Clustering Videos by Location

Florian Schroff[1]

C. Lawrence Zitnick[2]

Simon Baker[2]

[1] Visual Geometry Group
University of Oxford

[2] Microsoft Research
Microsoft Corp.

1 Overview

Location is a useful source of information for a variety of tasks. Just as users may want to tag and search their personal photo collections and videos for specific people, they may also want to specify a location to further narrow down the search. Users may also want to browse videos by location, annotate locations, or create location specific compilations.

We propose an algorithm that uses visual information to cluster video shots by the location in which they were captured. We demonstrate our algorithm on both home videos and professionally edited footage such as sitcoms [1, 8]. In the context of home movies, location generally means a specific room in the house, or a frequently visited place outside, such as in the garden, or at the local park. In the context of sitcoms, location means a film "set" such as the coffee shop in the sitcom "Friends."

Our algorithm first breaks the video into shots using a simple color histogram-based algorithm [3]. It is important to fully represent the visual varieties in each shot. We empirically compare three approaches: (1) Using a single keyframe, the middleframe of the shot. (2) Using multiple keyframes sampled uniformly in time from the video [7]. And (3) Stitching the frames into a mosaic [1]. We illustrate these three choices in Figure 1. We found the second approach to perform the best.

Next we need to measure the similarity between each pair of keyframes in the shot representation. We considered two approaches: (1) bag of visual words based [2]. (2) feature matching based [6]. We found the first approach to perform far better than the second.

The next step in the design of our algorithm is the core clustering algorithm. Again, we considered several choices: (1) k-means, (2) a "connected components" algorithm, (3) a spectral clustering algorithm [5], and (4) a model-based algorithm [4] using an energy function that is specifically designed to model the expected shape of clusters for the task at hand. We found the final approach (described in Section 2) to peform the best.

As subsequent shots in a video are likely to have been captured at the same location it is reasonable to incorporate this prior knowledge into the clustering process. The final component in our algorithm is to add a temporal prior, which significantly improves performance, particularly for professionally edited video.

We provide quantitative empirical evaluations on both home videos and professionally edited content (4 episodes of the sitcom "Friends") to justify each choice made in the design of our algorithm. These evaluations are performed using manually-specified ground-truth location labels.

2 Cluster Model

Given the shot representation and the texton- or feature-based similarity measure we would like to develop a clustering algorithm that allows us to model the likely cluster shapes, where each shot represents a node in the adjacency graph defined by the similarity measure. An approach that allows us to do this is the "model-based" approach of [4]. In this approach, almost any energy function can be used. This freedom to choose an energy function allows us to encode a preference for a certain shape of clusters. Our choice of energy function is based on the observation that each viewpoint is close to a number of others and encourages locally well connected clusters. Specifically, although clusters can be "elongated" in shape they need to be well connected. This differs from standard k-means where usually ball shaped clusters are assumed, or standard agglomerative clustering where also either very compact or very loosely connected clusters are assumed.

Assume that the graph has already been partitioned into a set of disjoint clusters $\{C_1, C_2, \ldots\}$. The cluster energy is then defined as a sum of

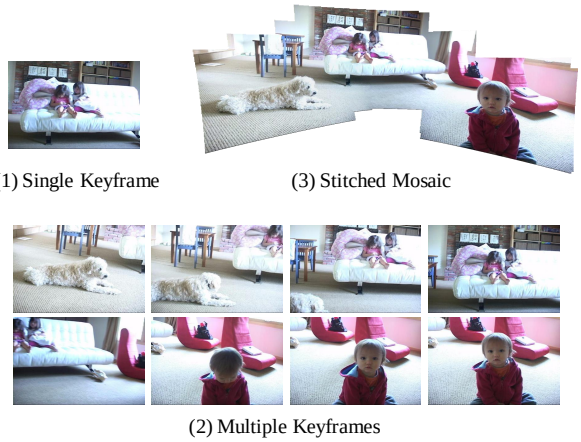

Figure 1: **Shot Representation**: (a) Three approaches to representing a video shot: (1) use a single keyframe, (2) use multiple keyframes, and (3) stitch the keyframes into a mosaic. (b) and (c) Empirical results show the multiple keyframe approach to perform slightly better than the other approaches.

energies, one for each cluster:

$$E_{\text{Cluster}} = \sum_i \text{MST}(C_i^N). \quad (1)$$

In this equation $\text{MST}(C_i^N)$ is the length of the minimum spanning tree (MST) of C_i^N and:

$$C_i^N = C_i^{N-1} - MST(C_i^{N-1}), \quad C_i^1 = C_i \quad (2)$$

is a recursive definition which says that C_i^N should be computed by removing all of the edges in the minimum spanning tree (MST) from C_i^{N-1}; *i.e.* C_i^N is the graph obtained after removing $N-1$ MSTs in sequence from C_i. In summary, the energy cost of a cluster $\text{MST}(C_i^N)$ is the length of its MST, after having previously removed $N-1$ MSTs. In this definition, $N = \alpha(|C_i| - 1)$ is a constant proportion of the size of the cluster $|C_i| - 1$.

[1] A. Aner and J. Kender. Video Summaries through Mosaic-Based Shot and Scene Clustering. In *European Conference on Computer Vision*, 2002.

[2] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004.

[3] B. Guensel, A. Ferman, and A. Tekalp. Temporal Video Segmentation Using Unsupervised Clustering and Semantic Object Tracking. In *SPIE*, 1998.

[4] S. Kamvar, D. Klein, and C. Manning. Interpreting and Extending Classical Agglomerative Clustering Algorithms using a Model-Based Approach. In *International Conference on Machine Learning*, 2002.

[5] A.Y. Ng, M. Jordan, and Y. Weiss. On spectal clustering: Analysis and an algorithm. *Neural Information Processing Systems*, 14, 2002.

[6] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition*, 2006.

[7] F. Schaffalitzky and A. Zisserman. Automated Location Matching in Movies. In *Computer Vision, Image Understanding*, 2003.

[8] M.M. Yeung and B.L. Yeo. Time-constrained clustering for segmentation of video into story units. In *International Conference on Pattern Recognition*, 1996.