COMPARING THE QUALITY OF MULTIPLE DESCRIPTIONS OF MULTIMEDIA DOCUMENT

N. Adami, M. Corvaglia & R. Leonardi
DEA - University of Brescia
Via Branze 38, 25123 Brescia, IT
Email:{adami,leon}@ing.unibs.it

***Abstract*—**

With the de£nition of the MPEG–7 standard, thanks to its interoperability behaviors, it is now possible for applications to use content descriptions of a same document, coming from different sources. This implies that the overall information available at the application can be highly redundant and mechanism for £ltering these informations are hence required. In this works a general approach to descriptions integration is considered. The idea is to manipulate information readily available from the considered descriptions to reach an accurate integration result, without having to reprocess the multimedia material. The proposed general has been applied at a real problem showing how segment decompositions can be integrated on the basis of *Dominant Color* descriptors series. The work introduce also the concept of reliability of descriptors extraction method demonstrating how it can be effectively used through the integration process.

I. INTRODUCTION

The ”quality” of multimedia information nowadays available is not only an intrinsic property. The real value of a document it is strongly related to how it can be retrieved and how it can be rapidly browsed. Focusing the attention to a subclass such is the one formed by audio-visual sequences, it points out that in the early past a lot of work has been done to de£ne suitable frameworks for ef£cient browsing through this material and for retrieving relevant information according to user speci£c requirements. Tools that can automatically parse video sequences, classify each segment, and thereby provide non-linear access capability based on the semantic content now can be provided not only to professional but also to generic users. To support the above mentioned tools a new standard for the description of the content of multimedia documents, called MPEG–7, has been de£ned by the International Standard Organization (ISO). However the introduction of this open standard, mainly thought with the objective of to guaranty inter-operability between multimedia applications, impose to solve new problems. Considering that different descriptions of the same multimedia document will be available to a given application the new question is now: how all this informations can be integrated together. Redundant information must be discarded while complementary ones must be integrated, in order to have a unique and richer description, tailored possibly to a speci£c need. This would enable an ordered organization of such content allowing quality information to be retrieved for any speci£c purpose.

In this paper, a general framework is proposed to compare and merge different visual Descriptors (D) and Description Schemes (DS) pertinent to a same video. A speci£c case study is considered where the objective is to obtain a better temporal segmentation of a video sequence by integrating two separate segment decompositions (in the MPEG–7 sense) in a single partition with a more accurate representation of the shot boundaries. The processed information to reach this result uses two *Dominant Color* series associated at two shot decomposition obtained applying different temporal segmentation method to a video sequence.

The presentation is organized as follows. Section II describes a general methodology for integration. In Section III some experimental simulation results are provided. Finally section IV concludes the presentation.

II. INTEGRATION OF DIFFERENT DESCRIPTION

The generic integration process is described in the ¤ow chart of Fig 1. In order to show the applicability of the proposed method, the integration of two DSs will be consider. However it is important to stress that it can be applied to a general number of DSs.

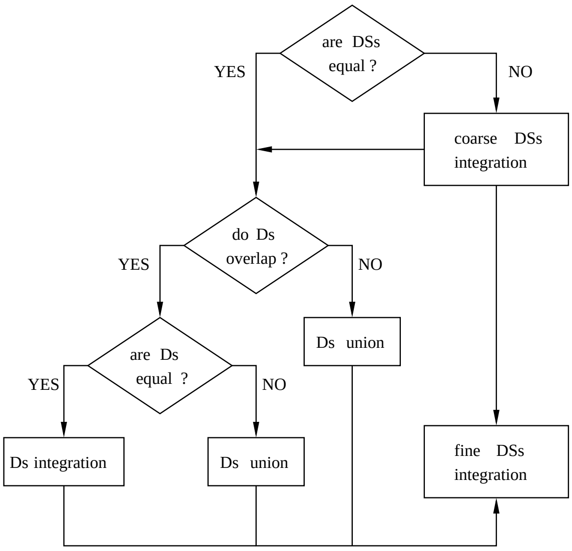

Fig. 1. Comparison between distances.

If two given description are equal there is no need to integrate them at the description level and only the associated descriptors has to be merged. Vice versa, if the composition of the descriptions do not perfectly match three main operation has to be performed:

- Coarse DSs integration

- Descriptors integration
- Fine DSs integration

Clearly the effective operations involved in each step are specialized with respect to the type of DSs and D that will be considered. For this reason the explanation will be given by mean of un example. In the subsequent part of the work we will consider the integration of two different segmentation into shots of a video sequence.

A. Preliminary de£nition and assumption

It is assumed to have two description schemes DS_1 and DS_2 describing the structural decomposition of the same video sequence but obtained applying different extraction algorithms. Each of them include two sub-DSs one derived from the *Video Segment DS* type, used to describe the temporal decomposition of a video and the other derived from *Time Series DS* type, used to associate the *Dominant Color D* to speci£ed frames of a sequence. These sub-DSs will be indicated respectively with $SEG_{1,2}$ and $DC_{1,2}$. It will be assumed that all the DC instances have been extracted from the video sequences by using the same method in order to guaranty interoperability as speci£ed by the MPEG–7 standard. Finally it is assumed that it is possible to have reliability indexes, which can assume values normalized between 0 and 1, of the extraction method used to produce the descriptions. For the DC extraction algorithm considered in this work (de£ned in the non normative part of the MPEG–7 standard) it is not useful to de£ne any reliability index because is would be the same for both the description. Regarding the temporal segmentation method two indexes has been used the probability of miss detection de£ned as

$$p^{miss} = \frac{N_{miss}}{N_{RT}} \qquad \text{and} \qquad p^{false} = \frac{N_{false}}{N_{CT} + N_{false}}$$

where N_{miss} is the number of missed transitions and N_{RT} the number of real transition that characterize the video. N_{false} is the number of false alarms and N_{CT} the correctly estimated editing effects. To allow an effective comparison these parameters has to be estimated on the top of the same ground truth.

B. Coarse Segment Decomposition DSs integration

In this phase it is evaluated how the two decomposition into shots are related to each other. The comparison can lead at two different results.

- A shot transition T is present in both the descriptions. In this case a transition has been recognized but especially for the gradual ones the associated shot boundaries could not perfectly mach in the two description. Assuming b and e represent respectively the beginning and the end of a shot a possible solution for the integration can given by:

  $$b = \frac{p_1 b_1 + p_2 b_2}{p_1 + p_2} \qquad \text{and} \qquad e = \frac{p_1 e_1 + p_2 e_2}{p_1 + p_2}$$

  where $p_{1/2} = 1 - p_{1/2}^{false}$ is the probability of correct recognition of the segmentation method. If the hypothesis of statistical independence between the two method is satis£ed the new reliability values $p = 1 - p_1^{false} \cdot p_2^{false}$.
- A shot transition T is present only in one decomposition. In this case a new interval is added to the temporary £nal segmentation with the following values:

  $$\begin{array}{ll} b = b_1, e = e_1, p = p_1 \quad \text{if} & T \in SEG_1 \\ b = b_2, e = e_2, p = p_2 \quad \text{if} & T \in SEG_2 \end{array}$$

After this step a new DS for the segmentation into shots, that integrate the information of the starting ones, is obtained. This description is characterized by a number of miss and false detection that in the worst case will be the sum of the initial ones. A re£nement of this intermediate result is then required.

C. Time Series DSs integration

The *Time Series DS* it is used in this work to associate a DC to a speci£ed frame. It is possible to use to different time series:

- Regular Time Series: where DC descriptor is associated to frames that has been obtained temporally sub-sampling the original video by a £xed factor.
- Irregular Time Series: a DC descriptor is associate to a general frame specifying every single gap between two consecutive frame which have associated a DC.

In this case the integration is not dif£cult and in general is given by the union od DC_1 and DC_2.

D. Fine Segment Decomposition DSs integration

After the integration of the *Dominant Color* descriptors we can expect to have a more dense serie of DC. This serie can be used to re£ne the initial temporal segmentation evaluating the distance between consecutive DC across a transition point. This is a critical operation but has it is shown in the next section a better segmentation can be obtained when the density of DC is suf£ciently high.

III. Experimental results

A real simulation of the previously described integration has been implemented showing the effectiveness of the proposed method. In the £rst part of this section the *Dominant Color* descriptor is de£ned. A study on distance measures for establish robust correlation between instances of this D are then reported At the end of the experiment the £ne integration of the two DSs is performed showing how the the £nal number of miss and false detection can decrease with respect to the initials one and how they are related to the DCFG.

A. Dominant Color Descriptor

Given a certain color space, the *Dominant Color* Descriptor represents a set of dominant colors that characterize a frame or one of its arbitrarily-shaped regions [3]; for any color in the *Dominant Color* descriptor, three parameters are used in the computation of the distance measure: *variance, probability, coherence*. The minimum number of dominant colors is 1, while the maximum one is 8. In general, in a video, the *Dominant Color* D is associated only to selected frames in the sequence. For simplicity, we have computed such D a subset of frames obtained by down–sampling the original sequence of frames.

B. Dominant Color distance measure

Two distance measures have been considered: an Euclidean distance and the Earth mover's distance.

Euclidean distance

In this case, the RGB color space has been selected. The distance between two *Dominant Color* P and Q is defined by:

$$D(P,Q) = \sqrt{\sum_{k=1}^{3} \sum_{j,i=1}^{N} (p_{i_k} - q_{j_k})^2} \tag{1}$$

where N indicates the number of dominant colors forming the D of each frame, p_{i_k} and q_{j_k} correspond to the *i-th* dominant color of P and the *j-th* one of Q, respectively (index *k* refers to the color component (R,G,B)).

We can use the Euclidean distance only if some hypothesis are satisfied:

- each *Dominant Color* D must have the same number of dominant colors;
- the set of dominant colors follows the same order for both D's: for example, the first element of the first *Dominant Color* is compared with the first element of the second *Dominant Color*.

It can be observed that the RGB color space appears inadequate as it does not consider any of the visual proprieties of the human eye.

Earth mover's distance

The Earth mover's distance (EMD) [4] allows to establish a distance measure between two probability density functions. Since the *Dominant Color* represents is a color distribution information, the EMD seems to be a good candidate with respect to the Euclidean distance.

Defining a distance between two distribution requires a notion of distance between the basic features that are aggregated in the distribution. This distance is called *ground distance*. For the *Dominant Color* D, the ground distance is the distance between each color; the ground distance used is the Euclidean distance in the CIE-Lab color space, since this color space is especially designed so that the Euclidean distance strongly correlates with the human ability to discriminate color information.

The two *Dominant Color* can be seem as two distribution: $P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), \cdots, (p_{N_P}, w_{p_{N_P}})\}$ is the first *Dominant Color*, where p_i is an element of the *Dominant Color* (a color), w_{p_i} its weight (probability), N_P the number of dominant colors; $Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), \cdots, (q_{N_Q}, w_{q_{N_Q}})\}$ is the second *Dominant Color*; let $\mathbf{D} = [d_{ij}]$ the *ground distance matrix* with d_{ij} the ground distance (Euclidean distance) between p_i and q_j:

$$d_{ij} = \sqrt{\sum_{k=1}^{3} (p_{i_k} - q_{j_k})^2}. \tag{2}$$

The EMD is given by:

$$EMD(P,Q) = \frac{\sum_{i=1}^{N_P} \sum_{j=1}^{N_Q} d_{ij} f_{ij}}{\sum_{i=1}^{N_P} \sum_{j=1}^{N_Q} f_{ij}}.$$

where f_{ij} represents the flow between p_i and q_j, that minimizes the overall cost [4].

An alternative ground distance is proposed by [2]; this distance can be used only for the *Dominant Color* D:

$$\begin{aligned} d_{ij} &= \sqrt{\sum_{k=1}^{3} (p_{i_k} - q_{j_k})^2 + \sum_{k=1}^{3} (\sigma_{p_{i_k}} - \sigma_{q_{j_k}})^2} \\ &+ \sqrt{(ch_{p_i} - ch_{q_j})^2} \end{aligned} \tag{3}$$

where p_{i_k} and q_{j_k} are the *k-th* color component respectively of the *i-th* element of *Dominant Color* P and of the *j-th* one of Q, σ_{p_i} and σ_{q_j} the *i-th* and *j-th* element variance, ch_{p_i} and ch_{q_j} their respective coherence.

For simplicity, we shall use EMD to indicate the EMD using the Euclidean distance (2) ground distance and EMDdc to indicate the EMD using equation (3) ground distance.

C. Comparison between distances

The distance measures introduced previous subsection are compared through the following experiment. Two different shot boundaries segmentations of a video (1400 frames) are considered: SEG_a is a shot segmentation extracted by the algorithm in [1], while SEG_h is a segmentation extracted by hand. Also, a *dominant color* D with 8 elements is computed every 10 frames of the video sequence, that is at frame 10, 20, etc.

By, comparing SEG_a and SEG_h, we observe that $sSEG_a$ is characterized by some false shot transitions. We then try to use the information given by the *dominant color* D to correct SEG_a. Specifically, for each shot transition of SEG_a, we compute the distance between the *dominant color* D just before the transition and the one just after it. For instance, if there is a transition that starts al frame 51 and ends to frame 52, we compute the distance between the *dominant color* D of the frame 50 and the one of the frame 60.

The results are reported in Table I where T and NT stand for shot Transition and No shot Transition, respectively.

TABLE I

COMPARISON BETWEEN DISTANCES.

$n.$	$\mathbf{seg_a}$	**Eucl. dist.**	**EMD**	**EMD dc**	$\mathbf{seg_h}$
1	51–52	69	73	852	T
2	81–82	27	29	573	NT
3	150–151	12	9	392	NT
4	184–185	5	16	308	NT
5	217–218	30	25	474	NT
6	233–260	65	106	788	T
7	295–316	56	41	893	T
8	622–633	80	59	763	T
9	839–840	18	10	260	T
10	904–905	20	8	264	T
11	956–964	1	1	45	NT
12	1013–1014	26	9	237	T
13	1063–1064	78	19	518	T
14	1156–1157	40	27	380	T
15	1258–1259	29	26	261	T
16	1359–1375	57	69	651	T

From the results in Table I, we can see that, the EMDdc is not a good distance since it is difficult to fix a threshold that can discriminate a shot transition from non existing shot transition. As shown in Figure 2, the EMD and the Euclidean distance provide better results: for example, setting the threshold to $d_{th} = 15$, the Euclidean distance identifies 3 false shot transitions (n.3, n.4, n.11) while the EMD identifies 2 false shot transitions (n.3, n.11) and 3 false no shot transition (n.9, n.10, n.12). The Euclidean distance seems to offer the optimal trade-off, but it can be used only if there are two *dominant color* Ds with the same number of elements ($N_P = N_Q$); if *dominant color* Ds with different number of elements ($N_P \neq N_Q$) are compared, the EMD represents a good trade-off.

D. Comparison between shot boundaries

Two different shot boundaries segmentations (SEG_1 and SEG_2) of a same video (6000 frames) are compared and integrated. Suppose a *dominant color* D with 8 elements has been used for both segmentations; since $N_P = N_Q$, we use the EMD with Euclidean distance, as ground distance.

In order to evaluate the performance, we create a ground truth by annotating by hand the correct shot boundaries. It is thus possible to extract from SEG_1 and SEG_2 the number of missed shot transitions and the number of false transitions:
– seg_1: 2 miss, 26 false;
– seg_2: 5 miss, 35 false.
The integration performance is indicated in Figure 2.

As it can be seen in the Figure, for a large sub-sampling factor in the assignment of the *dominant color* information (more than 10), the number of missed shot boundaries goes to zero, but the number of false alarms remains approximately constant (to about 25). For a small sub-sampling factor (less than 10), the number of missed boundaries is not zero while the number of false ones is reduced to about 15. With a small sub-sampling factor the distance measure is smaller than the one that would have been obtained with a larger sub-sampling factor, since the distance is computed between nearer frames. Thus, in the first case, we tend to cancel more easily false transitions while not avoiding the identification of misses. The number of false transition does not reduce to zero as:

- a boundary can be a false transition even if it is present in both seg_1 and seg_2;
- if a shot has a lot of motion activity, near frames are very different; so, the distance measure between the *dominant color* associated to two consecutive frames could be above the selected threshold.

Therefore, a better performance can be expected only with the use of additional Ds. In some cases satisfactory results can be obtained only by processing the original video material.

IV. Conclusion

In this work, a general approach to integration of different visual descriptions of a same video is explained; this is an ill-posed problem for many reasons: the definition of distance measure for each Descriptor used in the comparison, the definition of a reliability of the descriptors, the dependency of the integration results to the specific needs of a given user. While a general framework has been suggested, a specific case study has been implemented: namely the comparison and merging of two different shot boundary decomposition of a video sequence, using the *dominant color* information. The results of the methodology appear promising in this context, and it seems that they can be improved by using additional Descriptors.

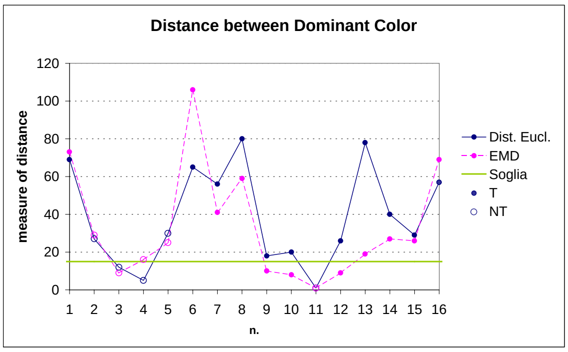

Fig. 2. Comparison between distances.

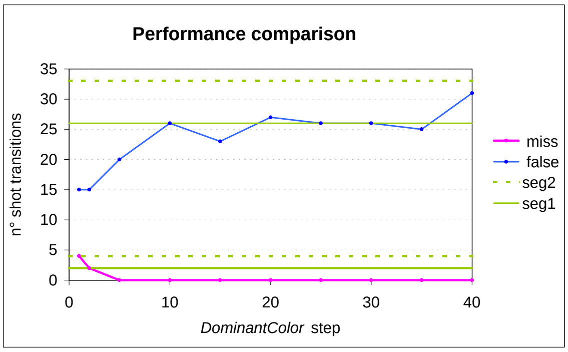

Fig. 3. Miss and false VS DC Frame Gap

References

[1] N. Adami and R. Leonardi. Identification of editing effects in image sequences by statistical modelling. In *Proc. of PCS99*, Picture Coding Symposium, pages 157–160, Portland - OR, USA, April 1999.

[2] N. Adami, R. Leonardi, and Y. Wang. Evaluation of different descriptors for identifying similar video shots. In *Proc. of ICME2001*, International Conference on Multimedia and Expo, pages 948–951, Tokyo, Japan, August 2001.

[3] MPEG-7 Video Group. Multimedia content description interface – part 3: Visual. proposal to the 56th mpeg meeting, ISO/IEC JTC1/SC29/WG11 MPEG01/N4062, Singapore, March 2001.

[4] C. Tomasi Y. Rubner and L. J. Guibas. A metric for distributions with applications to image databases. pages 59–66, Bombay, India, January 1998. Proc. ICCV 1998.