Assessing Scene Structuring in Consumer Videos

Daniel Gatica-Perez[1], Napat Triroj[2], Jean-Marc Odobez[1],
Alexander Loui[3], and Ming-Ting Sun[2]

[1] Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland
[2] University of Washington, Seattle WA, USA
[3] Eastman Kodak Company, Rochester NY, USA

Abstract. Scene structuring is a video analysis task for which no common evaluation procedures have been fully adopted. In this paper, we present a methodology to evaluate such task in home videos, which takes into account human judgement, and includes a representative corpus, a set of objective performance measures, and an evaluation protocol. The components of our approach are detailed as follows. First, we describe the generation of a set of home video scene structures produced by multiple people. Second, we define similarity measures that model variations with respect to two factors: human perceptual organization and level of structure granularity. Third, we describe a protocol for evaluation of automatic algorithms based on their comparison to human performance. We illustrate our methodology by assessing the performance of two recently proposed methods: probabilistic hierarchical clustering and spectral clustering.

1 Introduction

Many video browsing and retrieval systems make use of scene structuring, to provide non-linear access beyond the shot level, and to define boundaries for feature extraction for higher-level tasks. Scene structuring is a core function in video analysis, but the comparative performance of existing algorithms remains unknown, and common evaluation procedures have just begun to be adopted.

Scene structuring should be evaluated based on the nature of the content. (e.g. videos with "standard" scenes like news programs [3], or created with a storyline like movies [10]). In particular, home videos depict unrestricted content with no storyline, and contain temporally ordered scenes, each composed of a few related shots. Despite its non-professional style, home video scenes are the result of implicit rules of attention and recording [6, 4]. Home filmmakers keep their interest on their subjects for a finite duration, influencing the time they spend recording individual shots, and the number of shots captured per scene. Recording also imposes temporal continuity: filming a trip with a non-linear temporal structure is rare [4]. Scene structuring can then be studied as a clustering problem, and is thus related to image clustering and segmentation [9, 5].

The evaluation of a structuring algorithm assumes the existence of a ground-truth (GT) at the scene level. At least two options are conceivable. In the first-party approach, the GT is generated by the content creator, thus incorporating specific context knowledge (e.g. place relationships) that cannot be automatically extracted by current means. In contrast, a third-party GT is defined by a subject not familiar with the content [4]. In this case, there still exists human context

understanding, but limited to what is displayed. Multiple cues ranging from color coherence, scene composition, and temporal proximity, to high-level cues (recognition of objects/places) allow people to identify scenes in home video collections.

One criticism against third-party GTs is the claim that, as different people generate distinct GTs, no single judgement is reliable. A deeper question that emerges is that of consistency of human structuring of videos, which in turn refers to the general problem of perceptual organization of visual information[1]. One could expect that variations in human judgement arise both from distinct perceptions of a video scene structure, and from different levels of granularity in it [5]. Modeling these variations with an appropriate definition of agreement would be useful to compare human performance, and to define procedures to evaluate automatic algorithms. Similar goals have been pursued for image segmentation [5] and clustering [9], but to our knowledge work on videos has been limited.

We present a methodology to evaluate scene structuring algorithms in consumer videos. We first describe the creation of a corpus of 400 human-generated video scene structures extracted from a six-hour video database (Section 2). We then present a set of similarity measures that quantify variations in human perceptual organization and scene granularity (Section 3). The measures can be used to assess human performance on the task (Section 4), but they are also useful to evaluate automatic algorithms, for which we introduce an evaluation protocol (Section 5). Finally, the protocol is applied to compare the performance of two recent methods (Section 6). Section 7 provides some concluding remarks.

2 Video Scene Structure Corpus

2.1 Home video database

The data set includes 20 MPEG-1 videos, each with duration between 18-24 min. [4]. While relatively small (six hours), the set is representative of the genre, depicting both indoor (e.g. family gatherings and weddings), and outdoor (e.g. vacations) scenes. A manual GT at the shot level resulted in 430 shots. The number of shots per video substantially varies across the set (4-62 shots); see Fig. 2(a)).

2.2 Tools for scene structuring

We define a video structure as composed of four levels (video clip, scene, shot, and subshot) [4]. Home video shots usually contain more than one appearance, due to hand-held camera motion, so subshots are defined to be intra-shot segments with approximately homogeneous appearance. A shot can then be represented by a set of key-frames (thumbnails) extracted from each of its subshots.

The amount of time required for human scene structuring is prohibitive when subjects deal with the raw videos. Providing a GUI with video playback and summarized information notably reduces the effort, but remains considerable for long

[1] Perceptual organization is "a collective term for a diverse set of processes that contribute to the emergence of order in the visual input" [2], and "the ability to impose structural organization on sensory data, so as to group sensory primitives arising from a common underlying cause" [1]. In computer vision, perceptual organization research has addressed image segmentation, feature grouping, and spatio-temporal segmentation, among other problems, using theories from psychology (e.g. Gestalt).

videos due to video playing. In this view, we developed a GUI in which users were not displayed any raw videos, but only their summarized information (Fig. 1). Subshots and key-frames were automatically extracted by standard methods [4], and thumbnails were arranged on the screen in columns to represent shots. As pointed out in [8], images organized by visual similarity can facilitate location of images that satisfy basic requirements or solve simple tasks. In our case, the natural temporal ordering in video represents a strong cue for perceptual organization. In the GUI, a scene is represented by a list of shot numbers introduced by the user via the keyboard, so the scene structure is a partition of the set of all shots in a video, created by the user from scratch. Finding the scenes in a video depends of its number of shots, and it takes a couple of minutes in average.

2.3 The task

A very general statement was purposedly provided to the subjects at the beginning of the structuring process: "group neighboring shots together if you believe they belong to the same scene. Any scene structure containing between one and as many scenes as the number of shots is reasonable". Users were free to define in their own terms both the concept of scene and the appropriate number of scenes in a video, as there was not a single correct answer. Following [5], such broad task was provided in order to force the participants to find "natural" video scenes.

Fig. 1. Scene structuring tool (detail). Each column of thumbnails represents a shot.

2.4 Experimental setup

Participants. A set of 20 university-level students participated in the experiments. Only two of the subjects had some knowledge in computer vision.
Apparatus. For each video in the database, we created the video structures as described in Section 2.2, using thumbnails of size 88×60 pixels. We used PCs with standard monitors (17-inch, 1024×768 resolution), running Windows NT4.
Procedure. All participants were informed about the purpose of the experiment and the GUI use, and were shown an example to practice. As mentioned earlier, no initial solution was proposed. Each person was asked to find the scenes in all 20 videos, and was asked to take a break if necessary. Additionally, in an attempt to refresh the subjects' attention on the task, the video set was arranged so that the levels of video complexity -as defined by the number of shots- was alternated. A total of 400 human-generated scene structures were produced in this way.

3 Measuring Agreement

If a unique, correct scene structure does not exist, how can we then assess agreement between people? Alternatives to measure agreement in image sets [9] and video scenes [4] have been proposed. By analogy with natural image segmentation [5], here we hypothesize that variations in human judgement of scene structuring can be thought of as arising from two factors: (i) distinct perceptual organization of a scene structure, where people perceive different scenes altogether, so shots are grouped in completely different ways, and (ii) distinct granularity in a scene structure, which generates structures whose scenes are simply refinements of each other. We discuss both criteria to assess consistency in the following subsections.

3.1 Variations in perceptual organization

Differences in perceptual organization of a scene structure, that is, cases in which people observe completely different scenes, are a clear source of inconsistency. A definition of agreement that does not penalize granularity differences was proposed in [5] for image segmentation, and can be directly applied to video partitions. Let S_i denote a scene structure of a video (i.e., a partition of the set of shots, each assigned to one scene). For two scene structures S_i, S_j of a K-shot video, the *local refinement error* (LRE) for shot s_k, with range $[0, 1)$, is defined by

$$LRE(S_i, S_j, s_k) = ||R(S_i, s_k) \backslash R(S_j, s_k)||/||R(S_i, s_k)||, \tag{1}$$

where $\backslash$ and $||\cdot||$ denote set difference and cardinality, respectively, and $R(S_i, s_k)$ is the scene in structure S_i that contains s_k. On one side, given shot s_k, if $R(S_i, s_k)$ is a proper subset of $R(S_j, s_k)$, $LRE = 0$, which indicates that the first scene is a refinement of the second one. On the other side, if there is no overlap between the two scenes other than s_k, , $LRE = (||R(S_i, s_k)|| - 1)/||R(S_i, s_k)||$, indicating an inconsistency in the perception of scenes.

To obtain a global measure, LRE has to be made symmetric, as $LRE(S_i, S_j, s_k)$ and $LRE(S_j, S_i, s_k)$ are not equal in general, and computed over the entire video. Two overall measures proposed in [5] are the *global* and *local consistency errors*,

$$GCE(S_i, S_j) = \frac{1}{K} \min \left\{ \sum_k LRE(S_i, S_j, s_k), \sum_k LRE(S_j, S_i, s_k) \right\}, \tag{2}$$

$$LCE(S_i, S_j) = \frac{1}{K} \sum_k \min \{LRE(S_i, S_j, s_k), LRE(S_j, S_i, s_k)\}. \tag{3}$$

To compute GCE, the LREs are accumulated for each direction (i.e. from S_i to S_j and vice versa), and then the minimum is taken. Each direction defines a criterion for which scene structure refinement is not penalized. On the other hand, LCE accumulates the minimum error in either direction, so structure refinement is tolerated in any direction for each shot. It is easy to see that $GCE \geq LCE$, so GCE constitutes a stricter performance measure than LCE [5].

3.2 Variations in structure granularity

The above measures do not account for any differences of granularity, and are reasonably good when the number of detected scenes in two video scene structures

is similar. However, two different scene structures (e.g. one in which each shot is a scene, and one in which all shots belong to the same scene) produce a zero value for both GCE and LCE when compared to any arbitrary scene structure. In other words, the concept of "perfect agreement" as defined by these measures conveys no information about differences of judgment w.r.t. the number of scenes. In view of this limitation, we introduce a revised measure that takes into account variations on the number of detected scenes, by defining a weighted sum,

$$GCE'(S_i, S_j) = \alpha_1 GCE(S_i, S_j) + \alpha_2 C(S_i, S_j), \tag{4}$$

where $\sum_i \alpha_i = 1$, and the correction factor $C(S_i, S_j) = \frac{|N(S_i) - N(S_j)|}{N_{max}}$, where $N(S_i)$ is the number of scenes detected in S_i, and N_{max} is the maximum number of scenes allowed in a video (K). A similar expression can be derived for LCE'.

4 Human Scene Structuring

The discussed measures were computed for all pairs of human-generated scene structures for each video in the data set. Note that, as shots are the basic units, partitions corresponding to different videos are not directly comparable. Fig. 3(a) shows the distributions of GCE and LCE between pairs of human-generated scene structures of the same video. All distributions show a peak near zero, and the error remains low, with means shown in Table 1. It is also clear that GCE is a harder measure than LCE. Given the measures that only penalize differences in perceptual organization, people produced consistent results on most videos on the task of partitioning them into an arbitrary number of scenes.

However, the variation in performance with respect to the number of detected scenes -not directly measured by GCE/LCE- is considerable. Fig. 2(b) displays the mean and standard deviation of the number of detected scenes for each video. The videos are displayed in increasing order, according to the number of shots they contain (Fig. 2(a)). As a general trend, videos with more shots produce larger variation in the number of detected scenes. Referring to Fig. 3(a), the strong peaks near zero are somehow misleading, as it is obvious that human subjects did not produce identical scene structures. The distribution of the new performance measures (GCE' and LCE') for weights $\alpha_1 = 0.85, \alpha_2 = 0.15$ are shown in Fig. 3(b). The weights were chosen so that the weighted means of GCE and C approximately account for half of the mean of GCE'. For the new measures, the distributions no longer present peaks at zero. The errors are higher, as they explicitly penalize differences in judgement regarding number of scenes.

Overall, given the small dataset we used, the results seem to suggest that (i) there is human agreement in terms of perceptual organization of videos into scenes, (ii) people present a large variation in terms of scene granularity, and (iii) the degree of agreement in scene granularity depends on the video complexity.

5 Evaluation Protocol

To evaluate an automatic method by comparing it to human performance, two issues have to be considered. First, the original measures (GCE/LCE) are useful for comparison when the number of scenes in two scene structures is similar. This

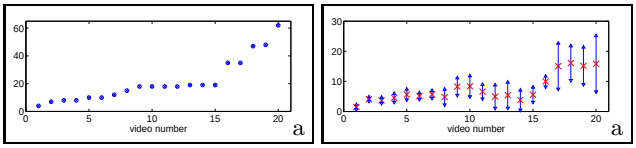

Fig. 2. (a) Number of shots per video in the database (in increasing order); (b) mean and standard deviation of number of scenes detected by people for each video.

is convenient when the number of scenes is a free parameter that can be manually set, as advocated by [5]. However, such procedure would not measure the ability of the algorithm to perform model selection. For this case, we think that the proposed measures (GCE'/LCE') are more appropriate. Second, the performance of both people and automatic algorithms might depend on the individual video complexity.

In this view, we propose to evaluate performance by the following protocol [7]. For each video, let S_A denote the scene structure obtained by an automatic algorithm, and S_j the j-th human-generated scene structure. We can then compute $GCE'(S_A, S_j)$ for all people, rank the results, and keep three measures: minimum, median, and maximum, denoted by $GCE'_{min}(S_A, S_j)$, $GCE'_{med}(S_A, S_j)$, and $GCE'_{max}(S_A, S_j)$, respectively. The minimum provides an indication of how close an automatic result is to the nearest human result. The median is a fair performance measure, which considers all the human responses while not being affected by the largest errors. Such large errors are considered by the maximum. An overall measure is computed by averaging the GCE' measures over all the videos. To compute the same measures among people, for each video, the three measures are computed for each subject against all others, and these values are averaged over all subjects. The overall performance is computed by averaging over all videos. To visualize performance, it is useful to plot the distributions of GCE' and LCE', obtained by comparing automatic and human-generated scene structures, as in Fig. 3. Finally, to compare two algorithms, the described protocol can be applied to each algorithm, followed by a test for statistical significance.

6 Assessing Automatic Algorithms

6.1 The algorithms

We illustrate our methodology on two recently proposed algorithms based on pair-wise similarity. For space reasons, we briefly describe the algorithms here.

The first algorithm is probabilistic hierarchical clustering (PHC) [4]. It consists of a sequential binary Bayes classifier, which at each step evaluates a pair of video segments and decides on merging them into the same scene according to Gaussian mixture models of intra- and inter-scene visual similarity, scene duration, and temporal adjacency. The merging order and the merging criterion are based on the evaluation of a posterior odds ratio. The algorithm implicitly performs model selection. Standard visual features for each shot are extracted from key-frames (color histograms). Additionally, temporal features exploit the fact that distant shots along the temporal axis are less likely to belong to the same scene.

The second method uses spectral clustering (SC) [7], which has been shown to be effective in a variety of segmentation tasks. The algorithm first constructs a

pair-wise key-frame similarity matrix, for which similarity is defined in both visual and temporal terms. After matrix pre-processing, its spectrum (eigenvectors) is computed. Then, the $\mathcal{K}$ largest eigenvectors are stacked in columns in a new matrix, and the rows of this new matrix are normalized. Each row of this matrix constitutes a feature associated to each key-frame in the video. The rows of such matrix are then clustered using K-means (with $\mathcal{K}$ clusters), and all key-frames are labeled accordingly. Shots are finally clustered based on their key-frame labels by using a majority vote rule. Model selection is performed automatically using the eigengap, a measure often used in matrix perturbation and spectral graph theories. The algorithm uses the same visual and temporal features as PHC, adapted to the specific formulation.

6.2 Results and discussion

Figs. 3(c-f) show the error distributions when comparing the scenes found by people and the two automatic algorithms. The means for all measures are shown in Table 1. Comparing the results to those in Figs. 3(a-b), the errors for the automatic algorithms are higher than the errors among people. The degradation is more noticeable for the GCE and LCE measures, with a relative increase of more than 100% in the mean error for all cases. These results suggest that the automatic methods do not extract the scene structure as consistently as people do. In contrast, the relative variations in the correction factor are not so large. Overall, the automatic methods increase the error for GCE' and LCE': 53.3% and 52.7% for GCE', and 27.8% and 41.0% for LCE', for PHC and SC, respectively.

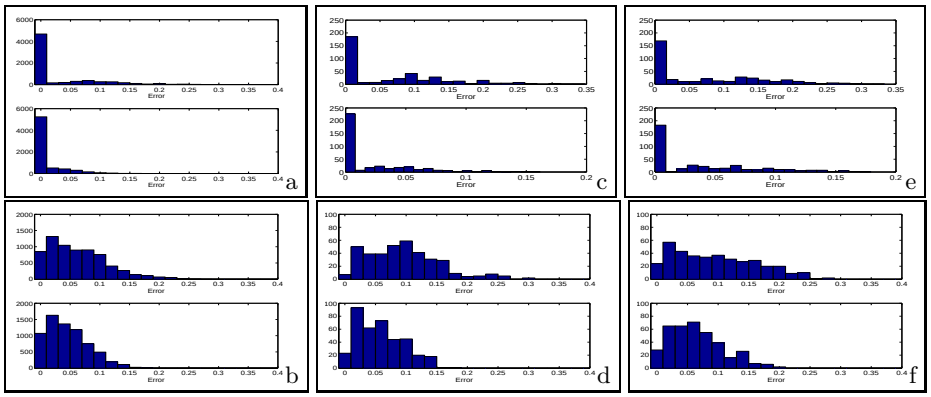

Fig. 3. (a-b) Human scene structuring; (a) distributions of GCE (top) and LCE (bottom) for all pairs of video scene structures (same videos) in the database; (b) distributions of GCE' and LCE'. (c-d) PHC vs. human: (c) GCE (top) and LCE (bottom); (d) GCE' and LCE'. (e-f) SC vs. human: (e) GCE and LCE; (f) GCE' and LCE'.

The results of our protocol appear in Table 2. Again, the error by automatic algorithms vs. people is higher than the errors among people, and the performance for both PHC and SC is quite similar. We used a two-tailed Wilcoxon signed-rank test analysis to detect significant differences between the two automatic algorithms for the min, med, and max performance over all videos. The obtained p-values

are 0.658, 0.970, and 0.881, respectively, so the difference in performance is not statistically significant. In contrast, the tests comparing human vs. SC produced p-values of 0.147, 0.030, and 0.004, respectively, which indicates that the difference in performance for the min is only significant at $p<0.15$ level, but the differences for med and max are significant at $p<0.05$ and $p<0.005$ levels, respectively. Similar results are obtained when comparing human vs. PHC with the Wilcoxon test. Examples of human- and computer-generated video scene structures can be seen at `www.idiap.ch/~gatica/homevideoassess.html`. Note that, although PHC and SC do not perform significantly different under this similarity measure, previous work using a different measure had favored SC [7].

Case	GCE	LCE	C	GCE'	LCE'
human/human	0.0321	0.0119	0.2416	0.0635	0.0463
PHC/human	0.0656	0.0216	0.2725	0.0966	0.0592
SC/human	0.0740	0.0377	0.2214	0.0962	0.0653

Table 1. Error means. Human vs. human and automatic vs. human.

Fig. 4 displays the results for each video for the two automatic algorithms. Figs. 4(a) and 4(c) show the number of detected scenes (red circles), and compare them to the mean number of scenes in the GT (blue crosses). The blue bar denotes the std in the GT. For both algorithms, the detected number of scenes matches well the GT, although somewhat underestimated. The number of scenes estimated by PHC (resp. SC) remain within one std of the mean human performance in 15 (resp. 17) of the 20 videos; in addition, in 14 (resp. 18) cases, the automatic method detected exactly the same number of scenes as at least one person did in the GT. These numbers are in agreement with the column for C in Table 1.

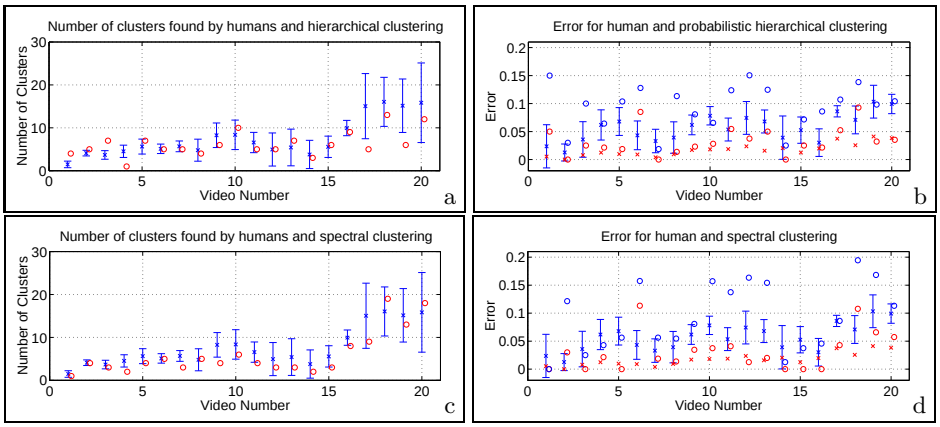

Fig. 4. Automatic (circles) vs. human (crosses) scene structuring. Top row: PHC. Bottom row: SC. (a-c) Number of detected scenes. (b-d) GCE' error. The bar is the spread of human performance (see text for details).

Figs. 4(b) and 4(d) show GCE' compared to the average of human performance. The circles denote the measures obtained with PHC/SC, the crosses denote human performance. Distinct colors represent different measures (minimum

Case	GCE'_{min}	GCE'_{med}	GCE'_{max}
human/human	0.0168	0.0563	0.1436
PHC/human	0.0333	0.0941	0.1827
SC/human	0.0308	0.0932	0.1870

Table 2. Error means over individual performance.

in red, median in blue, maximum ommitted for space reasons). The median performance of PHC (resp. SC) stays within or below one std of the median human performance (i.e., blue circles within or below blue bars) in 9 (resp. 12) videos.

7 Conclusions

We presented a methodology to benchmark scene structuring algorithms in home videos, using human performance on the task as the baseline. The agreement measures, adapted from work on natural image segmentation, attempt to model two concepts in perceptual organization. On a small but diverse data set, our experiments suggest that there exists human agreement in terms of organization of video scenes, but that there is a considerable variation w.r.t. scene granularity, which seems to depend on the visual content complexity. The comparison of two techniques with our methodology suggested that both performed similarly well, but still not as well as people. A comprehensive study that compares other agreement measures [9, 4] and structuring algorithms remains as a future goal.

Acknowledgements. We thank the Swiss NCCR on Interactive Multimodal Information Management IM2 for support, and Eastman Kodak for the home video database.

References

1. K. Boyer and S. Sarkar, "Perceptual Organization in Computer Vision: Status, Challenges, and Potential," *CVIU*, Vol. 76, No. 1, Oct. 1999.
2. S. Edelman, "Visual Perception", in *Encyclopedia of Artificial Intelligence*, 2:1655-1663, S. Shapiro, (ed)., Wiley, 1992.
3. S. Eickeler and S. Muller, "Content-based Indexing of TV News Using HMMs," in *Proc. IEEE ICASSP*, Phoenix, 1999.
4. D. Gatica-Perez, A. Loui, and M.-T. Sun, "Finding Structure in Home Videos by Probabilistic Hierarchical Clustering," in *IEEE T-CSVT*, Vol. 13, No. 6, Jun. 2003.
5. D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," in *Proc. IEEE ICCV,* Vancouver, Jul. 2001.
6. J.R. Kender and B.L. Yeo, "On the Structure and Analysis of Home Videos," in *Proc. ACCV*, Taipei, 2000.
7. J.-M. Odobez, D. Gatica-Perez, and M. Guillemot, "Spectral Structuring of Home Videos," in *Proc. CIVR*, Urbana, Jul. 2003.
8. K. Rodden, W. Basalaj, D. Sinclair, and K. Wood "Does Organization by Similarity Assist Image Browsing?," in *Proc. SIGCHI 2001*, Seattle, Apr. 2001.
9. D.M. Squire and T. Pun, "Assessing Agreement Between Human and Machine Clusterings of Image Databases," *Pattern Rec.,* Vol. 31, No. 12, 1998.
10. J. Vendrig and M. Worring, "Systematic Evaluation of Logical Story Unit Segmentation," *IEEE Trans. on Multimedia*, Vol. 4, No. 4, Dec. 2002.