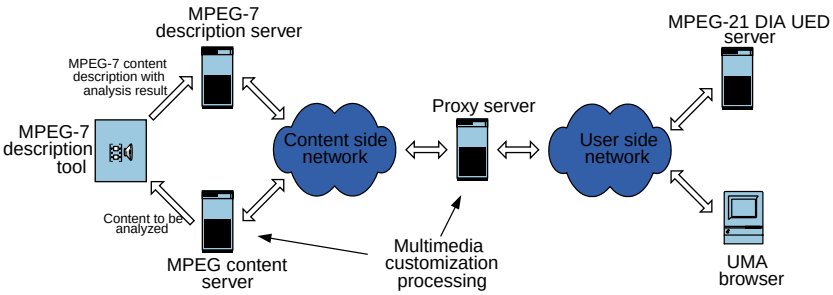

Fig. 9. UMA system simplified architecture.

content to build MPEG-7 descriptions that will be used later by the customization server. The generated descriptions can be saved locally or posted (through an HTTP command POST) to an MPEG-7 description server. The MPEG-7 description tool implements the media description tools presented in Section 2.1.

- *UMA browser*: Browser used to access content and allowing the user to create and manage his/her usage environment description. The implemented UMA browser is a Web browser test application, which allows creating, changing and managing several usage environment descriptions (all of its four dimensions), thus simulating several content consumption conditions. The UMA browser sends the MPEG-21 DIA usage environment description to the UED server, which keeps a database of UEDs to be used by the content customization engine when it needs information regarding a certain usage environment.
- *UED server*: Stores the MPEG-21 DIA usage environment description (UED) received from the UMA browser; provides the Customization server with the UED for the usage environment in question.
- *Customization server*: Includes the application implementing the content customization engine (and the needed interfaces) required to provide the best experience to the user for the content he/she asked.

In the application developed, the content server, the MPEG-7 description server and the UED server are Apache Web servers, which implement the HTTP POST command allowing other applications to store descriptions in the server. The MPEG-7 description tool, the customization server, and the UMA browser were implemented specifically for this UMA system.

The customization server implements the content customization engine plus the required modules to interface with a network. Several modules compose the implemented customization server as can be seen in the architecture presented in Fig. 10. The major modules in the customization server are

- *Network interface manager*: Responsible for the network communications between the customization engine, and the other UMA system modules. It is also responsible for retrieving the descriptions, both for content and usage environment. When the processing of a request is complete, this module updates the caching tables with the customized new content and the corresponding descriptions. The caching tables are useful to enable the repurposing of previous adaptations in future similar (or almost similar) requests. It is this module that retrieves the content, either from the local disk (in a server configuration) or from an URL (in a proxy configuration).
- *User request processor*: This module translates into the MPEG-21 DIA usage environment descriptions other description formats such as WAP UAProf (Wireless Application Protocol—User Agent Protocol) descriptions, and the PocketPC description (PocketPC uses HTTP headers to carry terminal related information, e.g. display resolution, decoding capabilities).
- *Multimedia content description analyzer*: Parses and validates the MPEG-7 descriptions into the platform internal data memory structures.
- *User environment description analyzer*: Parses and validates the MPEG-21 DIA UED descriptions into the platform internal data memory structures.
- *Customization action decision*: Matches the content description with the usage environment description and decides which customization solution must be adopted for the content in question, taking into account the available content processing operations.
- *Content customization*: Performs the content customization operations decided by the previous module. It has several Content customization modules to be used depending on the type of adaptation to be performed (see Section 3).

The *Content action decision* and the *Content customization* modules are interdependent since the first one must be aware of the adaptation methods available in the second in order to know which decisions can be executed. The content customization modules implemented perform

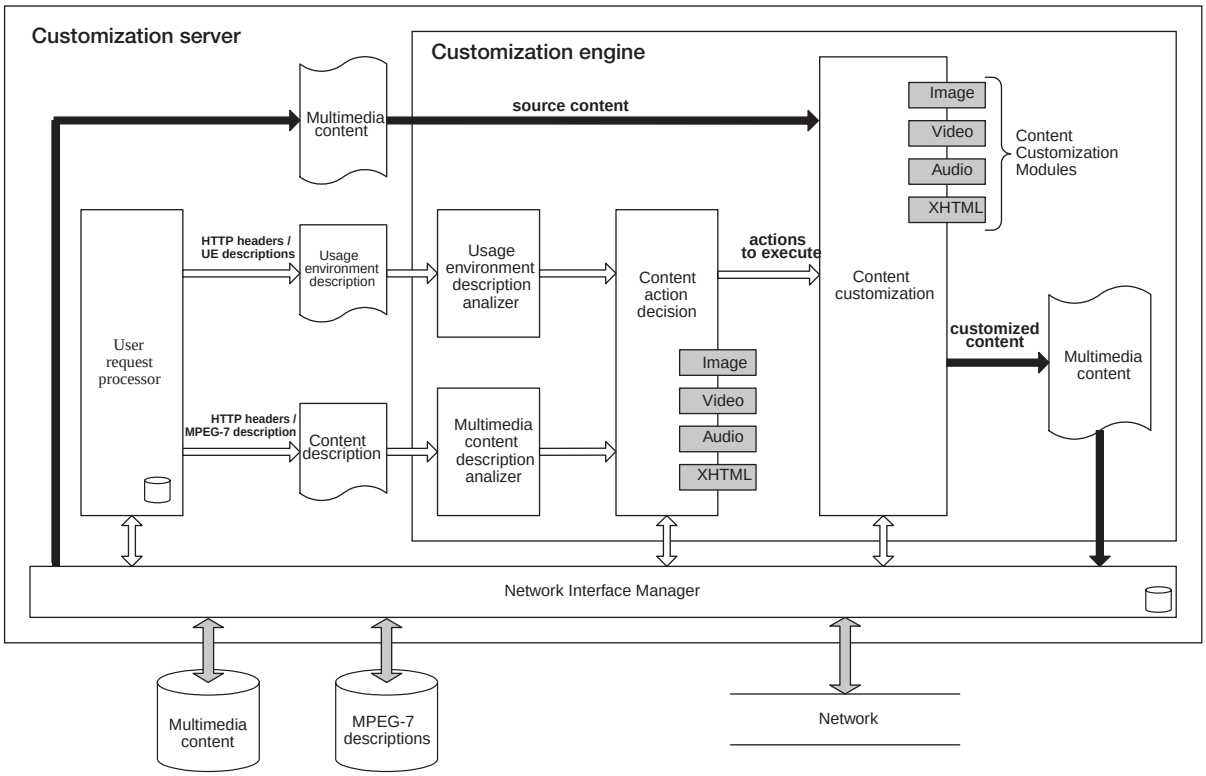

Fig. 10. Customization server architecture.

transcoding as well as perceptual transformations for images and video content; these adaptation tools are not presented here in detail because they just reproduce state-of-the-art technology. However, the customization server is prepared to receive other customization modules for other data types or with other adaptation capabilities. The perceptual transformation implemented regards a color temperature transformation in the uncompressed domain [38]; this customization capability is based on the corresponding MPEG-7 low-level visual descriptor, which expresses a human perceptual preference in terms of color temperature [16]. This is one of the first MPEG-based adaptation systems using MPEG-7 low level descriptions since most available systems are only based on textual descriptions, e.g. the format for the transcoding, or the preferred genre for the summarization.

5. Experiments and conclusions

To test the MPEG-based multimedia customization system, the server can be configured as a proxy server or as a content server. The experiments here reported consider transcoding and color temperature perceptual transformation of images and videos.

5.1. Test-bed architecture

Fig. 11 presents the test-bed used, in a content server configuration, operated with Universal Mobile Telecommunications System (UMTS) and General Packet Radio Service (GPRS) terminals. The two used configurations present the following characteristics:

- *Content server configuration*: Has the advantage to ease the dynamic (on-the-fly) content

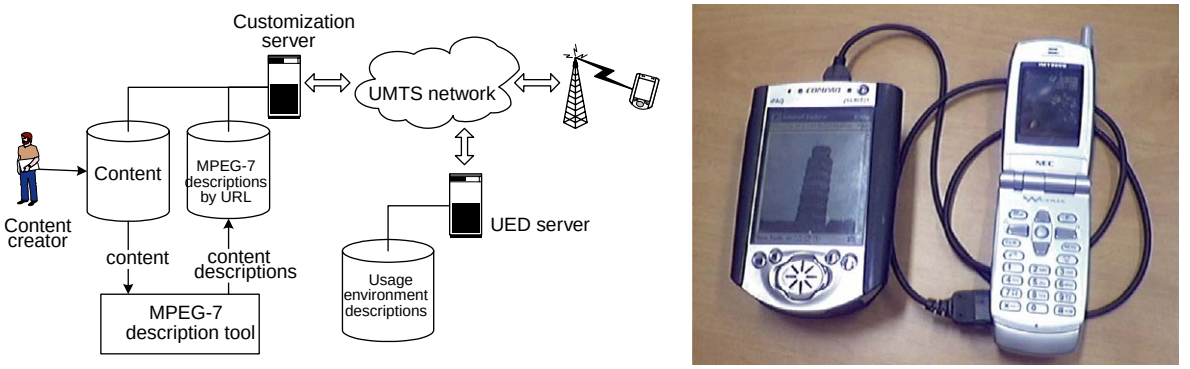

Fig. 11. Experimental test-bed (left) and PDA connected to the customization server through a UMTS terminal (right).

adaptation because content is locally available. If the author knows that the content will be adapted, and he/she has access to the content server, he/she can have control on the adaptation results. Consequently, the author can tune the adaptation process either by providing transcoding hints in the content description (MPEG-7 allows this), or by providing appropriate variations. It is also necessary to consider content rights management and its implications: since the content has to be changed, the right owner must allow for it and may be even approve the "re-authored" content.

- *Proxy-based configuration*: Since the content is not locally available, this configuration poses more technical difficulties. From the content server to the usage environment, several content customization proxies, with distinct behaviors, may exist in the network. Therefore the tuning that an author could perform with the previous configuration may become impossible to perform here, since the author may have little control over the proxies. Content rights management may become even more critical with a proxy configuration.

5.2. Test material and usage environment scenarios

In order to evaluate the influence of large content (high spatial resolution and data size) in the content customization processing, the selected image content ranged from 320×240 with 8 bit/pixel to 1600×1200 with 24 bit/pixel and the video content ranged from 176×144 pix-pixels/frame at 64 kbit/s to 352×288 pixels/frame at 256 kbit/s. On the user side, several UEDs were used to simulate (using the UMA browser) WAP terminals, and Pocket PCs which should represent terminals and connections with different characteristics.

5.3. Lessons from the experiments

The terminal and network driven adaptation tests performed with the Pocket PC, WAP terminal and UMA browser showed that the same content could be delivered after adaptation to different terminals, thus reducing the maintenance and storage costs of having one content variation for each terminal type. Fig. 12 illustrates the same high-resolution image delivered to a PocketPC and a GPRS WAP terminal. In the PocketPC case, the source image is too big to fit its display (see Fig. 12(a)), resulting in a poor user experience and a large data transfer time. The customized content clearly increased the user experience (see Fig. 12(b)), and only the required data was transferred. For the WAP terminal in Fig. 12(c), the source content could not even be consumed by the terminal since it is only able to consume black and white images, thus the customization was the only way to access to that piece of content. Similar processing is performed with video: screen size and network

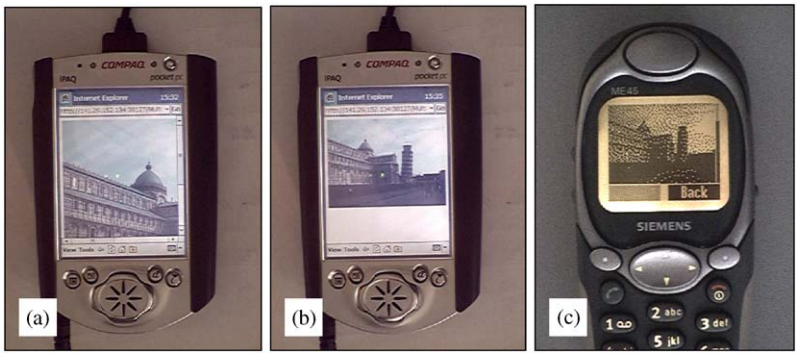

Fig. 12. Examples: (a) image not customized for a Pocket PC display resolution (only part of the image is seen); (b) image customized for a Pocket PC display resolution; (c) image customized for a black & white WAP terminal.

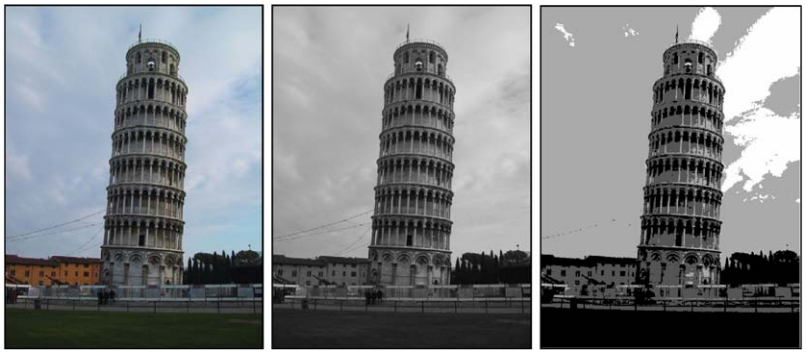

Fig. 13. Content variations, from left to right: RGB color space, 24 bit per pixel; gray color space, 8 bit per pixel; gray color space, 2 bit per pixel.

bandwidth are the input parameters for content transcoding.

When a big gap exists between a high resolution source content and the usage environment consumption capabilities, the CPU was taken up for a long period for content processing. In these situations, customization efforts may be reduced by making available strategic variations with characteristics closer to the most popular terminal/network capabilities. Also, the reuse of previous adaptations inserted in the content variations list allows the customization server to take more benefit of previous adaptation efforts; this policy may decrease the CPU occupation at the cost of a larger storage space. For example, a customization server may have several content variations at its disposal (e.g. Fig. 13 illustrates a source image and its variations in terms of bit depth) and, depending on the terminal, the best variation is selected to perform the remaining adaptation (if needed).

As referred in Section 3.2, customization can also be employed to offer different sensations to the user, notably in terms of hearing and sight. Fig. 14 presents some color temperature adaptations, which provide images able to stimulate different visual sensations, e.g. a warmer or colder sensation. The user specifies his preferred color temperature (e.g. by selecting an example image), which will be used in later adaptations. Other perception-based adaptations can provide different sensations, or even help handicap people to better access the content.

Looking at the relative costs of a bit, the (low) bit storage cost, the (relatively low) bit processing cost and the (relatively high) bit distribution cost,

Fig. 14. Adaptation of an image to different color temperatures, from left to right: $T = 4135\,\mathrm{K}$, $T = 7924\,\mathrm{K}$, $11711\,\mathrm{K}$.

the above results assume a different shape. These relative costs show that the existence of several content variations as well as off-line content processing is not critical, since storage and computational power (notably off-line) are not that expensive. However, the distribution cost is relatively high, which means that any typical UMA system should explore intensively the storage and processing capabilities to compensate the distribution costs and on-the-fly content processing, thus just distributing the essential content and not any content that cannot be adequately consumed, ideally not generated in real-time.

6. Final remarks

This paper navigated through standards, state-of-the-art tools, algorithms and systems for multimedia customization, concluding with a rather generic system architecture fitting in the MPEG-21 multimedia framework. Now, under the MPEG-21 umbrella, a short discussion on the evolution of the multimedia customization most important parts, content coding tools, content description and usage environment description, will be made.

Content delivery in heterogeneous environments, notably through terminal and network driven customization, may take great benefit of scalable audiovisual coding techniques. There are today a variety of scalable coding standards available, notably JPEG and JPEG 2000 for images, and MPEG-4 scalable profiles for audio and video. MPEG-4 fine granularity scalability (FGS) solution for video is especially relevant due to the very fine adaptation capabilities it provides. The recently initiated MPEG-21 Part 13, called Scalable Video Coding, targeting the development of an advanced scalable video coding solution where unlike in the past solutions coding efficiency is not affected, confirms the importance that the industry and MPEG keeps giving to this functionality.

Content description shows a different situation, since the MPEG-7 standard is available and providing a vast variety of description tools and no significant changes are foreseen. The great technical challenge is now associated to the automatic low-level to high-level mapping; multimodal processing and semantic networks are also very promising topics. On other fronts, the standard is still missing licensing conditions and may be because of that the industry did not yet take this powerful and still very much unexplored standard with full heart.

And finally, the last link in the content chain: the usage environment and the human user. Technical consumption restrictions in the usage environment (network, hardware or software) may be overcome with content customization although the author's intended usability and subliminal message associated to his/her original creation may be different. It is also true that physical limitations such as bandwidth and processing power may become less relevant as times goes by. The usage environment and its description define multimedia customization limitations and boundaries—the richest they are, the better adaptation results can be achieved.

Usage environments also affect user's predisposition, patience, and most important the time for consuming multimedia. So, according to the user's surrounding environment, different usability/interactivity/information designing rules are used in the content/services creation. These

designing rules take into account that the user looks for different experiences in different places—the author creates content with a certain quality of experience (QoE) in mind, content to be consumed in a certain experiential environment [9]. With time, the elements through the content chain will give more relevance to human features (semantic and sensory) than before. The content chain will evolve to a higher level of abstraction: the author intended experience (provided through the content) and the experiential environment will be the two ends of the future multimedia chain.

In the content delivery chain, UMA will evolve to a new arena where human factors (expressed through a QoE) will be the most conditioning aspect. UMA will then evolve to Universal Multimedia Experiences (UME) [27], where customization has to create the best experience for each experiential environment. The final goal in multimedia customization will then be the maximum preservation of the same QoE, going beyond physical access and entering the world of human experiences.

References

[1] W.H. Adams, G. Iyengart, C.-Y. Lin, M.R. Naphade, C. Neti, H.J. Nock, J.R. Smith, Semantic indexing of multimedia content using visual, audio and text cues, EURASIP J. Appl. Signal Process. 2003(2) (February 2003) 170–185.

[2] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, J. Cornelis, Scalable wavelet video-coding with in-band prediction implementation and experimental results, Proceedings of the IEEE International Conference on Image Processing, Vol. 3, Rochester, New York, September 2002, pp. 729–732.

[3] A.B. Benitez, S.F. Chang, Multimedia knowledge integration, summarization and evaluation, Proceedings of the 2002 International Workshop on Multimedia Data Mining in conjunction with the International Conference on Knowledge Discovery & Data Mining, Edmonton, Alberta, Canada, July 2002.

[4] S.F. Chang, The holy grail of content-based media analysis, IEEE Multimedia Mag. 9(2) (April–June 2002) 6–10.

[5] S.-F. Chang, D. Zhong, R. Kumar, Real-time content-based adaptive streaming of sports videos, IEEE Workshop on Content-Based Access of Image/Video Libraries, Hawaii, USA, 2001.

[6] A. Ekin, A.M. Tekalp, R. Mehrotra, Automatic soccer video analysis and summarization, IEEE Trans. Image Process. 12(7) (July 2003) 796–807.

[7] K. Homayounfar, Rate adaptive speech coding for Universal Multimedia Access, IEEE Signal Process. Mag. 20(2) (March 2003) 30–39.

[8] Y. Hwang, J. Kim, E. Seo, Structure-aware Web transcoding for mobile devices, IEEE Internet Comput. 7(5) (September–October 2003) 14–21.

[9] R. Jain, Quality of experience, IEEE Multimedia 11(1) (January-March 2004) 95–96.

[10] D.W. Joyce, P.H. Lewis, R.H. Tansley, M.R. Dobie, W. Hall, Semiotics and agents for integrating and navigating through multimedia representations, Proc. Storage Retrieval Media Databases 3972 (January 2000) 120–131.

[11] Z. Lei, N.D. Georganas, A rate adaptation transcoding scheme for real-time video transmission over wireless channels, Signal Process.: Image Commun. 18(8) (September 2003) 641–658.

[12] W.Y. Lum, F.C.M. Lau, A context-aware decision engine for content adaptation, IEEE Pervasive Comput. 1(3) (July–September 2002) 41–49.

[13] W.Y. Ma, I. Bedner, G. Chang, A. Kuchinsky, H.J. Zhang, A framework for adaptive content delivery in heterogeneous network environments, Proceedings of the SPIE/ACM Conference on Multimedia Computing and Networking 2000, San Jose, USA, February 2000, pp. 86–100.

[14] J. Magalhães, Universal access to multimedia content based on the MPEG-7 standard, M.Sc. Thesis, Instituto Superior Técnico, Lisbon, Portugal, June 2002.

[15] J. Magalhães, F. Pereira, Enhancing user interaction in UMA applications, Doc. ISO/MPEG, M7312, MPEG Sydney Meeting, Australia, July 2001.

[16] J. Magalhães, F. Pereira, MPEG-7 based color temperature customization, ConfTele 2003, Aveiro, Portugal, June 2003, pp. 481–484.

[17] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG 7: Multimedia Content Description Language, Wiley, New York, 2002.

[18] R. Mohan, J. Smith, C.-S. Li, Adapting multimedia internet content for universal access, IEEE Trans. Multimedia 1(1) (March 1999) 104–114.

[19] MPEG MDS Group, Multimedia framework Part 7: Digital item adaptation, Final Draft International Standard, Doc. ISO/MPEG, N6168, MPEG Waikaloa Meeting, USA, December 2003.

[20] MPEG Requirements Group, MPEG-21 Multimedia framework, Part 1: Vision, technologies and strategy, Proposed Draft Technical Report, 2nd Edition, Doc. ISO/MPEG N6269, MPEG Waikaloa Meeting, USA, December 2003.

[21] F. Nack, M. Windhouwer, L. Hardman, E. Pauwels, M. Huijberts, The role of high–level and low–level features in style-based retrieval and generation of multimedia presentations, New Rev. Hypermedia Multimedia 7 (2001) 7–37.

[22] K. Nagao, Digital content annotation and transcoding, Artech House, 2003.
[23] K. Nagao, Y. Shirai, K. Squire, Semantic annotation and transcoding: making web content more accessible, IEEE Multimedia 8(22) (April–June 2001) 69–81.
[24] Naphade, M.R., T.S. Huang, A probabilistic framework for semantic video indexing filtering and retrieval, IEEE Trans. Multimedia 3(1) (March 2001) 141–151.
[25] PacketVideo streaming server, http://www.packetvideo.com.
[26] G. Panis, A. Hutter, J. Heuer, H. Hellwagner, H. Kosch, C. Timmerer, S. Devillers, M. Amielh, Binary multimedia resource adaptation using XML bitstream syntax description, Signal Processing: Image Communication, Special Issue Multimedia Adaptation 18(8) (September 2003) 721–747.
[27] F. Pereira, I. Burnett, Universal multimedia experiences for tomorrow, IEEE Signal Process. Mag. 20(2) (March 2003) 63–73.
[28] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta,R. Jain, Content-based image retrievals at the end of the early years, IEEE Trans. Pattern Anal. Machine Intell. 22(12) (December 2000) 1349–1380.
[29] C.G.M. Snoek, M. Worring, Multimodal video indexing: A review of the state-of-the-art, to appear in Multimedia Tools and Applications, 2004 (http://carol.wins.uva.nl/~worring/pub/papers/snoek-review-mmta.pdf).
[30] Special Issue on Universal Multimedia Access, IEEE Signal Processing Magazine 20(2) (March 2003).
[31] Special Issue on Multimedia Adaptation, Signal Processing: Image Communication 18(8) (September 2003).
[32] R. Tansley, C. Bird, W. Hall, P. Lewis, M. Weal, Automating the linking of content and concept, Proceedings of the ACM Multimedia, Los Angeles, USA, November 2000.
[33] P. van Beek, J.R. Smith, T. Ebrahimi, T. Suzuki, J. Askelof, Metadata-driven multimedia access, IEEE Signal Process. Mag. 20(2) (March 2003) 40–52.
[34] N. Vasconcelos, A. Lippman, A Bayesian framework for semantic content characterization, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, USA, June 1998, pp. 566–571.
[35] A. Vectro, C. Christopoulos, H. Sun, Video transcoding architectures and techniques: an overview, IEEE Signal Process. Mag. 20(2) (March 2003) 18–29.
[36] Y. Wang, Z. Liu, J.C. Huang, Multimedia content analysis using both audio and visual clues, IEEE Signal Process. Mag. 17(11) (November 2000) 12–36.
[37] D. Wu, Y.T. Hou, W. Zhu, Y.-Q. Zhang, J.M. Peha, Streaming video over the Internet: approaches and directions, IEEE Trans. Circuits Syst. Video Technol. 11(3) (March 2001) 1–20.
[38] G. Wyszecki, W.S. Stiles, Color science: concepts and methods, quantitative data and formulae, Wiley Series in Pure and Applied Optics, 1982.