Clustering Geo-tagged Photo Collections Using Dynamic Programming *

Matthew Cooper
FX Palo Alto Laboratory
Palo Alto, CA USA
cooper@fxpal.com

ABSTRACT

This paper describes methods for clustering photos that possess both time stamps and geographical coordinates as metadata. We present a two part method that first analyzes photos' time and location information to independently partition the photos into multiple clusterings. A subset of the detected clusters is then selected for the final photo clustering using an efficient dynamic programming procedure that optimizes a clustering fitness score. We propose fitness measures to produce clusterings that are coherent in space, time, or both. One group of scores directly measures within-cluster inter-photo distances. A second set of scores measures clusters' consistency with the reference clusterings. We present experiments that validate our method using multiple data sets.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL—*Content Analysis and Indexing Indexing Methods*

General Terms

Algorithms, Management

Keywords

Digital photo organization, event clustering, digital libraries

1. INTRODUCTION

As digital photography continues its explosive growth, personal photo collections require more powerful management tools. A common organizational step for consumers prior to either sharing photos with others or creating personal albums is the grouping of photos, often around events. Automatic processing plays an increasingly important role here as personal photo collections grow in scale. The growing availability of geographic information recorded at photo capture represents an opportunity to enhance automatic management tools. Newer digital cameras and more commonly smart phones record latitude and longitude coordinates with photos.

In this paper, we process location and time metadata for event-based photo clustering within a dynamic programming (DP) framework. We order the photos by time and segment the photo collection into contiguous partitions by analyzing inter-photo similarity in both time and space independently. This generates a set of reference clusterings that each characterize the collection's structure in either time or space at a specific scale. The clusters comprising the reference clusterings are candidates for inclusion in the final photo clustering. We examine multiple clustering fitness scores for optimization by DP. The first scores measure within-cluster dissimilarity in either time or space. We next take an ensemble clustering perspective to assimilate the structural information within the reference spatial and temporal clusterings. We propose a score to quantify the shared information between a candidate cluster and the reference clusterings. Using this score, the DP procedure selects clusters to maximize this shared information. We describe our method in detail below and present experiments on several datasets demonstrating a high level of performance.

2. RELATED WORK

Events provide a natural framework for organizing both digital photos and more general multimedia collections [15]. A number of time-based approaches to event-based photo clustering have been presented [7, 11]. [5] presented a dynamic programming method for temporal photo clustering that we extend here to incorporate location information. Naaman, *et al.* [9] performed early work on clustering combining time and location information. Initially, time alone is used to over-segment the photos. Recorded locations are independently hierarchically grouped in clusters. In a third pass, temporal segments that belong to the same location cluster are merged. Pigeau, *et al.* [3, 10], also present a multi-pass system. The first pass performs clustering using mixture models learned jointly on the time and location data with a variational approach to determine model order. In a second pass, clusters are grouped using information measures and the mixture parameters. Cao *et al.* [4] studied hierarchical image annotation using event clustering of photos. Their method first processed time and then location using mean shift clustering. In contrast, we compute tempo-

*Area chair: Daniel Gatica-Perez

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

ral and spatial reference clusterings of the photo collection separately, but select a subset from the union of detected clusters. Our approach is non-parametric and cluster selection is performed via DP to directly optimize cluster fitness measures.

Ensemble clustering refers to the class of problems in which a final clustering must be determined given a set of available clusterings. In this context, various information theoretic criteria for clustering assessment have been studied [13]. We require that a clustering S labels each of the N photos as a member of exactly one cluster $s \in S$, and that each cluster s is contiguous in time order.[1] The mutual information between two clusterings R and S is:

$$I(R;S) = \sum_{r\in R, s\in S} P(r,s)\log\left(\frac{P(r,s)}{P(r)P(s)}\right) \quad (1)$$

$$\text{where } P(r) = \frac{|r|}{N} \quad \text{and} \quad P(r,s) = \frac{|r\cap s|}{N}$$

Direct application of mutual information favors over segmentation. To counter this, normalized forms have been proposed [12, 13]. We use the normalized mutual information (NMI):

$$NMI(R;S) = \frac{I(R;S)}{\sqrt{H(R)H(S)}} \quad (2)$$

where $H(R) = -\sum_{r\in R} P(r)\log(P(r))$ is the entropy of R.

Becker *et al.* [2] used NMI for event-based grouping of shared multi-user photo collections in a setting analogous to the TREC event detection and tracking task [14]. Given a number of information streams, the goal is to identify events and then group photos belonging to each event. Becker *et al.* compiled photos from Flickr with user annotated event tags from upcoming.org for testing. Their approach combined supervised classification with NMI measures for clustering. We use dynamic programming to construct a photo clustering with maximum average NMI with a set of reference clusterings. In contrast to [2], we perform unsupervised clustering of *single user* photo collections. We assume non-overlapping event clusters that partition the user's photo stream in time order.

3. PHOTO CLUSTERING

We cluster photos in two steps. The first step assembles a set of *reference clusterings* to characterize the structure of the photo collection. The second step builds a final clustering as a subset of candidate clusters from the reference clusterings. A DP procedure constructs the final clustering to share maximal information with the reference clusterings by optimizing a mutual information score.

3.1 Building reference clusterings

We assemble reference clusterings by segmenting the time-ordered photos into clusters with high pairwise similarity. This process is performed independently using both temporal and spatial similarity at multiple scales. Denote the time and location of the n^{th} indexed photo by t_n and l_n, respectively. For temporal clustering, we follow [5] which builds a hierarchical temporal partitioning using an exponential set of inter-photo similarity measures:

$$s_\tau(i,j) = \exp\left(-\frac{|t_i - t_j|}{\tau}\right) \quad , \quad (3)$$

where τ is varied to produce a set of temporal reference clusterings across a range of scales. For spatial reference clusterings, we use the distance between photo locations to construct a similar hierarchical spatial partitioning:

$$s_\sigma(i,j) = \exp\left(-\frac{d_g(l_i,l_j)}{\sigma}\right) \quad . \quad (4)$$

d_g is the approximate geodesic distance.[2] Each value of σ generates a spatial reference clustering. We add each clustering calculated using either (3) or (4) for a specific value of σ or τ into the set of reference clusterings $\mathcal{R}$. The set of all constituent clusters in $\mathcal{R}$ are candidates for potential selection by the DP procedure described below.

3.2 Cluster selection

We efficiently select the subset of candidate clusters that comprise the final photo clustering using a standard DP procedure for grouping an ordered set of objects [6, 8]. We represent each candidate cluster by the indices of its boundaries, and denote the set of all boundaries by $\mathcal{B}$. Generally, $\beta = |\mathcal{B}| \ll N$, the number of photos. We present two sets of scores for optimization via DP.

3.2.1 Content-based selection

The content-based cost of the cluster with boundary indices b_i and b_j is the total pairwise distance between the cluster's photos:

$$C_F(b_i,b_j) = \sum_{m,n=b_i}^{b_j-1} d(m,n) \quad . \quad (5)$$

We consider two distance measures:

$$d(m,n) = \begin{cases} |t_m - t_n| & \text{for temporal selection} \\ d_g(l_m,l_n) & \text{for spatial selection} \end{cases} . \quad (6)$$

DP successively builds minimum cost partitionings with k clusters based on the minimum cost partitioning with $k-1$ clusters. $E_F(b_j,k)$ is the optimal partitioning of the photos indexed $1,\cdots,b_j$ with cardinality k and is computed iteratively:

$$E_F(b_j,k) = \min_{b_i\in\mathcal{B}, k\le b_i\le b_j} E_F(b_i,k-1) + C_F(b_i,b_j) \ , \quad (7)$$

The result is a set of minimum cost partitionings with cardinality[3] $3 \le K \le \beta$. A traceback step identifies the clusters in each optimal clustering. The total clustering cost $E(N,K)$ decreases monotonically with K. We heuristically select the optimal cardinality, K^*, based on the ratio of change in the total clustering cost:

$$K^* = \underset{2\le k\le\beta-1}{\text{argmax}} \frac{E_F(N,k)}{E_F(N,k+1)} \quad . \quad (8)$$

3.2.2 Using normalized mutual information

An advantage of the DP framework is that it can select from the combined set of candidate clusters detected using

[1] Throughout, we refer to a labeling S of all N photos as a clustering, and to any of its constituent elements $s \in S$ as a cluster that labels some subset of the photos.

[2] See http://en.wikipedia.org/wiki/Great-circle_distance

[3] We define $b_1 = 1$, and $b_\beta = N$.

either spatial or temporal similarity. However, it is not clear how to effectively fuse the spatial and temporal distances of (6) into a combined distance in (5). Instead, we use NMI to integrate the spatial and temporal information in the reference clusterings $\mathcal{R}$. Using DP, we construct the clustering S to maximize the average shared information with each clustering $R \in \mathcal{R}$:

$$\frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} NMI(R;S) \quad . \tag{9}$$

This score prefers candidate clusters that are consistent in both time and space or across multiple scales. To apply DP, we decompose this total score into terms associated with a single candidate cluster $s \in S$. Define:

$$I(s;R) = \sum_{r \in R} P(r|s) \log \left(\frac{P(r|s)}{P(r)} \right) \quad , \tag{10}$$

and notice that

$$NMI(R;S) = \frac{1}{\sqrt{H(S)}} \sum_{s \in S} P(s) \frac{I(s;R)}{\sqrt{H(R)}} \quad . \tag{11}$$

Let s_{ij} be the cluster of photos between indices b_i and b_j. We then define a score for maximization by DP as in (5):

$$C_{NMI}(b_i, b_j) = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} P(s_{ij}) \frac{I(s_{ij};R)}{\sqrt{H(R)}} \quad , \tag{12}$$

This score is inserted into (7), replacing minimization (of within-cluster dissimilarity) with maximization (of average NMI with $\mathcal{R}$). Note we have neglected the $H(S)$ term from (11) in the score of (12). The DP procedure thus maximizes a scaled form of the average NMI:

$$E_{NMI}(S) = \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} \frac{I(R;S)}{\sqrt{H(R)}} \quad . \tag{13}$$

The $H(R)$ terms implicitly weight each reference clustering R, generally emphasizing clusterings with fewer clusters. This is consistent with the intuition that clusters detected at coarser scales are more important.

As before, the DP procedure produces a set of clusterings of different cardinalities, and we select the final cardinality K^* to maximize the average NMI of (9) directly. We consider the range of possible values, $3 \leq k \leq \beta$ and perform the traceback to select the clusters that maximize the score of (13). We denote the resulting clustering with cardinality k by S_k, and compute its entropy. We then scale the clustering score $E_{NMI}(S_k)$ to determine the average NMI. We select the final clustering with cardinality:

$$K^* = \underset{3 \leq k \leq \beta}{\operatorname{argmax}} \left(\frac{1}{\sqrt{H(S_k)}} \, E_{NMI}(S_k) \right) \quad . \tag{14}$$

4. EXPERIMENTS

We test our clustering methods using multiple data sets. We assess performance with the B^3 score [1, 2] which is the geometric mean of the per-photo precision and recall averaged over the collection. The first data set consists of two users' personal photographs accumulated over 15 months. The two sets contain 1036 and 413 photos respectively and include timestamps *without* location information. In this case, our method integrates structural information across temporal scale for photo clustering. We compute reference clusterings from the photos' timestamps by varying τ in (3) using the similarity-based method of [5]. We then apply the DP methods of Section 3.2.1 and 3.2.2 and compare performance with the confidence measure method described in [5] as a baseline. Table 1 shows that the DP method using the NMI score described in Section 3.2.2 achieves superior performance.

Table 1: Summary statistics for temporal photo clustering. Per-photo precision and recall are computed across the two single user data sets.

Clustering Method	Average Precision	Average Recall	B^3
[5]	0.714	0.855	0.778
Sec. 3.2.1	0.887	0.628	0.735
Sec. 3.2.2	0.951	0.915	0.932

The next data set is described in [2] and publicly available[4]. Photos are collected from Flickr with user annotated event tags from the upcoming.org website. We filtered the dataset, first removing photos without geotags and then grouping photos by user. We retained per-user photo sets with at least 350 photos and 10 events, producing 32 single-user data sets with a total of 37,434 photos. We compute temporal reference clusterings as above, and compute spatial reference clusterings similarly by varying σ in (4).

We examine several variants of each DP method for testing. We compare using temporal reference clusterings (TIME), spatial reference clusterings (SPACE), and both (COMBO). Each DP method uses the constituent clusters of $\mathcal{R}$ as candidates for selection. The content-based DP method of Section 3.2.1 selects clusters to minimize the cost of (5) using either temporal (TIME) or spatial (SPACE) distance as in (6). The NMI-based DP method of Section 3.2.2 selects clusters according to the NMI score of (12) using temporal reference clusterings (TIME), spatial reference clusterings (SPACE), or both (JOINT).

Experimental results appear in Table 2. The strictly temporal baseline of [5] achieves $B^3 = 0.680$ ($P = 0.943, R = 0.531$), underperforming the proposed DP methods. Content-based DP, described in Section 3.2.1, shows substantial variation in performance with different usage of spatial and temporal information for both building reference clusterings and cluster selection. While this approach improves on the performance of the temporal baseline, matching the configuration to a given data set is a critical practical consideration which may require a training stage or prior knowledge of the data or photographers.

In the absence of training data, it's often desirable to use all available information. An advantage of the NMI-based DP procedure of Section 3.2.2 is that it integrates available information in a principled framework both for building reference clusterings and also for cluster selection. Although combining the temporal and spatial information in both steps does not provide optimal performance (COMBO/JOINT $B^3 = 0.852$), it does essentially match the best performing system for content-based DP (SPACE/TIME, $B^3 = 0.879$). Also, the variation in performance with configuration is relatively limited using the NMI score. More gener-

[4]http://www.cs.columbia.edu/∼hila/wsdm-data.html

Table 2: Summary statistics for photo clustering using a subset of the upcoming data from [2].

Content-based DP: Section 3.2.1				
Reference Clusterings	Cluster Score	Average Precision	Average Recall	B^3
TIME	TIME	0.727	0.865	0.790
TIME	SPACE	0.560	0.651	0.602
SPACE	SPACE	0.487	0.918	0.636
SPACE	TIME	0.859	0.900	0.879
COMBO	TIME	0.742	0.811	0.775
COMBO	SPACE	0.497	0.690	0.578
NMI-based DP: Section 3.2.2				
Reference Clusterings	Cluster Score	Average Precision	Average Recall	B^3
TIME	TIME	0.944	0.733	0.825
TIME	SPACE	0.902	0.905	0.904
TIME	JOINT	0.946	0.761	0.844
SPACE	SPACE	0.905	0.881	0.893
SPACE	TIME	0.903	0.932	0.917
SPACE	JOINT	0.906	0.926	0.917
COMBO	TIME	0.943	0.736	0.826
COMBO	SPACE	0.904	0.880	0.892
COMBO	JOINT	0.940	0.780	0.852

ally, the NMI-based DP methods perform at a higher level than the content-based DP methods. The best performing NMI variations achieve the highest level of performance of all tested methods.

5. SUMMARY

In this paper, we present methods for clustering photos with timestamps and geographic information using dynamic programming. Motivated by ensemble clustering, we propose an NMI score for selecting clusters that share maximal average information with a set of reference clusterings. The reference clusterings characterize the photo collection's structure in space and time across multiple scales. The NMI-based method shows the best experimental performance in large scale tests.

Our present approach applies to any ordered data with heterogeneous attributes that can characterize the data's structure. In future work, we hope to incorporate other modalities including content-based similarity of photos and text tag-based similarity which is increasingly available. We expect these extensions of this method to further enhance our automatic clustering results. Additionally, we aim to extend the method to unordered collections by looking beyond clustering via dynamic programming.

6. REFERENCES

[1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12:461–486, August 2009.

[2] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proc. of the ACM Intl. Conf. on Web Search & Data Mining*, WSDM '10, pages 291–300, New York, NY, USA, 2010. ACM.

[3] P. Bruneau, A. Pigeau, M. Gelgon, and F. Picarougne. Geo-temporal structuring of a personal image database with two-level variational-bayes mixture estimation. In *Adaptive Multimedia Retrieval Workshop (AMR'08)*, 2008.

[4] L. Cao, J. Luo, H. Kautz, and T. S. Huang. Image annotation within the context of personal photo collections using hierarchical event and scene models. *IEEE Trans. on Multimedia*, 11:208–219, February 2009.

[5] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Trans. Multimedia Comput. Commun. Appl.*, 1(3):269–288, 2005.

[6] W. D. Fisher. On grouping for maximum homogeneity. *J. of the American Statistical Association*, pages 789–798, December 1958.

[7] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd. Time as essence for photo browsing through personal digital libraries. In *Proc. of the ACM/IEEE-CS Joint Conf. on Digital libraries*, JCDL '02, pages 326–335, New York, NY, USA, 2002. ACM.

[8] J. Hartigan. *Clustering Algorithms*. John Wiley and Sons, 1975.

[9] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *JCDL '04: Proc. of the ACM/IEEE-CS Joint Conf. on Digital libraries*, pages 53–62, New York, NY, USA, 2004. ACM.

[10] A. Pigeau and M. Gelgon. Building and tracking hierarchical geographical & temporal partitions for image collection management on mobile devices. In *MULTIMEDIA '05: Proc. of the ACM Intl. Conf. on Multimedia*, pages 141–150, New York, NY, USA, 2005. ACM.

[11] J. Platt, M. Czerwinski, and B. A. Field. Phototoc: automatic clustering for browsing personal photographs. In *Information, Communications and Signal Processing, 2003 and the Pacific Rim Conf. on Multimedia*, volume 1, pages 6–10, 2003.

[12] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Machine Learning Research*, 3:583–617, March 2003.

[13] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Machine Learning Research*, 11:2837–2854, October 2010.

[14] C. L. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Intl. Conf. on Language Resources and Evaluation*, 2000.

[15] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE MultiMedia*, 14:19–29, January 2007.