Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 19 (2004) 685–699
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Automatic video summarization based on MPEG-7 descriptions

Pedro Miguel Fonseca*, Fernando Pereira

Instituto Superior Técnico, Instituto de Telecomunicações, Portugal

Received 11 June 2003; received in revised form 29 March 2004; accepted 8 April 2004

Abstract

The ever-growing amount of audiovisual content available has raised the need to develop systems allowing each of us to consume the information considered 'essential', adapted to our tastes, our preferences, our time and also to our capacities of receiving, consuming and storing that information. In other words, there is an increasing need to develop systems able to automatically summarize audiovisual information. This paper proposes a novel query-based summary creation mechanism using a relevance metric and a constraints schema implemented in the context of an automatic video summarization system based on MPEG-7 descriptions. In the context of the same system, this paper also proposes a human skin filter allowing to build summaries based on the presence or absence of human skin. This skin colour filter is based solely on the MPEG-7 Dominant Colour descriptor, which means that the content is skin filtered with a rather small amount of processing, without accessing and processing the video data.

1. Introduction

Technical advances in the area of telecommunications have led, in the past few years, to a boom in the production, transmission and availability of multimedia content. In fact, the advent of digital television, the success of digital photograph and video cameras and the generalization of Internet usage around the globe have led altogether to an ever-increasing amount and panoply of audiovisual content. In the near future, with the arrival of the third generation of mobile networks, all this multimedia content will virtually be accessible anytime, anywhere. As such, the increasing amount of multimedia information available highlights the need to develop systems able to automatically describe this information for more efficient filtering, retrieval and, in general, management. So that the applications describing the content and the applications using the corresponding descriptions can interoperate, it is necessary to define a standard that specifies the syntax and semantics of these multimedia descriptions. With this goal in mind, the moving picture experts group (MPEG) group of International Standards Organization (ISO) developed a standard called "Multimedia Content Description Interface", also known as MPEG-7. This standard specifies a set of

*Corresponding author. Tel.: +351-966624628.
E-mail addresses: pmf@lx.it.pt (P.M. Fonseca), fp@lx.it.pt (F. Pereira).

0923-5965/$ - see front matter
doi:10.1016/j.image.2004.04.005

tools that allow the description of several types of multimedia information. The final objective is to allow that the huge amount of multimedia content available can be filtered, searched for, managed and consumed in a thoughtful, flexible, fast and efficient way [2].

While describing all this multimedia information would allow for more efficient management, the need still exists for the development of automatic systems that allow us to find and consume the information considered 'essential' to each of us, adapted to our tastes, our preferences, our time and our capacities of receiving, consuming and storing that information—in other words there is an increasing need to summarize and personalize audiovisual content.

Automatic video summarization may be defined as the selection and representation of a sequence of still images or video segments (with or without the corresponding audio content) expressing the full content available in such a way that only concise and relevant information (according to some criteria) is presented to the user [5]. The automatic video summarization system proposed in this paper uses previously created MPEG-7 descriptions of the video content to summarize it, rather than directly analysing the content in question. The main novelty of the proposed system lays in new selection and constraining strategies where a query relevance metric is used to create audiovisual summaries in the context of an MPEG-7 based querying application. The relevance metric proposed allows for the creation of summaries based on the automatic choice of the 'most relevant' keyframes in the video content; a constraints schema is also proposed, allowing for the creation of summaries to be based on constraints specified by the user in terms of number of keyframes and/or summary duration time. Also, a very low complexity solution for creation of skin colour based summaries using the same constraints schema is proposed.

In order to be useful for the ordinary user, a summarization system like the one proposed must typically be able to perform some kind of semantic mapping between low-level audiovisual content features (e.g., colour, shape, and motion for video) and relevant high-level features that will allow the 'recognition' of certain events or features that are semantically important in some context. For example, consider a system allowing the identification of faces in movies or documentaries, the detection of the *pivot* in a news service, the identification and translation of dialogues in movies or even real time adult video content filtering. All these examples require the detection of an important feature: the presence of human skin to which can be associated some semantic value. Following this scenario, this paper will also propose a skin colour filter, i.e., a filter that will allow the detection of images or video segments containing a significant amount of skin colour regions. The identification of this feature will allow the creation of summaries in which only parts with human skin are present (or, in opposition, absent). This filter can also be integrated in content filtering mechanisms (e.g., to inhibit the exhibition of scenes with nudity content) or in more advanced semantic queries in which the detection of images with skin can be important. The novelty of the proposed skin colour filter is that it is very simple since it does not act directly on the video data but rather on its MPEG-7 description.

The organization of this paper is as follows: next section describes the automatic video summarization system, the summarization application and the summary creation process, notably the relevance metric and the constraints schema; Section 3 describes the proposed skin colour filter; Section 4 provides some results and, finally, Section 5, concludes the paper.

2. Automatic video summarization system

Fig. 1 illustrates the architecture of the automatic video summarization system proposed. The summarization process basically consists on the selection of the elements (video segments, or keyframes) considered 'essential' to represent the *original video content* according to some criteria. This selection is based on the analysis of the video content's features which is performed by accessing a previously created MPEG-7 description of the content's visual features. The process of describing

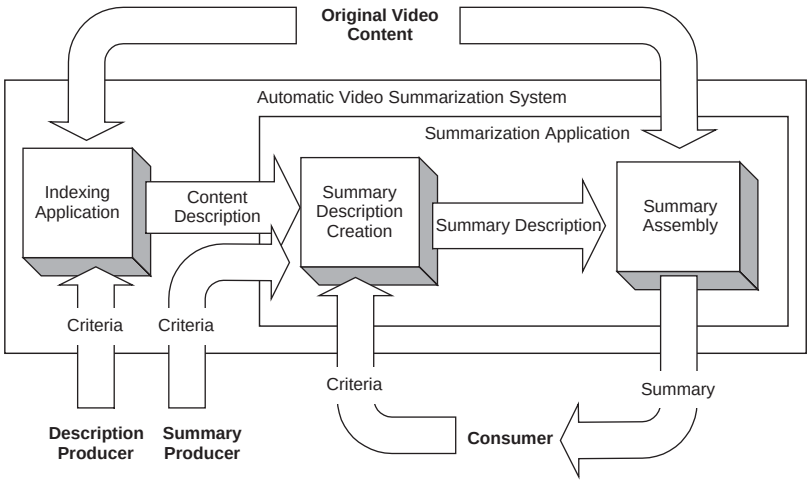

Fig. 1. Global architecture of the summarization system.

a given video content asset is called description or *indexing* and is performed by a so-called *indexing application*, which is controlled by an outside entity, the *description producer*. This entity determines which are the most relevant features and therefore the features that shall be instantiated in the content description using adequate (MPEG-7) descriptors for later use in the creation of summaries. The summarization is performed by a *summarization application*, which can be divided in two parts: (i) the *summary description creation* part, which creates an MPEG-7 description of the summary, based only on the MPEG-7 description of the content and on criteria specified by a *summary producer* or by the *consumer* himself; and (ii) the *summary assembly* part which, based on the created description and on the original video content, creates the summaries and presents or delivers them to the user. It should be emphasized that the original video content is not used in the process of summarization (i.e., in the selection of the elements that represent the original content) but only in the actual assembly of the summary. Since many times the creation of the summary description does not happen at the same time or even at the same place as the assembly of the summary based on the description, this split between the summary description and summary assembly process is important.

The indexing application is rather trivial and serves only the purpose of creating MPEG-7 descriptions for the original video content. For that reason, this paper will focus solely on the presentation of the summarization application for which an implementation architecture is proposed, including a novel query-based summary creation mechanism.

2.1. MPEG-7 based summarization application

The goal of the summarization application is to create summary descriptions for a given video content asset based only on an available MPEG-7 description. Besides creating summary descriptions using MPEG-7 summary description tools, this application may also be used to present to the user the summary created, immediately or later on. Fig. 2 illustrates the architecture of the summarization application.

The summarization application takes as an input an MPEG-7 description of the content to create, according to criteria determined by the user, descriptions of summaries that represent or replace the original video content. As was previously said, when describing the global architecture of the summarization system, the summarization application can be divided in two parts: the *summary description creation* part and

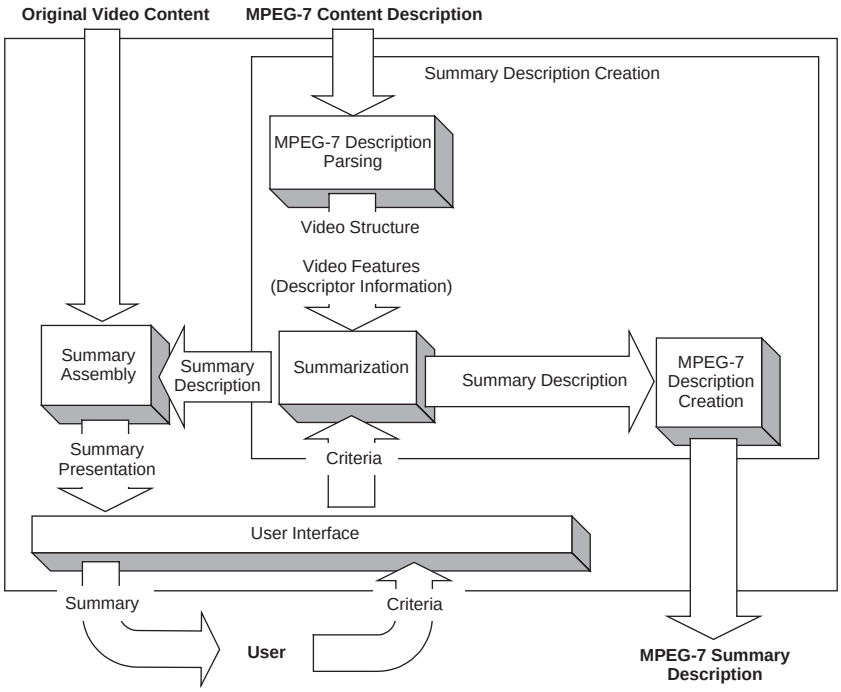

Fig. 2. Architecture of the summarization application.

the *summary assembly* part. The summary description creation part can be further sub-divided into several parts with distinct functionality: after the *MPEG-7 description parsing* block parses the description of the content, the *summarization* block stores in its data structures the information about the video structure and features provided by the description schemes and descriptors instantiated in the description. Based on this information and on criteria entered by the user through the user interface, this block creates a description of the summary. This description can then be stored in an MPEG-7 compliant file by the *MPEG-7 description creation* block. Based on the description of the summary created by the first part and on the original video content, the *summary assembly* module can then or later on assemble the summary using chosen frames and/or video segments from the original content, presenting or delivering it to the user. Note that this assembly process can be performed at a different time or location as the summary description creation process.

2.2. Summary creation

The summarization process is determined by criteria specified by the user. In the proposed system, these criteria assume the form of queries: queries are thus the starting point of the summary creation process—available queries will be described in the next section. After the query (and thus a criterion) is specified, a set of adequate functions will assign to each shot and/or keyframe of the video in question a relevance value depending on the information which defines the query/criteria. The higher the similarity between the content description of each shot/keyframe and the query description derived from the query information specified by the user, the higher the relevance of that shot/keyframe will be in terms of the performed query/criteria. Finally, according to the set of constraints (determined by the user) that should limit the duration of created summaries, the shots/keyframes with the highest relevance will be chosen to create the summary.

The proposed MPEG-7 based summarization application allows the creation of two types of summaries:

- *Keyframe summary*—Consists in a set of still images—keyframes—extracted from the original video content; the keyframes are presented in the same order as they appear in the original content.
- *Video summary*—Consists in a subset of video segments—shots (and associated audio segments) from the original video content, presented in the same order as they appear in the original video content.

For both cases, MPEG-7 provides tools to describe a summary created based on some criteria so that it can be easily (re)created when the original content is at hand.

A keyframe summary is described using the Sequential Summary description scheme (DS). Each keyframe's location (relative to the original video content) is described using a Visual Summary DS. On the other hand, a video summary is described using a Hierarchical Summary DS. Each shot's location and duration (relative to the original video content) are described using the Summary Segment and Key Audio Visual Clip DSs. All these description schemes are specified in the multimedia description scheme (MDS) part of the MPEG-7 standard [3]. The creation process of both types of summaries will be now explained.

2.2.1. Types of query

The summary resulting from the process of summarization depends naturally on the query(ies)/criteria chosen by the user. Each type of query differs in terms of the structural elements targeted by the query (i.e., the query may apply to shots, keyframes, or both), the type of information entered by the user, and the descriptor(s) used in the query.

The descriptors used in the queries correspond to descriptors defined by the MPEG-7 standard; these descriptors must have been instantiated in the available MPEG-7 description of the original video content. The visual descriptors used describe the frames' colour properties (notably Scalable Colour, Dominant Colour and Colour Structure descriptors) and the video's motion properties (through the Motion Activity descriptor). All these descriptors are specified in the visual part of the MPEG-7 standard [4]. Additionally, keywords present in keyframes and shots' descriptions can also be used in the queries; keywords are described using the Textual Annotation data type (specified in the MDS part of the MPEG-7 standard [3]).

In case multiple descriptors are involved in a single query, a combined relevance is determined for each shot/keyframe as the weighted sum of partial relevancies that result from the comparison of the query description and content description for each descriptor in question. Table 1 summarizes the available queries and the query information and query descriptors associated with

Table 1
Types of query available

Query	Summary targeted elements	Query information	Query descriptors
Keywords	Shots and keyframes	Set of keywords specified by the user	Keywords
Colour	Keyframes	Colour specified by the user	Scalable Colour, Dominant Colour
Example	Shots and keyframes	Descriptor information of an example frame	Keywords, Motion Activity, Scalable Colour, Dominant Colour, Colour Structure
Action	Shots	Motion activity level specified by the user	Motion Activity
Skin colour	Keyframes	Minimum spatial coherency specified by the user	Dominant Colour

each type of query for the summarization application developed.

In a *query by keywords*, the user specifies one or more keywords which determine the query—the 'closer' the shots and keyframes' set of keywords is from the set of keywords specified by the user, the higher their relevance will be for the query in question. In a query by colour, the user specifies a colour—the higher the percentage of the chosen colour in each of the keyframes, the higher their relevance will be for the query. In a *query by example*, the user provides an example frame—the 'closer' the shots' and keyframes' descriptor information is to the example query's descriptor information (for the specified descriptors), the higher their relevance will be for the query. In a *query by action*, the user selects an action level—the closer to the specified level each shot's Motion Activity descriptor information is, the higher its relevance will be for the query. Finally, the *query by skin colour* uses the skin colour filter proposed in this paper (see Section 3)—the stronger the presence of skin colour in the video keyframes, the higher their relevance will be for the query.

Obviously, a descriptor that is not instantiated for the MPEG-7 description of the video content in question cannot be used in the queries.

2.2.2. Relevance determination

By definition, a keyframe is (one of) the most representative frame(s) in a shot; so, for a given query, the keyframe's relevance is naturally dependant on the relevance of the shot it belongs to. By adding the relevance of each shot to the relevancies of its keyframes, summary creation (for keyframe summaries as well as for video summaries) can be based only on the choice of the 'most relevant' keyframes after a query is performed. For an example, consider a query by action—the query descriptor used in this query is the Motion Activity descriptor, which describes the motion properties of the shots. Although this query does not use any keyframe query descriptors, the keyframes will still have a relevance value relative to the relevance of the shot they belong to—in this sense, each keyframe represents its shot. Following this approach, summaries will be always created by selecting the most relevant keyframes, even if a video summary is being created; in this case, the summary corresponds to the shots associated to the most relevant keyframes.

The 'query relevance' is measured by a set of one or more elementary relevancies each one expressing the similarity between the query information (specified by the user) and the keyframe's/shot's description for a single descriptor. If more than one descriptor is involved in the query, a combined relevance measure has to be determined for each keyframe. The combined relevance for each keyframe, RK, is given by

$$\mathrm{RK} = \mathrm{RS} + \sum_{k=1}^{4} r_k \cdot qf_k \text{ with } \sum_{k=1}^{4} qf_k = 1, \tag{1}$$

where RS is the combined relevance of the shot to which the keyframe in question belongs to, r_k are the elementary relevancies associated with each keyframe descriptor ($k = 1$: keyframe's keywords, $k = 2$: Scalable Colour, $k = 3$: Dominant Colour, $k = 4$: Colour Structure) and qf_{k} are query factors (specified by the user) specifying the weight of each descriptor in the performed query (the query factor of an unused descriptor is zero). The combined relevance for each shot, RS, is given by

$$\mathrm{RS} = \sum_{s=1}^{2} r_s \cdot qf_s \text{ with } \sum_{s=1}^{2} qf_s = 1, \tag{2}$$

where r_s are the relevancies associated with the shot descriptors available ($s = 1$: shot's keywords, $s = 2$: Motion Activity) and qf_s are query factors (specified by the user) specifying the weight of each descriptor in the performed query (again, the query factor for an unused descriptor is zero).

The addition of the shot's relevance to each keyframe's relevance is needed to base the creation of summaries solely on the choice of keyframes which in this case represent the corresponding shot.

2.2.3. Summary length constraints

For summary creation purposes, the user can specify a set of constraints that will define the length of the created summaries:

- *Keyframe percentage*—Based on the specification of a keyframe percentage factor which determines the maximum number of keyframes

that will be used to create the summary. This maximum number is determined by

$$n_{\max} = kfp \cdot n_k, \quad (3)$$

where $n_{\max}$ is the maximum number of keyframes that will be included in the summary, *kfp* is the keyframe percentage specified by the user and n_k is the number of keyframes present in the MPEG-7 description of the video content.

- *Number of keyframes*—Based on the direct specification of the maximum number of keyframes that will be used to create the summary.

Both keyframe percentage and number of keyframes set a limit on the maximum number of keyframes in a summary—number of keyframes constraint. In case both are specified, the condition that sets the lowest number of keyframes in the summary prevails. However, another type of constraint may apply:

- *Summary duration*—Based on the specification of the maximum temporal length of the created summary — duration constraint. The choice of keyframes is such that the sum of the durations of the shots to which they belong cannot exceed the specified total duration.

Number of keyframes and duration constraints can be specified simultaneously. In this case, the duration constraint is applied to the subset of keyframes chosen after applying the number of keyframes constraint.

2.2.4. Choice of keyframes for summary creation

After each keyframe's combined relevance is computed and the constraints are set, the summaries are finally created by selecting the keyframes with the highest relevance. If a number of keyframes constraint of *n* keyframes if specified, the choice of keyframes is simply performed by choosing the *n* keyframes with the highest relevance.

If a duration constraint is specified, the keyframes are chosen according to the duration of the shot they belong to—the keyframes are chosen by decreasing order of relevance until one of the two following conditions are verified:

- There are no more keyframes left to be chosen or,
- the sum of the durations of the shots to which the keyframes already chosen belong is such that no further keyframes can be chosen without exceeding the summary duration limit specified.

3. Summaries based on a skin colour filter

It is well known that the presence of persons is a very important factor for the selection of the most relevant parts in a certain video content asset. A typical automatic way to detect the presence of human beings is by detecting the presence of skin colour regions. Having this in mind, a rather simple but efficient skin colour filter only based on the analysis of a single MPEG-7 descriptor is proposed in the following. The skin colour query here proposed uses a skin colour filter to determine which keyframes from the video content have important regions of exposed human skin. For that purpose, each keyframe in the content will be analysed to evaluate the presence of regions with skin colour. Their relevance will indicate if they have small or large regions of colour identified as skin colour. In the next sections, the proposed skin colour filter will be presented.

3.1. Skin colour characterization

Before the skin colour filter is described, it is necessary to characterize skin colour. It should be noted that the colours associated with skin suffer the influence of several factors such as pigmentation (varying from person to person), concentration of blood, and illumination. There are many studies available in the literature characterizing skin colour. In [1], Ascenso et al. present a study on the skin colour distribution for a population of Caucasian males and females (this is a limitation of the skin colour characterization; however the proposed skin colour filter will work in the same way if a more general skin colour characterization is used). As it can be easily seen from the

histograms (presented in [1] and reproduced in Fig. 3), the distribution of each skin colour component in the HSV colour space corresponds, approximately, to Gaussian distributions, especially for the hue and saturation components.

Based on these histograms, a set of probability distribution functions corresponding to Gaussian distributions that closely reproduce the histograms has been obtained. The Gaussian approximations are: for the hue component:

$$p_{\mathrm{h}}(h) = \frac{15.040}{\sqrt{2\pi 6^2}} \mathrm{e}^{\frac{(h-18)^2}{2\cdot 6^2}}, \tag{4}$$

for the saturation component:

$$p_{\mathrm{s}}(s) = \frac{22.560}{\sqrt{2\pi 9^2}} \mathrm{e}^{\frac{(s-47)^2}{2\cdot 9^2}}, \tag{5}$$

and for the value component:

$$p_{\mathrm{v}}(v) = \left(0.08 + \frac{39.896}{\sqrt{2\pi 17.3^2}}\right) \mathrm{e}^{\frac{(v-64)^2}{2\cdot 17.3^2}}, \tag{6}$$

where h represents the hue, in degrees, in the interval [0; 360] (°), s represents the saturation, in percentage, in the interval [0; 100] (%), and v represents the value, also in percentage, in the interval [0; 100] (%).

The graphical representations of the obtained Gaussian distribution functions for the hue, saturation and value components are illustrated in Fig. 4. Since the hue component's range of values different from zero is small (when compared to the other components), in order to make it clear and easily visible, the graphical representation of the distribution function for this component is clipped to the range [−30; 60].

It should be noted that the Gaussian distribution for the value component presented in Fig. 4 is slightly different from the one presented in [1]—also in Fig. 3—which corresponds in fact to a sum of several Gaussian distributions. To reduce the complexity of the filter so that it can easily work in real time, only one Gaussian was considered—it was confirmed later that a single Gaussian approximation is a good enough solution for the purpose at hand.

The error between the value component histogram and the proposed Gaussian approximation is only relevant when the value component is high. However this error is not that relevant since a skin

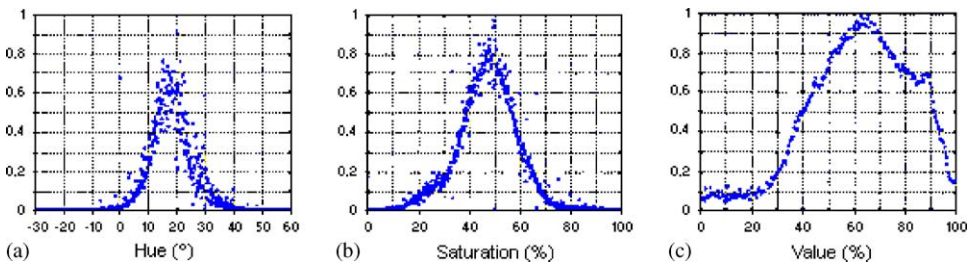

Fig. 3. Skin colour histograms for (a) hue (b) saturation and (c) value [5].

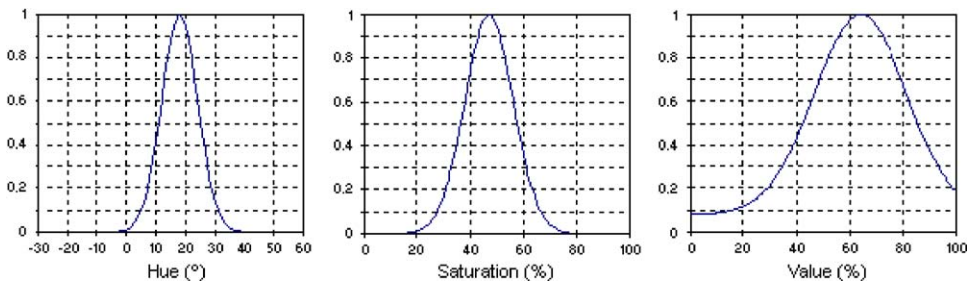

Fig. 4. Gaussian approximations of the hue, saturation and value histograms in regions of skin.

colour in those conditions corresponds to a bright colour which can be easily confused anyway with any other bright colour that may exist in the image.

A colour will be considered as skin colour only if its hue, saturation and value components lay simultaneously within certain limits. As in [1], a given colour will be considered as skin colour if:

- Its hue (h) component lays within the following limits:

$$L_{h1} \le h \le L_{h2} \quad \text{with} \quad \begin{matrix} L_{h1} = 0^\circ, \\ L_{h2} = 50^\circ, \end{matrix} \tag{7}$$

- its saturation (s) component lays within the following limits:

$$L_{s1} \le s \le L_{s2} \quad \text{with} \quad \begin{matrix} L_{s1} = 20\%, \\ L_{s2} = 85\%, \end{matrix} \tag{8}$$

- and, finally, its value (v) component is greater than the following limit:

$$v \ge L_v \quad \text{with} \quad L_v = 35\%. \tag{9}$$

Naturally, as the distance between the HSV components for a certain colour and the centre (mean value) of each of the skin colour Gaussian distributions increases, the less likely it is for that colour to correspond to a skin colour.

3.2. The MPEG-7 Dominant Colour descriptor

The skin colour filter proposed in this paper is based on the information provided by the Dominant Colour descriptor, specified in the visual part of the MPEG-7 standard [4]. The Dominant Colour descriptor, DCD, offers a compact description of the most representative colours in an image or in a region of an image. DCD is defined as

$$\text{DCD} = \{(c_i, p_i, v_i)_{i=1,2,\ldots,N}, s\}, \tag{10}$$

where N is the number of dominant colours in an image or region of an image and c_i is the vector of dominant colours—c_i depends on the colour space being used, e.g., if the default RGB colour space is used, $c_i = (R_i, G_i, B_i)$ where R_i, G_i and B_i represent the red, green and blue components. The parameter p_i represents the fraction of pixels in the image or region corresponding to the colour c_i (although not necessarily equal to c_i since it is taken as the representative colour of a cluster of colours). The optional parameter v_i expresses the variance of colour values in the set of pixels represented by the dominant colour c_i. The spatial coherency, s, represents the weighted sum of the spatial coherency values associated to each dominant colour. For each dominant colour, the spatial coherency expresses how coherent in space the pixels corresponding to that dominant colour are, i.e. if they are scattered throughout the image (low coherency, left image on Fig. 5) or clustered together in a single region (high coherency, right image on Fig. 5).

The spatial coherency of an image is represented by a value in the interval [1; 31]. A value of 0 is used to indicate that the spatial coherency was not computed, while a maximum spatial coherency corresponds to a value of 31. Typically, images filmed from a long distance have low spatial coherency when compared to images that correspond to close-ups of objects. The reason for this is that, generally, the first type of images includes more objects of a smaller size while the latter type usually corresponds to larger, first plane objects and thus, large patches of the same colour.

The MPEG-7 Dominant Colour descriptor allows the specification of a maximum of eight dominant colours. The chosen dominant colours are those with higher presence in the image, this means those with higher p_i. The colour spaces that can be used by the MPEG-7 Dominant Colour descriptor are: *RGB*, YCbCr, HSV, HMMD, linear transformation matrix with reference to *RGB*, and monochrome. The transformations

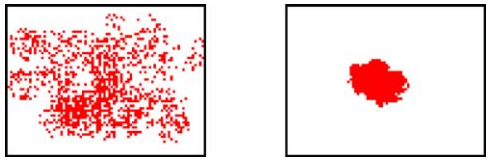

Fig. 5. Examples of high and low spatial coherency.

between any of these Colour spaces and the *RGB* colour space are also defined by the MPEG-7 Visual standard [4].

3.3. Skin colour filter proposal

The skin colour filter aims to associate to a given image (or, when used in a skin colour query, to a given keyframe) a value indicating the relevance of the presence of regions with skin colour in that image. As previously said, this filter is only based on the information about the colour of an image as given by the MPEG-7 Dominant Colour descriptor previously described, namely:

- For each dominant colour, its components in the used colour space and its percentage in the image;
- Global spatial coherency information (if available).

Therefore, the relevance of an image in terms of skin colour presence will be basically determined by the presence in the image of dominant colours that resemble to skin colour. In this paper, this resemblance is expressed in terms of the distance between the colour in question and the centre of the skin colour Gaussian distributions defined in Section 3.1. Naturally, the closest to the centre of the distributions the HSV components of each dominant colour are, the more resembling to skin colour that colour is.

Consider the dominant colour description of an image given by Eq (10) with $N \leqslant 8$. Since skin colour was characterized in the HSV colour space (see Section 4.1), it is convenient that the same colour space is now used for the filter. If any colour space other than HSV is used by the Dominant Colour descriptor, appropriate transformations must be performed, as defined in the MPEG-7 visual standard [4]. The resulting dominant colour vector, c_i, is defined as

$$c_i = (h_i, s_i, v_i), \tag{11}$$

where h_i, s_i and v_i represent the hue, saturation and value components for dominant colour i.

The relevance of an image in terms of skin colour, R_{sc}, is here expressed by the sum of the partial skin colour relevance factors, rf_i, for each dominant colour in the image, i.e.,

$$R_{sc} = \sum_{i=1}^{N} rf_i \tag{12}$$

with

$$\begin{cases} [p_h(h_i)]^{k_h} \cdot [p_s(s_i)]^{k_s} \cdot [p_v(v_i)]^{k_v} \cdot p_i \\ \quad (0^\circ \leq h_i \leq 50^\circ) \wedge (20\% \leq s_i \leq 85\%) \wedge (v_i \geq 35\%), \\ 0 \\ \quad \text{otherwise} \end{cases} \tag{13}$$

where $p_h(h)$, $p_s(s)$ and $p_v(v)$ are the Gaussian distributions of the hue, saturation and value components proposed in Section 4.1 and p_i is the percentage of pixels in the image that correspond to that dominant colour.

With rf_i depending on $p_h(h_i)$, $p_s(s_i)$ and $p_v(v_i)$, the attribution of a skin colour relevance value to an image will depend on the distance between the components of each dominant colour and the components of the colour range considered as characteristic of human skin. In practice, each one of these factors rf_i consist in a penalizing factor for the distance between a certain colour and the most common skin colour—the higher the distance, the lower the image's relevance in terms of skin colour. The parameters k_h, k_s, k_v, specified in the [0;1] interval allow to modify the shape of each Gaussian distribution (hue, saturation and value, respectively) so that the colours whose components are closer to the limits given by Eqs. (7)–(9) are not excessively penalized; a value of zero for one of these parameters indicates that a specific component (hue, saturation or value) is not used in the relevance attribution process. After performing exhaustive tests, a value of 0.05 for all parameters k_h, k_s, k_v was found to be adequate for the detection of skin colour in images.

In addition to the information about the most representative colours in an image—the dominant colours—information included in the MPEG-7 dominant colour description about the spatial distribution of the colours can also be used to improve the quality of the summary results. Relevant skin regions are typically characterized by a high coherency in terms of colour

distribution, i.e., the pixels of a certain colour appear to be spatially coherent, with small spatial spreading. Therefore, we may use the spatial coherency information given by the MPEG-7 Dominant Colour descriptor to eliminate images with skin colour like regions but with non-spatially coherent distribution. This results in the following additional condition:

$$R_{sc} = 0 \text{ if } s < s_{min}, \quad (14)$$

where s_{min} is the minimum (global) coherency necessary to consider that an image has any skin colour region. This condition can be applied only if (global) spatial coherency information is available in the MPEG-7 Dominant Colour descriptor, i.e., if s is different from zero (a spatial coherency value of zero indicates that spatial coherency was not computed). After performing exhaustive tests, values of s_{min} between 10 and 15 were found to be adequate.

4. Summarization examples

In this section some summarization examples using the system developed will be presented. It should be noted that since this is a video summarization application, the examples in this paper cannot illustrate its full potential in terms of creating video summaries. In the examples shown, each image has two indices: the first index indicates the ranking of the image in terms of the query performed (e.g., keyframe with number 1 is the most relevant, keyframe with number 2 is the second most relevant and so forth) and the second index indicates its final combined relevance computed as described in Section 4.

4.1. Example query

In an example query, a frame is given as example. This frame can either be an arbitrary video frame or a keyframe, i.e., a frame identified as such in the MPEG-7 description of the content. The relevance in terms of summaries will be determined by comparing each keyframe's and shot's descriptor instantiations with the query description for the same descriptor extracted from the example frame (if an arbitrary frame is used as example) or directly available from the MPEG-7 description of the content (if a keyframe is used as example). The user can specify which query descriptors shall be used among those which are instantiated in the MPEG-7 description of the video content (uninstantiated descriptors, obviously, cannot be used): textual annotation (keywords), Motion Activity, Scalable Colour, Dominant Colour and Colour Structure.

In the first example, the frame presented in Fig. 6 is given as the example. This is an arbitrary video frame which is not a keyframe of the video content according to the MPEG-7 description at hand. Fig. 7 illustrates the keyframe summary obtained if the Dominant Colour descriptor is the only descriptor used as query descriptor using a certain type of metric that expresses the similarity between the dominant colours of two images. Fig. 8 illustrates the keyframe summary obtained if the Dominant Colour, Scalable Colour and Colour Structure descriptors are specified simultaneously as query descriptors (to be extracted from the example frame). Both summaries were generated with a number of keyframes constraint of eight keyframes. As it can be seen, the simultaneous use of three colour descriptors dramatically improves the quality of the results if the goal is to find all the keyframes which are similar to the example frame not only in terms of the global set of colours but also in terms of their structure and spatial distribution. In fact, the Dominant Colour

Fig. 6. Example frame.

Fig. 7. Keyframe summary after an example query using only the Dominant Colour descriptor.

Fig. 8. Keyframe summary after an example query using the Dominant Colour, Scalable Colour and Colour Structure descriptors.

descriptor does not express the way the colours are spatially distributed in the image; for this reason, the keyframes in Fig. 7 have some similar colours to the example frame (notably those with highest presence, in this case colours around the grey) but the colours are not spatially distributed in the same way as happens for Fig. 8 after using colour descriptors which also express the colour spatial distribution.

4.2. Skin colour query

The example presented in Fig. 9 corresponds to the set of the most relevant keyframes selected by applying the skin colour filter on a video that consists of a documentary (with close-ups on interviewed persons) and a football match. Initially, the spatial coherency verification is disabled. As it can be seen in Fig. 9, some of the selected keyframes have colours that resemble skin colour but have uncharacteristic (when compared to typical skin colour regions) spatial distributions, namely the frames with indices 4, 5 and 6 (the first, third and fourth images on the second row). By enabling spatial coherency verification, with a minimum threshold $s_{\min}$ of 15, these keyframes are eliminated, as can be seen in Fig. 10, since they have very low spatial coherency and better images

Fig. 9. Applying the proposed skin colour filter without spatial coherency verification.

Fig. 10. Applying the proposed skin colour filter with spatial coherency verification $s_{\min} = 15$.

have been selected—the eliminated images correspond to images filmed from a long distance thus with a lower spatial coherency value for the whole image (remind that MPEG-7 spatial coherency reflects the image's global spatial coherency and not the spatial coherency of each dominant colour).

This simple verification condition allows to significantly reduce the number of keyframes which are identified as having a significant amount of skin colour like regions but typically are less semantically relevant, e.g., crowds versus close-ups or even some objects with a colour close to skin colour.

Although the skin colour filter was extensively applied to videos with significant nudity parts, for obvious reasons examples will not be shown here. However, the results were very satisfactory taking into account the simplicity of this filter. For adult content filtering, the proposed filter may be improved in terms of reducing the number of false alarms (parts declared as adult content while this is not the case) if other MPEG-7 descriptors are also taken into account such as shape and texture descriptors. The proposed skin colour filter may easily be implemented at a home set-top box for adult content filtering of content not explicitly declaring this fact. This highlights one of the first

cases where MPEG-7 low-level descriptions (in this case, colour) show its importance also for broadcasting type environments.

5. Conclusions

Although there has been an enormous research effort in the area of video summarization, there are not many applications available that allow the creation of summaries from generic video content. In fact, the vast majority of the applications developed so far were developed aiming specific content types like sports, news [6] and movies [5]. Besides, these applications do not allow any kind of interoperability in terms of descriptions/metadata since they use proprietary content description solutions.

The summarization application proposed in this paper can be used to create summaries for any given content and can work with MPEG-7 descriptions created by any indexing application provided that they are MPEG-7 compliant. The approach to use MPEG-7 descriptions as the base for the summarization process avoids that feature information extraction is repeatedly performed whenever a summary is created. Finally, since the summary description creation process is based on a description that can exist physically separated from the video content, this type of summarization process has the advantage of not requiring the availability of the content itself.

This paper proposes a novel query-based summary creation mechanism using a relevance metric and a constraints schema implemented in the context of an automatic video summarization system based on MPEG-7 descriptions. The proposed relevance metric allows, based on query information specified by the user, the creation of summaries to be based on the choice of the 'most relevant' keyframes; on the other hand, the proposed constraints schema influences the summary creation process with constraints specified by the user in terms of the number of keyframes and/or summary duration time.

In order for a summarization system like the one proposed to be even more useful for the ordinary user, it must perform some kind of semantic mapping between low-level audiovisual content features (e.g., colour, shape, motion) and high-level features. The detection of the presence of human skin in a video content can be useful to perform this kind of semantic mapping, for example in the detection of faces in movies or documentaries, pivots in news services or nudity content filtering. For that purpose, much work has been done in the area of skin colour detection, both in the context of face recognition as well as targeting content classification and filtering. However, all the existing solutions require the analysis of the image data in terms of colour, shape or texture rather than relying on previously created standard descriptions of these features. Unlike existing solutions, the skin colour filter proposed in this paper is based solely on the information given by the MPEG-7 Dominant Colour descriptor. Although it is rather simple and relies only on about 30 bits of information per image/keyframe, it shows nevertheless good performance in the detection of large regions of exposed skin in images. The inclusion of spatial coherency verification improves the filter by eliminating images that have regions with colour similar to skin colour but with low spatial coherency. This simple verification test can be, in some cases, a fallible one since the (global) spatial coherency of an image as expressed in the MPEG-7 Dominant Colour descriptor depends on the coherency associated with all dominant colours in the image and not only on the coherency of each dominant colour separately.

The skin colour filter's accuracy could probably be increased if shape and texture information were also used. However, this would render the filter more complex and highly dependent of the existence of this kind of information in the MPEG-7 description of the content—dominant colour information is relatively easy to extract and therefore highly likely to be present in MPEG-7 descriptions; on the contrary, texture and specially shape information is more complex to extract and maybe less generally useful and therefore less likely to appear in MPEG-7 descriptions. This is anyway a good topic for further work.

Another important topic for future work is the development of summary evaluation methodologies able to measure the efficacy of a certain summarization application, depending on previously identified summarization criteria. These methodologies may be subjective or objective and would be rather helpful in comparing the performance of competing summarization engines.

References

[1] J. Ascenso, P. Correia, F. Pereira, A Face Detection Solution Integrating Automatic and User Assisted Tools, RECPAD 2000, Porto-Portugal, May 2000.

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

[3] MPEG MDS Group, Multimedia Content Description Interface—Part 5: Multimedia Description Schemes, ISO/IEC FDIS 15938-5, ISO/IEC JTC 1/SC 29/WG 11/N4242, Sydney, Australia, July 2001.

[4] MPEG Video Group, Multimedia Content Description Interface—Part 3: Visual, ISO/IEC FDIS 15938-3, ISO/IEC JTC 1/SC 29/WG 11/N4358, Sydney, Australia, July 2001.

[5] S. Pfeiffer, R. Lienhart, S. Fischer, W. Effelsberg, Abstracting digital movies automatically, J Visual Commun. Image Represent. 7 (4) (1996) 345–353.

[6] H.D. Wactlar, Informedia—Search and Summarization in the Video Medium, In: Proceedings of Imagina 2000 Conference, Monte Carlo, Monaco, January/February 2000.