Pattern Recognition Letters xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Intelligent photo clustering with user interaction and distance metric learning

Meng Wang [a,*], Dinghuang Ji [b], Qi Tian [c], Xian-Sheng Hua [d]

[a] AKiiRA Media Systems Inc., Palo Alto 94301, USA
[b] Institute of Computing Technology, Beijing 100190, PR China
[c] University of Texas at San Antonio, USA
[d] Microsoft Research Asia, Beijing 100080, PR China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Photo clustering
Distance metric learning
Online learning
Interactive computing

ABSTRACT

Photo clustering is an effective way to organize albums and it is useful in many applications, such as photo browsing and tagging. But automatic photo clustering is not an easy task due to the large variation of photo content. In this paper, we propose an interactive photo clustering paradigm that jointly explores human and computer. In this paradigm, the photo clustering task is semi-automatically accomplished: users are allowed to manually adjust clustering results with different operations, such as splitting clusters, merging clusters and moving photos from one cluster to another. Behind users' operations, we have a learning engine that keeps updating the distance measurements between photos in an online way, such that better clustering can be performed based on the distance measure. Experimental results on multiple photo albums demonstrated that our approach is able to improve automatic photo clustering results, and by exploring distance metric learning, our method is much more effective than pure manual adjustments of photo clustering.

1. Introduction

With the popularity of digital cameras, recent years have witnessed a rapid growth of personal photos. More and more people capture photos to record their lives and share them on the web. For example, Facebook and Flickr are declared to host 10 billion and 4 billion personal photos, respectively.[1,2] Clustering is an effective approach to helping users manage, browse and annotate photos. Here a "cluster" is referred to as a batch of photos that are visually and semantically consistent. Album summarization is a typical application. By grouping photos into clusters, we can select some photos from each cluster and these photos together form a good summarization of the whole album, which can be useful in album visualization and photo sharing (Sinha et al., 2009; Papadopoulos et al., 2010). Another application is batch annotation. Given a photo set, manually annotating each photo will be a labor-intensive process and there is also a waste of efforts as many photos are close to some others. For example, many photos are continuously captured and they usually describe the same scene or object. Batch annotation is an effective approach to reducing the cost by directly assigning a set of tags to a batch of photos. Therefore, if the images in a set can be effectively clustered, then users can easily adopt batch annotation and tags only need to be assigned once for each cluster, such as the work in Liu et al. (2009).

Extensive efforts have been dedicated to automatic image clustering (Moellic et al., 2008; Wang et al., 2007; Goldberger et al., 2006; Mei et al., 2006; Yang et al., 2010; Nie et al., 2009). However, although great progress has been made, automatic photo clustering usually can hardly achieve satisfactory results due to the large variation of photos' content. Actually photo clustering also suffers from the "semantic gap" problem, which is the main challenge in multimedia content understanding. In comparison with photo semantic understanding tasks that are usually associated with labeled training data, getting unsupervised clustering results that are consistent with human's perception is more difficult, as it lacks supervision information. For example, the effective features and distance metric for photo clustering may vary across different albums. Without supervision, we can hardly verify the effectiveness of different features and distance metrics in the clustering process.

To address the problem, in this paper we introduce an interactive photo clustering paradigm. Instead of automatic clustering, users are supported to manually adjust clustering results, and supervision information thus becomes available. Three kinds of operations are supported, namely, *Move*, *Split*, and *Merge*. A set of equivalence and inequivalence constraints can be generated under the operations. An online distance metric learning is performed to construct a better distance measure for photo pairs. Consequently, we implement clustering with the updated metric. This process repeats until satisfactory clustering results are obtained. Fig. 1 illustrates the main scheme of our approach. We will show that

* Corresponding author. Tel.: +1 650 996 0471.
E-mail addresses: eric.mengwang@gmail.com (M. Wang), robotshanxi@gmail.com (D. Ji), qitian@utsa.edu (Q. Tian), xshua@microsoft.com (X.-S. Hua).

[1] http://en.wikipedia.org/wiki/Flickr
[2] http://www.facebook.com/press/info.php?statistics

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.02.012

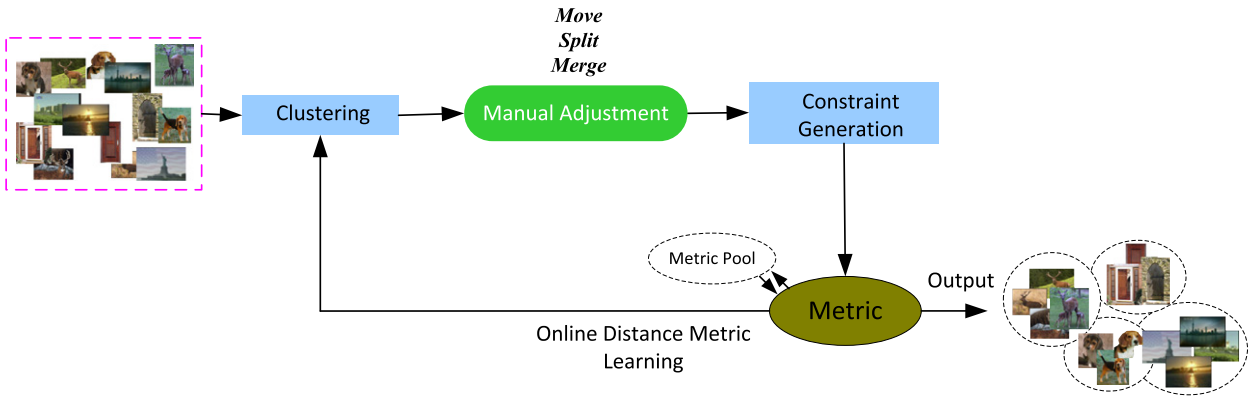

Fig. 1. A schematic illustration of semi-automatic photo clustering. Users are allowed to manually adjust photo clustering results with different operations and a distance metric is learned accordingly.

our approach can obtain better results than automatic clustering or pure manual adjustments of clustering results.

Our contribution can be summarized as follows:

(1) We propose an interactive photo clustering approach, in which users are allowed to adjust clustering results and a distance metric is learned based on users' interactions such that better clustering can be performed.
(2) We introduce an online distance metric learning algorithm that updates distance metric based on a new set of constraints, which can maintain close performance with the traditional method while significantly reducing computational cost.

The organization of the rest of this paper is as follows. In Section 2, we provide a review on the related work. In Section 3, we introduce our semi-automatic photo clustering approach based on distance metric learning. Experiments are presented in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

There is extensive research on automatic image clustering. Some of them explore textual information that is associated with the images, such as surrounding text or tags (Moellic et al., 2008; Cai et al., 2004; Wang et al., 2007). Goldberger et al. (2006) proposed an image clustering approach based on information theoretic scheme. Kennedy and Naaman (2008) and Simon et al. (2007) proposed two clustering methods to extract representative images from an image set. In Platt (2000), an AutoAlbum scheme is proposed, which adopts a photo clustering approach with the help of temporal information and the order of photo creation. Platt et al. (2003) then further proposed PhotoTOC, which supplies a new interface to assist users efficiently get the photos they want. Sinha et al. (2009) proposed a clustering-based method to select a set of photos as the summarization of an album. Papadopoulos et al. (2010) employed clustering to select representative photos corresponding to landmarks and events in a city. Cooper et al. (2003) and Mei et al. (2006) both integrated photos' visual content and temporal information to cluster them into different events. Photo clustering technology has also been widely explored in face tagging. By performing clustering on photos based on the features extracted from face regions, it is expected that photos with the same faces can be grouped into a cluster and thus the tagging cost can be reduced as the photos in such clusters only need to be tagged once. Suh and Bederson (2004) adopted such an approach. They first group faces into clusters and then label each cluster. However, as previously mentioned, automatic photo clustering is a difficult task and it will suffer from the lack of supervision. Tian et al. (2007) proposed a method that integrates user interaction in order to boost clustering performance. First, only very close faces are grouped with partial clustering, and the other faces are regarded as a background cluster. Then in the labeling process of clusters, information is learned and several faces in the background cluster are further clustered for labeling. However, there is no investigation of allowing users to directly adjust clustering results. In this work, we support users to make different operations on photo clustering results and learn distance measurements between photos based on users' interactions. By iteratively learning from users' operations, significant performance improvement of photo clustering can be achieved.

Distance metric learning is intended to construct an optimal distance metric for the given learning task based on the pairwise relationships among samples. A number of algorithms have been proposed for distance metric learning. Bar-Hillel proposed a Relevant Components Analysis (RCA) method to learn a linear transformation from the equivalence constrains, which can be used directly to compute the distance between two examples (Bar-Hillel et al., 2005). Xing et al. (2003) formulated distance metric learning as a constrained convex programming problem by minimizing the distance between the data points in the same classes under the constraint that the data points from different classes are well separated. Neighborhood Component Analysis (NCA) Goldberger et al. (2004) learned a distance metric by extending the nearest neighbor classifier. Weinberger et al. (2006) proposed the maximum-margin nearest neighbor (LMNN) method that extends NCA through a maximum margin framework. Alipanahi et al. (2008) show a strong relationship between distance metric learning methods and Fisher Discriminant Analysis (FDA). Davis et al. (2007) proposed an information-theoretic metric learning algorithm which learns a Mahalanobis distance by minimizing the differential relative entropy between two multivariate Gaussians under constraints on the distance function. Hoi et al. (2008) proposed a semi-supervised distance metric learning method that integrates both labeled and unlabeled examples. In this work we employ the information-theoretic metric learning algorithm as its superiority over many other distance metric learning methods has been shown in Davis et al. (2007). However, the complexity scales linearly with the number of constraints and the time cost will dramatically increase when there is a large number of constraints. Therefore, in this work we proposed an online learning approach that keeps updating distance metric based on recent user interactions. This approach can achieve close performance in comparison with the conventional method while significantly reducing much time cost.

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012

Input:

Photos $\mathcal{F} = \{x_1, x_2, \cdots, x_n\}$;

The number of clusters m; //the number can be specified by users

Output:

Clusters $C_1, C_2, \cdots, C_m$;

1. Initialize distance metric $\mathbf{M}$ to be Euclidean distance;
2. Cluster $\mathcal{F}$ into m clusters with the distance metric $\mathbf{M}$;
3. Manually adjust clustering results;
4. Generate constraints based on the manual adjustments;
5. Update distance metric with the constraints to obtain a new $\mathbf{M}$;
6. Go to 1 and repeat the process, until user is satisfied with the clustering performance.

Fig. 2. The implementation process of the semi-automatic photo clustering approach.

3. Interactive photo clustering

As introduced in Section 1, our interactive photo clustering works as follows. First, photos are grouped into a certain number of clusters (the number can be specified by users). Then users can view the clustering results and make manual adjustments. Based on the adjustments, equivalence and inequivalence constraints among photos are generated and a distance metric learning algorithm is performed. Consequently, we perform clustering again with the learned distance metric and the process can repeat until a satisfactory clustering performance is achieved. Fig. 2 illustrates the implementation process.

3.1. Photo clustering with a distance metric

We adopt spectral clustering approach. It is a technique that explores the eigenstructure of a similarity matrix to partition samples into disjoint clusters with samples in the same cluster having high similarity and points in different clusters having low similarity (von Luxburg, 2007; Bach and Jordan, 2003; Ding, 2004). The spectral clustering can be viewed as a graph partition task. The simplest and most straightforward way to construct a partition of the similarity graph is to solve the Min-cut problem, and Normalized-cut can also be used to replace Min-cut (Shi and Malik, 2000; Chan et al., 1994) in order to obtain more stable clusters. Existing studies have shown that spectral clustering outperforms many conventional clustering algorithms such as *k*-means algorithm. Here we adopt the method proposed in Chan et al. (1994). But in our scheme we have learned a distance metric from users' operations of clustering results, thus we adopt the distance metric in the computation of the similarity matrix, i.e.,

$$W_{ij} = \exp\left(-\frac{(x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j)}{\sigma^2}\right) \tag{1}$$

The clustering process is illustrated in Fig. 3.

3.2. Manual adjustment and constraint generation

Clustering results can be adjusted in different ways, and here we consider three basic operations: *Move*, *Merge* and *Split*. Clearly, with these operations, we can change a set of clustering results to any other clustering results with finite steps. Based on each operation, we can generate a set of constraints among photos. Here we describe the three operations and the constraint generation strategies as follows:

(1) Moving a photo from one cluster to another. We denote the operation as $Move(x, C_i, C_j)$, which indicates removing a sample x from C_i and adding it to C_j. For this operation, we can assume that x forms an inequivalence constraint with each remained samples in C_i as it has been moved out from C_i. On the contrary, we can generate an equivalence constraint for x and each sample in C_j. Therefore, if the sizes of C_i and C_j are u and v respectively, we obtain $u - 1$ inequivalence and v equivalence constraints.

Input:

Photos $\mathcal{F} = \{x_1, x_2, \cdots, x_n\}$;

The number of clusters m; //the number can be specified by users;

Distance metric $\mathbf{M}$;

Output:

Clusters $C_1, C_2, \cdots, C_m$;

1. Construct a similarity graph based on Eq. (1);
2. Compute the unnormalized Laplacian $\mathbf{L}$;
3. Compute the first m generalized eigenvectors $u_1, \cdots, u_m$ of the generalized eigenproblem $\mathbf{L}u = \lambda \mathbf{D}u$;
4. Let $\mathbf{U} \in \mathcal{L}^{n \times m}$ be the matrix containing the vectors $u_1, \cdots, u_m$ as columns.
5. For $i = 1, \cdots, n$, let $y_i \in \mathcal{R}^m$ be the vector corresponding to the i-th row of $\mathbf{U}$;
6. Cluster the points $\{y_i\}_{(i=1,\cdots,n)}$ *in* $\mathcal{R}^k$ with the k-means algorithm into clusters $C_1, \cdots, C_m$.

Fig. 3. The implementation process of the spectral clustering algorithm with distance metric $\mathbf{M}$.

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012

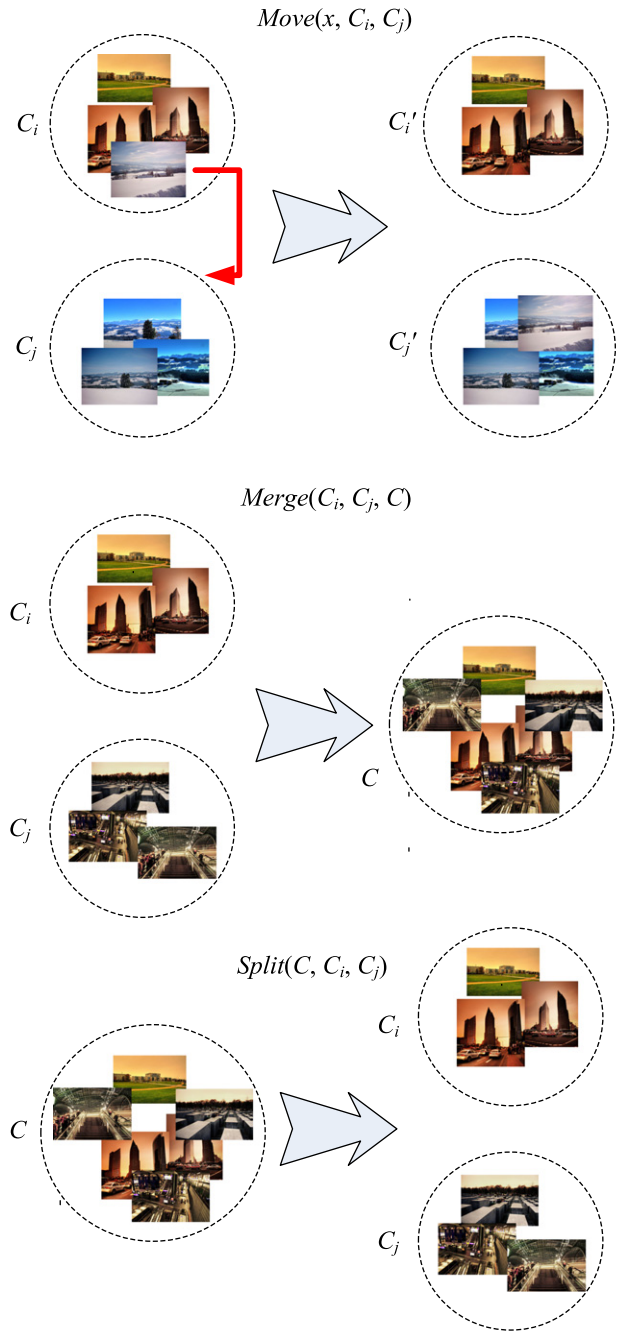

Fig. 4. The illustration of three operations of clustering results. *Move*(x,C_i,C_j) means removing a sample x from C_i and adding it to C_j, *Merge*(C_i,C_j,C) means merging two clusters C_i and C_j to obtain a new cluster C, and *Split*(C,C_i, C_j) means splitting a cluster C into C_i and C_j.

(2) Merging two clusters. We denote the operation as *Merge*(C_i,C_j,C), which means merging two clusters C_i and C_j to obtain a new cluster C. For this operation, we can assume that the sample pairs across C_i and C_j have equivalence constraints. Therefore, we obtain $u \times v$ equivalence constraints from the manipulation, where u and v are the sizes of C_i and C_j respectively.

(3) Splitting a cluster. We denote the operation as *Split*(C,C_i,C_j), which means splitting a cluster C into C_i and C_j. For this operation, we can assume that the sample pairs across C_i and C_j have inequivalence constraints. Therefore, we obtain $u \times v$ inequivalence constraints from the manipulation, where u and v are the sizes of C_i and C_j respectively.

Fig. 4 illustrates the examples of the three manipulations. We generate constraints among the photos, and distance metric learning is performed with these constraints.

3.3. Online distance metric learning

Based on the constraints among photos that are generated from users' operations, we employ distance metric learning to learn a Mahalanobis distance measure. In comparison with Euclidean distance or a globally learned distance metric, our learned distance metric reflects user's intention on the clustering, such as which kinds of photos should be clustered together, and thus using it can greatly boost clustering performance.

We employ the Information-Theoretic Metric Learning (ITML) algorithm in Davis et al. (2007). The most intuitive approach is to implement metric learning based on all the accumulated constraint information, but it can be analyzed that the cost of the ITML algorithm scales as $O(ld^2)$, where l is the number of constraints and d is the dimensionality of feature space. Thus the computational cost will increase dramatically if there is a large number of constraints. To address this problem, here we formulate an online distance metric learning algorithm that updates the metric with only the newly generated constraint in each round. First, we consider the Information-Theoretic Metric Learning (ITML) algorithm proposed in Davis et al. (2007). Denote by $\mathcal{S}$ and $\mathcal{D}$ the sets of sample pairs with equivalence and inequivalence constraints respectively, the ITML algorithm is formulated as

$$\begin{aligned} &\min_{\mathbf{M}} \; D(\mathbf{M},\mathbf{M}') \\ &\text{s.t. } \mathbf{M} \succeq 0 \\ &\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \leqslant T_1, \quad (i,j) \in \mathcal{S} \\ &\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \geqslant T_2, \quad (i,j) \in \mathcal{D} \end{aligned} \tag{2}$$

where $D(\mathbf{M},\mathbf{M}')$ is a divergence measure between $\mathbf{M}$ and $\mathbf{M}'$. Here $\mathbf{M}'$ is a distance metric that reflects certain prior knowledge, such that the learning of $\mathbf{M}$ can be regularized to be reasonable. For example, $\mathbf{M}'$ is usually set to $\mathbf{I}$, i.e., Euclidean distance. T_1 and T_2 are two predetermined parameters that satisfy $T_2 > T_1 > 0$. To be clear, we list all the notations and the definitions used in the distance metric learning algorithm in Table 1. We adopt Bregman divergence to measure the difference between $\mathbf{M}$ and $\mathbf{M}'$ as

$$D(\mathbf{M},\mathbf{M}') = g(\mathbf{M}) - g(\mathbf{M}') - < \nabla_g(\mathbf{M}'), \quad \mathbf{M} - \mathbf{M}' > \tag{3}$$

where $g(\cdot)$ is a strict convex and continuously differentiable function. We define $g(\cdot)$ as $-\log\det(\cdot)$ and get the log-determinant divergence between $\mathbf{M}$ and $\mathbf{M}'$

Table 1
Notations and their descriptions.

Notation	Description
$\mathbf{M}$	Mahalanobis distance metric to be learned
$\mathbf{M}_k$	The distance metric learned in kth round
$\mathcal{M} = \{\mathbf{M}_0, \mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_{k-1}\}$	Metric pool that contains the distance metric learned in the first k rounds
$\mathcal{S}$	Set of sample pairs with equivalence
$\mathcal{D}$	Set of sample pairs with inequivalence
w	The weights for integrating the existing metrics in online distance metric learning
ξ	Slack variables for softening constraints in distance metric learning (see Eq. (5))
T_1, T_2	Pre-determined parameters used in the ITML algorithm
ζ	Weighting parameter of the L_2 regularizer of w (see Eq. (8))
p, δ, α, β	Variables used in the iterative solution process of ITML (see Fig. 5)

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012

$$D(\mathbf{M},\mathbf{M}') = \mathrm{tr}(\mathbf{M}'\mathbf{M}^{-1}) - \log\det(\mathbf{M}'\mathbf{M}^{-1}) - n \tag{4}$$

By introducing slack variables, a soft version of the algorithm can be written as

$$\begin{aligned}
&\min_{\mathbf{M}} \; D(\mathbf{M},\mathbf{M}') + \gamma D(Diag(\xi), Diag(\xi')) \\
&\text{s.t. } \mathbf{M} \succeq 0 \\
&\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \leqslant \xi_{ij}, \quad (i,j) \in \mathcal{S} \\
&\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \geqslant \xi'_{ij}, \quad (i,j) \in \mathcal{D}
\end{aligned} \tag{5}$$

In the above equation, ξ is a vector of slack variables, and ξ' is set to T_1 and T_2 for the sample pairs with equivalence and inequivalence constraints respectively.

The most intuitive approach for realizing online distance metric learning is to employ the above formulation and set $\mathbf{M}'$ to the distance metric obtained in the last step. But here we change the formulation of ITML such that it considers all the previously obtained distance metrics. The rationality lies on the fact that the constraints for metric learning can be noisy and metric learning also may suffer from overfitting. Thus the metrics learned in different rounds are able to share complementary information. By using adaptively learned weights, integrating all previous metrics can be more robust than only updating the most recent one.

We change the objective function to $\sum_{\mathbf{M}'\in\mathcal{M}} w_k D(\mathbf{M},\mathbf{M}') + \gamma D(Diag(\xi), Diag(\xi)')$, where $\mathcal{M} = \{\mathbf{M}_0, \mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_{k-1}\}$ is the metric pool that contains the metrics obtained in the previous rounds and w_k is the weight for kth metric. Here $\mathbf{M}_0$ is set to $\mathbf{I}$, i.e., in the first round the distance metric is updated from Euclidean distance. We add a 2-norm regularizer on the weigh vector and then optimize the weights simultaneously. Thus the formulation becomes

$$\begin{aligned}
&\min_{\mathbf{M},w} \sum_{k=0}^{K-1} w_k D(\mathbf{M},\mathbf{M}_k) + \gamma D(Diag(\xi), Diag(\xi)') + \zeta\|w\|^2 \\
&\text{s.t.} \quad \mathbf{M} \succeq 0 \\
&\quad \sum_{k=0}^{K-1} w_k = 1 \\
&\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \leqslant \xi_{ij}, \quad (i,j) \in \mathcal{S} \\
&\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \geqslant \xi'_{ij}, \quad (i,j) \in \mathcal{D}
\end{aligned} \tag{6}$$

We adopt alternating optimization approach to solve the problem. First, we fix w and optimize $\mathbf{M}$, and thus the problem becomes

$$\begin{aligned}
&\min_{\mathbf{M}} \sum_{k=0}^{K-1} w_k D(\mathbf{M},\mathbf{M}_k) + \gamma D(Diag(\xi), Diag(\xi)') \\
&\text{s.t. } \mathbf{M} \succeq 0 \\
&\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \leqslant \xi_{ij}, \quad (i,j) \in \mathcal{S} \\
&\quad (x_i - x_j)^{\mathrm{T}}\mathbf{M}(x_i - x_j) \geqslant \xi'_{ij}, \quad (i,j) \in \mathcal{D}
\end{aligned} \tag{7}$$

Following the approach in Davis et al. (2007), we solve the above optimization problem by repeatedly projecting the solution onto each single constraint. The solution process is illustrated in Fig. 5.

We then fix $\mathbf{M}$ and update compute w, and it can be derived that

$$\begin{aligned}
&\min_{\mathbf{w}} \sum_{k=0}^{K-1} w_k D(\mathbf{M},\mathbf{M}_k) + \zeta\|w\|^2 \\
&\text{s.t.} \sum_{k=0}^{K-1} w_k = 1
\end{aligned} \tag{8}$$

It can be derived that

$$w_j = \frac{1}{K} + \frac{\sum_{k=0}^{K-1} D(\mathbf{M},\mathbf{M}_k) - KD(\mathbf{M},\mathbf{M}_j)}{K\zeta} \tag{9}$$

Since each step reduces the objective function in Eq. (5), the convergence of this iterative process is guaranteed. After updating the distance metric $\mathbf{M}$, we can perform spectral clustering again using the new metric, as described in Section 3.1.

Table 2
The numbers of photos in the albums used in our experiments.

Album	Photo number
Germany	163
China	500
HongKong	136
London	204
Korea	493
LongExposure	185
Manasquan	333
Nature	183
New Book	175
New York	117
Widelife	1221

1. Initialize $\mathbf{M}$ to be $\sum_{\mathbf{M}_k\in\mathcal{M}} w_k\mathbf{M}_k$;
2. Initialize t_{ij} to be 0 for every $(i,j) \in \mathcal{S}$ or $\mathcal{D}$;
3. Initialize ξ to be T_1 and T_2 for $(i,j) \in \mathcal{S}$ and $(i,j) \in \mathcal{D}$ respectively;
4. Repeat the following steps for each $(i,j) \in \mathcal{S}$ or $\mathcal{D}$ until convergence
   - 4.1. $p \longleftarrow (x_i - x_j)^T\mathbf{M}(x_i - x_j)$;
   - 4.2. $\delta \longleftarrow 1$ if $(i,j) \in \mathcal{S}$, -1 otherwise;
   - 4.3. $\alpha \longleftarrow \min(t_{ij}, \frac{\delta}{2}(\frac{1}{p} - \frac{\gamma}{\xi_{ij}}))$;
   - 4.4. $\beta \longleftarrow \frac{\delta\alpha}{1-\delta\alpha p}$;
   - 4.5. $\xi_{ij} \longleftarrow \frac{\gamma\xi_{ij}}{\gamma+\delta\alpha\xi_{ij}}$;
   - 4.6. $t_{ij} \longleftarrow t_{ij} - \alpha$;
   - 4.7. $\mathbf{M} \longleftarrow \mathbf{M} + \beta\mathbf{M}(x_i - x_j)(x_i - x_j)^T\mathbf{M}$;

Fig. 5. The iterative solution process of Eq. (7).

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012

4. Experiments

4.1. Experimental settings

We conduct experiments with 11 different personal albums that are collected from Flickr. These photos are captured at different locations around the world and contain diverse content, including the records of cityscape, landscape, wide life, etc. Table 2 illustrates the number of photos in these albums.

Many of the photos are of high resolution. To speed up feature extraction, we resize each photo such that its width is 240 pixels. Then we extract the following features from each photo in order to comprehensively describe its color, edge and texture: (1) 64-dimensional HSV color histogram; (2) 75-dimensional edge

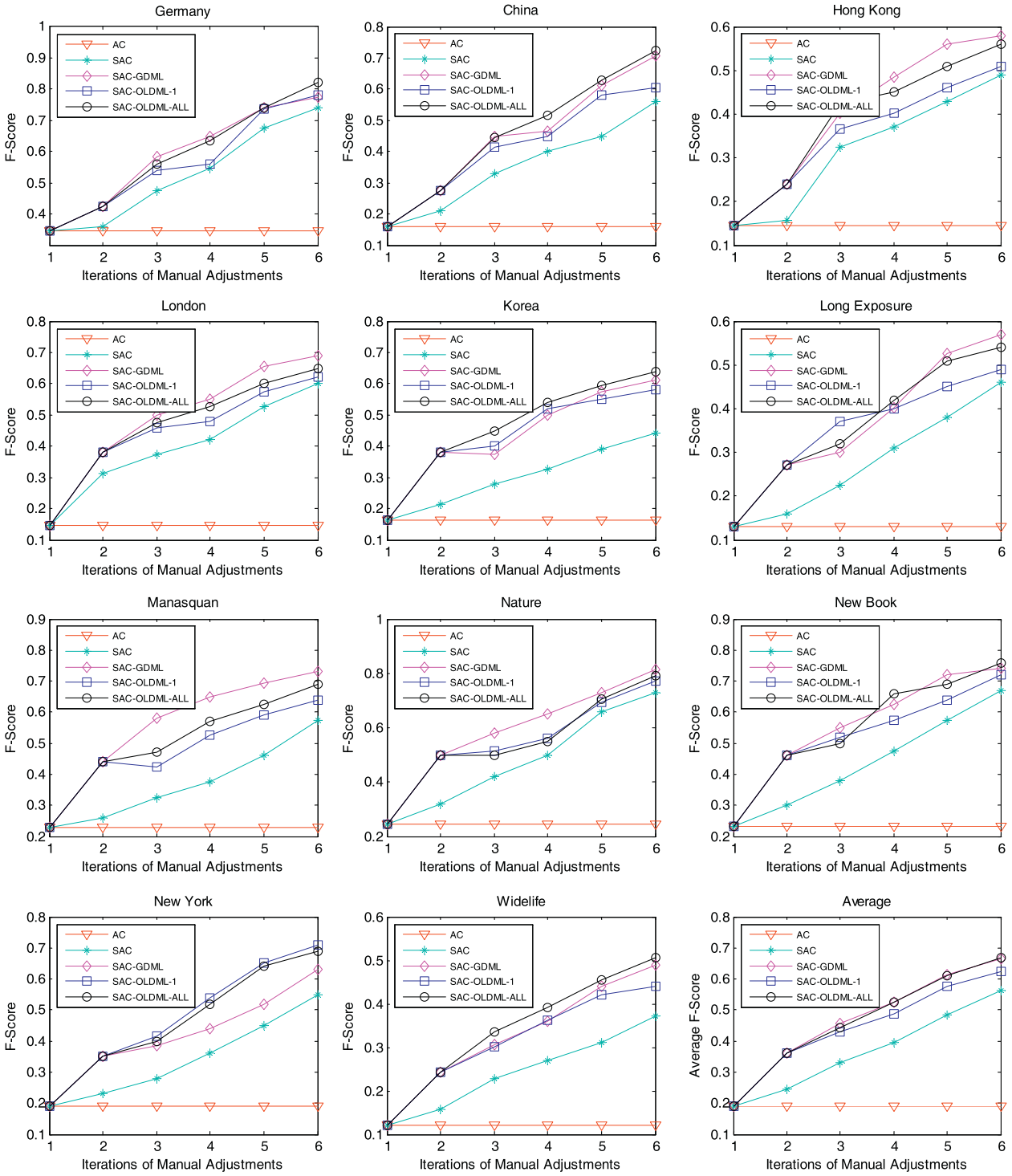

Fig. 6. Clustering performance comparison of different methods. We can see that the clustering performance can be significantly improved by adopting manual adjustments and distance metric learning. Our proposed online distance metric learning, which progressively updates distance metric based on all the previously obtained metrics, performs quite closely to the method that uses all accumulated constraints in each round, and it is better than the online method that only updates the most recent metric in most cases.

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012

Table 3
The left part illustrates the average rating scores and variances converted from the user study on the clustering performance comparison of SAC-OLDML-ALL and SAC-OLDML-1. The right part illustrates the ANOVA test results. The better result is illustrated in bold. The *p*-values show that the difference of the two learning methods is significant and the difference of users is insignificant.

SAC-OLDML-ALL vs. SAC-OLDML-1		The factor of methods		The factor of users	
SAC-OLDML-ALL	SAC-OLDML-1	*F*-statistic	*p*-value	*F*-statistic	*p*-value
1.9 ± 0.737	1.0 ± 0.0	14.878	0.0039	1.0	0.5

histogram; (3) 225-dimensional block-wise color moment features generated from 5-by-5 partition of the image; and (4) 128-dimensional wavelet texture features. These visual features have shown effectiveness in many image and video recognition tasks (Wang et al., 2009a,b; Wang and Hua, 2011).

How to evaluate our approach is a problem. Generally, if ground truth is available, many performance evaluation metrics are available for clustering, such as precision-recall, rand index and mutual information (Jain and Dubes, 1988). Here our first challenge is how to determine the ground truths for photo clustering. In some works, images are selected from exclusive classes and thus their ground truths can be naturally obtained. In several other works, such as Cooper et al. (2003) and Mei et al. (2006), the target is to group photos based on different events, and thus the photo clustering ground truths can be established based on the events. But for more general cases that photos do not belong to exclusive classes, subjective evaluation or user study will be involved (Cai et al., 2004; Wang et al., 2007; Kennedy and Naaman, 2008). Here we have designed a method to establish the subjective ground truths as follows. A user is asked to manually adjust clustering results until the best clustering results are achieved (the number of clusters is also specified by the user as there are *Split* and *Merge* operations). Then we evaluate different clustering approaches based on the ground truths. This is reasonable because the interactive clustering approach is aiming at the best clustering results in the user's opinion. There are three users involved in this process, two responsible for four albums each and the other one responsible for three albums. Based on the ground truths, we compute precision and recall measurements as follows

$$precision = \frac{|\mathcal{V} \cap \mathcal{V}_0|}{|\mathcal{V}|} \tag{10}$$

$$recall = \frac{|\mathcal{V} \cap \mathcal{V}_0|}{|\mathcal{V}_t|} \tag{11}$$

where

$\mathcal{V} = \{(x_i, x_j) | x_i \text{ and } x_j \text{ are in the same cluster after clustering}\}$

$\mathcal{V}_t = \{(x_i, x_j) | x_i \text{ and } x_j \text{ are in the same cluster in ground truths}\}$

Then F-score is computed as $2 * (precision * recall)/(precision + recall)$ and we adopt F-score as our clustering performance evaluation metric.

In order to validate the effectiveness of our approach on a large dataset, we also conduct experiments on a handwritten digit dataset (Hull, 1994), which contains 11,000 images and each image describes a digit from '0' to '9' (note that for this dataset, the ground truths for clustering are naturally available, that is, the digit labels). Each image is described with 256 pixels, and thus a 256-dimensional feature space is employed.

4.2. Experimental results

We compare the following five methods:

(1) Automatic clustering (AC), i.e., we perform spectral clustering with Euclidean distance without any adjustment of the clustering results.

Table 4
The left part illustrates the average rating scores and variances converted from the user study on the clustering performance comparison of SAC-OLDML-ALL and SAC. The right part illustrates the ANOVA test results. The better result is illustrated in bold. The *p*-values show that the difference of the two learning methods is significant and the difference of users is insignificant.

SAC-OLDML-ALL vs. SAC		The factor of methods		The factor of users	
SAC-OLDML-ALL	SAC	*F*-statistic	*p*-value	*F*-statistic	*p*-value
2.8 ± 0.422	1.0 ± 0.0	182	$7.21 \times e^{-11}$	1.0	0.5

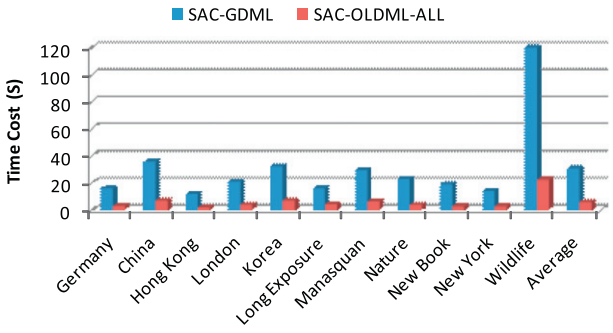

Fig. 7. The comparison of average response time of the SAC-GDML and SAC-OLDML-ALL methods in each round. We can see that SAC-OLDML-ALL is much rapid due to its less involved constraints in the distance metric learning.

(2) Semi-automatic clustering without distance metric learning (SAC), i.e., we perform spectral clustering with Euclidean distance and then adopt pure manual adjustments on the clustering results.
(3) Semi-automatic clustering with global distance metric learning (SAC-GDML), i.e., we learn a new distance metric after users adjust clustering results and then re-implement spectral clustering, and the metric learning is performed with all accumulated constraints in each round.
(4) Semi-automatic clustering with online distance metric learning using only the most recent metric (SAC-OLDML-1). We employ online distance metric learning but only using the most recently obtained metric, i.e., we adopt Eq. (2) and $\mathbf{M}'$ is set to $\mathbf{M}_{k-1}$ in kth round.
(5) Semi-automatic clustering with online distance metric learning using all previous metric (SAC-OLDML-ALL). This is our proposed method.

In our experiments, the interactive clustering of each album is test by the user that determines its ground truths. In each round we allow the user to make 10 manual adjusts operations (including *Move*, *Merge* and *Split* manipulations) to achieve a trade-off of the smoothness of the process and user's experience[3]. The parameter σ

[3] There is a dilemma here: users will prefer low tolerance and the response time is the shorter the better, but each time after re-performing the clustering results, users will need to further observe the results before making adjustments and there is actually implicit cost here. The number of manipulations set here can be viewed as a trade-off: we try to make the response time tolerable while keeping the re-clustering too frequent.

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012

Table 5
The left part illustrates the average rating scores and variances converted from the user study on the clustering performance comparison of SAC-OLDML-ALL and SAC-GDML. The right part illustrates the ANOVA test results. The better result is illustrated in bold. The *p*-values show that the difference of the two learning methods is significant and the difference of users is insignificant.

SAC-OLDML-ALL vs. SAC-GDML		The factor of methods		The factor of users	
SAC-OLDML-ALL	SAC-GDML	*F*-statistic	*p*-value	*F*-statistic	*p*-value
2.3 ± 0.823	1.1 ± 0.316	13.5	0.005	0.46	0.87

in spectral clustering is set to the median value of the pairwise distances of all photos. Following the strategy in Davis et al. (2007), the parameters T_1 and T_2 are set to the 5th and 95th percentiles of the pairwise distances of all photos in the album. As it lacks a good method to tune the parameters γ and ζ, they are empirically set to 1 and 100 respectively in our experiments.

Fig. 6 illustrates the performance comparison of different methods on the 11 albums as well as the handwritten digit image dataset. We also illustrate the average results. We can see that semi-automatic methods can effectively improve clustering performance in comparison with automatic clustering. By adopting distance metric learning, the clustering performance can be significantly improved in comparison with relying on pure manual adjustments. For most albums, the proposed SAC-OLDML-ALL method performs closely to the method of SAC-GDML, i.e., employing all constraints for metric learning. From the average results we can also see that the two performance curves are quite close. But the online method is quite faster due to the less involved constraints in each round. Fig. 6 illustrates the comparison of the average time costs of the SAC-GDML and SAC-OLDML-ALL, and we can see that the SAC-OLDML-ALL method is several times faster. All the time costs mentioned in this paper are recorded on a PC with Pentium 3.40 G CPU and 2 G memory. Comparing the SAC-OLDML-ALL and SAC-OLDML-1 methods, we can see that the SAC-OLDML-ALL method is better for most albums, including "Germany", "China", "Hongkong", "London", "Korea", "Manasquan", "New Book" and "Widelife". As introduced in Section 3, this is because integrating all previous metrics with adaptively learned weights can be more robust than using only the most recent metric. The superiority of SAC-OLDML-ALL over SAC-OLDML-1 can be clearly observed from the average F-score curves.

We also conduct a user study to compare different methods. There are 10 users involved in this study. We first compare the SAC-OLDML-ALL method with SAC-OLDML-1 and SAC in terms of clustering performance. The users are demonstrated the clustering results of different methods for each album, and they are then asked to give the comparison results using "$>$", "$\gg$" and "=", which indicate "better", "much better" and "comparable", respectively. To quantify the results, we convert the results into ratings. We assign score 1 to the worse scheme, and the other scheme is assigned a score 2, 3 and 1 if it is better, much better and comparable than this one, respectively. We perform an ANOVA test (King and Minium, 2003) to statistically analyze the comparison. The comparison results of SAC-OLDML-ALL versus SAC-OLDML-1 and SAC-OLDML-ALL versus SAC are demonstrated in Tables 2 and 3 respectively. The results demonstrate the superiority of our approach over the other methods. ANOVA test shows that the superiority is statistically significant and the difference of the evaluators is not significant. For the comparison of SAC-OLDML-ALL and SAC-GDML, the users are asked the experience the two methods and then compare them considering both the clustering results and response time. We quantize the comparison using the above method and perform ANOVA test as well. The results are demonstrated in Table 4. We can see that, users prefer the SAC-OLDML-ALL method, as it performs very closely with the SAC-GDML method while needing much less computational cost.

4.3. Discussion

It can be analyzed that the computational costs of the spectral clustering and online distance metric learning scale as $O(n^3 + n^2 d)$ and $O(Tld^2)$, respectively. Here n, d, l and T indicate the number of photos, the dimension of features, the number of constraints, and the iteration time for alternating optimization in Section 3, respectively. As shown in Fig. 7, we can see that, for most albums, the time cost can be less than 5 s. But for the album "Wildlife" that contains 1200 photos, the response time becomes much longer. Handling very large photo sets, such as those that contain tens of thousands of photos, will be difficult for our current scheme. For example, even for the spectral clustering algorithm, the computational cost will become unacceptable if the dataset is too large. Presenting and manipulating such large number of photos also become a problem. We leave the investigation approaches for handling extremely large albums to our future work. For computational cost, we can employ graph sparsification methods, such that the computational cost of spectral clustering can be significantly reduced. For presentation and manipulation, we can explore hierarchical structure.

5. Conclusion

This paper introduces a semi-automatic photo clustering approach. Different from automatic method that directly applies a clustering algorithm, our approach allows humans to manually adjust clustering results. An online distance metric learning algorithm is employed under the human's adjustments. Based on users' different operations, a set of equivalence and inequivalence constraints are generated for distance metric learning, and then the clustering can be re-implemented with the updated metric. This process can repeat until satisfactory performance is obtained. We have conducted experiments on different albums and encouraging results demonstrate the effectiveness of our approach.

In this work we mainly focus on the clustering algorithms and the distance metric learning approach that learns users' interaction, but user interface is also important in an interactive photo clustering system. We leave this aspect, i.e., how to facilitate users' adjustments of clustering results and better visualize them to our future work. We will also investigate methods for handling very large albums, including reducing computational cost and efficient presentation and manipulation with hierarchical structure (see Table 5).

References

Alipanahi, B., Biggs, M., Ghodsi, A., 2008. Distance metric learning versus fisher discriminant analysis. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI).

Bach, F., Jordan, M., 2003. Learning spectral clustering. In: Advances in Neural Information Processing Systems 16 (NIPS). MIT Press, Cambridge, MA, pp. 305–312.

Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D., 2005. Learning a Mahalanobis metric from equivalence constraints. Journal of Machine Learning Research (JMLR) 6.

Cai, D., He, X., Li, Z., Ma, W.Y., Wen, J.R., 2004. Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information, ACM International Conference on Multimedia.

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012

Chan, P.K., Schlag, M.D.F., Zien, J.Y., 1994. Spectral K-way ratio-cut partitioning and clustering. IEEE Transactions on CAD 13 (9), 1088–1096.
Cooper, M., Foote, J., Girgensohn, A., Wilcox, L., 2003. Temporal event clustering for digital photo collections. In: Proceedings of ACM Multimedia.
Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S., 2007. Information-theoretic metric learning, International Conference on Machine Learning.
Ding, C., 2004. A tutorial on spectral clustering. Talk presented at ICML.
Goldberger, J., Roweis, S.T., Hinton, G.E., Salakhutdinov, R., 2004. Neighbourhood components analysis. In Advances in Neural Information Processing Systems (NIPS).
Goldberger, J., Gordon, S., Greenspan, H., 2006. Unsupervised image-set clustering using information theoretic Framework. IEEE Transactions on Image Processing 15 (2), 449–458.
Hoi, S.C.H., Liu, W., Chang, S.-F., 2008. Semisupervised distance metric learning for collaborative image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Hull, J.J., 1994. A dataset for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Jain, A.K., Dubes, R.C., 1988. Algorithms for clustering data. Prentice-Hall, Inc..
Kennedy, L., Naaman, M., 2008. Generating diverse and representative image search results for landmarks. International World Wide Web Conference.
King, B.M., Minium, E.W., 2003, Statistical reasoning in psychlogy and education. John Wiley &.
Liu, D., Wang, M., Hua, X.-S., Zhang, H.-J., 2009. Smart batch tagging of photo albums. In: Proceedings of ACM Multimedia.
Mei, T., Wang, B., Hua, X.-S., Zhou, H.-Q., Li, S., 2006. Probabilistic multimodality fusion for event based home photo clustering. In: Proceedings of International Conference on Multimedia and Expo.
Moellic, P.A., Haugeard, J.E., Pitel, G., 2008. Image clustering based on a shared nearest neighbors approach for tagged collections, ACM International Conference on Image and Video Retrieval.
Nie, F., Tsang, I.W., Zhang, C., 2009. Spectral embedded clustering. International Joint Conference on Artificial Intelligence
Papadopoulos, S., Zigkolis, C., Kapiris, S., Kompatsiaris, Y., Vakali, A., 2010. ClustTour: City exploration by use of hybrid photo clustering. In: Proceedings of ACM Multimedia.
Platt, J.C., 2000. Proc. AutoAlbum: Clustering digital photographs using probabilistic model merging. In: Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 96–100.
Platt, J.C., Czerwinski, M., Field, B., 2003. PhotoTOC: automatic clustering for browsing personal photographs. Fourth IEEE Pacific Rim Conference on Multimedia.
Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Transactions on PAMI 22 (8), 888–905.
Simon, I., Snavely, N., Seitz, S.M., 2007. Scene summarization for online image collections. International Conference on Computer Vision.
Sinha, P., Pirsiavash, H., Jain, R., 2009. Personal photo album summarization. In: Proceedings of ACM Multimedia.
Suh, B., Bederson, B.B., 2004. Semi-automatic image annotation using event and torso identification. Technical report, HCIL-2004-15, Computer Science Department, University of Maryland, MD.
Tian, Y., Liu, W., Xiao, R., Wen, F., Tang, X., 2007. A face annotation framework with partial clustering and interactive labeling. IEEE International Conference on Computer Vision and Pattern Recognition.
von Luxburg, U., 2007. A tutorial on spectral clustering. In: Statistics and Computing, Vol. 17(4), pp. 395–416, December.
Wang, M., Hua, X.-S., 2011. Active learning in multimedia annotation and retrieval: a survey. ACM Transactions on Intelligent Systems and Technology 2 (2).
Wang, S., Jing, F., He, J., Du, Q., Zhang, L., 2007. IGroup: presenting web image search results in semantic clusters, ACM International Conference on Human Factors in Computing Systems.
Wang, M., Hua, X.-S., Tang, J., Hong, R., 2009a. Beyond distance measurement: constructing neighborhood similarity for video annotation. IEEE Transactions on Multimedia 11 (3).
Wang, M., Hua, X.-S., Hong, R., Tang, J., Qi, G.-J., Song, Y., 2009b. Unified video annotation via multi-graph learning. IEEE Transactions on Circuits and Systems video Video Technology 19 (5).
Weinberger, K., Blitzer, J., Saul, L., 2006. Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems (NIPS).
Xing, E.P., Ng, A.Y., Russell, S., 2003. Distance metric learning with application to clustering with side information. In: Proceedings of NIPS.
Yang, Y., Xu, D., Nie, F., Yan, S., Zhuang, Y., 2010. Image clustering using local discriminant models and global integration. IEEE Transactions on Image Processing.

Please cite this article in press as: Wang, M., et al. Intelligent photo clustering with user interaction and distance metric learning. Pattern Recognition Lett. (2011), doi:10.1016/j.patrec.2011.02.012