IMAGE SIMILARITY FOR AUTOMATIC VIDEO SUMMARIZATION

Itheri Yahiaoui, Bernard Merialdo and Benoit Huet
Institut Eurecom
Departement Communication Multimedia
BP 193 – 06904 Sophia – Antipolis- France
{Itheri.Yahiaoui, Bernard.Merialdo, Benoit.Huet}@eurecom.fr

ABSTRACT

Image similarity is a key issue for many multimedia applications. Video summarization is no exception. We have recently proposed a number of methodologies for creating visually significant summaries of videos. Our approach relies heavily on the metric which decides on whether two video key-frames are similar or not. In this paper, we compare a number of histogram representations and possible distance measures with the objective of improving the quality of video summaries.

1. INTRODUCTION

As multimedia data become increasingly available the requirements for the implementation of efficient manipulation and presentation tools becomes of critical importance. Automatic video summarization tools aim at creating with little or no human interaction shorter versions which contains the salient information of original video. The key issue here is to select what should be kept in the summary and how the relevant information can be automatically extracted. To perform this task we have developed a number of alternative algorithms [12][13]. However, our focus in this paper is on the comparison of image. Basically, we are trying to answer the following question: which representation and which measure is the most suitable for deciding on the similarity of a pair of images (or video frames).

A number of approaches have been proposed to define and identify what is the most important content in a video. However, most have two major limitations. First, evaluation is difficult, in the sense that it is hard to judge the quality of a summary, or, when a performance measure is available, it is hard to understand what its interpretation is. Secondly, while summarization of a single video has received increasing attention [3][7][10][14], little work has been devoted to the problem of multi-episode video summarization [12] which raises other interesting issues.

Existing video summarization approaches can be classified in two categories. The rule based approaches combine evidences from several types of processing (audio, video, text) to detect certain configuration of events to include in the summary. Examples of this approach are the “video skims” of the Informedia Project [7], and the movie trailers of the MoCA project [3]. The mathematically oriented approaches, on the other hand, use similarities within the video to compute a relevance value of video segments or frames. Possible relevance criteria include segments duration, inter-segment similarities, and combination of temporal and positional measures. Examples of this approach include the use of Singular Value Decomposition [14], and shot-importance measure [10]. The methods we have proposed in previous papers falls [12][13] in the later category.

The literature, and more specifically the field of content based image retrieval, provides numerous alternative methods for comparing images. These methods can be divided in three categories according to the type of feature that are employed for comparison; color, texture or shape features. In this paper we are solely concerned with approaches based on color attributes. Among the most popular we find basic color histogram [9], color constant [1], color moments [8], color tuple histograms [5], color correlograms [2], local color histogram [6] and blob histograms [4]. Some of the methods have the drawback of discarding all spatial relationship between color pixels, while others encapsulate limited spatial relationships. The later methods are either difficult to use in practice or require computationally expensive preprocessing of the image data. Blob histograms have been proposed for integration in the MPEG7 standard and have been shown to outperform most other color representation for content-based image retrieval thanks to the embedding of spatial and size information within conventional color histograms.

In this paper, we propose to study various image representation alternatives along with a number of distance measures in order to improve the quality of video summaries created automatically based on the construction method described in [12] and [13]. Section 2 presents the image representation and the distance measures under consideration. In section 3, we briefly describe the algorithms used to construct multi-episode summaries and the setting in which the experimentation took place. Experimental results are reported in section 4. Conclusions and future extensions are presented in section 6.

2. IMAGE SIMILARITY APPROACHES

We are considering two histogram representations for capturing the color distribution of the video frames as well as two distance measure for measuring the similarity between pairs of frame. There are many approaches in the literature for image similarity determination [9][1][8][4]. It seems that there are no methods which perform ideally and that each method presents some advantages which are directly dependant on the context in which the similarity is required. For example, there is a difference between comparing image for locating scene changes in a video and for identifying whether two images depict the same object or person. Here, we compare a simple region based histogram representation and a recently proposed representation for content based image retrieval called blob histograms [4].

2.1. Color Histograms

Color histograms are employed to capture the color distribution characteristics of each key-frame. The similarity between any pair of shots is computed by comparing their corresponding color histograms. This is a similar approach to the one of Swain and Ballard for content-based image retrieval [9] but with the addition of a locality constraint. In order to capture some locality information key-frames are divided in nine equal regions from each of which a color histogram is computed. As a result, characteristic key-frames are represented using a vector based on the concatenation of the nine histograms. The size of the resulting histogram is 256x9 bins.

2.2. Blob Histograms

As an alternative to the region color histogram representation, we are considering the use of blob histograms. Quian et al. [4] have recently proposed a histogram representation which instead of encoding the frequency distribution of single pixels color, uses a structuring element in order to include locality information in the histogram. The structuring element, a square in our experiments, is moved over the image and groups of pixel with uniform color within that element are called blobs. We construct blob histograms using the HSV color space. Image color pixels values are quantized into 166 HSV values and the percentage of pixel of each color within any blob of size nxn is quantized into three values: {0-33%; 34-66%; 67-100%}. The size of a blob is therefore 166x3 bins.

2.3. Manhattan Distance

The familiar L1 norm can be written as follows:

$$L_1(P_D, P_M) = \sum_i \left| P_D(i) - P_M(i) \right|$$

with the normalized histograms $P_D(i)$ and $P_M(i)$ composed of i distinct bins each.

2.4. Euclidean Distance

Similarly the L2 norm is defined as follows:

$$L_2(P_D, P_M) = \sqrt{\sum_i \left(P_D(i) - P_M(i) \right)^2}$$

3. VIDEO SUMMARY CONSTRUCTION AND EVALUATION

A key issue in automated summary construction is the evaluation of the quality of the summary with respect to the original data. Since there is no ideal solution a number of alternative approaches are available. With user based evaluation methods, a group of user is asked to provide an evaluation of the summaries. Another method is to ask a group of users to accomplish certain tasks (i.e. answering questions) with or without the knowledge of the summary, and measure the effect of the summary on their performance. Alternatively, for summaries created using a mathematical criterion, the corresponding value can be used directly as a measure of quality. However, all these evaluation techniques present drawbacks; User-based one's are difficult and expensive to set-up and their bias is non trivial to control, whereas mathematically based one's are difficult to interpret and compare to human judgment.

Our approach for the automatic creation and evaluation of summaries is based on the Simulated User Principle [13]. This method addresses the problem related to the evaluation of the summary and is applicable to both cases of single video and multi-episode videos. The summary construction and evaluation are both inspired from the following scenario:

- The user views all the summaries,
- He is shown a randomly chosen excerpt of a randomly chosen video,
- He is then asked to guess which video this excerpt was extracted from.

The simulated behavior of the user is the following:

- If the excerpts contains images which are similar to one or several images in a single summary, he will provide the corresponding video as an answer (but it is not certain that this is the correct answer),
- If the excerpt contains images which are similar to images in several summaries, the situation is ambiguous and the user cannot provide a definite answer,
- If the excerpt contains no image which is similar to any image in any summary, the user has no indication and cannot provide a definite answer.

The performance of the user in this experiment is the percentage of correct answers that he is able to provide when he is shown all possible excerpts of all videos. Note that only in the first case described above is the user able to identify a particular video. But this answer might not be necessarily correct, because an image in an excerpt of one video can be similar to an image in the summary of another video.

4. EXPERIMENTAL RESULTS

In order to determine a representation of the entire set of images extracted from a video, we have decided to compare the feature vectors constructed from the region histograms on one side and those build using the blob approach on the other side. In parallel with this study we evaluated which of either the L1 norm and Euclidean distance is the most appropriate measure given the representation under consideration. Additionally, this study aims at, given the most suitable representation (blob or region histogram) associated with the adequate norm (L1 or L2), determining the threshold for which two video frames (images) are similar.

Our experimental setup was the following. First, we compute the feature vectors of video frames using region histograms and various histogram blob sizes. For our experiments we have tested the following dimension for square blobs: 3, 5, 7, 9, 11, 13, 15, 20, 30, 40, 50, and 100. Then, four hundred image pairs (video frames) are randomly selected from videos with as only constraint that selected pairs are evenly distributed over a number of histogram distance ranges (i.e. 0-100, 101-200, etc…).

For all image pairs previously selected, a small number of users (4) are asked to determine whether both images are similar or not. That is to say that a human judgment of similarity is associated with all 400 video frame pairs. Then, the various distances between the selected pairs of video frames are computed, based on all possible combination of representations (region and blob histograms) and histogram distance measures (Manhattan and Euclidean distances).

Image pairs are then sorted according their distances and assigned to categories corresponding to a fixed number of distance ranges, for the various representation and measures, in order to determine the most suitable threshold for similarity.

It is then possible to compute the error rate corresponding to each threshold range category. This error rate corresponds to the number of non similar images for which the distance is below the corresponding category threshold and the number of similar pairs for which have a distance greater than the current threshold with respect to all the pairs under consideration. This is repeated for a number of threshold values.

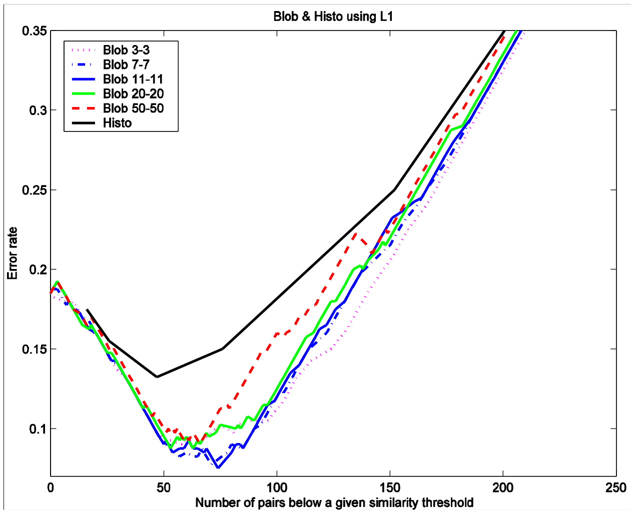

Figure 1

The plot shown in figure 1 represents the error rate for the task of image classification according to similarity as a function of the number of image pairs for which the distance is below a given threshold. It is worth pointing out that the blob representation provides lower error rate than standard region histogram comparison. Additionally, the best performance for the comparison of video frame of 320x240 is obtained for blob histograms constructed with a blob “window” size of 11x11 pixels.

The second figure represents the distance comparison results. In the plot the minimum error rate for blob sizes varying from 3x3 to 100x100 for both the L1 and L2 norms. The region histogram results are presented as blob of size 0x0 (0 on the horizontal axis) on this figure. Thanks to this study we have determined that for blob sizes smaller or equal to 40x40 the best results are obtained with the L1 norm. In addition, this plot shows that it is recommended to represent video frames using color blob histograms based on a window size of either 11x11 or 13x13. The thresholds corresponding to the minimal error of 0.075 for blob sizes of 11-11 and 13-13 are respectively 455 and 520. Out of the 400 image pairs 74 have a distance below 455 for blobs 11x11 and 76 have a distance below 520. We have opted for the smaller size solution as it is the least computationally expense of the two during histogram construction.

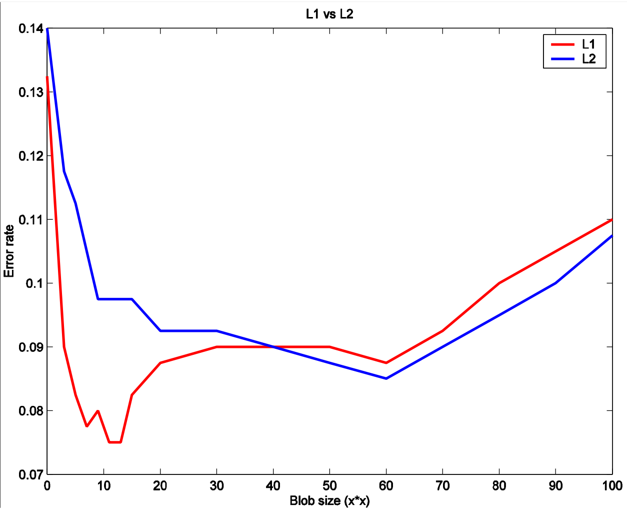

Figure 2

Our simulated user principle, upon which the multi-video based construction is based, relies principally on the image similarity measure. In order to validate our finding on the various image representation and comparison metrics, we study further the quality of the classification (similar or not) over all consecutive frames of the video. It is obvious that consecutive frames within a shot are very similar, and should therefore help us in determining the most appropriate threshold for image similarity. Figure 3 shows the distance of all consecutive frames for three of our test videos. From this plot we can see that most consecutive image have a distance of approximately 200. However this is too strict for judging similarity between image pairs. A threshold value of 300 or 400 is probably far more suitable. In order to refine the threshold choice, we have randomly selected pairs of images for which the blob histogram distance is close to various prospective threshold values. Those image pairs were then evaluated by real users in order to withdraw inappropriate threshold values. Finally, we focused on an image similarity threshold of 350 for multi-video summary construction.

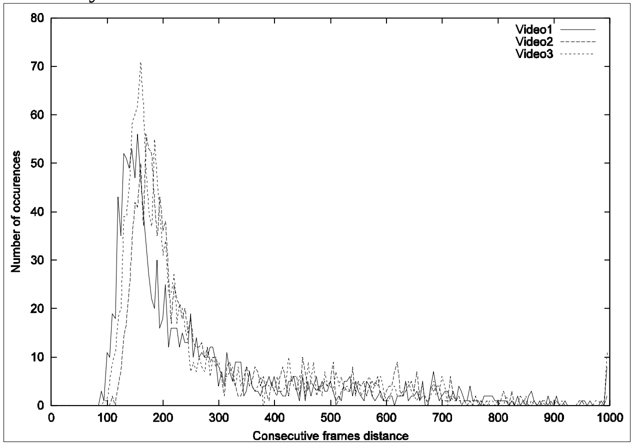

Figure 3

So far, our study of the most appropriate representation and distance measure for image similarity has led us to the use of color blob histogram of size 11x11 in conjunction with the Manhattan distance. Our aim is the improvement of the visual quality of video summaries. In order to determine the performance amelioration of the new representation we have build and evaluated summaries for the various approaches described in this paper.

Table 1, provides the evaluation results of our algorithms using either blob histograms or the region histogram. Please note that although evaluation is performed using various excerpt duration (1 to 40 seconds) the construction was effected based on excerpts of 4 seconds. Independently of the length of the excerpts, blob histograms perform better than the region histograms and that this difference is non negligible. Theses results show that there is a significant improvement our summary's quality thanks to the use of blob histograms as underlying representation for video frames.

	Excerpt duration in seconds						
	1	4	6	8	10	20	40
Histo	17.84	25.25	29.87	33.36	36.82	46.70	54.06
Blobs	30.98	43.84	50.31	55.98	60.43	69.20	73.25

Table 1

5. CONCLUSION

We have presented a study which aims at providing a better understanding about the importance of both the representation and distance for determining image similarity. This comparison was performed in the particular context of image similarity for video summary construction. Our results showed that the blob histograms of size 11x11 used in combination with the Manhattan distance produced improved video summaries compared with conventional histograms with Euclidean distance. We have recently initiated a serie of experiments aimed at identify the "semantic" limit of color histogram based approaches for selecting the appropriate frame for video summarization. Thanks to the results of this study we expect to be able to identify which image processing and computer vision techniques should be employed to further enhance the quality of automated video summary construction techniques.

6. ACKNOLEDGEMENTS

This research was supported by Eurecom's industrial members: Ascom, Bouygue, Cegetel, France Telecom, Hitachi, ST Microelectronics, Motorola, Swisscom, Texas Instruments and Thales.

7. REFERENCES

[1] Funt B. V. and G. D. Finlayson, "Color constant color indexing", IEEE trans. PAMI, 17(5), pp 522-529, 1995.

[2] Huang J., S. R. Kumar, et al., "Image indexing using color correlograms", Proc. CVPR, pp 762-768, 1997.

[3] Lienhart R., S. Pfeiffer and W. Effelsberg. "Video Abstracting". In Communications of ACM, vol. 40, no. 12, pp 54-62, December 1997.

[4] Qian, R.J., P.J.L. Van Beek and M.I. Sezan. Image retrieval using blob histograms. IEEE International Conference on Multimedia and Expo, ICME 2000, Vol: 1, pp 125 -128, 2000.

[5] Rickman and J. Stonham, "Content-based image retrieval using color tuple histograms", SPIE Proc. Storage and Retrieval of Image and Video DB, vol. 2670, pp 2-7, 1996.

[6] Rubner Y. et al., "The earth mover's distance, multi-dimensional scaling, and color-based image retrieval", Proc. ARPA Image Understanding WorkShop, 1997.

[7] Smith M.A. and T. Kanade. "Video skimming and characterization through the combination of image and language understanding". IEEE Int. Workshop on Content-Based Access of Image and Video DB, pp. 61-70, 1998.

[8] Stricker M. and M. Orengo, "Similarity of color images", SPIE Proc. Storage qnd Retrieval of Image and Video Databases, vol. 2420, page(s) 381-392, 1996.

[9] Swain, M. and G. Ballard. Color Indexing. International Journal of Computer Vision, 7(1), page(s) 11-32, 1991.

[10] Uchihashi S. and J. Foote. "Summarizing Video Using a Shot Importance Measure and a Frame-Packing Algorithm". IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 3041-3044, 1999.

[11] Vasconcelos N. and A. Lippman. "Bayesian Modeling of Video Editing and Structure: Semantic Features for Video Summarisation and Browsing". IEEE Int. Conference on Image Processing, vol. 3, pp. 153-157, 1998.

[12] Yahiaoui I., B. Merialdo and B. Huet, "Automatic Summarization of Multi-episode Videos with the Simulated User Principle", Workshop on MultiMedia Signal Processing, Oct. 3-5, 2001, Cannes, France.

[13] Yahiaoui I., B. Merialdo and B. Huet, "Generating Summaries of Multi-Episodes Video", International Conference on Multimedia & Expo, August 22-25, 2001 Tokyo, Japan.

[14] Yihong Gong; Xin Liu. "Generating Optimal Video Summaries". IEEE International Conference on Multimedia and Expo, vol. 3, pp. 1559-1562, 2000.