INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONLE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11/m3107
MPEG 97

Title: **Suggestions to the Draft of MPEG-7 Requirements**

Source: **Yong Rui, Thomas S. Huang and Sharad Mehrotra**
University of Illinois at Urbana-Champaign, USA

Status: **Proposal**

The suggestions in this document are based on the modified N1921 by Frank Nack [1] and recent discussions on the MPEG-7 reflector [2-5]. (We are aware that there exists a more recent version of [1]. But because of the time limitation, our discussions are still based on [1]. It is therefore highly possible that some suggestions and questions in this document have already been addressed in the new version of [1])

1. Introduction

2. MPEG-7 Framework

3. MPEG-7 Terminology

1. Data

Suggestions: "... or technology" seems to be redundant.

Questions: What are the differences and relationships between data, material and media?

Questions: When we are talking about multimedia, we normally are talking about 5 media, i.e. text, image, graphics, video and audio. When we see a paper book, we will consider it as a text medium. However, if we digitize it, do we still consider it as a text medium? If so, how about shooting a video of this paper, page by page? Will we consider the video as video medium or text medium?

2. Feature

Based on Frank Nack[1] and Richard Qian[5]'s discussions on the MPEG-7 reflector, Fernando Pereira proposed the following definition for feature[3].

A feature is any prominent part or characteristic of the data, which stands to somebody for something in some respect or capacity.

Suggestions: Compared with previous definitions of feature, the above definition is a much better one. Here we just want to emphasize the scope of the features.

Consider the video-frame-object hierarchy. At the video level, we can talk about salient object trajectory feature or video activity feature. But at frame level, we will talk about the global image features such as global color or texture. At the object level, we can talk about local features such as the color or texture properties of the objects. Features have different scopes at different levels. Some features are only defined at a particular level.

3. Descriptors

The most recent definition is given by Fernando Pereira[3] based on Frank Nack[1] and Richard Qian[5]'s discussions on the MPEG-7 reflector.

A descriptor is a representation associated to one or more features.

Qian made a good point in terms of distinguishing features and descriptors in [5]. That is, "One can discuss the concept of features without specifying any metric or form of representation. For example, discussing color without limiting to a specific color space or a reference white color. When we want to be specific, i.e., quantify things in most cases, we use descriptors." We just have the following suggestions.

Suggestions: Instead of asking a descriptor to support more than one feature, we can divide the features into two categories, say intra-feature and inter-feature. Intra-features are atomic features such as color, texture, etc. On the other hand, inter-features characterize the inter-relationships between atomic features, as proposed in [5]. (However, it is not very clear to us what is the importance of inter-features, and what is an example) Basically, we are trying to avoid using a descriptor to support more than one feature. To us, one feature can have multiple descriptors, characterizing the feature from different perspectives. For example, the color feature can be characterized by both color histogram and color moments descriptors. We think it might be confusing if one descriptor can support more than one feature [6].

4. Description scheme

5. Description

6. Coded description

Suggestions: Compression and indexing are normally contradictory to each other. This coded description should achieve a good trade-off.

4. MPEG-7 Requirements

4.1. MPEG-7 Common Audio and Visual Requirements

4.1.1 Descriptors and Description schemes

1. Description Classes

Questions: Are the items in this section descriptions (both descriptors and description schemes) or just descriptors [3], or both features and descriptors?

Questions: Why statistical information is so important that they should be emphasized separately from other features?

Questions: "Objective attributes" and "Subjective attributes"
Are we talking about features? Since *attribute* is not defined yet.

Questions: "Composition Information ..."
What is a scene? It has not been defined yet. For some AV material, say a still image, there is no *scene*.

Questions: *Concept* is too specific. For a still image, what are the concepts?

2. Description Models for Multimedia Material

Suggestions: We should have a title which can reflect better the content of this section [3].

3. Cross-modality

4. Types of features

Suggestions: We have primitive features vs logic features. We also have objective features and subjective features. What are the relationships of the above concepts?

5. Feature priority

6. Feature hierarchy

7. Feature scalability

Suggestions: Items 5, 6 and 7 seem to be related. They are also related to item 2.

8. Description temporal range

4.1.2 Functionality of Descriptors and Description Schemes

1. Content-based retrieval

Suggestions: Currently, content-based retrieval means low-level feature based retrieval in multimedia information retrieval research community [6-8]. In MPEG-7, however, we should clearly state that content includes not only the low-level features but also high-level information, such as concepts, etc. Features are defined at different levels.

2. Similarity-based retrieval

3. Associated information

Questions: Text is also a feature. Why we want to differentiate one feature from another? What we want is to support queries based on more than one feature or descriptor [6].

4. Streamed and stored descriptions

5. Distributed multimedia databases

Suggestions: Although this is related to MPEG-7, it is more towards the search engine end.

6. Linking

Suggestions: The first sentence is not quite related to the title.

7. Prioritization of Related Information

Suggestions: 11 does not exist [3].

8. Browsing

Suggestions: Are we talking about the data itself or the structure of the data collection. If it is the former, "visualization of summary" is more appropriate; if it is the latter, then this functionality is more related to search engine than to MPEG-7.

9. Associate Relations

Questions: What is a component and what is a representation scheme? They are not defined.

4.1.3 Coding of Descriptors and Description Schemes

1. Description efficient representation

2. Robustness to information errors and loss

3. Copyright information

Suggestions: We should clearly indicate if the copyright refers to the data described or to the description [3].

4.2 MPEG-7 Visual Requirements

1. Description Classes

Questions: What are *visual objects* and what is *volume*?

Suggestions: "Still and moving images" are more AV materials than features or descriptors. Is it appropriate to put them here?

Questions: There are many ways of characterizing "Motion". What do we mean by motion here? It is too vague.

Suggestions: "Deformation" is more related to similarity matching than a to a feature or descriptor.

2. Description Visualization

Questions: Are we talking about supporting the visualization of the data itself or the structure of the index.

3. Visual data formats

Questions: What are the differences and relationships between digital *video* and *film*?

4. Visual data classes

Questions: What are the differences and relationships between "visual data formats" and "visual data classes"?

4.3 MPEG-7 Audio Requirements

Questions: We have visual and audio requirements in this and previous sections. However, do we want to support text documents as well? If so, where should we fit the corresponding requirements?

4.4 MPEG-7 Application-dependent Requirements

Questions: "Support of full-text descriptions as well as structured fields (database descriptions)"
Text description can also be structured, such as a technical paper (with sections) and a HTML file (with tags).

Questions: "Language independence"
Are we talking about natural language or computer language? If the former, what is a feasible solution?

Questions: "Support of feature-based and concept-based queries at segment level"
What are feature-based and concept-based queries and what is a segment? They have not been defined.

Questions: "Strong database scalability for access to large databases and distributed databases"
Does this mean MPEG-7 will also explore multi-dimensional indexing techniques (such as R tree) as required in large databases?

Questions: "Support of a combination of feature-based and concept-based queries ..."
What are feature-based and concept-based queries? In general, what are the differences and relationships between *content-based*, *feature-based*, and *concept-based* queries?

Suggestions: "Support for interactive queries ...."
Some techniques have been developed in this aspect [6].

Questions: "the ability to perform on-line annotations and mark regions-of-interest ..."
This is more related to database population or feature extraction than to MPEG-7 itself.

Suggestions: "Ability to easily export existing databases into the MPEG-7 format"
Redundant with previous items (items 3 and 4 in this section).

Annex A Examples

A.1 Hierarchical model for a video description scheme

Frame

Questions: What are the *features* (not just descriptors) associated with Frames? In general, in this section we are only talking about descriptors. Should we also talk about features, since it is already a defined concept.

Microsegment

Questions: Do we really want this entity and what are the advantages? It is largely overlapped with shots. If a shot consisting of two parts, one is a pan and the other is a zoom, it is relatively easy to find its corresponding microsegments (but this is a very rare case). However, if the shot consists of a simultaneous panning and zooming, should we say there is only one microsegment? In general, in most shots, it is difficult to classify a port as a pure focus, pan, tilt, or zoom. Then, how can we find the microsegments?

Shot

Suggestions: A more common definition of shot is the following:

> *A shot consists of a sequence of unbroken frames obtained from a single camera* [6, 7].

It is a physical entity. But in current definition ("associated streams") it is not a physical entity but a semantic entity. If it is a semantic entity, automatic detection of shots then becomes almost impossible.

Sequence

Topic

Questions: What are the differences and relationships between sequence and topic? Do we really want to introduce more both concepts?

Program

Document

Questions: What are the differences and relationships between program and document? It seems to us that the only difference between the two is that the former is delivered while the latter is not.

Questions: Sequence and topic are "subject-based" while program and document are "object-based". What are the relationships between the former and the latter?

A.2 Other category models for video description schemens

Shot cluster

Keyframe

Questions: What is a mosaic image? Is it an image consisting of multiple sub-images? If so, is there any specification of each sub-image's size and their spatial relationship? In multimedia information retrieval research community, keyframe is more an image than a mosaic image consisting of multiple sub-images.

A.3 Category models for video ontology description schemes

Character

Action

Object

Questions: What is a "sign"? It is not defined.

Script
Space
Time

Annex B Issues

B1 Open Issues

Questions: "Role of sub-titling information in MPEG-7"
How is sub-titling defined and what is the significance?

Questions: "Temporal relation between the MPEG-7 descriptions and coded data ..."
Why we want to emphasize *temporal* relation, not other types of relations?

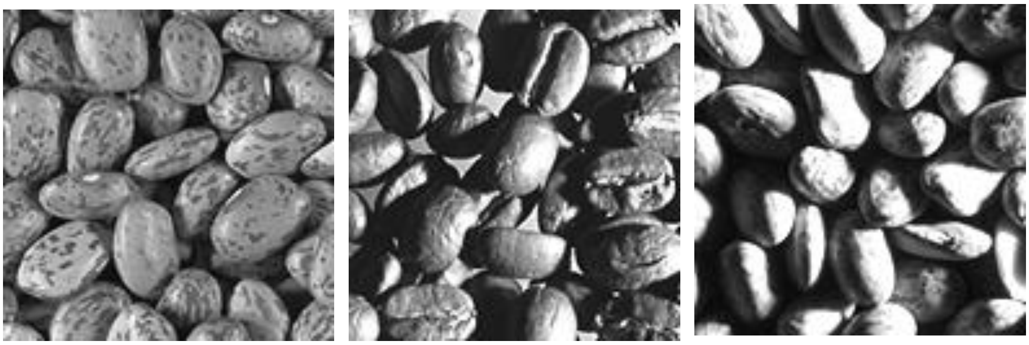
Suggestions: For the same AV data, different people or the same people under different circumstances may perceive it differently. This is called *perception subjectivity* of AV data.

Example1: Suppose we have three shapes. The first one is a perfect circle. The second one is a perfect circle corrupted by continuous small random noise around the whole circle. The third one is exactly the same as the first one except that it has a single "bump" at one point of the circle. If we ask people which of the last two shapes is more similar to the first one, different people may give us different answers. This is an example of perception subjectivity of AV data.

Example2: Look at the following 3 texture images:

If we ask people which of the last two textures is more similar to the first texture, again, different people may give us different answers.

A major characteristic of AV data is its perception's subjectivity. How to effectively support this is thus essential to the success of MPEG-7. One possible solution is to use relevance feedback techniques [6].



B2 Issues resolved during the San Jose

References:

[1] Frank Nack, V1 of 4. Draft of Req. Doc. - finally (Modified version of N1921), MPEG-7 reflector

[2] Ibrahim Sezan and Richard Qian, MPEG-7 Req. Doc. and Terminology Clarification, MPEG-7 reflector

[3] Fernando Pereira, Re: V1 of 4. Draft of Req. Doc. - finally, MPEG-7 reflector

[4] Yong Rui, An Object Model, MPEG-7 reflector

[5] Richard Qian, Re: V1 of 4. Draft of Req. Doc. - finally, MPEG-7 reflector

[6] Yong Rui, Thomas Huang and Sharad Mehrotra, Relevance Feedback techniques in Interactive Image Retrieval, SPIE Conf on Storage and Retrieval of Image and Video Databases VI.

[7] P. Aigrain, HongJiang Zhang and D. Petkovic, Content-based representation and retrieval of visual media: A state-of-the-art review, Multimedia Tools and Applications, vol. 3, Nov 1996

[8] Yong Rui, Thomas S. Huang and Shih-Fu Chang, Image Retrieval: Past, Present and Future, submitted to Journal of Visual Computing and Image Representation. An early version is published as an invited paper at International Symposium of Multimedia Information Processing, Taipei, Taiwan, 1997