

A Robust Distance Measure for the Retrieval of Video Objects

Berna Erol
RICOH California Research Center
2882 San Hill Rd, Suite 115
Menlo Park, CA, USA
{berna_erol@rii.ricoh.com}

Faouzi Kossentini
University of British Columbia
2356 Main Mall
Vancouver, BC, Canada
{faouzi@ece.ubc.ca}

Abstract

In this paper, we propose a new method for measuring the similarity between two arbitrarily shaped video objects. Our method is based on comparing the low-level still features of the representative planes of video objects. We demonstrate the performance of the proposed method in a shape retrieval system in which boundary- and region-based still shape features were employed to retrieve video objects. The experimental results show that i) the retrieval performance using the proposed similarity matching technique significantly outperforms the commonly used feature vector averaging technique and ii) the distance measure performs robustly when the content of the object changes in time.

1. Introduction

Object-based video retrieval is an emerging research area that has been driven by the increasing availability of arbitrarily shaped video content, due to the recent standardization of the MPEG-4 [1] object-based representation and the advancements in segmentation technology. Although there has been a significant amount of work in content-based retrieval of arbitrarily shaped still image objects, e.g., in the frameworks of trademark and 3-D object databases, the research in object-based video retrieval is still in its infancy.

Considering that digital video is a collection of still images, it is intuitive to employ still image retrieval techniques for video retrieval as well. However, the vast amount of still image data associated with even short video sequences makes this adaptation non-trivial. In the literature, there were several attempts to employ still image features for object-based video retrieval. For example, in VideoQ, color histograms, shape descriptors,

(i.e. eccentricity, first and second moments), area of the video object, and texture descriptors, (e.g., coarseness, contrast and orientation), are employed to describe video objects [2]. In another retrieval system, NeTra-V, color histograms, Gabor texture feature set, and Fourier descriptors of the curvature, centroid distance, and complex coordinate functions are used to represent the sub-objects in video scenes [3].

In the existing systems, the color, texture, and shape feature vectors of the video objects are obtained by averaging the feature vectors that belong to each temporal instant of these video objects. Then the distance between two video objects is measured by simply computing the distance between their corresponding mean feature vectors. While such a technique works well for video objects that have a fixed color, texture, and shape properties during their lifespans, it is insufficient in situations where the object properties may change significantly – for example, where an object enters or exits from a scene, or when rotation, changing light conditions, occlusion, or high motion are present. Here, we address this issue by proposing a similarity measure that is based on matching the representative frames of video objects. Our similarity matching technique is robust to the significant variations that an object may have during its existence. In the next section, we give an overview of our proposed video object matching technique. In Section 3, we present the retrieval performance obtained using the proposed technique versus the commonly used averaging technique in a shape retrieval framework.

2. A Distance measure based on the comparison of representative object planes

In a typical retrieval system, a feature vector is formed for each object and then the similarity between objects

are found by computing the distance between objects, using a measure such as Euclidian or Block distance. Similar to a video sequence being a collection of two-dimensional still images (frames), an arbitrarily shaped video object is a collection of two dimensional video object planes (VOPs). VOPs describe the shape and texture that the object has in a particular instant. Associating feature vectors for each of these object planes to capture the video object content may not be feasible considering that a one-minute sequence may contain more than 1500 video object planes. In the existing systems, this problem is addressed either by employing the feature vector of one key object plane to represent the whole video object sequence or by computing the average of the feature vectors belonging to all object planes. Such techniques work well for representing the some content of video objects that tend to remain somewhat consistent during an object’s lifespan. However, these techniques may fail to accurately represent the object’s shape content, considering the object’s shapes may change significantly during its existence, due to occlusion, motion of the articulated parts, etc. We propose to overcome this problem by employing a similarity measure that is based on matching a selected subset of video object planes that capture the different shapes that an object has in its lifespan. A straightforward way to obtain such a subset would be to temporally sample the VOPs of a video object. However, using this method, some important changes of the video object content could be missed. A much more efficient way of obtaining a subset of VOPs would be to employ one of the key VOP extraction algorithms described in the literature [4][5].

We propose to compute the similarity of two video objects, VO_A and VO_B , via comparing the feature vectors of their representative object planes. The proposed distance measure requires that for every VOP of VO_A , we find the smallest distance to any key VOP in VO_B . Then the summation of these distances is divided by the number of VOPs in VO_A to obtain the distance between two video objects as follows:

$$d(VO_A, VO_B) = \frac{1}{N} \sum_k \min_{VOP_b \in VO_B} \{d_{vop}(VOP_{a_k}, VOP_b)\},$$

where N is the number VOPs of VO_A , VOP_{a_k} is the k^{th} representative video object plane of VO_A , $d_{vop}(VOP_a, VOP_b)$ is the Euclidian distance between VOP_a and VOP_b , and computed as

$$d_{vop}(VOP_a, VOP_b) = \left\| \vec{R}_a - \vec{R}_b \right\|,$$

where \vec{R}_a and \vec{R}_b are the feature vectors of the VOP_a and VOP_b , respectively.

This distance measure is asymmetric, i.e. the distance from VO_A to VO_B , $d_{vop}(VO_A, VO_B)$, is not equal to the distance from VO_B to VO_A , $d_{vop}(VO_B, VO_A)$. Therefore, we define the final distance between two VOs by

$$D_{VO}(VO_A, VO_B) = \max\{d_{vop}(VO_A, VO_B), d_{vop}(VO_B, VO_A)\}.$$

2.1 Reducing the video object content redundancies

Since the above distance measure requires finding the distance between each key VOP of video object A to that of video object B, it could be computationally intensive, especially if the number of key VOPs is large. The number of representative VOPs obtained by the algorithms in the literature depends on how much change there is in a video object during its existence. For example, a rigid video object with no occlusion and motion could be represented with one object plane, where a high motion video object with many articulated parts would be represented with many object planes. Nevertheless, in most cases it might be desired to have an upper limit to the number of VOPs to be compared in order to limit the number of computations. We propose to reduce the number of key VOPs to a small fixed number via K-means clustering prior to computing the similarity distance. We employ the shape feature vectors for classifying the key VOPs.

Figure 1 shows the representative VOPs of two arbitrarily shaped video objects and the results of the K-means clustering where the upper limit for the number of VOPs was set to 3. The computational overhead introduced by summarization of the video content could be ignored, considering that it need only be performed once when adding a video object to the database.

3. Experimental results

Here, we demonstrate the performance of our proposed distance measure in a shape retrieval framework, where we employ a boundary based Fourier shape descriptor [6], and a region based descriptor, ART descriptor [7][8]. Our database contains 20 arbitrarily shaped video objects, coded in 2 to 3 different spatial resolutions each. The evaluation of the retrieval performance is achieved by utilizing the Normalized Modified Retrieval Rank (NMRR) measure used in the MPEG-7 standardization activity [9]. NMRR not only indicates how many of the correct items are retrieved, but

also how highly they are ranked among the retrieved items. The NMRR is in the range of [0 1] and smaller values represent a better retrieval performance. The NMRR values are smaller than 0.1 correspond to excellent retrieval performance, where 90% or more items are correctly retrieved. When the NMRR takes a value between 0.1 and 0.2, the retrieval performance is still very good, as in average more than 80% of the items are correctly retrieved. The NMRR values above 0.4 correspond to a poor retrieval performance.

The NMRR values presented in this section are obtained by averaging the retrieval results of the 4 query video objects that are shown in Figure 2. The representative VOPs of the video objects are found by using the method described in [4]. Note that the key VOP selection is performed separately on the each different resolution of the video object, resulting in a different set of key VOPs for the each resolution. Then the averaging and the propose distance measures are computed using the shape masks associated with these VOPs.

Table 1 compares the retrieval results obtained using our proposed distance measure with the distance measure commonly used in the current systems, i.e. averaging of the feature vectors. As can be seen from the table, our proposed method consistently results in better retrieval performance, i.e. lower NMRR rate. It is also possible to observe from the table that when the "News 1" object is given as a query, the results of averaging and our proposed method are identical. This is mainly due to the fact that this video object is very low motion and only one key VOP is used to summarize its content.

Figure 3 and Figure 4 present some example query results obtained by using the Fourier and ART descriptors together. Note that each of the query and database video objects contains approximately 300 VOPs and only one representative VOP for each object is shown in the figures. As can be seen from the figures, the shapes of the retrieved objects match to those of the query video objects.

Similar retrieval improvement results should be expected when the low level features, other than the shape features, are employed as well. One important point to note is that if a particular feature of a video object remains the same during its existence (e.g. its color), then the averaging technique would perform as well as the proposed method.

4. Conclusions

In this paper, we proposed a method to compute the distance between two video objects based on the distances

of their representative video object planes. Employing this measure in a shape retrieval framework resulted in a retrieval performance improvement over the averaging technique.

5. Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada.

6. References

- [1] N. Brady, "MPEG-4 Standardized Methods for the Compression of Arbitrarily Shaped Video Objects", IEEE Transactions on Circuits and Systems for Video Technology, pp. 1170-1189, 1999.
- [2] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "A Fully Automated Content-based Video Search Engine Supporting Spatiotemporal Queries", IEEE Trans. on Circuits and Systems for Video Technology, pp. 602-615, September 1998.
- [3] Y. Deng, and B. S. Manjunath, "NeTra-V: Toward an Object-based Video Representation", IEEE Transactions on CSVT, vol.8, no.5, pp. 616-627, 1998.
- [4] B. Erol and F. Kossentini, "Automatic Key Video Object Plane Selection Using the Shape Information in the MPEG-4 Compressed Domain", IEEE Transactions on Multimedia, vol. 2, no. 2, pp. 129-138, June 2000.
- [5] C. Kim, J.-N. Hwang, "Object-based video abstraction using cluster analysis", IEEE International Conference of Image Processing, 2001.
- [6] Del Bimbo, A., "Visual Information Retrieval", Morgan Kaufmann Publishers, California, 1999.
- [7] ISO/IEC JTC1/SC29/WG11, "Multimedia Content Description Interface - Part 3 Visual". Publicly available at http://mpeg.telecomitalia.com/working_documents.htm, March 2001.
- [8] B. Erol and F. Kossentini, "Similarity Matching of Arbitrarily Shaped Video Objects by Still Shape Features and Shape Deformations", IEEE Int. Conference on Image Processing, 2001.
- [9] B.S. Manjunath, J.R. Ohm, V.V. Vandevan, K. Yamada, "Color and Texture Descriptors", IEEE, Trans. on CSVT, vol. 11, no. 6, pp. 703-715, June 2001.

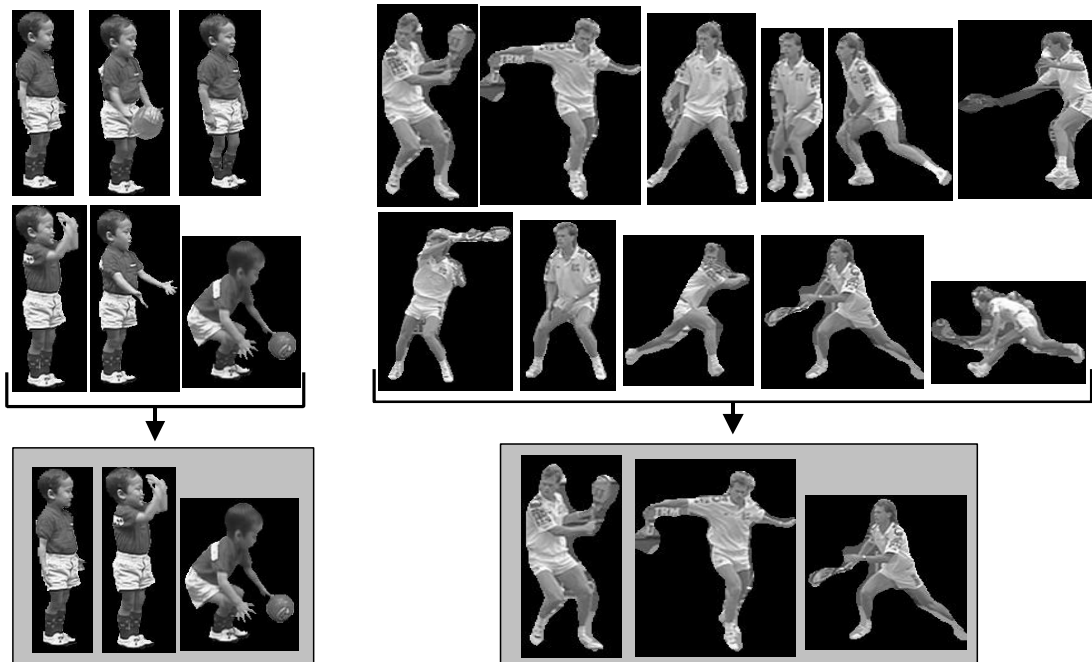


Figure 1. Further summarization of the video objects into 3 VOPs with K-means clustering.

VO name	Averaging Measure			Proposed Measure		
	NMRR Boundary based shape descriptor	NMRR Region based shape descriptor	Combined descriptors	NMRR Boundary based shape descriptor	NMRR Region based shape descriptor	Combined descriptors
Fish 1	0.45	0.60	0.51	0.10	0.30	0.18
Hall Monitor 1	0.30	0.45	0.36	0.28	0.40	0.22
News 1	0.0	0.03	0.0	0.0	0.03	0.0
Children 1	0.31	0.18	0.15	0.22	0.03	0.08
Average NMRR	0.27	0.32	0.26	0.15	0.19	0.12

Table 1. The retrieval performance employing Fourier, ART, and combination of the descriptors.

Video Object Representative Video Object Planes

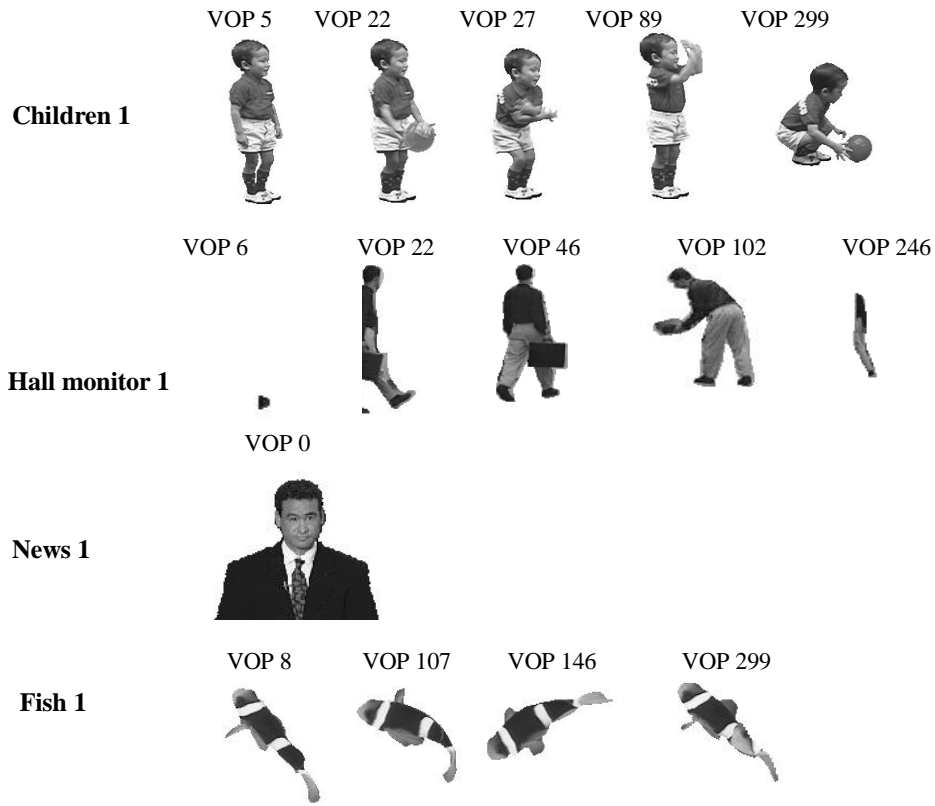


Figure 2: Examples of query video objects.



Figure 3. The shape retrieval results for the News 1 video object query.

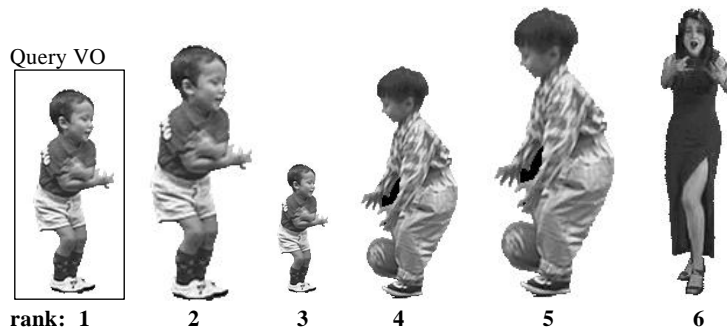


Figure 4. The shape retrieval results for the Children 1 video object query.