Image Content Clustering and Summarization for Photo Collections

Cheng-Hung Li, Chih-Yi Chiu, Chun-Rong Huang, Chu-Song Chen, and Lee-Feng Chien

Institute of Information Science, Academia Sinica, Taiwan
{chli, cychiu, nckuos, song, lfchien}@iis.sinica.edu.tw

ABSTRACT

Rapid growth of digital photography in recent years spurred the need of photo management tools. In this study, we propose an automatic organization framework for photo collections based on image content, so that a novel browsing experience is provided for users. For each photograph, human faces, together with corresponding clothes and nearby regions are located. We extract color histograms of these regions as the image content feature. Then a similarity matrix of a photo collection is generated according to temporal and content features of those photographs. We perform hierarchical clustering based on this matrix, and extract duplicate subjects of a cluster by introducing the contrast context histogram (CCH) technique. The experimental results show that the developed framework provides a promising result for photo management.

1. INTRODUCTION

Advances in capturing devices spur the proliferation of digital photographs. Since users can easily obtain digital versions of their photo collections, managing and accessing these photo collections becomes an increasingly difficult task. There is thus a need for users to organize and browse these collections efficiently. Several commercial systems (e.g., flickr, picasa, etc.) and academic frameworks have been developed. Even so, it is desirable to apply some automatic techniques to reduce the effort in photo management.

For general consumer photographs, photos can be organized through the "four Ws" model: where, when, who, and what. GPS and timestamp metadata and face recognition techniques have shown the effectiveness for "when," "where," and "who" models [1-2]. However, for the "what" model, it is still full of challenges to understand photo content by computers. Current research mainly focuses on image classification and event detection. Image classification employs low-level feature to infer high-level semantic concepts in an image [3], while event detection partitions a sequence of photos into contiguous clusters corresponding to the underlying events [4-5].

Let us take a look at a photo collection taken during a trip. Suppose that the photo collection records several events. Events are usually associated with specific time and image content. Besides, an event often contains duplicates that are taken in a number of photographs from different viewpoints or compositions. For example, in Kodak's consumer image database [6], 19% of the images are perceived to be either "duplicates" or similar "non-duplicates." Therefore, a good way to organize and browse a photo collection can be realized by clustering photographs into events, and summarizing duplicates appeared in these events.

Based on the above observations, we propose an automatic organization framework for photo collections based on image content. This study is confined to general consumer photographs, which are usually composed by people, subjects, and scenery. For each photograph, human faces, together with corresponding clothes and nearby regions are located. We extract color histograms of these regions as the image content feature. Then a similarity matrix of a photo collection is generated according to temporal and content features of those photographs. We perform hierarchical clustering based on this matrix, and extract duplicate subjects of a cluster by using the contrast context histogram (CCH) technique. Our framework can provide a novel experience for users to browse their photo collections.

This paper is organized as follows. Sections 2, 3, and 4 detail our framework. Section 5 shows experimental results. Section 6 presents our conclusions and future work.

2. CONTENT ANALYSIS

The first step is to analyze the image content for feature extraction. In a photograph, we define region of interests (ROIs) dependent on people locations in the photograph. For example, if people are located in the bottom-left of a photo, the top and right regions of the photo will be the ROIs of the photo. For each ROI, we extract its color histogram as its feature representation. Details are described in the following.

2.1 ROI Extraction

Given a photo sequence $P = \{p_i: i = 1, \ldots, N\}$, the first step is to locate people in p_i. We use the AdaBoost learning algorithm, a powerful face detection tool, to locate face position first. Denote the detected faces in p_i as a face set $F_i = \{f_{ij}\}$, where j is the index of the face set. If the area of f_{ij} is too small, or it is cut by image margins, it will be omitted from F_i. Next, for each $f_{ij} \in F_i$, its corresponding clothes region, c_{ij}, is defined based on the face locations and scales. In this study, we assume c_{ij} is beneath f_{ij}. Then the nearby regions of f_{ij} and c_{ij} are regarded as ROIs. We denote the ROIs of the photo p_i as $B_i = \{b_{ik}: k = 1, 2, 3\}$, including the top, left, and right parts. Note that if nobody is detected, the whole image is denoted as b_{i1}. Also, if the area of b_{ik} is smaller than a certain threshold, we consider that b_{ik} is not representative enough and omit it from B_i. Figure 1 shows some examples of ROIs extracted from the photos. In Figure 1*a*, the left ROI is ignored, while in Figure 1*b,* ROIs are extracted based on the people group.

(a) *(b)*

Figure 1. The solid lines rectangles indicate ROIs.

2.2 Feature Representation

We use the color histogram for each region in ROIs as the feature representation. In color histogram, first the image color space is transformed from RGB to YUV. U and V components that store chrominance information are used. Since the color distribution is non-uniform, histogram quantization should not use m equally spaced bins. We apply fuzzy c-means clustering to obtain appropriate resolutions of U and V components. As a example, we use the background region $b_{ik} \in B_i$ and set the bin number $m = 15$, where b_{ik} takes two 15-length histogram vectors of U and V components and concatenates them into a 30-length vector, denoted as $h^{(b)}{}_{ik}$. Finally $h^{(b)}{}_{ik}$ is normalized to remove the size effect:

$$h^{(b)}{}_{ik} / (2 \cdot \text{area}(b_{ik})),$$

where area(x) returns the area size of the region x. The normalized vector $h^{(b)}{}_{ik}$ is denoted as the feature of b_{ik}.

3. PHOTO CLUSTERING

A photo sequence is partitioned into contiguous clusters as corresponding events. First, a similarity matrix of the photo sequence is generated by computing temporal and content-based cues of any two photos. We then analyze the similarity matrix iteratively to split the photo sequence in a top-down way.

3.1 Similarity Matrix Generation

The similarity matrix stores the feature similarities of all possible pairs of a photo sequence. Let p_i and $p_{i'}$ be the i-th and i'-th photos in the photo sequence P, and S be the similarity matrix. The (i, i') element of S, i.e., $S(i, i')$, quantifies similarity between p_i and $p_{i'}$. $S(i, i')$ is computed based on both temporal and content-based factors. For the temporal factor, a time decay function is defined as follows:

$$\text{temp}(p_i, p_{i'}) = \exp(-\alpha \,|\, i - i' |),$$

where α is a parameter controlling the time effect. In our case, we set α=1. The concept of the above function is simple: if the two photos were taken close together, they would be associated with the same theme, otherwise their themes would be different.

For the content-based factor, we measure the similarity among the background regions in p_i and $p_{i'}$. Recall that $B_i = \{b_{ik}\}$ is the set of background regions of p_i, and $H^{(b)}{}_i = \{h^{(b)}{}_{ik}\}$ is the corresponding histogram feature set. If the two photos were taken in the same location, their background regions should be somewhat similar. A content similarity function is defined as follows:

$$\text{cont}(p_i, p_{i'}) =$$
$$\max\{\text{intersect}(h^{(b)}{}_{ik}, h^{(b)}{}_{i'k'}) : \forall h_{ik} \in H^{(b)}{}_i, \forall h_{i'k'} \in H^{(b)}{}_{i'}\},$$

where intersect(a, b) returns the intersection area of the two histograms a and b, and max$\{Y\}$ returns the maximum of the set Y. Finally the above two functions are combined to evaluate the similarity between p_i and $p_{i'}$:

$$S(i, i') = \text{temp}(p_i, p_{i'}) \cdot \text{cont}(p_i, p_{i'})$$

3.2 Top-Down Clustering

We analyze the intensity distribution in the similarity matrix to select a partition point that separates the photo sequence into two non-overlapped subsequences. Let the i-th photo p_i be a partition candidate of the first level hierarchy. The partition produces two squares in the similarity matrix S(1:i, 1:i) and S(i+1:N, i+1:N), where N is the number of the photo sequence. We define the intra-class intensity as the average intensity of the two squares:

$$\text{intra}(i) = \frac{1}{i^2} \sum_{u=1}^{i} \sum_{v=1}^{i} S(u,v) + \frac{1}{(N-i)^2} \sum_{u=i+1}^{N} \sum_{v=i+1}^{N} S(u,v).$$

In the above function we want to reward the subsequence whose backgrounds are very similar, i.e., the two squares

with high intensity values. In addition, we define the inter-class contrast as the difference between the i-th row and the (i+1)-th row of S:

$$\text{inter}(i) = \sum_{v=1}^{N} |S(i,v) - S(i+1,v)|.$$

That is, we hope the partition point is different from its neighborhood so that it forms a good boundary between two subsequences. The best partition point of the sequence is selected by a linear combination of intra- and inter-class evaluation:

$$i^* = \arg_i \max\{w \cdot \text{intra}(i) + (1-w) \cdot \text{inter}(i) : i \in [1,2,...,N]\},$$

where w is the weight parameter. In our case, we set w=0.5. The partition generates two subsequences as its children: $\{p_1, p_2, \ldots, p_{i^*}\}$ and $\{p_{i^*+1}, p_{i^*+2}, \ldots, p_N\}$, which are continually partitioned until the intra-class intensity of a successor is higher than a certain threshold.

4. SUMMARIZATION

After generating the underlying clusters of the photo collection, for user browsing conveniently, we have to select representative photo for each cluster. Thus we apply the contrast context histogram (CCH) technique to identify duplicates in a cluster as summarization. Hence users can only browse the result of summarization from thousands of photo collection. The CCH technique is described follows.

Given an image $I(e)$, we first apply Gaussian kernels to obtain a smoothed image L. Then, salient corner points are extracted from a multi-scale Laplacian pyramid by detecting Harris corners. Denote the salient point $e_c = (x_c, y_c)$ in the smoothed image, centered at this point we establish a log-polar coordinate system (r,θ) defined as follows:

$$r = \log_{10}(\sqrt{(x-x_c)^2 + (y-y_c)^2})$$

and

$$\theta = \tan^{-1}\left(\frac{y-y_c}{x-x_c}\right).$$

With the log-polar system, the feature can be more sensitive to positions of nearby points. After obtaining the salient points, the descriptor is computed using contrast of a salient point sp with respect to the salient point sp_c is defined as follows:

$$C(sp, sp_c) = L(sp) - L(sp_c),$$

When the log-polar coordinate of the neighborhood points are computed, the contrast information between neighborhood points and salient point is used to build discriminating features.

In our framework, to increase the discriminative ability of the descriptor, we compute positive and negative contrast-value histogram bins for each local region based on the location and orientation of sample points with respective to the salient points. For each pixel e in e_c's neighbor region R, the positive difference histogram bin of location bin r_i and the orientation bin θ_j respect to e_c is defined as follows:

$$CCH_{r_i\theta_j+}(e_c) = \frac{\sum\{C(e,e_c) \mid e \in R \text{ and } L(e) - L(e_c) > 0\}}{\#_{r_i\theta_j+}},$$

where $r_i = 0,...,k, k = \log(\sqrt{2n^2})$, $\theta_j = \frac{j\pi}{4}, j = 0,...,l,$ and $\#_{r_i\theta_j+}$ is the number of points satisfying that $L(e)$–$L(e_c)$>0. Similarly, the negative difference histogram by considering all the points satisfying that $L(e)$–$L(e_c)$<0. Then the contrast context histogram CCH of the salient point e_c is defined as:

$$CCH(e_c) = (CCH_{r_1\theta_1+}, CCH_{r_1\theta_1-}, CCH_{r_1\theta_2+}, CCH_{r_1\theta_2-}, ..., CCH_{r_k\theta_l+}, CCH_{r_k\theta_l-}),$$

where $CCH(e_c)$ is a $2kl$-length vector. As a result, we obtain CCH features to describe local invariance of the salient points.

For any two photographs p_a and p_b in a cluster, we find correspondences between p_a's and p_b's salient points, as shown in Figure 2. The correspondence is built by finding the minimum Euclidean distance between two CCH vectors of p_a's and p_b's salient points.

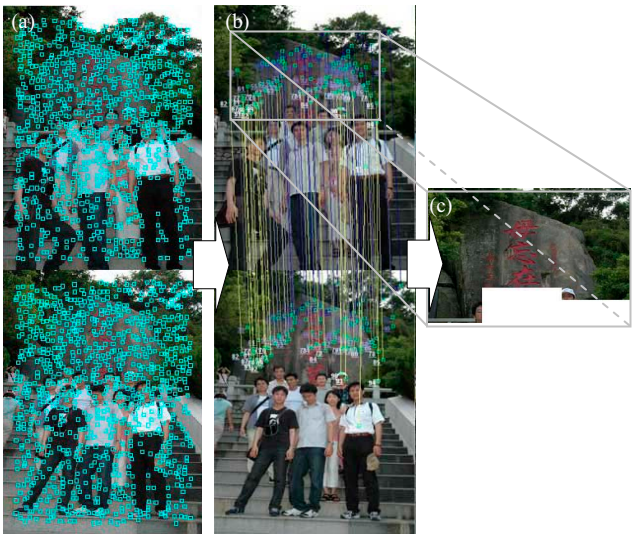

Figure 2. Summarization process of a cluster within album. **(a)** The points show the salient points in each photo; **(b)** The points show the matched salient points between two photographs, and the line connects corresponding points between two photographs; **(c)** The result of scene summarization for a cluster.

5. EXPERIMENTAL RESULTS

We took three collections of consumer photographs as our test collections to demonstrate the effectiveness of the proposed framework. Each collection was taken during trips. A ground truth is built by manually classifying all the photographs according to their events. The F measure is adopted to evaluate the generated hierarchy:

$$\text{F measure}\ (a,b) = \frac{2 \cdot \text{precision}\ (a,b) \cdot \text{recall}\ (a,b)}{\text{precision}\ (a,b) + \text{recall}\ (a,b)}$$

where a is the class in the ground truth, b is the cluster in the generated hierarchy, precision(a, b) and recall(a, b) are defined as:

$$\text{precision}\ (a,b) = \#\{a \cap b\} / \#\{b\},\ \ \text{recall}(a,b) = \#\{a \cap b\} / \#\{a\}.$$

For the entire cluster hierarchy, the F measure of any class is the maximum value it attains at any node in the tree, and an overall F measure is computed by taking the weighted average of all the values of the F measure as follows:

$$\text{F measure} = \sum_{a} \frac{\#\{a\}}{N} \cdot \max\{\text{F-measure}(a,b)\},$$

where N is the number of photographs in the collection.

We list the clustering performances under four different configurations in Table 1. The first row is the clustering result using intra-class only, and the second row is inter-class only. The third and fourth rows are the combination of 50% intra-class and 50% inter-class. They are different in their content analysis method, where the third row result is generated by extracting the whole region of an image, and the fourth row result is generated by extracting ROIs in an image. The fourth row that shows the performance of the proposed framework (i.e., ROIs) can approximate 88% F-measure rate overall. It reveals that the proposed framework is feasible for organizing consumer photographs. Summarization examples of photo clusters are shown in Figure 3.

Table 1. The F-measure results of hierarchical clustering.

	Coll. 1 *(135 photos)*	***Coll. 2*** *(291 photos)*	***Coll. 3*** *(373 photos)*
intra-class only	*0.3639*	*0.2280*	*0.1247*
inter-class only	*0.6841*	*0.6538*	*0.5311*
whole region	*0.8483*	*0.8569*	*0.8476*
ROIs	***0.8794***	***0.8880***	***0.8770***

6. CONCLUSIONS AND FUTURE WORK

In this study, we propose an automatic photo organization framework. Region of Interests are extracted to cluster photo collections. A summarization mechanism is applied to detect duplicates for a cluster. Thus a novel photo management and browsing experience is provided for users. Experimental results show the effectiveness of the introduced framework.

For future study, we will integrate GPS information to improve the clustering performance. Another direction is to integrate automatic and semi-automatic annotation mechanisms for photo clusters and summarization results.

Figure 3. Examples of a photo collection summarization.

7. REFERENCES

[1] M. Naaman, S. Harada, Q.Y. Wang, H. Garcia-Molina, and A. Paepcke, "Context data in geo-referenced digital photo collections," *ACM International Conference on Multimedia*, New York, USA, pp. 196-203, Oct. 10-16, 2004.

[2] M. A. Mottaleb and L. Chen, "Content-based photo album management using faces' arrangement," *IEEE International Conference on Multimedia and Expo,* 2004.

[3] J. Luo, A. E. Savakis, and A. Singhal, "A Bayesian network-based framework for semantic image understanding," *Pattern Recognition*, Vol. 38, No. 6, pp. 919-934, 2005.

[4] A. Jaimes, A.B. Benitez, S.F. Chang, and A.C. Loui, "Discovering recurrent visual semantics in consumer photographs," *IEEE International Conference on Image Processing*, Vancouver, Canada, pp. 528-531, 2000.

[5] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," *ACM International Conference on Multimedia*, 2003.

[6] A. Loui, and A. E. Savakis, "Automatic Image Event Segmentation and Quality Screening for Albuming Applications," *IEEE International Conference on Multimedia and Expo*, 2000.