Sub-Shot Summarization for MPEG-7 based Fast Browsing

Costantino Grana, Rita Cucchiara
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Modena e Reggio Emilia
Email: {grana.costantino,cucchiara.rita}@unimore.it

***Abstract*—In this paper, we propose a system for automatic video summarization at sub-shot level. Our work covers two main aspects: the first is the sub-shot detection, which is performed without a priori constraints on the number or length of the shots. The algorithm is based on color histograms and motion features, and employs fuzzy c-means with variable number of clusters. The second aspect is an in depth discussion on the annotation of summaries with the MPEG-7 standard. Results on mixed genres TV material, from TRECVID videos, are reported.**

I. Introduction

With the wide diffusion of camera equipped mobile phones, low cost consumer video cameras and Internet technologies, there is nowadays a huge amount of audio-visual data which needs to be stored and organized before its use. Efficient Video Data Management Systems must be designed to index the video according to its content and allow effective browsing and access. Automating large part of the tiresome process of annotation at the low level feature analysis stages to aid human operators would be an advantageous and achievable goal.

Video summaries, in which the whole of a video is abstracted by a general view, result in very compressed representation of the video without losing crucial contents. The aim of a video summary is efficient browsing as well as a fast overview of the original contents by dropping the time spent on tedious operations such as fastforward and rewind. In this aspect, video summarization is popularly regarded as a good approach to the content-based representation of videos [1].

Different approaches have been developed to this aim, focusing on either shot clustering approaches or sub-shot information descriptions. The first group often focuses on providing a compact and static short view of the video, to allow the user a fast comprehension of the video topics. Examples are provided in [2] and [3] in which similar scenes clustering is performed, without regard for the temporal aspects, using respectively k-means (ISODATA) and SVD of the feature space, while in [4] a Scene Transition Graph is automatically built by combining similar shots clusters into temporal neighborhoods. On the other hand, if the aim is letting the user move quickly through the video, getting a compact view of *all* the video contents, shot level information should be provided, enriched by sub-shot description too. In [5] a pairwise clustering generates a binary tree structure which is used as the representation for the shot. Here it is pointed out that sometimes a single key frame is not able to capture all the shot information. In [6] partitional clustering with cluster-validity analysis, employing Minkowski metric, produces a hierarchy of video abstractions, dependent on the number of shots in the video. Recently, [7] proposed a more complex approach which allows testing different number of clusters without trying all the possible choices. The features used are alpha trimmed histograms, as in the GoF descriptor of MPEG-7, which is also employed in the article as a mean of specifying the user preferences.

In order to store and deliver the information on the tailored summaries, an effective and interoperable scheme should be used, and MPEG-7 was designed specifically to this aim [1].

This paper presents an effective summarization algorithm which employs both color and motion features for sub-shot partition, in which the temporal ordering of the obtained clusters is respected, allowing coherent temporal browsing. An in depth description on how summaries should be reported based on the MPEG-7 Summarization Description Scheme [8] is provided, with some discussion on issues of the standard, along with results to assess performances.

II. Sub-Shot Summarization

Given a shot *s*, the designed sub-shot extraction method is composed by 4 consecutive steps: 1) extraction of visual features for each frame of *s*, 2) clustering with Fuzzy C-Means algorithm, 3) Cluster Validity Analysis, 4) Temporal segmentation of clusters. Each shot is analyzed independently.

Both color and motion information are used as visual features. For the first one we adopted 3-Dimensional HSV histograms, quantized following the MPEG-7 recommendation (16 bins for H, 4 bins for S and V), which are quite simple and computationally inexpensive color descriptors. As a distance metric between histograms, we used the normalized L_1 distance:

$$d(H_a, H_b) = \frac{1}{2N} \sum_{i=1}^{B} \left| H_a(i) - H_b(i) \right| \qquad (1)$$

```
<Mpeg7>
  <Description xsi:type="SummaryDescriptionType">
    <Summarization>
      <Summary xsi:type="HierarchicalSummaryType"
        components="keyFrames" hierarchy="dependent">
        <SourceLocator>
          <MediaUri>file://video.mpg</MediaUri>
        </SourceLocator>
        <SummarySegmentGroup level="0">
          <SummarySegment>
            <!-- Here we include the main (dummy) key frame -->
          </SummarySegment>
          <SummarySegmentGroup level="1">
            <!-- Description of the first shot key frames -->
          </SummarySegmentGroup>
          <!-- All other shots follow… -->
        </SummarySegmentGroup>
      </Summary>
    </Summarization>
  </Description>
</Mpeg7>
```

Fig. 1. MPEG-7 description of a hierarchical summary containing different levels of key frames.

Here, H_a and H_b are the histograms, B the number of histogram bins (in our case 256), and N the frame area. For motion characterization, we chose to provide a frame level motion description by dividing the image in 4 quadrants, and for each one we computed three motion features based on the average motion vector: magnitude, angle and deviation of motion vectors from the average. Details of motion descriptor can be found in [9].

After extracting the features for each frame in the shot, and given a number of clusters c, the clustering is then performed using fuzzy c-means. The result is a set of c cluster centroids, $V = (\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_c})$, and a membership matrix, $u_{ik} \in [0,1]$, with $\sum_{i=1}^{c} u_{ik} = 1$. The centroids represent the average color and motion of the clusters.

To determine the degree of aptness of the clusters representation to the data, a cluster validity test is performed using the Xie-Beni index [10]. This is a function of the partition U, the centroids V, the data set X, and the distance measure d:

$$v_{XB}(U,V,X;d) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^2 \left\|\mathbf{x_k} - \mathbf{v_i}\right\|^2}{n\left[\min_{i\neq j}\left(\left\|\mathbf{v_i} - \mathbf{v_j}\right\|^2\right)\right]}, \qquad (2)$$

where n is the number of elements in X.

A small value of this index suggests a good quality of clustering: this occurs when the distances between data points and centroids $\left\|\mathbf{x_k} - \mathbf{v_i}\right\|^2$, weighted with their membership, are low, and the centroids are well-separated, so that the value of $\min_{i\neq j}\left(\left\|\mathbf{v_i} - \mathbf{v_j}\right\|^2\right)$ is high. This index only provides a relative measure of compactness and separation of clusters, but they could be very close in the feature space. To prevent this to happen, the distance between the closest pair of centroids is compared against a threshold T_c. If the distance, for any value of c, falls below T_c, the cluster is marked as invalid.

Then in order to establish the appropriate number of cluster, the clustering is performed using different values of c, and the validity index is computed for each (U,V) pair. As proposed in [7], instead of computing the clusters for every value of c, between 1 and the number of frames in the shot, we use an iterative method: at each iteration the validity indices for three successive values of c are calculated and compared, starting from 2. The algorithm terminates when the best partition is produced by the lowest value of c. If a valid index is not found (because the minimum distance between centroids falls below T_c for every value of c), the number of clusters is set to $c=1$.

Differently from other approaches, the temporal segmentation of clusters is taken into account in the last step of the algorithm. Since the clustering is not aware of the temporal location of each frame, if the shot contains similar frames, but separated in time, they shall be clustered together. As we want to detect sub-shots with their temporal location, we use a post-processing filter that splits the clusters into temporal-coherent "slices" of frames, i.e. groups of adjacent frames belonging to the same cluster. Slices with too few frames (less than 3) are pruned and they don't constitute valid sub-shots.

III. MPEG-7 Summaries

Within the Multimedia Content Description Interface (MPEG-7), the results of a summarization process are considered to enable fast and effective browsing and navigation of multimedia content. Each multimedia summary is used to convey the essential information about the original multimedia content. The different Summarization Tools are presented in clause 13 of Part 5 of the standard.

Two different tools are described by MPEG-7: Sequential Summary and Hierarchical Summary. The first one describes summaries of time-varying audio, video or audiovisual data that support sequential navigation; an example of Sequential Summary could be a movie trailer, which is a shortened version of the movie with synchronized audio and speech. It is specifically oriented to the first approach mentioned in the introduction, in which the aim is to produce a compact continuous view of the video content. In our case we need a structured indexing representation of the video content, to provide the user with fast navigation that also allows different levels of detail. This is exactly what MPEG-7 provides with the HierarchicalSummaryType Descriptor Scheme (DS). This tool is a container for one or more SummarySegmentGroup DS which are the containers for the real content, and may have a SummaryThemeList, a *components* attribute that describes which type of content is used in the summary (keyFrames, keyAudioClips, etc.), and a mandatory *hierarchy* attribute which specifies if every level in the

```
<SummarySegmentGroup numOfKeyFrames="1" level="1"
  id="SummaryShot_1197">
  <SummarySegment order="3">
    <KeyFrame>
      <MediaUri>Seg_001037-001242__2.jpg</MediaUri>
      <MediaTimePoint>T00:00:39:28197F30000</MediaTimePoint>
    </KeyFrame>
  </SummarySegment>
  <SummarySegmentGroup numOfKeyFrames="2" level="2">
    <SummarySegment order="1">
      <KeyFrame>
        <MediaUri>Seg_001037-001242__0.jpg</MediaUri>
      <MediaTimePoint>T00:00:35:27076F30000</MediaTimePoint>
      </KeyFrame>
    </SummarySegment>
    <SummarySegment order="2">
      <KeyFrame>
        <MediaUri>Seg_001037-001242__1.jpg</MediaUri>
      <MediaTimePoint>T00:00:37:20129F30000</MediaTimePoint>
      </KeyFrame>
    </SummarySegment>
  </SummarySegmentGroup>
</SummarySegmentGroup>
```

Fig. 2. MPEG-7 description of the 3 key frames of a shot, distinguishing between the main key frame and two sub shots.

HierarchicalSummary is dependent on the previous level or not, that is if the summary in a level should employ elements in the previous level.

The SummarySegmentGroup DS is the container for either SummarySegment DSs, or other SummarySegmentGroup DSs, which would start a more detailed version of this particular summary. The strength of this approach is that at every level of the hierarchy it is possible to add more detail, or to start an alternative view of the same content. For example, two SummarySegmentGroup within a HierarchicalSummaryType could be used the first to summarize all the shots in a video in which a person is in the foreground, the other to summarize those in which a landscape is present. Apart from the real content (KeyVideoClip, KeyAudioClip, etc.), the SummarySegment DS can have an *order* attribute which allows to specify which element should be presented first. Even if this is not explicitly stated in the standard, we assume that the order should be respected not only within a SummarySegmentGroup, but also in the whole hierarchy, so that elements in higher levels can be correctly inserted between elements of lower levels, in case of dependent hierarchies.

In Fig. 1 an example of a complete summary described with MPEG-7 syntax is shown. Attributes of the Mpeg7 tag have been removed for space constraints. The top-level type (subclause 4.4 of Part 5 of the Standard), used to convey summaries, is the SummaryDescriptionType, which can contain one or more Summarization DS, which in turn can contain one or more Summaries (Hierarchical or Sequential).

Three aspects of the MPEG-7 standard are problematic, to our opinion, and cause doubtful interpretations or redundant

TABLE I
RESULTS OBTAINED WITH THE ALGORITHM, IN NUMBER OF MACRO SHOTS (MS) CORRECTLY SEGMENTED (ALL SHOTS EXACTLY IDENTIFIED), UNDER SEGMENTED, AND OVER SEGMENTED IN A MEANINGFUL WAY OR REALLY EXCESSIVELY.

#MS	Correct	Under	Over	
			Correct	Wrong
308	238 (77%)	0	47 (15%)	23 (8%)

data. First of all, a minor constraint that we feel problematic is the fact that it is mandatory to include at least one SummarySegment into every SummarySegmentGroup. The choice makes sense to avoid empty SummarySegmentGroups, but in case of a key frame hierarchy we are forced to have a level zero SummarySegmentGroup with a *main* key frame for the whole video followed by one SummarySegmentGroup for every shot, with its key frames. Inside these SummarySegmentGroups the sub-shot key frames are listed into a third level of SummarySegmentGroups. This choice is also followed by the examples included in the informative part of the standard. In this case the first (top level) key frame is, by all means nearly useless, since it is very unlikely that the single key frame could represent the whole video. Nonetheless by this approach it is possible to keep all the key frames of a single shot together in a key frame tree structure.

Another point which is more critical is the fact that MPEG-7 provides two different tools for shot/transition video description and for summarization, and it is not possible to establish a link between a shot segmentation and a summarization. We employ the AnalyticEditedVideoType DS to describe how the video is structured in shots and transitions, but there is no way to specify that the key frames listed in the Summary have a corrispondence with the shots. This requires a redundant information storage, since we are forced to include the key frame references also in every shot. This leads to the third limitation of the standard.

The key frame is an ImageLocatorType, so we should choose if we want either to include a reference to the video (using the MediaTimePoint element) or if we want to refer to an external file (to enable, for example, a quicker loading or remote transimission of the summary). Differently from other descriptors, here we cannot specify different choices. In this case we bent the standard a little, by including two elements which refer to different things: a reference to an external image and a time point in the video file. Of course, since it is not possible to have a time point in a JPEG image, a reader which doesn't support this notation, would just skip the MediaTimePoint element.

Apart these considerations, the HierarchicalSummary DS allows an extensible description of the video summary, which could also span in more than two levels, if shot aggregation is required. In the example of Fig. 2, we show a shot summary which contains 3 sub shots, of which number 3 is the most significant and is selected as the shot representative.

a) 

b) 

Fig. 3. Example of two panning sequences (respectively 163 and 155 frames long), automatically summarized in 2 and 3 subshots.

IV. Results

The algorithm has been tested on different kind of videos, and it has shown promising results. For space constraints, we include here only numerical result of a single video of 50950 frames (28:20 min), taken from the TRECVID 2005 Data Set. The video contains a commercial TV programming composed of news, forecast, cartoons, soap operas and a sitcom. We exploited the shot annotation provided in MPEG-7 together with the data set. In the TRECVID data set creation phase the video was segmented in shots, which were aggregated, if necessary, until the current shot was at least 2 seconds in duration, thus obtaining a set of "macro-shots", composed of one or more shots. A difficult aspect of summarization results reporting is that it is sometimes very subjective to evaluate if a summary contains enough information. To cope with this, we considered that the sub shots could be regarded as "context change", that is definitely one of the aims of a summarization process. We thus tried to detect the sub-shots as if they were just different situations in the same shot.

Given only the macro-shot segmentation as input, our algorithm was able to extract one key-frame for every shot contained in each macro-shot and, in 23% of cases, some key-frames more (Table I). Each macro-shot was marked as true positive if the number of key-frames extracted in it was coherent with the degree of change in the visual content or false positive in presence of redundant frames. In the example of Fig. 3a, a panning sequence leads to the correct extraction of two key-frames.

A macro-shot is correctly segmented when the number of key-frames extracted equals the number of shots included in it. When a macro-shot is over-segmented, key-frames in excess can be due to real visual content changes, as for instance the political talk scene in Fig. 3b (the summarization is considered correct since the key frames provide a view of the different actors and their positioning), or to an excessive segmentation. Using our approach only the 8% of over segmentation cases were wrong, while no sub-shots were missed.

V. Conclusion

We presented a system for sub-shot segmentation, in order to provide the user with a two levels of detail view of video material. Complete MPEG-7 summarization was presented and issues of the standard were discussed.

Acknowledgement

The work is supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies Program of the European Commission (Contract G038-507618).

A demonstrator of the algorithm is available for download at http://astral.ced.tuc.gr/delos/ under the *demonstrators* section.

References

[1] J.G. Kim, H.S. Chang, Y.T. Kim, K. Kang, M. Kim, J. Kim, and H.M. Kim. Multimodal approach for summarizing and indexing news video. ETRI Journal, 24(1):1-11, Feb 2002.

[2] Di Zhong, HongJiang Zhang, Shih-Fu Chang, Clustering Methods for Video Browsing and Annotation, Proc. SPIE, Storage and Retrieval for Still Image and Video Databases IV, vol. 2670, pp. 239-246.

[3] Yihong Gong; Xin Liu, Video summarization using singular value decomposition, Proceedings of the IEEE Conf on Computer Vision and Pattern Recognition, Vol. 2, 2000

[4] Minerva Yeung, Boon-Lock Yeo, Bede Liu, Segmentation of video by clustering and graph analysis, Computer Vision and Image Understanding Volume 71 , Issue 1 (July 1998) 94 – 109.

[5] Taskiran, C.; Jau-Yuen Chen; Albiol, A.; Torres, L.; Bouman, C.A.; Delp, E.J., ViBE: a compressed video database structured for active browsing and search, IEEE Transactions on Multimedia, Vol.6, Iss.1, Feb. 2004.

[6] Hanjalic, A.; HongJiang Zhang, An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis, IEEE Transactions on Circuits and Systems for Video Technology, Vol.9, Iss.8, Dec 1999, 1280-1289.

[7] Ferman, A.M.; Tekalp, A.M., Two-stage hierarchical video summary extraction to match low-level user browsing preferences, IEEE Transactions on Multimedia, Vol.5, Iss.2, June 2003, 244- 256.

[8] MPEG MDS Group, Text of ISO/IEC FDIS 15938-5 Information Technology Multimedia Content Description Interface Part 5 Multimedia Description Schemes, ISO/IEC JTC1/SC29/WG11 N4205, July 2001.

[9] G. Tardini, C. Grana, R. Marchi, R. Cucchiara, Shot Detection and Motion Analysis for Automatic MPEG-7 Annotation of Sports Videos, in press in Proceedings of the 13th ICIAP, 2005.

[10] X.L. Xie, G.A. Beni, A validity measure for fuzzy clustering, IEEE Trans. on PAMI, vol. 3, n. 8, 841-846, 1991