COMPARING DESCRIPTIONS OF MULTIMEDIA DATA FOR SIMPLIFICATION AND INTEGRATION

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

Abstract

In this paper, different description of a same audio-visual material are considered. A general formulation for their integration is proposed. The descriptions are considered to be compliant with the specifications of the MPEG7 standard. 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. An example is shown in the context of a video sequence separation into shots. The proposed solution enables an integration of two separate temporal segmentations of the video sequence, leading to a possible automatic comparison of shot cut detection algorithms. The method uses the *dominant color* Descriptor information associated to the two original descriptions to reach the integration result.

1 Introduction

As the amount of digital multimedia material is growing extremely rapidly, it has become well understood that there is a need to build adequate summaries and accurate descriptions of any new content. This would enable an ordered organization of such content enabling quality information to be retrieved for any specific purpose. Clearly a lot of attention has been devoted to the possibility to automatically extract semantic information. In this work, the focus is placed instead on the ability to integrate descriptions of a same material provided by different sources. Indeed it is important to improve the quality of a description by combining different sources of descriptions taking into account, if possible, also their reliability.

In the MPEG7 standard, the content of multimedia documents is expressed by a set of standard Descriptor (D) and Description Schemes (DS), expressed according to a Description Definition Language (DDL). The reliability requirement is fundamental when more sources provide heterogeneous Descriptors about the same audio-visual material; in fact, redundant informations must be discarded while complementary ones must be integrated, in order to have a unique and richer description, tailored possibly to a specific need.

In this paper, a general framework is proposed to compare and merge different visual Descriptors and Description Schemes pertinent to a same video. A specific case study is considered so as to enable 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 the *dominant color* Descriptor associated to a subsampled set of frames of a video sequence.

The presentation is organized as follows. Section 2 describes a general methodology for integration. Section 3 outlines how distance measures between *dominant color* Descriptors can be used to combine two temporal segmentations of a video sequence. In section 4, several distance measures to compare *dominant colors* are proposed and compared. Section 5 provides some experimental simulation results. Finally section 6 concludes the presentation.

2 Comparing different descriptions

The MPEG7 standard defines a set of Descriptors and Description Schemes which provide information about the content of multimedia documents. In this work, only Descriptors pertinent to visual features described in the Visual part of MPEG7 [1] are considered.

Description Schemes can organize Descriptors in order to provide structural information about the content. *Characteristic* Ds give instead effective representation of the visual features in the document (for examples *dominant color*, *contour shape*, *motion trajectory*, etc.).

Therefore, a description integration process can occur at two levels: First, DSs can be compared and merged; then, it may be the turn of their *characteristic* Ds (Figure 1). However an optimal result should take into account jointly both DSs and Ds information so as to derive an "optimal" integration result.

The meaning of the flowchart in Figure 1 is explained with an example.

Consider two sources A and B, that have generated two different descriptions for a specific video frame j. For simplicity, we have limited consider the following description information: *dominant color* (DC), *contour shape* (CS), and *grid layout* (GL).

As it can be seen in Figure 2a, source A has created a segmentation of frame j, in two parts (GL1) while the description produced by source B, has divided frame j in four parts (GL2). GL1 and GL2 which represent two segment decompositions are different. In

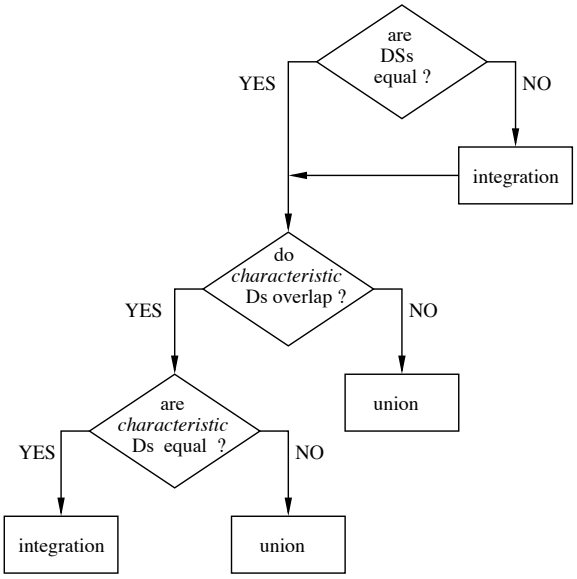

Figure 1: Integration process

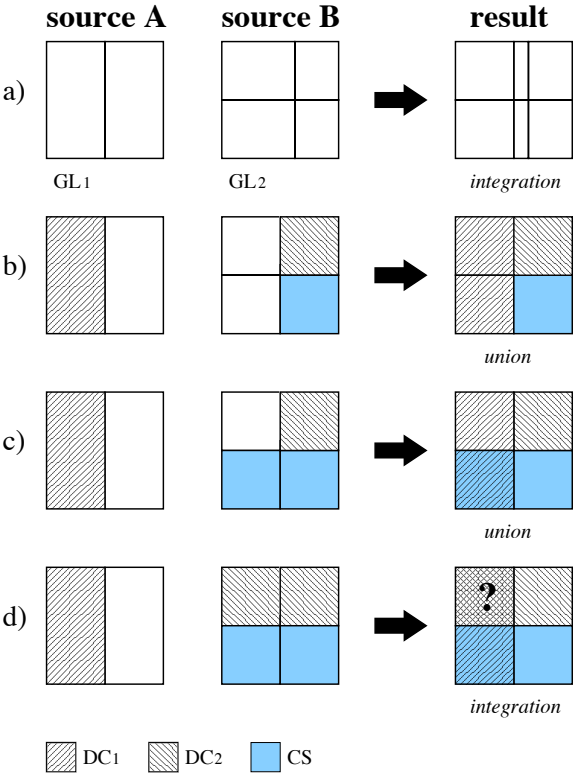

Figure 2: Example showing how the integration process takes place.

this case, their merging is simple (see Figure 2a), as there is a consistency of the segment boundaries, so that one segmentation can be perfectly included in the other one. In general, the integration between DSs providing structural information is immediate if there is a direct correspondence between segment boundaries. Clearly what represents a boundary will depend on the type of decomposition (spatial, temporal or spatio-temporal), so that defining correspondences between the segmentation maps must be dimension dependent.

If instead the two segmentation maps differ so that none can be included in the other one, the integrated segmentation map contains as many regions as the segmentation generated by the overlaping of the two maps. If Descriptor information is associated to the original segmentation layouts, it may be possible by comparing individual Descriptors associated with the two initial segmentations, to rearrange this initial segmentation result so that a more accurate integration is reached.

When individual Descriptor information provided by the sources A and B refers to *characteristic* Ds (DC1, DC2, CS) corresponding to non-overlaping regions (Figure 2b), the integration can simply be achieved by a union of the D information. As shown in the example, after the integration, it has been possible to obtain a more description of frame j with respect to those of sources A and B. Clearly, there would be no change in the final segmentation result in such a case.

The union of D information can also be associated to each individual region of the integrated map if the *characteristic* Ds refering to the original regions provided by sources A and B are of a different type (Figure 2c).

On another hand, if the *characteristic* Ds refering to a same region are of the same type (Figure 2d), the integration is more complex. In the simple example shown in Figure 2, DC1 and DC2 are compared in order to extract a unique DC starting from the information provided in DC1 and DC2. However deriving the new DC is not easy to obtain, because of the following reasons:

1. there is a need to understand how much two *characteristic* Ds differ, i.e. there is the need to establish a distance measure between same type Ds;
2. there is a specific distance measure for any type of D: for example, the distance measure between two *camera motion* Ds is different from the distance measure between two *region shape* Ds;
3. any D should be assigned a reliability field; so an inaccurate D value should be appropriately weighted in the calculation of a distance with respect to a more accurately derived D;
4. the integration should be application independent;
5. sometimes, Descriptor information is not sufficient to ensure a proper integration, so it may be necessary to process the original multimedia information to obtain an adequate description.

How to jointly rearrange a segmentation obtained by overlaping multiple segmentation maps, while taking also into account associated Descriptor information of the same type will not be considered in this work in a general framework. Rather it will be explained through a special case study where two shot boundary decomposition are integrated through the use of associated *dominant color* Descriptor information. For this purpose, we shall describe in the next section how at first an adequate distance measure can be found to compare *dominant color* Descriptors.

3 Comparing two video sequence decompositions

3.1 *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 [1]; for any color

in the *dominant color* descriptor, three parameters are used in the computation of the distance measure:

- its associated *variance*: which measures the average deviation of color values in the frame/region which can be associated to this individual dominant color ;
- its associated *probability*: i.e., the relative number of pixels that can be associated to the considered dominant color;
- its *coherence*: which measures of the spatial distribution of the pixels associated to the considered dominant color.

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.

3.2 Establishing correspondences between the video decompositions

Suppose that the video decomposition has been obtained using two shot boundary segmentation algorithms: seg_1 and seg_2. According to the methodology introduced in Section 2, there are two possible situations:

1. a boundary is reported by both seg_1 and seg_2;
2. a boundary is reported either only by seg_1 or only by seg_2.

3.2.1 Boundary reported by both $\mathbf{seg_1}$ and $\mathbf{seg_2}$

If the shot boundary is provided by both seg_1 and seg_2, we assume with a good confidence that the boundary really exists. Note that a boundary between consecutive shots may have a certain extent because of the type of transitions used during the editing of the video material. Accordingly, assuming that boundaries of seg_1 and seg_2 start, respectively, at frames a_1 and a_2 and finish at frames b_1 and b_2. If the two intervals $[a_1, b_1]$ and $[a_2, b_2]$ overlap even partially, it is considered that both segmentations seg_1 and seg_2 report the presence of a boundary. In this case it is decided that the integrated segmentation result would adjust the final boundaries to the means of the originating interval limits:

$$a = \frac{a_1 + a_2}{2}; \quad b = \frac{b_1 + b_2}{2}.$$

Using the *dominant color* Descriptor, it would be difficult to improve the boundary accuracy provided by seg_1 and seg_2, since the *dominant color* Descriptor is associated to a subsampled set of the original video frames. Even if other Ds were used, a and b could rarely correspond to the exact locations of the shot boundaries. A better performance could only be expected through a complete use of the available video sequence information and with a design of a very robust boundary extraction algorithm.

3.2.2 Different boundaries reported by $\mathbf{seg_1}$ and $\mathbf{seg_2}$

If the boundary is given by either seg_1 or seg_2, there is a need to verify whether a shot transition really occurred (it is a miss of one of the two extraction methods) or it is a false detection of one of the two extractions methods. This is performed using the available *dominant color* information. For example, suppose a *dominant color* D with 8 color elements is provided each 10 frames and suppose that seg_1 reports a dissolve that starts at $a_1 = 83$ and finishes at $b_1 = 95$. In order to determine if the dissolve really occurs, the distance measure is computed between the *dominant color* D associated to frame 80, just before frame $a_1 = 83$, and the one just after frame $b_1 = 95$, that is frame 100. If such a distance is larger than a certain threshold, a dissolve is declared. Assuming that there indeed is a dissolve, not much more information can be used to determine the exact frames of the shot boundaries. Therefore, the shot boundaries are set to the initial $[a_1, b_1]$ interval.

The performance of this integration process will depend on the selection of the threshold,

but mainly on the design of a proper distance measure between the *dominant Color* Ds.

4 Dominant color distance measure

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

4.1 Euclidean distance

In this case, the RGB color space has been selected. The distance between two *dominant colors* x and y is defined by:

$$d(x,y) = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{N}(m_{x_{ij}} - m_{y_{ij}})^2} \quad (1)$$

where N indicates the number of dominant colors forming the D of each frame, $m_{x_{ij}}$ and $m_{y_{ij}}$ correspond to the *j-th* dominant color of x and y, respectively (index i 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.

4.2 Earth mover's distance

The Earth mover's distance (EMD) [2] [3] [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 EMD is a metric distance based on the minimal cost that must be paid to transform one distribution in another one; in other words, it is a particular solution to the *transportation problem* from linear optimization. Suppose that several suppliers, each with a given amount of goods, are required to supply several consumers, each with a given limited capacity; for each supplier-consumer pair, the cost of transporting a single unit of goods is given. The transportation problem is then to find a least-expensive flow of goods from the supplies to the consumers that satisfies the consumers' demand. The extension to the *dominant color* D is simple: the *dominant color* D of a frame is the set of supplies while *dominant color* D of an other frame is the set of consumers; the transportation cost is the ground distance between an element of the first *dominant color* and an element of the second *dominant color*. The problem can be formalized as a dynamic linear programming problem: let $P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), \cdots, (p_{N_P}, w_{p_{N_P}})\}$ be 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; let $Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), \cdots, (q_{N_Q}, w_{q_{N_Q}})\}$ be 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:

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

The aim is find the minimum flow $\mathbf{F} = [f_{ij}]$, where f_{ij} represents the flow between p_i and q_j, that minimizes the overall cost. After the transportation problem is solved and the best flow F is extracted [2], the EMD is given by:

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

An alternative ground distance is proposed by [5]; this distance can be used only for the *dominant color* D:

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

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.

5 Experimental results

5.1 Comparison between distances

The distance measures are compared using two different shot boundaries segmentation of a video (1400 frames). A *dominant color* with 8 elements has been associated to every 10-th frame of the video sequence. The results are reported in Table 1 (T=shot transition, NT=no shot transition).

The EDMdc is not a good distance since it is difficult to fix a threshold that can discriminate a shot transition from non existing shot transition. The EMD and the Euclidean distance provide better results: for example, setting the threshold to 15, the Euclidean distance identifies 2 wrong shot transitions while the EMD indentifies 2 wrong shot transitions and 3 wrong no shot transition. 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* with different number of elements ($N_P \neq N_Q$) are compared, the EMD represents a good trade-off.

Table 1: Comparison between distances

seg_1	seg_2	Eucl. dist	EMD	EMD dc	real seg.
51–52	50–51	69	73	852	T
81–82		27	29	573	NT
150–151		12	9	392	NT
184–185		5	15	308	NT
217–218		30	25	474	NT
259–260	233–258	65	106	788	T
295–316	298–313	56	41	893	T
622–633	619–631	80	59	763	T
839–840	838–839	18	10	260	T
904–905	903–904	20	8	264	T
	956–964	1	1	45	NT
1014–1015	1013–1014	26	9	237	T
1064–1065	1063–1064	78	19	518	T
1157–1158	1156–1157	40	27	380	T
1259–1260	1258–1259	29	26	261	T
1369–1372	1359–1375	57	69	651	T

5.2 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 Euclidean 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:
– $seg1$: 2 miss, 26 false;
– $seg2$: 5 miss, 35 false.
The integration performance is indicated in Figure 3.

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

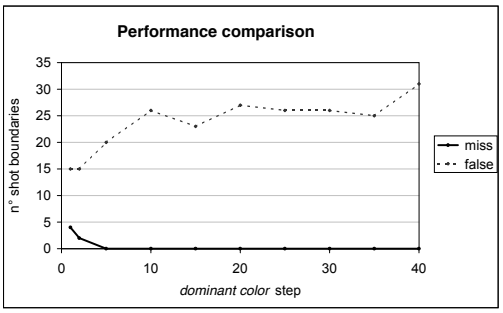

Figure 3: Comparison using *dominant color*

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.

6 Conclusion

In this work, a general approach to integration of different visual descriptions of a same video is explained; this is a ill-posed problem for a lots of 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 descomposition 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.

References

[1] MPEG7 Video Grup. Multimedia content description interface – part 3: Visual. Iso/iec jtc1/sc29/wg11/N4062. Singapore, March 2001.

[2] Y.Rubner, C.Tomasi, L.J.Guibas. *The earth mover's distance as a metric for image retrivial.* Stanford, CA.

[3] Y.Rubner, C.Tomasi, L.J.Guibas. *The earth mover's distance, multi–dimensional scaling and color–based image retrivial.* Stanford, CA 1997.

[4] Y.Rubner, C.Tomasi, L.J.Guibas. *A metric for distribution with application to image database.* Stanford, CA 1998.

[5] N.Adami, R.Leonardi, Y.Wang. *Evaluation of different descriptors for identifying similar video shots.* Brescia, IT.