Consumer Photo Management and Browsing Facilitated by Near-Duplicate Detection with Feature Filtering

Wei-Ta Chu and Chia-Hung Lin
Department of Computer Science and Information Engineering
National Chung Cheng University, Taiwan
wtchu@cs.ccu.edu.tw, lchu96m@cs.ccu.edu.tw

Abstract

Near-duplicate detection techniques are exploited to facilitate representative photo selection and region-of-interest (ROI) determination, which are important functionalities for efficient photo management and browsing. To make near-duplicate detection module resist to noisy features, three filtering approaches, i.e., point-based, region-based, and probabilistic latent semantic (pLSA), are developed to categorize feature points. For the photos taken in travels, we construct a support vector machine classifier to model matching patterns between photos and determine whether photos are near-duplicate pairs. Relationships between photos are then described as a graph, and the most central photo that best represents a photo cluster is selected according to centrality values. Because matched feature points are often located in the interior or at the contour of important objects, the region that compactly covers the matched feature points is determined as the ROI. We compare the proposed approaches with conventional ones and demonstrate their effectiveness.

Keywords: near-duplicate detection, representative selection, region-of-interest, feature filtering, photo management and browsing

1. Introduction

Creation, display, and management of digital photos have been important activities in the digital life and in the cyberspace. People are accustomed to record their daily life or journeys by digital cameras, and share their living/travel experience on the web. Due to large amounts of multimedia content and many variations of sharing, dissemination, and browsing manners, users urgently demand systems that provide intelligent management and browsing.

We undertake management and browsing issues from two perspectives. First, users often select one representative photo for each of their web albums so that visitors can preview the content inside the album at a glance. This function has been popularly provided by photo sharing websites. Photo owners can not only share their experience efficiently, but also easily recall their life or travel experience by seeing the representative photos. Second, nowadays browsing devices are not limited to high-definition PC monitors but also PDA or cell phones. Crudely resizing the representative photo to meet the limits of different devices would cause large information loss and diminish the advantage of “fast preview” from representative photos. Intelligent thumbnailing technologies are needed to enhance browsing experience in resource-constrained applications or on mobile devices.

In this paper, we address these two issues by developing (1) automatic selection of representative photos and (2) smart thumbnailing based on region-of-interest (ROI). We focus on photos in journeys because the number of this kind of photo increases explosively, and most users suffer difficulties of efficient management and browsing. Moreover, these photos have clear and specific themes so that we can objectively determine the representative photo and find the most prominent region (ROI) in it. Assume that we visit several scenic spots in a journey. Photos taken in the same scenic spot can be clustered together by a time-based clustering method [1]. Then, the goal of selecting the representative photo is to automatically determine a photo that best presents this cluster. After selecting representative photo, we further find the “representative region” of this photo to generate an information-rich thumbnail. The desired region can be viewed as a kind of region-of-interest (ROI), although our approach is based on a viewpoint different from conventional content-based ones.

In this work, we advocate that both the selection of representative photos and ROI determination can be achieved by utilizing the concept of near-duplicate detection [2] (NDD). It’s reasonable to assume that the most prominent landmark/view would appear several times in a time-based photo cluster. After finding the near-duplicate photos, we model the relationship between photos in the same cluster as a graph, and analyze its structure to select one photo as the best representation of this scene spot. Moreover, we exploit spatial distribution of local feature points in the representative photo to find the most prominent region, which often consists of the most important building or the most canonical view. Therefore, the result of NDD not only facilitates the selection in the inter-photo domain but also in the intra-photo domain.

Although the prescribed ideas have been proven reasonable in our previous work [3], large amounts of local feature points with varied characteristics may substantially influence the performance of near-duplicate detection, and therefore diminish the effectiveness of the proposed methods. To human beings, the concept “duplicate” often comes from that two images have the same objects, such as building, tower, and other artificial objects. Although pieces of grass or surface of waterfront in two images may be similar as well, they pose little impact in near-duplicate detection and extended applications. Therefore, it’s more reasonable to eliminate the influence of noisy feature points in near-duplicate detection, which is not extensively studied in the literature. Figure 1 shows examples of a photo marked with all feature points and only with feature points on artificial objects, respectively. In this case, if only the feature points on artificial objects are considered in near-duplicate detection, more robust results can be obtained. In this work, three different feature filtering methods are investigated, and comprehensive experiments are conducted.

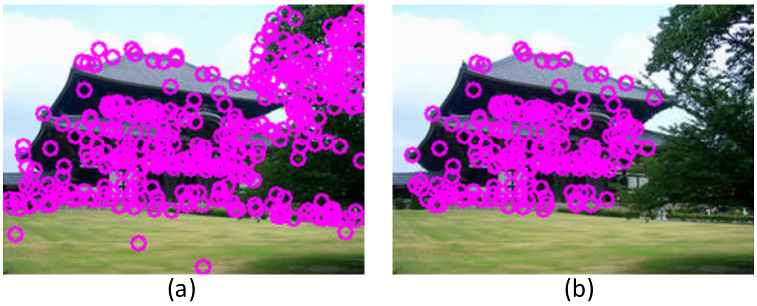


Figure 1. Examples of a photo marked with (a) all feature points and with (b) only feature points on artificial objects.

Contributions of this paper are summarized as follows:

- We advocate that near-duplicate detection techniques can be used to find representative photos and ROIs in travel photos. Any NDD technique using local feature points can be exploited in our framework. We demonstrate practicality of near-duplicate detection other than copy detection and multimedia retrieval.
- We point out that different feature points differently influence near-duplicate detection. Therefore, we model characteristics of local feature points and classify feature points to facilitate robust near-duplicate detection.
- We show that the proposed framework can be applied to other extensions, such as re-ranking of image search results, photo summarization, and image retrieval.

The rest of this paper is organized as follows. Section 2 reviews related studies. The whole system processes are described in Section 3. We first describe the system framework, and then propose three feature filtering methods to elaborate features fed to the near-duplicate detection module. With the results of near-duplicate detection, we model the relationships between photos as a graph, and automatically select the most representative one and perform region-of-interest determination. Section 4 provides extensive evaluation results. Extensions of the proposed framework and discussions are given in Section 5, and Section 6 concludes this work.

2. Related Works

2.1 Near-Duplication Detection

Many near-duplicate detection techniques have been proposed in recent years, and related studies have been applied in many fields, such as object recognition, copyright violation, video copy detection, image/video retrieval, and etc. Ke et al. [2] proposed one of the earliest image retrieval systems based on near-duplicate detection. They adopted an efficient variation of SIFT-based (Scale-Invariant Feature Transform) descriptor [5], i.e., PCA-SIFT [9], and proposed a hash-based structure to achieve efficient retrieval. Jing and Baluja [10] measured similarity between images based on SIFT-based matching, and applied the PageRank algorithm to achieve large-scale image search. From the idea of image search, Wang et al. [4] developed a system to automatically annotate images based on matching images with text descriptions. For video search, Wu et al. [11] proposed a hierarchical manner that first filters out unlikely video clips based on color information, and then performs expensive but accurate duplicate analysis to retrieve similar video data. With the constraints derived from the results of near-duplicate detection, Wu et al. [12] further developed a co-clustering algorithm to achieve news story clustering.

Near-duplicate detection plays the central role of this work. Although any variation of image near-duplicate detection technique can be applied, we adopt the method proposed in [6] due to its computation efficiency and satisfactory detection accuracy. Readers who are interested in the advances of near-duplicate detection can refer to the series of studied conducted by Ngo's group [13][14].

2.2 Local Feature Descriptor and Feature Classification

Local feature descriptors have been applied in many aspects. For example, Sivic and Zisserman [15] conducted object and scene retrieval based on visual words, which are constructed from clustering local feature descriptors. In related studies, images and videos are viewed as documents described by visual words, and techniques that are originally proposed for text retrieval can be modified for multimedia information retrieval [16]. Many researches then arise due to the success of visual words, such as video concept detection [17][18].

Although local feature descriptor and visual words are demonstrated to be effective in many studies, it draws relatively little attention about analyzing and classifying features such that different features provide different impacts to targeted applications. Dorko and Schmid [19] proposed a method for selecting the most discriminative features to allow robust part detection. This method was evaluated in car detection with varied viewing conditions. Monay et al. [20] used the fact that specific bags of visual words are correlated with the same semantic class. They modeled the context information based on the approach of probabilistic latent semantic analysis (pLSA) [23], and feature points are classified as in artificial regions or in natural regions.

In our work, we focus on intelligent management and browsing for photos taken in journeys, and would like to emphasize the impacts of artificial objects in image matching. In addition to develop a pLSA-based approach modified from [23], we further develop two approaches based on single feature points and region-based features [29]. We evaluate performance of these three feature classification approaches, for the applications of representative selection and region-of-interest determination.

3. Proposed Technique

3.1 Overview of Framework

Photos taken around the same place would include significant content variations. Some of them may include the most famous landmark or view, but some of them may include the shops around there, pedestrians, or something that is not directly related to this scenic spot. Figure 2 shows content variations in the photos taken in the famous Rokuonji temple in Kyoto. From this example and many other web-based albums, we found that most travelers incline to take the landmark or famous views several times. Moreover, tourists usually take photos at some specific locations such that they can capture the canonical view as that in postal cards. According to these observations, we propose that we can approach selection of representative photo based on near-duplicate detection, which finds near-duplicate pairs like the fifth to the eighth photos in Figure 2.

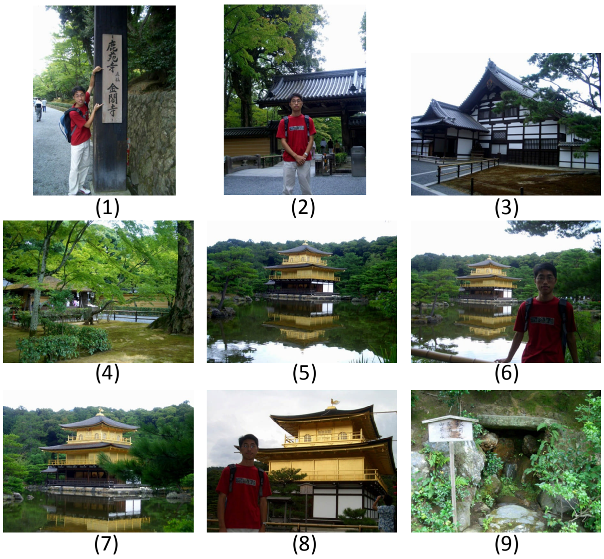


Figure 2. Photos taken around the same scenic spot.

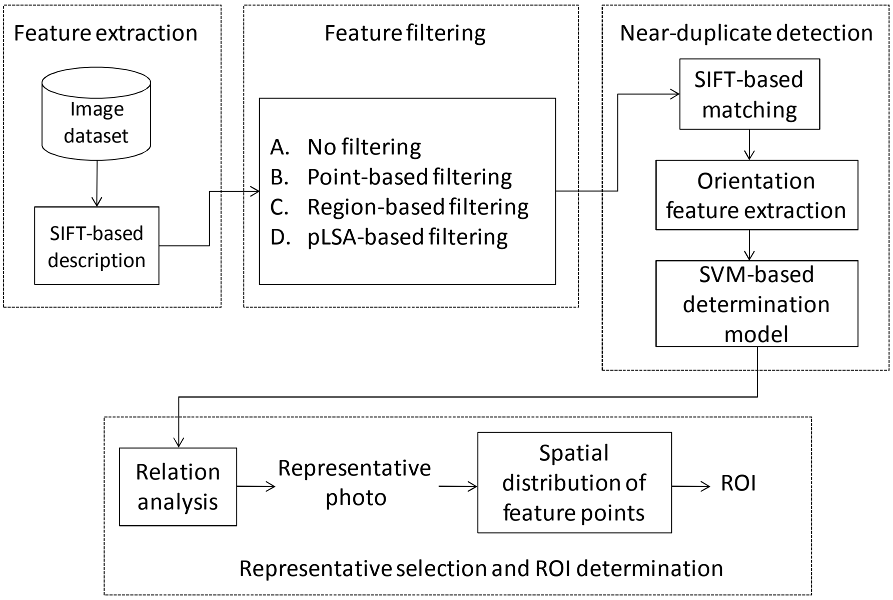


Figure 3. The proposed framework.

Figure 3 shows the four stages conducted in the proposed framework. First, we detect interest points based on DoG (difference of Gaussian) detector and describe them by SIFT descriptors. To filter out noisy feature points, we investigate the effects of three filtering approaches, including point-based filtering, region-based filtering, and pLSA-based (Probabilistic Latent Semantic Analysis) filtering. Among them, the pLSA-based method was proposed mainly for image classification or segmentation in the literature. To study the impacts of features' spatial relationships on feature classification, we develop point-based filtering and region-based filtering methods based on SVM-based classifiers.

At the near-duplicate detection stage, we basically follow the process proposed in [6], while any other NDD technique can be applied. Orientation of similar feature points between two photos is calculated and modeled by an SVM classifier. Therefore, whether two photos are near-duplicate is determined by checking the orientation characteristics of matched lines between them. This method largely reduces false alarms caused by conventional nearest-neighbor matching approaches and increases matching speed with a multidimensional index structure.

For a cluster of photos, we express duplicate relationships between photos as a graph. Link analysis is then performed to facilitate finding the most important node, which is the most representative photo in a cluster. Spatial distribution of matched feature points in the representative photo provides clues of determining the region-of-interest.

3.2 Feature Extraction and Filtering

Several kinds of features have been designed to characterize points or patches in images. For effectiveness, feature points should be distinct and robust to different viewing conditions. For efficiency, the number of features is preferred as small as possible, conforming to the constraint that it's enough to adequately describe the original data.

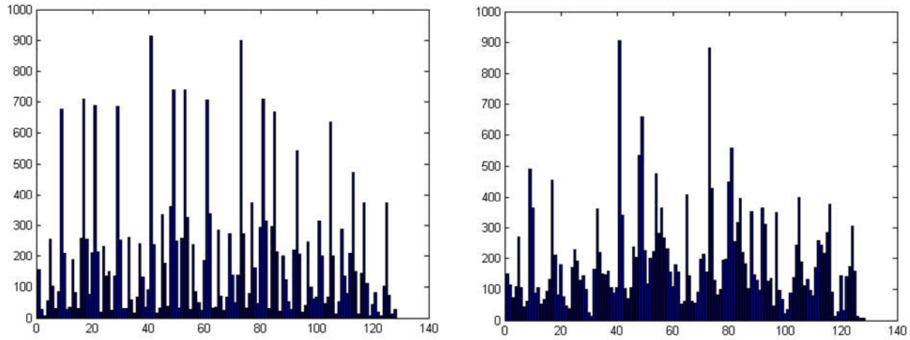
As regards the effectiveness issue, we apply a DoG detector [5] to find the location of feature points. For feature description, we utilize the SIFT descriptor [5] to describe each feature point as a 128-dimensional vector, which is robust to scale and orientation variation, and sort of illumination change. The work in [7] has demonstrated that the SIFT-based descriptor outperforms other local descriptors.

Relative few works discuss the efficiency of features to different applications. As illustrated in Figure 1, not all feature points pose positive influence in the sense of humans, although the detected points really present distinct properties at notable positions, such as corners of a building or tips of leaves. In this work, we consider photos taken in journeys, and put efforts on finding near-duplicate artificial objects, which are foundations of extended applications. The reason of putting focus on artificial objects is that people often recognize two photos as being near-duplicate if they consist of similar artificial objects.

After feature extraction, we would like to further classify feature points into that on artificial objects, such as buildings and towers, or that on natural scenes, such as tips of leaves or water surface [29]. SIFT-based feature points are further modeled and classified by the following process, and the ones being declared as on natural scenes are put aside from the applications described in Sections 3.3, 3.4, and 3.5.

3.2.1 Point-based Filtering

SIFT-based description is based on orientation information of small patches in different resolutions, centered by the feature point. Therefore, the 128-dim feature vector implicitly embeds local structure. Figure 4 shows SIFT-based description of feature points on artificial objects and natural scenes, which are respectively statistics of 1000 points from different objects. Each bin in the horizontal axis means an orientation at some resolution, and the value in the vertical axis means the number of feature points with such orientation. We can see that feature points on artificial objects generally have larger values in some specific orientations. This observation matches our intuition, because artificial objects often have strict geometric structure and common elements, while natural scenes have relatively random structure.



(a) (b)

Figure 4. SIFT-based description of feature points on (a) artificial objects and (b) natural scenes.

To model the characteristics of feature points, we conceptually need to construct a mapping function $f : \mathcal{R}^{|D|} \rightarrow \{1, 0\}$, which maps a SIFT descriptor $\boldsymbol{d}_i$, $\boldsymbol{d}_i \in \mathcal{R}^{|D|}$, to a binary value. The values 1 and 0 denote that a feature point is an artificial point or a natural point, respectively. The typical dimension of a SIFT descriptor, i.e., $|D|$, is 128 [5]. In this work, we respectively collect two types of feature points, and construct the mapping function by a binary SVM classifier [21]. At the filtering stage, each feature point is evaluated by the classifier, and is then categorized into an artificial point or a natural point.

3.2.2 Region-based Filtering

Although Figure 4 shows distinct characteristics on single SIFT descriptors, spatial correlation between feature points in neighborhood is not considered. Conceptually, examining a single feature point may unavoidably suffer an issue similar to the "aperture problem" in object tracking. The point-based approach just looks into a small patch of pixels (a feature point), and similar feature points may not necessarily present the same type of objects. For example, a corner of a building may be similar to a corner of a rock.

In order to consider the characteristics of SIFT descriptors in a locality, we instead construct a mapping function $f : \mathcal{R}^{|D|} \rightarrow \{1, 0\}$ that maps a SIFT descriptor $\bar{\boldsymbol{d}}_i$ to a binary value. We divide each image into regions, with each size 40×40, and represent each region by a vector $\bar{\boldsymbol{d}}_i$ that is the average of the descriptors in the same region. That is,

$$\bar{\boldsymbol{d}}_i(j) = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{d}_k(j), \qquad (1)$$

where $\bar{\boldsymbol{d}}_i(j)$ and $\boldsymbol{d}_k(j)$ denote the value of their *j*th bin, and N is the total number of feature points in the *i*th region.

Similarly, we respectively collect two types of feature points, and construct the mapping function by a binary SVM classifier. The only difference between the region-based approach and the point-based one is that the features put to training and testing are average values of feature points in the same region. At the filtering stage, each region is evaluated by the classifier, and is then categorized into an artificial region or a natural region.

3.2.3 pLSA-based Filtering

Another approach to consider context information between feature points was proposed in [23]. We modify their method to fit our needs as follows. Feature points specifically from artificial objects and natural objects are collected, respectively. For the set of artificial feature points, we apply the k-means algorithm to cluster them into a specific number of clusters. The set of clusters is called the *visual vocabulary for artificial objects*, denoted by $\mathcal{V}^a$. Centroid of each cluster is calculated by averaging SIFT descriptors in this cluster, and is called as a *visual word* that represents a cluster of features. By the same method, we construct the visual vocabulary for natural objects, denoted by $\mathcal{V}^n$.

Given a feature point s, we determine its corresponding visual word in $\mathcal{V}^a$ and $\mathcal{V}^n$ by quantizing it into one of the pre-trained visual vocabularies. That is,

$$s \rightarrow Q(s) = v_i^a \leftrightarrow i = \arg\min_{j=1,\dots,|\mathcal{V}^a|} dist(s, v_j^a), \tag{2}$$

$$s \rightarrow Q(s) = v_i^n \leftrightarrow i = \arg\min_{j=1,\dots,|\mathcal{V}^n|} dist(s, v_j^n), \tag{3}$$

where $Q(\cdot)$ denotes the quantization function, $dist(s, v_j)$ denotes the Euclidean distance between the feature point s and the visual word v_j, and $|\mathcal{V}^a|$ ($|\mathcal{V}^n|$) denotes the size of the visual vocabulary for artificial (natural) objects.

The probability of a visual word v_i^a corresponding to an artificial object is estimated based on co-occurrence information between visual words in a collection of artificial objects. We exploit probabilistic latent semantic analysis (pLSA) models to describe the joint probability over the image d_j and the visual word v_i^a:

$$P(v_i^a, d_j) = P(d_j) \sum_{\ell=1}^{N_A} P(z_\ell|d_j) P(v_i^a|z_\ell), \tag{4}$$

where $z_\ell \in \mathcal{Z} = \{z_1, \dots, z_{N_A}\}$ is a latent concept subtly embedded in the visual vocabulary $\mathcal{V}^a$. The pLSA model is defined by the conditional probabilities $P(v_i^a|z_\ell)$ that represent the probability of observing the visual word v_i^a given the concept z_ℓ, and the condition probability $P(z_\ell|d_j)$ of the occurrence of z_ℓ in the image d_j. The parameters of the model are estimated using the Expectation-Maximization (EM) algorithm [24], using a set of training data that includes feature points in artificial objects. Construction of the pLSA model for natural objects is in the same manner.

Given a feature point s in the image d_j, which corresponds to the visual word v_i^a with respect to artificial objects, we try to map the visual word to the most likely concept $z_{v_i^a}$. Based on the pLSA model, the mapping can be computed by

$$z_{v_i^a} = \arg\max_z P(z|v_i^a, d_j)$$
$$= \arg\max_z \frac{P(v_i^a|z)P(z|d_j)}{\sum_z P(v_i^a|z)P(z|d_j)}. \quad (5)$$

The same manner is applied to calculate the probability of the most likely concept $z_{v_i^n}$, based on the pLSA model for natural objects. Finally, the probability of the feature point s corresponding to the artificial concept $z_{v_i^a}$ is $P(z_{v_i^a}|v_i^a, d_j)$, and the probability of s corresponding to the natural concept $z_{v_i^n}$ is $P(z_{v_i^n}|v_i^a, d_j)$. The feature point s in the image d_j is claimed to be an artificial feature point if

$$\frac{P(z_{v_i^a}|v_i^a, d_j)}{P(z_{v_i^n}|v_i^n, d_j)} > \sigma, \quad (6)$$

where σ is a threshold that can adjusted to give different preference in feature classification. If the ratio is less than the threshold σ, the feature point s is claimed to be a natural point. In this work, we simply set the threshold σ as 1 so that no special preference is applied.

- Incorporating prior probability in pLSA

The eqn. (6) solely evaluates the probabilities of a visual word corresponding to an artificial concept $z_{v_i^a}$ and a natural concept $z_{v_i^n}$. We can further take prior probabilities of visual words v_i^a and v_i^n in the image d_j into account, and therefore classify feature points according to characteristics of different images. The feature point s in the image d_j is claimed to be an artificial feature point if

$$\frac{P(v_i^a)P(z_{v_i^a}|v_i^a, d_j)}{P(v_i^n)P(z_{v_i^n}|v_i^n, d_j)} > \sigma, \quad (7)$$

where the threshold σ is the same as that in eqn. (6).

Our filtering method is different from conventional pLSA approaches as in [23]. In [23], a pLSA model is constructed to jointly consider the conditional probability of a latent concept occurring in a specific image, and the condition probability of a visual word occurs when a latent concept presents. Various types of images containing different scenes are used to train the pLSA model. The discovered latent concepts may mix artificial objects and natural objects. Because there are enormous elements for artificial and natural objects, describing all these elements with a model seems to be impractical. In our work, we specifically construct a model for artificial objects and natural model, respectively. Therefore, the discovered latent concepts in artificial pLSA model, for example, may include windows, doors, tips of towers, stairs, and etc. A feature point is quantized into visual words in terms of artificial objects and natural objects Based on the likelihood ratio between the most probable concepts corresponding to artificial and natural pLSA models, we determine the category of this feature point. We believe that the pLSA models trained from separated data can more precisely describe the variations of artificial and natural objects.

3.3 Near-Duplicate Detection

3.3.1 Near-Duplicate Process

Given a set of photos $\mathrm{P} = \{p_1, p_2, ..., p_N\}$ that are clustered together by using the time-based clustering method [1], we first filter out feature points that are claimed as natural points. Then, whether a pair of photos (p_i, p_j), $i \neq j$, $i, j \leq N$, is near-duplicate is determined by the following steps, as shown in Figure 5.

- SIFT-based matching: For any pair of photos in this cluster, the method in [6] that embeds a one-to-one symmetric criterion to filter out false matches is applied. The one-to-one symmetry means that a pair of matched points (s_i, s_j) should be the nearest neighbor to each other. The white lines in Figure 6 are false matches. We can see that the one with one-to-one symmetric criterion effectively reduces false alarms (Figure 6(b)), as compared to the conventional approach (Figure 6(a)).
- Orientation feature extraction: Due to the characteristics of local coherence and spatial smoothness, the orientation of the link connecting matched points in two near-duplicate photos is coherent. We calculate the orientation of links and quantize it into 36 ranges, with each range representing ten degrees. A 36-bin orientation histogram is then constructed. In near-duplicate pairs, the values of the orientation histogram would apparently concentrate.
- SVM-based determination model: A binary SVM classifier is used to model the characteristics of orientation histograms. We estimate model parameters based on 40 near-duplicate pairs and 40 non-near-duplicate pairs. At the test stage, the orientation histogram of the matching situation between a pair of photos is put to the SVM classifier, and we determine whether this pair of photos is near-duplicate or not.

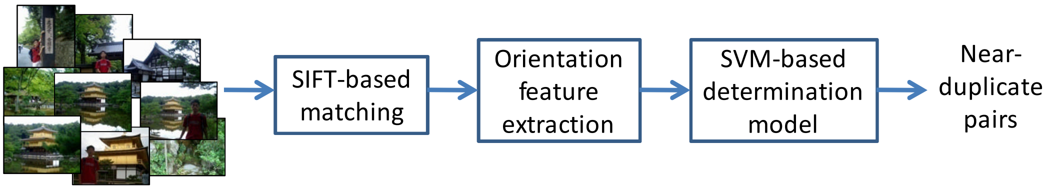


Figure 5. The process of near-duplicate detection.

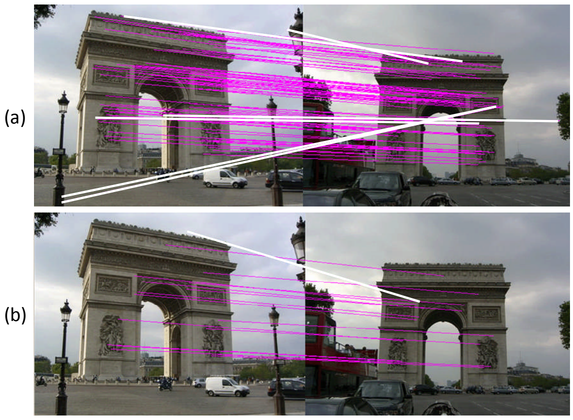


Figure 6. Sample results of (a) conventional SIFT-based matching and (b) one-to-one symmetric SIFT-based matching.

3.3.2 Sub-Clustering Before Matching

One of the critical issues in NDD is that there are tremendous pairs of photos to be examined. For example, if there are *N* photos in a set, totally $\binom{N}{2}$ different pairs of photo are needed to be checked. To reduce computation complexity, we first cluster the given set of photos based on content-based characteristics, and then perform NDD for each sub-cluster, i.e., any two photos that are in different sub-clusters would not be examined.

Because the representative landmark or view would have similar appearance, we assume that they would be categorized in the same sub-cluster. For example, if the set of *N* photos are categorized into *M* sub-clusters $\{C_1, C_2, ..., C_M\}$, the total number of pairs for NDD is

$$\sum_{i=1}^{M} \binom{|C_i|}{2}, \tag{8}$$

where $|C_i|$ is the number of photos in the *i*th sub-cluster. In the case of *N* = 10, *M* = 2, $|C_1| = 4$, and $|C_2| = 6$, we originally need to check $\binom{10}{2} = 45$ photo pairs. However, we only have to evaluate $\binom{4}{2} + \binom{6}{2} = 21$ photo pairs if we perform sub-clustering first. In this work, the sub-clustering process is implemented by the k-means algorithm, based on RGB histograms of photos.

3.4 Selection of Representative Photos

Without loss of generality, assume that the sub-cluster C^* in the set $\{C_1, C_2, ..., C_M\}$ contains near-duplicate photos. Now the problem is to select one of the photos in C^* to be the representative photo.

We represent the relationship between near-duplicate photos as a non-directed, non-weighted graph $G = \langle V, E \rangle$, where any node (photo) v_i in $V = \{v_1, v_2, ..., v_n\}$ is at least once determined as a near-duplicate to someone else. The edge e_{ij} is in E if v_i and v_j are detected as a near-duplicate pair. Figure 7 shows an illustrative example of this graphical representation.

Given this graph, we can determine the most important node by checking the “centrality value” of each node. From the idea of social network analysis, the person who is “closest” to all others plays the most important role. Similarly, we can say that the photo mostly near-duplicate to others is the most representative one. Therefore, the photo v_r is selected as the representative if

$$r = \arg\max_i \mathrm{centrality}(v_i). \tag{9}$$

There are various measurements to evaluate the centrality value of each node, including degree centrality, betweenness centrality, and closeness centrality [25]. The degree centrality of a node v_i is calculated by

$$\mathrm{degree\ centrality}(v_i) = \frac{\sum_{k=1}^{n} a(v_i, v_k)}{n-1}, \tag{10}$$

where $a(v_i, v_k) = 1$ if v_i and v_k are connected, and otherwise $a(v_i, v_k) = 0$. The node that has the most connected edges is the most central one in this graph.

In this work, because relationships between photos are often not complicated, we evaluate the centrality value of each node by the degree centrality. Therefore, in Figure 7, the second photo is selected as the representative photo. Details of other centrality values please refer to [25].

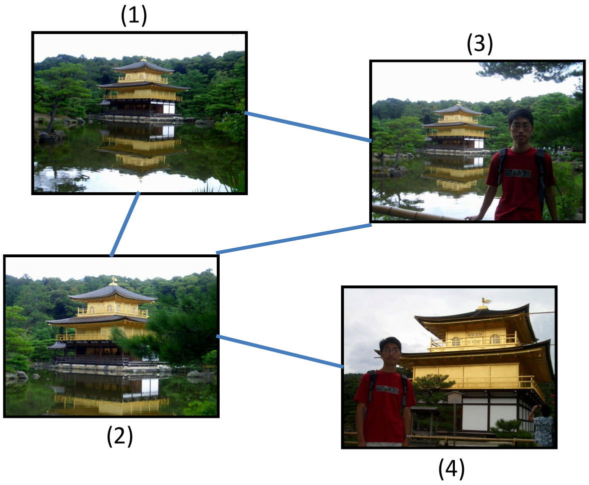


Figure 7. Relationship between near-duplicate photos.

3.5 ROI Determination

In order to ease users in browsing large amounts of albums at a glance, many photo sharing platforms facilitate users to manually select a representative photo and resize it to be the epitome of each album. We address the selection issue before. However, resized representative photos are often suffered from severe information loss, and we may only see rough appearance of the most important object. This situation becomes more critical as the rapid emergence of low-definition mobile devices.

In this section, we further determine the "representative region" in the selected representative photo. This task is similar to finding the region-of-interest in an image. After finding the ROI, we can just extract the region and generate a better thumbnail for the representative photo.

Currently, works on ROI determination are mostly based on the bottom-up approach proposed by Itti and Koch [26]. According to the human vision system, the idea is to compute contrast of color, intensity, and orientation, and then combines these factors to construct a saliency map that describes how a photo attracts humans. We develop the determination module from a different perspective. In photos taken in journeys, ROIs in representative photos are landmarks or specific views. Therefore, we advocate that it's more reasonable to find ROIs based on local feature points that contribute to near-duplicate detection, rather than color or intensity contrast.

On the basis of this idea, we take advantage of the byproducts produced in the process of NDD. As shown in Figure 8, we found that the matched points lie on or around the most important object in photos. These points provide the foundation of linking near-duplicate objects, and near-duplicate objects are often landmarks or specific views that should be in ROIs.

Let's consider the most representative photo p_R and its nearest duplicate p_Q. Let $\{\ell_1, \ell_2, ..., \ell_N\}$ be the set of lines connecting pairs of SIFT matched points in p_R and p_Q, respectively. The orientation of these lines $o(\ell_i)$, $1 \leq i \leq N$, are gathered to construct a 36-bin orientation histogram H. To determine the ROI in the most representative photo, we first find the SIFT points that confidently contributes to NDD. Based on the orientation histogram, the bin with the largest histogram value is:

$$j^* = \arg\max_j H(j), \quad j = 1, 2, ..., 36. \tag{11}$$

We select the lines which orientations fall into the j^*-th bin or its two adjacent bins:

$$\{\ell_i | (Q(i) = j^*) \vee (Q(i) = j^* - 1) \vee (Q(i) = j^* + 1)\}, \tag{12}$$

where $Q(i)$ denotes the bin where the orientation of the line ℓ_i is quantized into.

Let $\{(x_1, y_1), ..., (x_M, y_M)\}$, $M \leq N$, be the coordinates of the SIFT points that are in the representative photo and meet the eqn. (12). The left, right, top, and bottom boundaries (x_L, x_R, y_T, y_B) of the desired ROI are determined by

$$x_L = \min_k x_k, \qquad x_R = \max_k x_k,$$

$$y_T = \min_k y_k, \qquad y_B = \max_k y_k,$$

where $k = 1, 2, ..., M$.

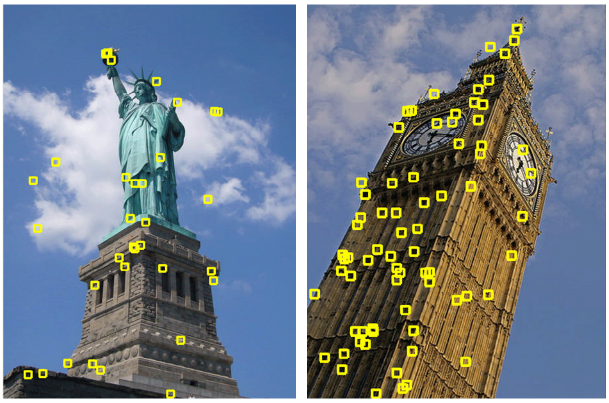


Figure 8. The matched SIFT points in representative photos.

4. Performance Evaluation

4.1 Model Training

For performance evaluation, we collect photos taken by amateurs, and taken in different places from Flickr and our lab's members. There are totally 1024 photos, which contain 52 different famous scenic spots around the world, such as Arc de Triomphe, Statue of Liberty, and Time Square. The resolution of each photo is normalized into 320×240 or 240×320.

To conduct feature filtering, we need to construct three determination models that correspond to the methods described from Sections 3.2.1 to 3.2.3. Training data for constructing these models are described as follows.

1) Training for the point-based filtering: We collected photos that solely include artificial feature points and natural feature points, respectively. Figure 9(a) shows some samples of the training data. There are totally 3483 artificial feature points and 6170 natural feature points for training. By labeling artificial points as positive samples and natural points as negative samples, we construct an SVM classifier to determine whether a feature point is artificial or natural. In this paper, all SVM classifiers are constructed by the package in [21].
2) Training for the region-based filtering: Photos that solely include artificial regions and natural regions are collected, respectively. Figure 9(b) shows some samples of the training data. Each photo is divided into 40×40 regions, and the feature vector for each region is extracted. There are totally 846 artificial regions and 921 natural regions. Similarly, we label artificial region as positive samples and natural regions are negative samples, and construct an SVM classifier to determine whether a region is artificial or natural. The feature points in an artificial region are claimed as artificial points, and vice versa.
3) Training for the pLSA-based filtering: Two hundred photos that solely contain

artificial feature points are used to construct the pLSA model for artificial objects. There are totally 53655 artificial feature points, which are clustered into 600 groups according to the k-means algorithm. The centroid of each group forms a visual word. Similarly, two hundred photos containing 63769 natural feature points are used to construct a visual vocabulary consisting of 600 visual words, and then the pLSA model for natural objects is constructed. In this work, we use the program in [22] to implement the proposed approach.

For near-duplicate detection, we model the matching patterns between photos by an SVM classifier. We collect sixty pairs of near-duplicate photos, and extract the orientation histograms of matching patterns to be positive training samples. From these 120 photos, we randomly select 120 pairs of photos that are not near-duplicate, and extract the orientation histograms of matching patterns to be negative training samples. An SVM classifier is then constructed to evaluate whether two photos are near-duplicate or not.

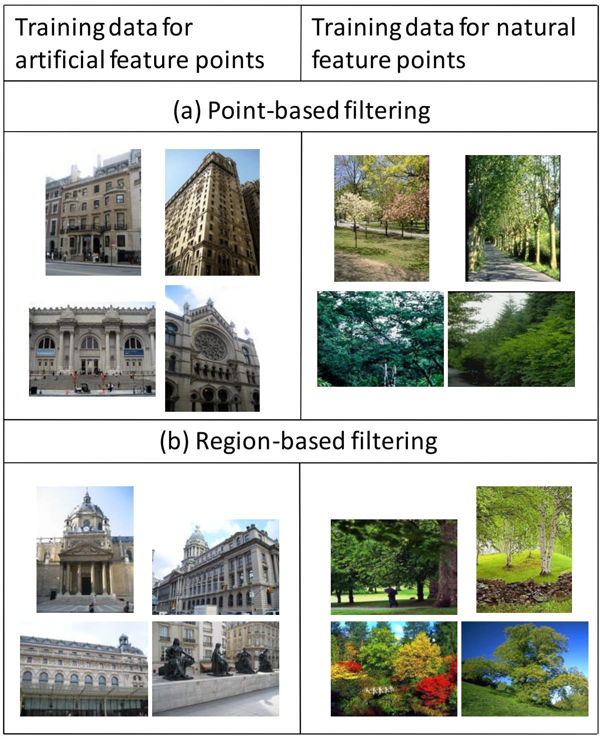

Training data for artificial feature points
Training data for natural feature points
(a) Point-based filtering
(b) Region-based filtering

Figure 9. Sample training data for (a) point-based filtering and (b) region-based filtering

4.2 Performance of Feature Classification/Filtering

To compare the performance of different feature classification methods, the most straightforward measurement is the precision rate. Based on manually labeled ground truths that include 3182 artificial feature points and 5173 natural feature points, we calculate precision rate for each method as follows:

$$\mathrm{Precision} = \frac{C_a + C_n}{N}, \qquad (13)$$

where C_a is the number of artificial feature points that are correctly classified by a classification method, and C_n is the number of natural feature points that are correctly classified by a classification method. The denominator N is always 3182+5173=8355 in this evaluation.

Table 1 shows the precision rates achieved by different methods. We can easily see that the point-based and region-based methods work much better than the pLSA-based approach. The pLSA method with prior probability consideration works slightly better than that without prior probability, but its classification performance is still far behind the methods based on SVMs.

Figure 10 gives some examples of the classification results. From the third to sixth columns, only the points that are classified as artificial feature points are marked. We can obviously see that the region-based method works better than others. In the results of the pLSA-based methods with or without prior probability, many feature points on trees are remained after filtering, which cause large amount of noise to near-duplicate detection.

Table 1. Comparison of classification methods in terms of precision and information gain.

	Precision
Point-based filtering	0.81
Region-based filtering	0.92
pLSA-based filtering without prior prob.	0.26
pLSA-based filtering with prior prob.	0.38

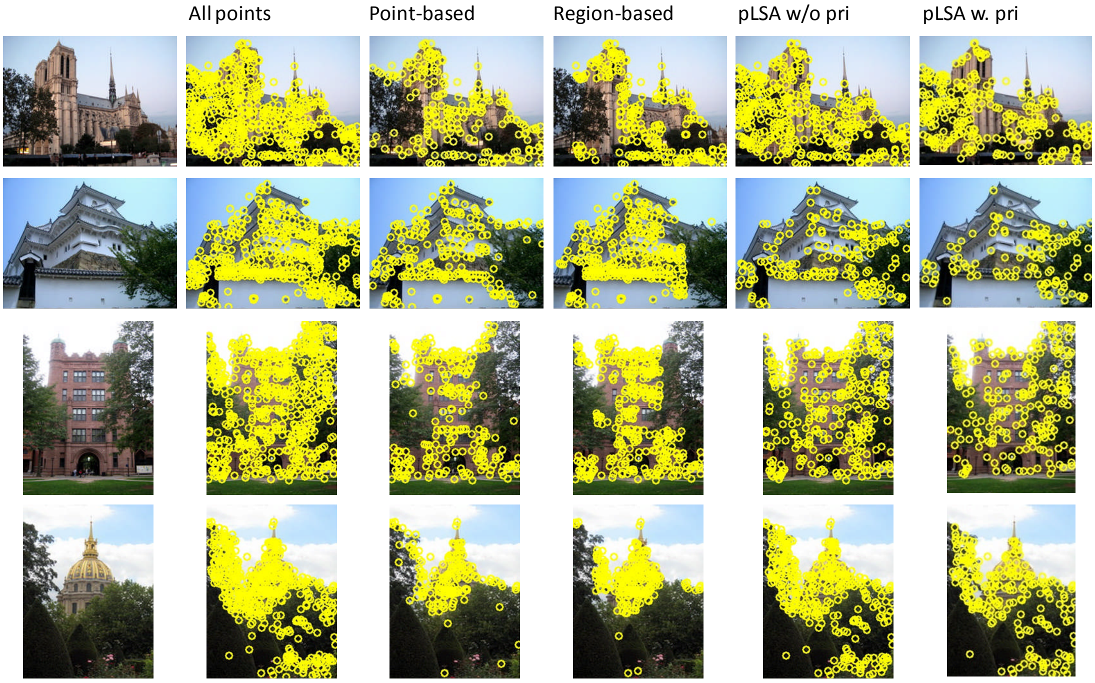


Figure 10. Comparison of different feature filtering approaches.

The pLSA-like methods have been widely applied in scene understanding and visual information retrieval [20][23]. However, this approach has weak performance of feature classification when limited numbers of training data are available. The work in [23] discovered latent concepts based on pLSA models, and used latent space representation to perform scene modeling. The global latent space represents the distribution of various aspects and is used to distinguish indoor scenes from outdoor scenes. The meaning of each latent concept, and which concept a feature point belonging to, are not concerns of this work. Similar to our work, Monay et al. [20] relied on likelihood ratio computation to classify each feature point into man-made or natural class. Different thresholds for likelihood ratio test are set to achieve different detection performance. Although promising performance has been reported in both [20] and [23], large amounts of training data were applied, and therefore enormous computation was needed in model training. For example, 6000 photos which may include more than 1 million feature points are used to construct a visual vocabulary consisting of 1000 visual words. A pLSA model that includes 60 concepts is then constructed based on 300 photos. On the other hand, our proposed region-based classification built by a discriminative model needs significantly smaller number of training data, and has superior performance in feature classification. Moreover, no specific threshold is needed in the point-based and region-based methods.

4.3 Performance of Near-Duplicate Detection

We perform feature classification and put only artificial feature points to the near-duplicate detection module in [6]. We don't propose a novel near-duplicate module, but elaborate the features fed to it. Any near-duplicate detection module based on local feature points can be used in our work. To evaluate the influence of feature filtering on near-duplicate detection, we compare the performance of NDD with feature filtering and without feature filtering. Based on the 1024 photos in 52 scenic spots, we manually define ground truths of near-duplicate photos and measure performance by precision and recall rates.

Overall, the precision and recall rates of NDD without filtering are 0.33 and 0.41, and that of NDD with filtering are 0.57 and 0.20, respectively. We clearly see that the method with feature filtering largely increases precision but decreases recall. Higher precision but lower recall means that we provide fewer but more accurate near-duplicate photo set for representative selection. With feature filtering, many matched pairs between natural objects are filtered out, and the number of detected near-duplicate photo pairs decreases. From the perspective of representation selection, simpler graph is constructed, and time for analysis decreases. From the perspective of photo management, fewer but more accurate presentation also benefits users in efficient browsing.

4.4 Performance of Representative Selection

To evaluate the performance of representative selection, which is involved with subjective judgment, we asked seven observers to give a score to each photo that is determined to be near-duplicate to others. The score ranges from one to five. A larger score is given if the observer thinks a photo better represents a scenic spot. In order to reduce the observers' indecisiveness in the judgment process, we provide them guidelines as in Table 2, although the observers were not forced to follow the guidelines strictly.

Table 2. Guidelines of giving a score to each photo.

Score	Description
5	The image shows the most representative object you know for this scenic spot.
4	Although the most representative object shows on the image, it's not good in shooting angles or in lighting conditions.
3	Although the image doesn't show the most representative object, some other buildings or specific objects are shown, e.g., a statue.
2	There are objects without specific topic in this image, e.g., a sign, or the quality of the image is bad.
1	I totally don't know the purpose of this image, e.g., crowd, grass, flower.

For each photo, the degree of representative is calculated by averaging the scores from seven observers. The selection performance of a selected representative photo is measured by the corresponding score. For example, if the second photo in Figure 7 is selected as the representative, and the average score given to this photo is 3.8, the score of this selection result is 3.8. Therefore, the automatic selection method obtains higher score when the selected photo better matches human's judgments. Table 3 lists the selection performance of photo sets (1) without feature filtering; (2) with the point-based feature filtering; (3) with the region-based feature filtering; and (4) with the pLSA-based feature filtering without consideration of prior probability. There are totally fifty-two photo sets, and contents in them mainly include building, statute, and cityscape.

Overall, the selection performance is over 3.3, i.e., at least photos containing specific buildings or objects are selected as representatives. Performance of representative selection is data-dependent, but performance of the ones with point-based or region-based filtering is generally better than that without feature filtering. This confirms the idea we introduced in Section 1. However, the one with pLSA-based feature filtering contrarily degrades the performance. The reason is that the pLSA-based approach doesn't have good feature classification performance, and therefore leaves many noises to harm the representative selection module. The worse performance for the pLSA-based filtering matches the trend shown in Table 1.

Although Table 3 seems to show that the methods with feature filtering have limited improvement over that without filtering, the effect of compact representation is not reflected by this table. When feature filtering is applied to near-duplicate detection, matching false alarms are largely filtered out, and the number of near-duplicate photo pairs decreases. For observers to judge performance or from the perspective of photo browsing, displaying fewer but important photos on screen greatly improves browsing efficiency. In the case without feature filtering, 819 out of 1024 photos are claimed to be at least once near-duplicate to other photos. That is, almost 80% photos are detected as near-duplicate to someone else, which shows extremely large ratio of false alarms. When we apply the region-based filtering, only 382 out of 1024 photos are claimed to be at least once near-duplicate to others.

Table 3. Performance of representative selection.

Scenic spot	(1)	(2)	(3)	(4)	Scenic spot	(1)	(2)	(3)	(4)
Arco di Tito	4.89	5	4.78	5	NotreDame1	2.33	2.33	2.22	2.44
Athens	3.44	3	3.67	2.89	NYBrooklyn	3.33	2.78	4.89	2
Back Bay	4.11	2.89	2.11	4.33	Paris-1st	3	4	3.78	3.44
Baltimore	4.33	3.33	3.78	4.44	Paris-6th	4.22	3.22	4	4.44
Basllique	3.89	4.78	3.89	3.78	Paris-7th	3.22	3	4.78	3.66
Bloomsbury	2.44	2.78	2.33	2.56	Paris-8th	4.78	4.56	4.67	2.22
Boston	3	2.67	3	3.22	Paris-9th	4.44	3.11	4.33	1.44
Columbia Univ.	2.56	2.56	3	3.22	ParisTower	4	3.44	3.67	4.78
Connecticut	2.11	3.89	4.33	4	Philadelphia	3.11	3.89	3.56	2.56
Eiffel Tower	4	4	3.11	3.89	Piccadilly	4	4.11	3.11	3.33
FartherAfield	1.44	2.67	2.44	1.89	Rokuonji	4.33	3	4.11	1.67
FreedomTrail	4.44	3.67	2.44	4.44	StatueofLiberty	4.33	4.67	4.56	2.22
GreenwichVillage	1.56	3.78	2	4.44	StJames's	2.78	2.44	3.78	4.44
Himejijo	4.44	3.44	4	4.33	Stonehenge	4.44	4.56	4.56	4.67
Hoboken	1.78	4	1.56	3	TheCity	4.78	3.89	4.67	3.11
JacksonSquare	4.56	4.67	4.44	2.22	TimeSquare	4	4.11	4	3.78
Kensington	4.22	2.56	2.89	2.22	Todaiji	4.11	4.11	4	4
Liberty	3.89	4.78	5	4.33	TowerHamlets	3.78	3.67	3.89	4.56
LouisCathedral	4	3.66	4.11	3.67	Trail	4.33	4.44	3.56	2.89
FrenchQuarter	3.33	3.44	3	2.67	UnionCounty	2.22	3.78	2.11	1.56
Millan	3.67	3.33	3.22	4.22	UWS	2.56	2.56	4	3.67
Millan1	3.78	4.78	3.11	4.67	ValleyForge	5	4.33	4.11	1.56
Morning	2.67	3.89	4	1.56	WestEnd	3.44	3.89	4.22	2.67
Mykonos	3.22	2.11	3	2.78	Westminster	3.78	3.89	4	3
Northside	3.44	3.67	3.67	3.56	Windsor	3.78	3.56	3.44	3.22
NortrDame	4.78	4.78	3.22	4.78	WrigleyField	2.22	2.33	4.44	3.22
Overall	3.58	3.61	3.63	3.32					

4.5 Performance of ROI Determination

We compare the proposed method with a saliency-based approach. We use the SaliencyToolbox [8] to calculate the saliency value of each pixel and generate a saliency map for an image. Saliency values are derived from the impacts of color contrast, intensity contrast, and orientation contrast [26]. We then find the centroid of the saliency map, which coordinate is denoted as (x_C, y_C) in the following. Using this centroid as the center of the desired ROI, boundaries of the saliency-based ROI is determined as

$$x'_L = x_C - \lfloor \tfrac{x_R - x_L}{2} \rfloor, \quad x'_R = x_C + \lfloor \tfrac{x_R - x_L}{2} \rfloor,$$
$$y'_T = y_C - \lfloor \tfrac{y_B - y_T}{2} \rfloor, \quad y'_B = y_C + \lfloor \tfrac{y_B - y_T}{2} \rfloor,$$

where (x_L, x_R, y_T, y_B) are determined by the process described in Section 3.5, which guarantees the size of a saliency-based ROI is the same as that of our proposed ROI. If the calculated boundaries exceed the boundaries of the image, we do appropriate shifts in horizontal or vertical directions to guarantee correct size of ROI.

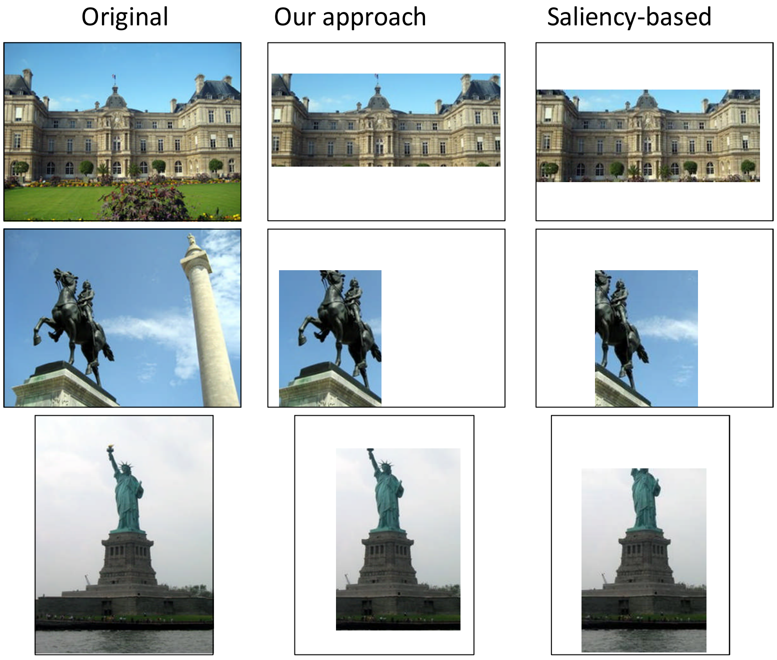


Figure 11. Performance comparison of ROI determination.

Figure 11 shows performance comparison of ROI determination based on our approach and the saliency-based method. We deliberately align the cropped ROI with the original image to facilitate intuitive interpretation. The second column of Figure 11 shows ROI determination results based on near-duplicate detection, with the consideration of region-based feature filtering. We see that our approach is notably better than the saliency-based method. The reason is that the saliency-based approach only considers contrast of color, intensity, and orientation. This information doesn't necessarily reflect important objects or embedded semantics in images. On the contrary, local feature points are directly related to the region of interest, i.e., the near-duplicate object. Thus we more elaborately find the contour of important objects and determine more accurate region-of-interest.

One of the purposes to find an ROI in an image is to efficiently display information in a resolution-limited environment, such as mobile phones and PDAs. Displaying only an ROI rather than the whole image enhances browsing experience because the effects caused by large-scale resizing are reduced. It also improves browsing efficiency because users have fewer necessities to zoom in an important object on the image. Figure 12 shows some examples of displaying ROIs on a smartphone. Note that the ROIs conform to the constraints that the aspect ratio of ROIs should match smartphone's screen. We can easily see that the important object can be displayed more clearly, comparing with the cases of resizing the whole image. These results demonstrate the impacts of ROI determination for efficient browsing.

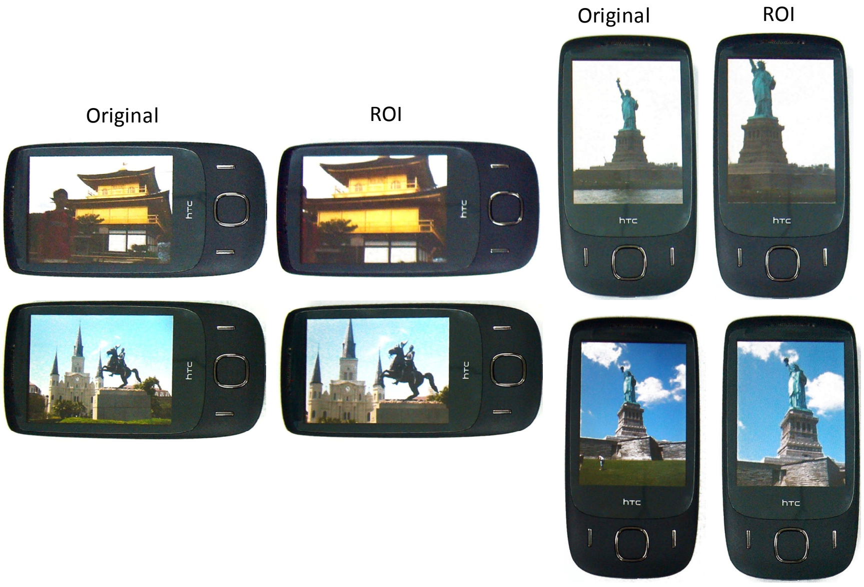


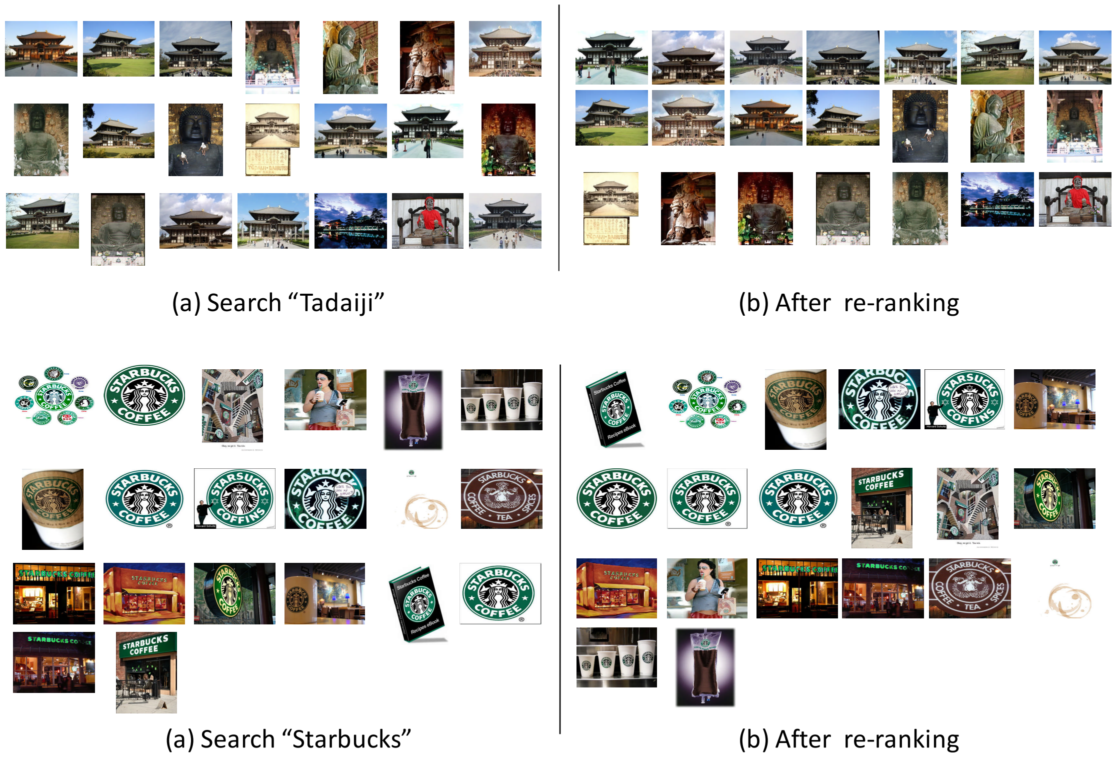
Figure 12. Displaying ROIs on mobile devices.

4.6 Complexity Reduction

Table 4 shows nine examples about the number of photo pairs needed to be checked in NDD with and without the sub-clustering process described in Section 3.3.2. We can see that the times of NDD is largely reduced if we cluster photos into smaller groups first. Note that the number of reduction depends on the content characteristics of a photo cluster. If photos in the same cluster have large variations, i.e., higher entropy in this cluster, there may be more sub-clusters with similar sizes, and the number of reduction is larger.

Table 4. Number of photo pairs needed for NDD.

Scenic spot	# photos in this cluster	# pairs without sub-clustering	# pairs with sub-clustering
Arco di Tito	14	91	52
Basllique du Sacre Coeur	18	153	29
Eiffel Tower	16	120	40
Himeji	20	190	117
Rokuonji	15	105	82
Statue of Liberty	24	276	83
Time Square	21	210	167
Todaiji	15	105	90
Westminster	21	210	99



(a) Search “Tadaiji”

(b) After re-ranking

(a) Search “Starbucks”

(b) After re-ranking

Figure 13. Comparison between search results (a) before re-ranking and (b) after re-ranking.

5. Extensions and Discussion

In this section, we provide some extensions of the proposed framework and demonstrate practicality of these extensions.

- Search Results Re-Ranking

We describe a few interesting extensions of the proposed representative selection method. Because we define the extent of representative by centrality values, the degree of representative for each photo can be quantitatively expressed. Therefore, we can rank photos according to the extent of representative.

Although many search engines have provided image search functions, search results are still not very accurate, and are often not appropriately ordered. Here, we apply the proposed representative selection process to re-rank image search results rather than just picking a representative photo. We invoke an image search in Google, collect the first twenty returned images, calculate the degree of representative for each photo, and re-rank search results according to centrality values. Figure 13 shows two real examples about search results before and after re-ranking. For the search results of “Tadaiji,” the largest wooden building in the world, we re-rank the most canonical views first (appearance of this temple) and then others (the statue of Buddha). For the search results of “Starbucks,” we re-rank the famous trademarks first. We can clearly see that the idea of re-ranking based on representative measurement is practical, though it may be able to be applied in topics having clear structure, e.g., trademark or buildings. Results of this pilot trial provide a possible direction for future research.

- Photo Summarization

After transforming near-duplicate relationships between photos into a graph, we can dynamically determine the number of representative photos for each scenic spot. Three cases would be considered: (1) if only one near-duplicate group exists in this cluster, i.e., one subgraph, only the photo with the largest centrality value is selected as the representative photo; (2) if there is only one subgraph, the nodes with the first few largest centrality values are selected; (3) if there are more than two subgraphs, one or a few nodes with the largest centrality values are selected for each subgraph.

After the process described above, a sequence of representative photo displayed in temporal order can be generated to present the progress of a journey. This summarization method takes advantage of NDD, and this system can effectively select the photos containing important objects. In contrast to conventional content-based method, we can select photos that convey clear semantics and provide more impact on recalling travel experience. A system demo can be found in [27].

- Image Retrieval Aided by Feature Classification

We have demonstrated that different features bring different impacts to near-duplicate detection. In this work, impacts of features are either applied to NDD or not, i.e., only artificial feature points are considered. However, impacts brought by different features can be “softly” fused to facilitate image retrieval. For example, matching between artificial features may be weighted larger than that between natural features.

Note that the proposed feature classification method is not limited to classify features into artificial or natural ones. Decision of feature classes depends on dataset and targeted problems. Recently, researchers start to pay attention to conduct soft fusion of local feature points [28], and we believe the proposed feature classification could play an important role in how to fuse different features’ impacts. Softly fusing feature points is still an open issue needed extensive investigation, and therefore we leave detailed experiments and fusion schemes in the future work.

6. Conclusion

We have presented an efficient photo management and browsing system, by exploiting the techniques of near-duplicate detection. To make near-duplicate detection process more robust to noisy features, we develop three approaches to accomplish feature filtering. The region-based filtering approach that considers spatial relationships between feature points and models feature characteristics by an SVM classifier is found to be the most effective method, under the condition of limited training data. On the basis of a discriminative model that describes matching patterns between photos, we find near-duplicate photo pairs, transform near-duplicate relationships into a graph, and determine the most representative photo by discovering graph structure. We design an evaluation method, based on subjective judgement, to quantitatively express the performance of representative selection. For ROI determination, we compare the proposed method with the saliency-based approach and demonstrate that our approach more adequately captures semantics or important views in images. Comprehensive experimental results are provided, and a few extended studies are described as clues to apply near-duplicate detection to other fields.

This paper focuses on photos taken in travels. However, the proposed approaches are able to be applied to other images that have specific themes or structure. For example, keyframes of news videos can be examined to facilitate video copy detection or topic tracking. In feature filtering, more types of feature points can be modeled and categorized to facilitate image segmentation or to lift the robustness of image concept detection. Finally, computational complexity of detecting near-duplicate is still a big issue to make related studies more practical in real-world usage.

Acknowledgements

This work was partially supported by the National Science Council of ROC under NSC 96-2218-E-194-005 and NSC 97-2221-E-194-050. The authors would like to thank anonymous reviewers for giving valuable comments.

References

[1] Platt, J.C., Czerwinski, M., and Field, B.A. 2003. PhotoTOC: automating clustering for browsing personal photographs. In Proc. of IEEE Pacific Rim Conference on Multimedia, 6-10.

[2] Ke, Y., Sukthankar, R., and Huston, L. 2004. Efficient near-duplicate detection and sub-image retrieval. In Proc. of ACM Multimedia, 869-876.

[3] Chu, W.-T. and Lin, C.-H. 2008. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In Proc. of ACM Multimedia, 829-832.

[4] Wang, X.-J., Zhang, L., Jing, F., and Ma, W.-Y. 2006. AnnoSearch: image auto-annotation by search. In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1483-1490.

[5] Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 2, 91-110.

[6] Zhao, W.-L., Ngo, C.-W., Tan, H,-K., and Wu, X. 2007. Near-duplicate keyframe identification with interest point matching and pattern learning. IEEE Trans. on Multimedia, 9, 5, 1037-1048.

[7] Milolajczyk, K. and Schmid, C. 2005. A performance evaluation of local descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence, 27, 10, 1615-1630.

[8] Walther, D., and Koch, C. 2006. Modeling attention to salient proto-objects. Neural Networks, 19, 1395-1407.

[9] Ke, Y. and Sukthankar, R. 2004 PCA-SIFT: A more distinctive representation for local image descriptors. In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, 506-513.

[10] Jing, Y. and Baluja, S. 2008. VisualRank: Applying pagerank to large-scale image search. IEEE Trans. on Pattern Analysis and Machine Intelligence, 30, 11, 1877-1890.

[11] Wu, X., Hauptmann, A.G., and Ngo, C.-W. 2007. Practical elimination of near-duplicates from web video search. In Proc. ACM Multimedia, 218-227.

[12] Wu, X., Ngo, C.-W., and Hauptmann, A.G. 2008. Multimodal news story clustering with pairwise visual near-duplicate constraint. IEEE Trans. on Multimedia, 10, 2, 188-199.

[13] Zhao, W.-L. and Ngo, C.-W. 2009. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. IEEE Trans. on Image Processing. 18, 2, 412-423.

[14] Wu, X., Ngo, C.-W., Hauptmann, A.G., and Tan, H.-K. 2009. Real-time near-duplicate elimination for web video search with content and context. IEEE Trans. on Multimedia, 11, 2, 196-207.

[15] Sivic, J. and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In Proc. of IEEE International Conference on Computer Vision, 2, 1470-1477.

[16] Sivic, J. and Zisserman, A. 2008. Efficient video search for objects in videos. Proceedings of the IEEE, 96, 4, 548-566.

[17] Wang, F., Jiang, Y.-G., and Ngo, C.-W. 2008. Video event detection using motion relativity and visual relatedness. In Proc. of ACM Multimedia, 239-248.

[18] Zhou, X., Zhuang, X., Yan, S., Chang, S.-F., Hasegawa-Johnson, M., and Huang, T.S. 2008. SIFT-bag kernel for video event analysis. In Proc. of ACM Multimedia, 229-238.

[19] Dorko, G. and Schmid, C. 2003. Selection of scale-invariant parts for object

class recognition. In Proc. of IEEE International Conference on Computer Vision.

[20] Monay, F., Quelhas, P., Odobez, J.M., and Gatica-Perez, D. 2006. Integrating co-occurrence and spatial contexts on patch-based scene segmentation. In Proc. of Conference on Computer Vision and Pattern Recognition Workshop.

[21] Chang, C.-C., and Lin, C.-J. (2001) LIBSVM: a library for support vector machine. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[22] Probabilistic Latent Semantic Analysis,
http://www.robots.ox.ac.uk/~vgg/software/

[23] Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T., and Van Gool, L. 2005. Modeling scenes with local descriptors and latent aspects. In Proc. of IEEE International Conference on Computer Vision.

[24] Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42, 177-196.

[25] Freeman, L.C. 1979. Centrality in social networks. Social Networks, 1, 3, 215-239.

[26] Itti, L., and Koch, C. 2001. Computational modeling of visual attention, Nature Rev. Neuroscience, 2, 3, 194-203.

[27] Chu, W.-T., and Lin, C.-H. 2009. Automatic summarization of travel photos using near-duplication detection and feature filtering, Proceedings of ACM Multimedia Conference, Multimedia Grand Challenges, pp. 1129-1130.

[28] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. 2008. Improving particular object retrieval in large scale image databases, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[29] Chu, W.-T., Lin, C.-H. and Yu, J.-Y. 2009. Feature classification for representative photo selection. Proceedings of ACM Multimedia Conference, pp. 509-512.