I. Quality Metrics

In this section we first overview some of the objective methods for summary evaluation proposed in the literature. Next we present a new quality metric that addresses the problems of the current metrics.

A. Literature Review

[1] describes existing evaluation methods for video summarization and groups them into three different categories: result description, objective metrics, and user studies. In the first category, the proposed technique is applied to a few video sequences and the generated summaries are presented without any comparison to other techniques. The second category of evaluation methods usually involve defining a quality function that is computed from the extracted key frames and the original sequence. These methods try to model the human perception of quality but there is no justification. Moreover, the proposed metric is closely related to the technique used for extracting the set of key frames. Finally, methods in the last category involve subjective studies where users are asked to judge the quality of the generated summary. These methods are regarded the most useful form of evaluation however they are not widely employed due to their difficult setup. In addition subjective studies cannot be automatized.

Authors in [2], define a distortion metric to evaluate the quality of the key frames extracted by their algorithms. Since in their algorithms, each key frame k_i represents a time segment $[\tau_{i-1}, \tau_i]$, all original frames within this time segment are compared against k_i. To compute the distance between two frames, EMD [3] is used for short sequences (≤ 1 min) due to its complexity and for long videos L_1 is applied. It is not clear how this method can be applied for comparing summaries generated by different algorithms since it relies on segment boundaries. In addition, EMD is computationally intensive while L_1 is very sensitive to small movements of the objects in the scene.

In [4], key frames in the summary sequence are repeated in a Zero Order Hold (ZOH) fashion before the distortion from the original sequence can be computed. Authors use a weighted combination of Color Layout Descriptors (CLD) [5] and Motion Activity Descriptors (MAD) [6] to compute the distance between individual frames. Same authors present a more general framework in [7]. Again key frames are repeated using ZOH. All frames in the original and the reconstructed sequence are projected onto a subspace of lower dimension through Principle Component Analysis (PCA) [8]. A Euclidian distance metric is used to compute the distance between individual video frames. By using ZOH the authors implicitly assumes that the key frames are the first frame in their respective shot. Thus, in effect for general videos the metric may compare frames from different shots which are totally different in content. Another drawback for the proposed method is the intensive computation involved in PCA.

A semi-Hausdroff distance metric is used in [9] to measure the fidelity of a set of key frames and the set of original frames. The Hausdroff distance [10] is used for finding the distance between two sets of points. Each point in set A is matched to its closes point in set B. The maximum over all these distances is reported as the distance between A and B. To measure the distance between points (frames in this case) different metrics can be used. Color histograms are proposed for this purpose due to their efficiency. However, this results in a metric which is not sensitive to object movements.

B. Metric Design

Based on the above discussion of current quality metrics for summarization applications, we first outline some of the desirable characteristics of a good quality metric:

- It should be easily computable with low cost.
- It should be independent of the method used for extracting the key frames.
- No extra information other than the original sequence and the extracted key frames should be required for computing the metric. For example it should not rely on shot boundary information.
- A good summary represents the original sequence with little redundancy thus conveying the most information. The quality metric should capture this redundancy in an effective way.
- The sensitivity of the quality metric to little changes in the scene (e.g. object movements) should to be adjustable according to the desired degree of detail (summary length).

To address the above issues, we present a quality metric that measures the redundancy R, for a set of extracted key frames S. To calculate redundancy, we measure the similarity between successive key frames using local color histograms. The basic idea is to augment the color histogram with some location information in order to make a motion sensitive metric. Assuming all frame are of size $W \times H$, we define a window of size $L \times H$, where L can be adjusted according to the summary length. Each frame is tiled with this window and a color histogram is calculated for each part resulting in W/L local histograms.

In addition to these local histograms, W/L average histograms are computed from all frames. These average histograms are then subtracted from their respective local histograms in order to remove the background effect. To compute the distance between two key frames K_i, and K_j, their local histograms are subtracted from each other and the results are added up:

$$d(K_i, K_j) = \sum_{k=1}^{W/L} \|H_k^i - H_k^j\|$$

The effective redundancy of a sequence is then computed as sum of the distances between successive key frames, where n is the number of key frames in the sequence:

$$R(S) = \sum_{i=1}^{n-1} d(K_i, K_{i+1})$$

The choice for window width L, depends on the summarization factor. For longer summaries a smaller value is chosen the metric is more sensitive to object movements.

REFERENCES

[1] B. Troung and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 3, no. 1, February 2007.

[2] T. Liu and J. Kender, "Computational approaches to temporal sampling of video sequences," *ACM Transactions on Multimedia Computing, Communication, and Applications*, vol. 3, no. 2, May 2007.

[3] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Proc. of the Sixth International Conference on Computer Vision*, Bombay, India, January 1998, pp. 59–66.

[4] Z. Li, K. Katsaggelos, and B. Gandhi, "Temporal rate-distortion based optimal video summary generation," in *Proc. of the International Conference on Multimedia and Expo (ICME'03)*, Baltimore, MA, July 2003, pp. 693–696.

[5] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, June 2001.

[6] S. Jeannin and A. Divakaran, "Mpeg-7 visual motion descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, June 2001.

[7] Z. Li, G. Schuster, K. Katsaggelos, and B. Gandhi, "Rate-distortion optimal video summary generation," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1550–1560, October 2005.

[8] D. Lay, *Linear Algebra and its Applications*. New York: Addison-Wesley, 2000.

[9] A. Divakaran, R. Radhakrishnan, and K. Peker, "Motion activity-based extraction of key-frames from video shots," in *Proc. of IEEE International Conference on Image Processing (ICIP'02)*, Rochester, NY, September 2002.

[10] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdroff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, Sep 1993.