Matching of Complex Scenes Based on Constrained Clustering

Alexander Loui and Madirakshi Das

Kodak Research Labs
Eastman Kodak Company, Rochester NY
alexander.loui@kodak.com

Abstract

This paper presents an algorithm and a software application for matching complex scenes in consumer image collections. By applying appropriate constraints to detected keypoint matches between a target and reference image, complex scenes from users' collections can be matched with very good precision. Experimental results confirm the effectiveness of this approach. The method can be further enhanced to retrieve similar scenes from different users by incorporating event information.

Motivation

In this work we present a constrained clustering algorithm for automatically matching regions of interest in complex consumer images. This leverages the fact that in consumers' personal image collections, it is possible to identify unique objects present in the images that can be tied to specific areas in the house, such as the living room or the children's playroom, or to specific locations, such as the office or grandma's house. Typical examples include pictures hanging on the wall, patterns on curtains, wallpapers or rugs, furniture upholstery, and children's play equipment. These unique objects in the scene can be reliably matched even with significant occlusion and viewpoint changes. A desirable characteristic of objects to be matched is a dense set of localized features that can be matched reliably resulting in fewer false matches. Our approach to determine whether two images were captured at the same location is based on automatically matching such reliable objects.

The applicability of our approach can be expanded to a significantly larger set of images when we take temporal information into account. Using the domain knowledge that an entire event [1] typically takes place at the same location in consumer image collections, when a pair of images (one from each event) matches, we can assign the same location tag to other images in the same events. In addition, we can use transitive reasoning to determine that if an image in event A matches an image in event B, and another image in event B matches an image in event C, then events A, B, and C are likely to be co-located, even though there is no direct match between images in event A and event C.

Our Approach

There has been recent work on matching feature-rich scenes using scale-invariant feature transform (SIFT) [2, 3]. However, these techniques have been applied mainly to matching and registering entire scenes. Parikh et al. [4] have used Lowe's SIFT [2] features to model objects in the scene, assuming the same objects are present in both candidate images. In the Photosynth [5] application, objects common between multiple images of the same scene are registered, and a 3D representation of the object is created to aid browsing. Matching SIFT features between two images using the method described by Lowe [2] produces a set of matched keypoints between the reference and target images. However, in cluttered scenes such as consumer images, false positives are quite common. False positives occur when points matched do not correspond to the same objects in the two scenes. To overcome this problem, we propose a number of constraints on the matched points to produce matches with high precision.

The first step in matching two images is to identify regions of interest in the images that have some likelihood of matching. This is achieved by spatially clustering the matched SIFT keypoints [2] independently in each image. This is followed by a filtering step that aims to apply constrains that will increase the certainty that the match is derived from an underlying common object. The clusters are analyzed to determine whether there are correlations between pairs of clusters. Because keypoints on the same object are spatially constrained to the region occupied by the object, it is expected that clusters in the reference image will correspond to cluster(s) in the target image for true matches. The next constraint ensures that the global trajectory of points from the reference to the target is consistent, i.e., all clusters in the scene move in the same general direction. This is expected because true objects in the scene are likely to maintain the same spatial configuration relative to each other. The final constraint aims to ensure that the clusters are compact, by removing outlier points. Finally, a match score is computed based on the keypoints that remain after the filtering process. This match score can be used in applications to determine the likelihood that two images could be co-located.

In summary, the proposed scene matching algorithm consists of the following steps [6]:

1. Matching SIFT Features

For each SIFT keypoint in the target image, the algorithm finds a reference keypoint that minimizes the Euclidean distance between the corresponding target and reference descriptors.

2. Clustering Matched Features

The Iterative Self-Organizing Data (ISODATA) algorithm is used for matching each keypoint to a specific partition in the target and reference images. ISODATA attempts to find the partition for each keypoint, which minimizes the mean distance of member keypoints from the partition's center.

3. Applying Constraints

The next step in filtering the matched keypoints is the determination of the best match for each cluster in the target image from among the clusters in the reference image. A pseudo-confusion matrix is created where target clusters form rows and clusters in the reference image form columns. A correlation score is determined for each cluster in the target image, which is computed as the proportion of points in this cluster that match points in the cluster's strongest match. Further filtering of spurious keypoint matches is obtained by checking the direction of movement of the matched point from the target to the reference. To ensure that the global trajectory is consistent, keypoints that form a Cartesian angle that is more than a certain standard deviation (empirically set to 1.0 σ) away from the average keypoint angle are eliminated from the pseudo-confusion matrix row or column.

Target and reference partitions are defined as having both a spatial center and a spatial size. The size of each cluster is determined by the result of the criterion function applied by the particular iterative optimization strategy (i.e., ISODATA). The current criterion leverages members' subpixel distance from a respective center so the size of each scene region is inversely proportional to the density of features distributed normally about the center. A spatial threshold (empirically set at 2.0 σ) is used to eliminate keypoints from a cluster that exceed a maximum distance from the partition center.

4. Matching Score

If there are keypoints that remain after the filtering process, this indicates a match between the reference and target image. Hence, the larger the number of keypoints left, the more reliable the match. The scene match score we currently use is simply a count of the remaining features that survive after the filtering process has removed clusters and points.

Performance Evaluation

To test the performance of scene matching in the consumer image domain, a software application (see Figure 1) was created using our approach. The application allows the detailed viewing of matches to show the keypoints matched between an image pair. It also allows the retrieval of matched images when provided with a query image, and has the capability of tagging retrieved image groups with location. The test images and events were gathered from personal image collections of 18 subjects. Each collection had one to two thousand images spanning a timeframe of one to two years. The images depict common consumer picture-taking events such as vacations and family gatherings. Table 1 shows the improvement in precision and recall when using our constrained clustering approach, compared to using SIFT features alone. Figure 1 shows the software application shown in our demonstration.

Table 1. Scene-matching application showing retrieval of matched images based on a query image.

Scene match using	Recall (%)	Precision (%)	Avg. rank of best match
SIFT only	81	21	1.95
SIFT + constrained clustering	85	85	1.05

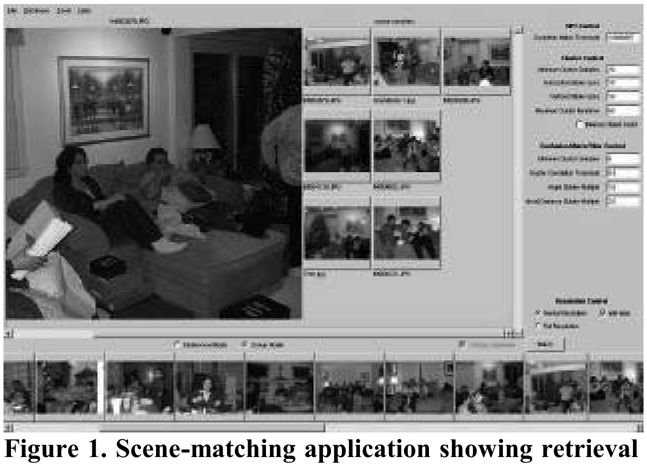

Figure 1. Scene-matching application showing retrieval of matched images based on a query image.

Figure 2 shows some examples of scene matches obtained using our approach, which illustrates some of the strengths of the approach. In the first example, only a small part of the painting is visible, while the rest of the image (the two guitars) looks very different in terms of color and texture. In the second example, a different part of the room is visible in each image, with the common object visible being a colorful window hanging.

Figure 3 illustrates event co-location determined using transitive reasoning as mentioned earlier. Each image is from a different event (which typically has 10-20 images). Note that the reference and target images are taken many months apart, as can be seen by the fact that the same baby has grown considerably, and the change of season from summer to winter in the second set. Note that A can also be matched to C as the same rug appears in both images. Therefore, event B can be matched to events C and D though no matching image is found between them.

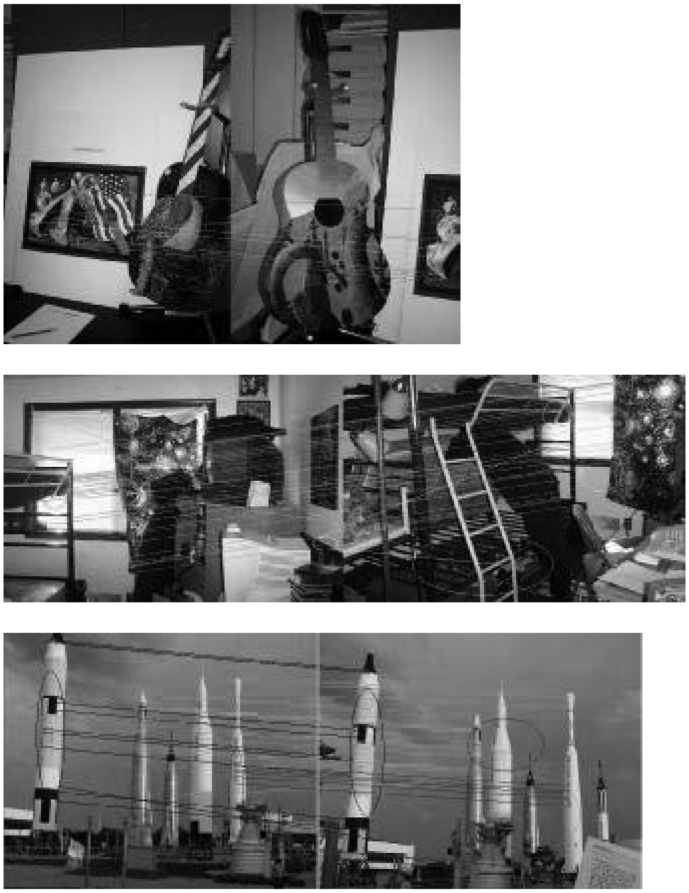

Figure 2. Matched mage pairs showing corresponding keypoints between reference and target images.

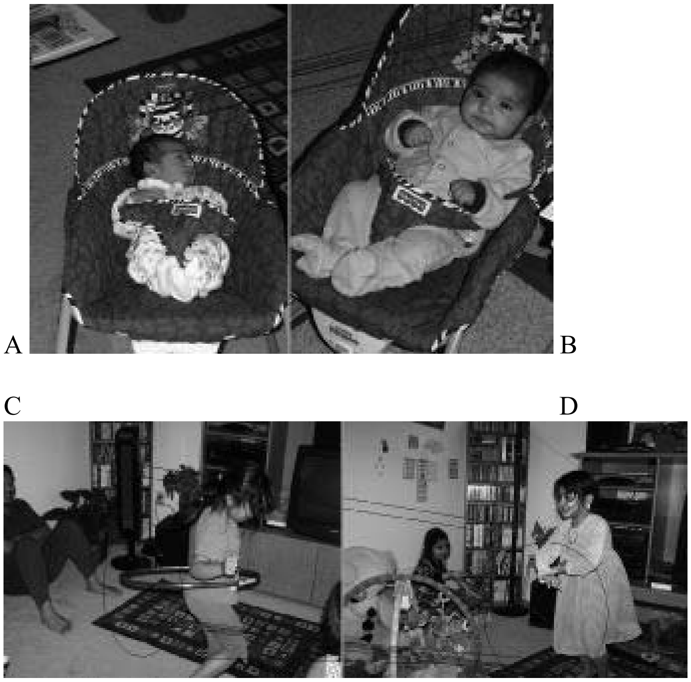

Figure 3. Matched image pairs that can be used for event co-location determination. Events from which the images are extracted are denoted by block letters.

Acknowledgment

The authors thank Jens Farmer for his contributions on developing the software application.

References

[1] Loui, A., and Savakis, A. 2003. Automated event clustering and quality screening of consumer pictures for digital albuming. IEEE Trans. Multimedia, 390-402.

[2] Lowe, D. 2004. Distinctive image features from scale invariant features. Intl. J. Comput. Vision (IJCV), 60(2) 91-110.

[3] Bay, H., Tuytelaars, T., and Van Gool, L. 2006. Surf: Speeded up robust features. In Proceedings of the 9th European Conference on Computer Vision (ECCV).

[4] Parikh, D. and Chen, T. 2007. Hierarchical semantic objects. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[5] Snavely, N., Seitz, S., and Szeliski, R. 2006. Photo tourism: Exploring photo collections in 3D. ACM Trans. Graph. 25(3) (Aug. 2006).

[6] Das, M., Farmer J., Gallagher, A., Loui, A. 2008. Event-based location matching for consumer image collections. To appear, Proc. ACM Intern. Conf. on Image and Video Retrieval (CIVR).