DISCOVERING RECURRENT VISUAL SEMANTICS IN CONSUMER PHOTOGRAPHS

Alejandro Jaimes [*], *Ana B. Benitez* [*], *Shih-Fu Chang* [*], *and Alexander C. Loui* [Ψ]

* Electrical Engineering Department
Columbia University
New York, NY 10027, USA

[Ψ] Imaging Science and Technology Laboratory
Eastman Kodak Company
Rochester, NY 14650-1816, USA

ABSTRACT

We present techniques to semi-automatically discover *Recurrent Visual Semantics* (*RVS*) -the repetitive appearance of visually similar elements such as objects and scenes- in consumer photographs. First, we introduce the detection of "bracketing" (very similar photographs) using an edge-correlation metric, which outperforms color histogram. Then, we use color and novel composition features (based on automatic region segmentation) to perform scene-level clustering of images. We use a novel sequence-weighted technique, which uses the structure of standard film (only image sequence information), to perform hierarchical clustering. We show performance results of bracketing, explore clustering evaluation, and discuss *STELLA*, an interactive albuming and story telling application that uses these techniques to assist users in building digital albums. The *STELLA* system uses a new approach to album creation: instead of automatically creating albums, it provides an interactive environment that assists users in digital album creation.

1. INTRODUCTION

Digital and APS cameras are becoming very popular. Many of these cameras have the capability of recording simple annotations such as a category ("indoor", "party"), time (date and time stamp), and location (GPS). The quality of these cameras, however, is very limited in comparison to 35 mm film. Because of this, most amateur/professional photographers continue to use film, although most also welcome having their photographs in digital format (by scanning their film). In addition, images produced with digital cameras seldom include annotations, even if the cameras provide such capabilities.

As a result, very large collections of digital images without annotations continue to grow. These collections are often recorded in Photo CDs in which the original sequence of the pictures is maintained, with no other information available.

One important aspect of having photographs in digital form is the organization of such collections. In particular, there has been a strong interest in the creation of albums [4][5][6] and picture sharing mechanisms. Given the large amount of images scanned, and produced by digital cameras, it is desirable to apply automatic or semi-automatic techniques that organize the digital images and ultimately lead to the creation of digital albums.

In this paper, we present techniques that can be used to organize images and create digital albums semi-automatically. We use the concept of *Recurrent Visual Semantics* [3] as the basic organizing principle. First, we introduce the detection of "bracketing"[1] using a simple edge-correlation technique that outperforms color histogram. Then, we propose a novel sequence-weighted clustering technique that uses a modified version of Ward's hierarchical clustering algorithm [2]. Finally, we present a system (*STELLA*), which, using the proposed techniques, provides the user with a hierarchical organization of the contents of individual rolls of film. The user interactively modifies the clusters to create digital albums.

For clustering, images are first segmented automatically based on color and edge information [9]. The sequence-weighted clustering algorithm uses color features (i.e., color histogram), and novel composition features (i.e., based on the segmentation).

Our work differs from the albuming application presented in [4][5] and others in several aspects: our system uses image sequence information (not time-stamps); new composition features (not block-matching) are used; and the detection of "bracketing" is introduced.

In section 2 we discuss the framework for discovering *Recurrent Visual Semantics* in consumer photographs. In Section 3 we present the bracketing detection mechanism, the features used, and the clustering algorithm. In section 4 includes experimental results. In section 5 we present the *STELLA* system, and we conclude with a summary.

2. RECURRENT VISUAL SEMANTICS

The concept of *Recurrent Visual Semantics* (*RVS*) was first introduced in [3] as the *repetitive appearance of elements (e.g., objects, scene) that are visually similar and semantically meaningful within a specific context*. *RVS* occurs in many domains including sports video and consumer photographs.

2.1. RVS in Consumer Photographs

RVS in consumer photographs occurs at several levels.

[1] Several pictures are taken, without moving the camera, but using different exposure settings.

For example, in photos of a trip to Paris, it is common to find photographs of the Eiffel tower. Likewise, birthday photographs taken by different individuals often include similar scenes.

Images can present *RVS* due to the following:

- *Bracketing:* several pictures are taken using different exposure settings, without a change of subject or composition.
- *Scene:* a scene is repeated in different images (e.g. party scene).
- *Object:* the same object appears in several photographs (e.g. Eiffel Tower).

Since people often photograph the same subject more than once, *RVS* frequently occurs in the same roll of film. In those cases, it is common for similar images to appear close to each other in the film sequence.

In order to detect *RVS* it is necessary to use appropriate features (i.e., for scenes, or objects) and exploit the sequential structure of the film. We aim at discovering *RVS* in professional/amateur photographs, by detecting bracketing, and by performing scene-level image clustering.

3. BRACKETING AND CLUSTERING

As will be discussed in section 5, one of our goals is to provide the user with a preliminary organization of the images in rolls of film. In order to achieve this, first, bracketing is detected. Then, the images in the roll are clustered using color and composition features. We apply the techniques to individual rolls, although they can also be used for multiple rolls of film.

3.1. Bracketing Detection

Bracketing refers to taking several pictures of the same subject using different exposure settings, without modifying the camera angle or composition. In amateur/professional photography, this technique is quite common.

In many of those cases, color histograms are unsuitable for capturing such differences. For example, the images in Figure 1 are identical in composition with only a variation in the exposure (i.e., aperture and shutter speed). The color histograms of the two images are considerably different, although their semantic content is identical.

Figure 1: Example of bracketing where the images have the same composition, but very distinct histograms.

In order to capture similarities between these types of images, we use the edge correlation technique presented in [1]. Each image is first normalized (to a resolution of 64x64) and an edge map is generated. The edge map is then divided into 8x8 blocks and the summation of local measures of correlation among the blocks results in an edge correlation measure. We detect bracketing by comparing the edge correlation measure for each pair of contiguous images in each roll of film with a threshold. A detection occurs when the edge correlation of the two images is above the detection threshold. The threshold can be computed automatically or pre-determined using a training set (see section 4).

3.2. Color and Composition Features

An image is automatically segmented based on color and edges using [9]. Since images with different compositions will yield different segmentation results, we use the regions obtained in the segmentation to compute each image's composition. First, we extract the following set of visual features [3], for each region of the image: d = {extent, roundness, aspect ratio, orientation, location, dominant color, minmax difference}. These features were selected to represent the overall shape, orientation, location, color, and texture of image components. For the same image, we then compute the weighted average (*CF*) of each of the features extracted, as follows:

$$CF = \frac{\sum_{i=0}^{n}(f_i * a_i)}{n}$$

Where the image has n regions, and f_i, and a_i are the feature value (i.e., for a feature from set d above), and area corresponding to region i. This average is weighted by each region's area, because larger regions have a stronger impact on the composition. To perform clustering, we create a feature vector that concatenates the averages for each of the features in d, with a "number of regions" feature, and a color histogram. The number of regions is important because it may vary considerably for images with distinct compositions, and the histogram is useful in representing the image's color distribution.

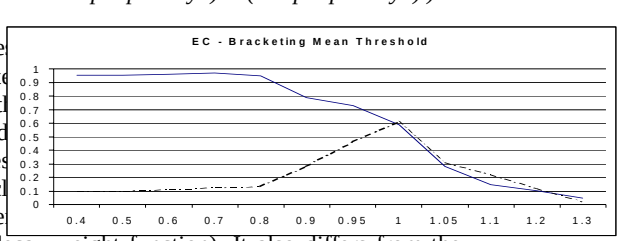
3.3. Sequence-weighted Clustering

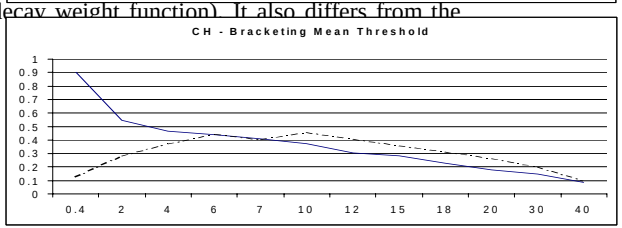
In consumer photography, visually similar images tend to appear often, in contiguous blocks (or in proximity) in the film. We exploit the structure of the film by performing sequence-weighted clustering on each roll of film. We use a variation of Ward's hierarchical clustering algorithm- several comparative studies have shown that Ward's method outperforms other hierarchical clustering methods [2]. In that algorithm, clusters that minimize the within-cluster square error (*sqerr*) are merged. The squared error for a cluster is the sum of squared distances to the centroid for all patterns in the cluster.

We have modified Ward's algorithm to be sequence-weighted so that it accounts for the location of each image within the roll of film. In particular, we penalize images that are far apart in the film. When a cluster is being formed, we use two additional variables *nskip*, number of times a skip occurs, and *tskip*, total number of frames skipped (see Figure 2):

*sqerr += ((sqerr + tskip * penalty1) + (nskip * penalty2))*

These value
resulting in cluste
the film. Note t
method presented
technique, images
be in the same cl
criterion used the
an exponential decay weight function). It also differs from the





work in [6], in which an image is added to a cluster only if it is contiguous (in the film) to an image already in the cluster.

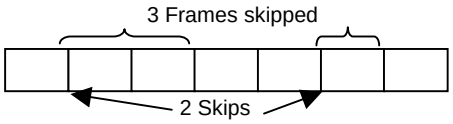


Figure 2: Number of times a skip occurs and total number of skips in a cluster.

The approach we propose has strong benefits over the previous ones because it allows *very similar* images to be in the same cluster (even if they are very far apart in the film), at the same time that it discourages non-contiguous images in the same cluster.

4. EXPERIMENTAL RESULTS

To test our techniques, we obtained a set of approximately 1,700 professional/amateur travel photographs (over 45 rolls of film) to be part of the test-bed [2].

4.1. Bracketing

To evaluate bracketing detection, we used only the rolls of film that included at least one case of bracketing. Out of the 45 rolls of film, 20 rolls included bracketing. For those, the average number of bracketing cases per roll was 3.25.

We experimented with several mechanisms to automatically determine the bracketing detection threshold. In the first mechanism, we used a threshold proportional to the average edge correlation of the images in the roll of film being processed (i.e., roll-dependent threshold). The second one was proportional to the average edge correlation, of a bracketing training set (i.e., roll-independent threshold).

The average recall/precision results in detecting bracketing for edge correlation and color histogram metrics, using a threshold based on statistics from a bracketing set are shown in Figure 3. An independent training set of 30 bracketing cases (excluded from the test set of Figure 3) was selected randomly and used to calculate the roll-independent detection threshold (the roll-dependent threshold was inferior because more false alarms occurred). Each case contains 2 images. The recall decreases when the threshold increases because less bracketing cases are detected. The tendency of the precision is to increase with the threshold for low threshold values because non-bracketing examples are discarded. However, for high threshold values, true bracketing cases are not missed, then the precision tends to decrease. The optimal operation point is the one that maximizes the product of recall/precision (factor~=1). Figure 3 confirms that edge correlation outperforms color histogram with a higher recall/precision product.

Figure 3: Recall (continuous line) and precision (dash line) results in detecting bracketing for edge correlation (EC) and color histogram (CH) metrics. A factor (hor. axis) of the training set mean value was used.

4.2. Clustering

In order to evaluate the automatically generated clusters produced by our system, we explored three different strategies: (1) comparison with a human-constructed clustering ground truth; (2) automatic evaluation; and (3) subjective evaluation.

For the first task, two of the authors independently clustered the same 16 rolls from the collection (creating a ground truth similar to the one performed by a single person in [6]). No similarity or clustering guidelines were set: the authors clustered each roll of film independently and subjectively. Although we did not quantitatively compare individual clusters, we found that the clusters and the number of clusters generated by the two authors for the same rolls differed substantially. For a set of 8 rolls, for example, the average number of clusters for authors A and B was 19 and 10, respectively. As discussed in [7], clustering of images is a subjective task, in which often there is little agreement between users (in the experiments reported there, 34.6% higher than agreement expected by chance). The experiment we performed ratified the difficulties of using a ground truth to evaluate our algorithm- with the data set we are using. In some cases, however, (e.g., [5]) creating a ground truth is possible and very useful, particularly if the goal is to evaluate fully automatic techniques.

The second strategy was to use automatic cluster validation techniques [2]. External criteria (i.e., does the cluster hierarchy match an expected hierarchy?) are difficult to implement, particularly for hierarchical clusters, because expected hierarchies are not usually available (as is the case here). Internal (i.e., does the hierarchy fit the data well?), and relative criteria (i.e., which of two hierarchical clusterings fits the data better?), on the other hand, usually require a baseline distribution. Such distribution is difficult to obtain here, since a typical roll of film has a maximum of 37 images.

The third strategy, subjective evaluation, consisted of performing several clustering experiments, using different features and examining the results at different levels of the clustering hierarchies. Although in some cases it is possible to quantify the results and score clusters, we found that many factors are involved in the grouping of consumer photographs (as occurred in the collection we used), making this a very difficult task. In particular, we found that the evaluation is task-specific, user-specific, and content-specific. Images can be grouped in multiple ways, based on different levels of similarity: visual (e.g., colors, textures, etc.) or semantic (e.g., events, objects, etc.).

In spite of the difficulties encountered in evaluating our clustering approach, we can make the following subjective observations: (1) the combination of histogram and composition features provides better results than either one of those features alone; (2) sequence-weighted clustering produces better results than it's non-weighted equivalent.

Next, we discuss how evaluation can be performed in our framework, and how our techniques benefit the application we envision.

[2] MPEG-7 content, Photographs by Philip Greenspun, and by Alejandro Jaimes (taken independent of this project). The authors wish to thank Kodak for providing the equipment necessary to scan part of this photography collection.

5. THE STELLA SYSTEM

Grouping images can be a difficult task because very often the groupings will rely on specific knowledge about the subjects, events, or locations that were photographed. Based on this, we propose a new paradigm to album creation, in which the goal is not to automatically construct an album, but rather, to automatically organize the images as much as possible so that a user can easily construct an album.

The *STELLA* (Story TELLing and Albumimg) system we are building provides an interactive environment for album creation. The system receives as input one or more rolls of film, and creates clusters based on the techniques discussed in earlier sections. A considerable advantage of using hierarchical methods is that the results can be easily used for browsing. This is an important quality in our framework- when users are creating albums they are expected to work with the results provided by the system. As the user browses a hierarchy, he/she will be able to perform several operations to construct the digital album. In the process, the user can perform *organization* and *selection* of images. For organizing images, the user can perform the following operations: (1) circle satisfactory clusters at different levels; (2) make corrections to existing clusters; and (3) cross-out unsatisfactory clusters (see Figure 4).

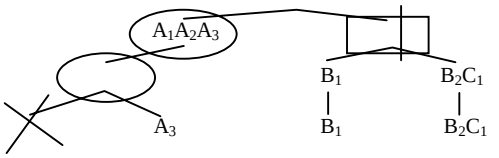


Figure 4: User operations on the output produced by *STELLA*. "Good" clusters are circled, "bad" clusters are crossed out, and others edited (square).

The album can be constructed based solely on the operations above (e.g., include only satisfactory clusters), or be more elaborate. In a second scenario, the user may wish to select satisfactory clusters, and then select the best images from those clusters to build an album. The environment we wish to provide in *STELLA* will be flexible enough to facilitate such operations. The success of the system will be measured in terms of the user's interaction (not the clustering accuracy). For example, in building a digital album, the user will perform some of the operations described above. In order to evaluate the system, scores can be assigned to each operation (e.g., corrections to existing clusters), using measures that are task, content, and user-specific.

6. CONCLUSIONS AND FUTURE WORK

We have presented techniques to semi-automatically discover *Recurrent Visual Semantics* in consumer photographs. For bracketing detection, we use an edge correlation technique that produces superior results to those obtained using color histograms. To cluster the images, we use a novel sequence-weighted approach based on Ward's algorithm. We outlined the difficulties of evaluating clustering, and proposed a different (semi-automatic) approach to digital album creation. We are constructing *STELLA*, an interactive environment in which the techniques we described are used to produce clusters that can be modified by the user.

Evaluation techniques for creating albums are of great importance. In future work, we will develop techniques for this purpose, that will be specific to the content of given rolls, the user, and the task (i.e. album creation). Instead of measuring the clusters created automatically, we will measure the user's interaction (based on the operations we outlined) with the system as he/she creates digital albums.

REFERENCES

[1] K. Hirata and T. Kato, "Query by Visual Example", *Proceedings of the Intl. Conf. On Extending Database Technology EDBT92*, pp. 56-71, March, 1992.

[2] A. Jain and R. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, 1988.

[3] A. Jaimes and Shih-Fu Chang, "Automatic Selection of Features and Classifiers", *SPIE Vol. 3972, Storage & Retrieval for Media Databases 2000*, San Jose, Jan. 2000.

[4] A. Loui, and M. Wood, "A software system for automatic albuming of consumer pictures," *Proc. ACM Multimedia 99*, pp. 159-162, Orlando, Fl., Oct. 30-Nov. 5, 1999.

[5] A. Loui and A. E. Savakis, "Automatic Image Event Segmentation and Quality Screening for Albuming Applications", *ICME 2000*, New York City, July 2000.

[6] J. C. Platt, "AutoAlbum: Clustering Digital Photographs Using Probabilistic Model Merging", *IEEE Workshop on Content-Based Access of Image and Video Libraries CBAIVL-2000*, Hilton Head Island, June 2000.

[7] D. M. Squire, T. Pun, "Assessing Agreement Between Human and Machine Clusterings of Image Databases", *Pattern Recognition*, Vol.31, No. 12, pp. 1905-1919, 1998.

[8] M. M. Yeung, B.-L. Yeo and B. Liu, "Segmentation of Video by Clustering and Graph Analysis", *Computer Vision and Image Understanding*, V. 71, No. 1, July 1998.

[9] D. Zhong and S.-F. Chang, "Video Object Model and Segmentation for Content Based Video Indexing", *IEEE Int Symp. on Circuits and Systems*, Hong Kong, June 1997.