Feature Classification for Representative Photo Selection

Wei-Ta Chu
Department of CSIE
National Chung Cheng University, Taiwan
wtchu@cs.ccu.edu.tw

Chia-Hung Lin
Department of CSIE
National Chung Cheng University, Taiwan
lchu96m@cs.ccu.edu.tw

Jen-Yu Yu
Info. and Comm. Research Labs
Industrial Technology Research Inst. Taiwan
KevinYu@itri.org.tw

ABSTRACT
This paper points out that different local feature points provide different impacts to near-duplicate detection and related applications. Aiming to automatic representative photo selection, we develop three feature classification methods, i.e., point-based, region-based, and pLSA-based classification, to differentiate local feature points described by SIFT descriptors. We investigate the performance of these classification methods, and discuss how they influence near-duplicate detection and extended applications. Experiments show that, with effective feature classification, more accurate representative selection results can be achieved.

Categories and Subject Descriptors
I.5.2 [**Pattern Recognition**]: Design Methodology – *classifier design and evaluation, feature evaluation and selection.* H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods.*

General Terms
Algorithms, Measurement, Experimentation.

Keywords
Feature classification, near-duplicate detection, probabilistic latent semantic analysis, representative selection.

1. INTRODUCTION
Efficiently browsing and managing large amounts of digital photos have been significant issues in recent years; especially capturing, storing, and disseminating photos become extremely popular and easy. Therefore, related researches have been explosively proposed from many perspectives. For example, some studies work on automatic tagging or semantic concept detection to facilitate media management, and some works deal with finding regions of interest or generating vivid presentation to enhance user's browsing experience.

In our previous work [1], we exploit near-duplicate detection (NDD) to describe the relationships between photos, and then discover the relationships to find the most representative photo that conveys the most canonical landmark or view of a scenic spot.

MM'09, October 19–24, 2009, Beijing, China.

Besides, we bring up an interesting idea that the matched feature points, which are the intermediate information in the NDD process, often lie on the contour or inside of the representative object. This characteristic inspires us that the most prominent region in a photo can be determined by utilizing the spatial information of matched feature points.

Although promising results were reported in [1], we found that different local feature points with varied characteristics may discrepantly influence the performance of near-duplicate detection. To human beings, the concept "duplication" often comes from that two images have the same artificial objects, such as building, tower, and statute. This characteristic is especially convincing in consumer photos taken in journeys. 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 paper, we investigate three different feature classification methods and conduct comprehensive experiments.

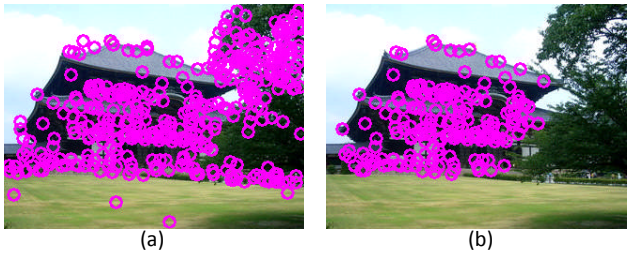
(a) (b)

Figure 1. An example of showing (a) all feature points and (b) only the feature points on artificial objects.

The rest of this paper is described as follows. Section 2 gives the system overview. Section 3 describes the major contribution of this paper, i.e., feature extraction and classification. One application, i.e., representative selection, that benefits efficient photo management and browsing is described in Section 4. Section 5 reports the experimental results, and Section 6 provides concluding remarks.

2. SYSTEM OVERVIEW
The 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. However, 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 representative photo can be automatically determined on the basis of near-duplicate detection [1].

Figure 2 shows the four stages conducted in the proposed framework. First, we detect interest points based on a DoG (difference of Gaussian) detector and describe them by SIFT (scale-invariant feature transform) descriptors [2]. To filter out noisy feature points, we investigate the effects of three feature classification methods, including point-based, region-based, and pLSA-based (Probabilistic Latent Semantic Analysis) methods.

At the near-duplicate detection stage, we basically follow the process proposed in [3], 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. For a cluster of photos, we express duplicate relationships between photos as a graph, and then perform relation analysis facilitate finding the most representative photo in a cluster.

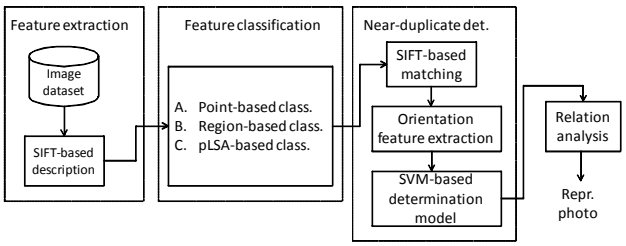

Figure 2. The proposed system framework.

3. FEATURE CLASSIFICATION

Several kinds of features have been designed to characterize interest points or image patches. 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 to find the location of feature points. For feature description, we utilize the SIFT descriptor [2] to describe each feature point as a 128-dimensional vector, which is robust to scale and orientation variations, and sort of illumination changes.

Relatively fewer works discuss the efficiency issue of features to different applications. In this work, we consider photos taken in journeys, and devote to find near-duplicate artificial objects. The reason of putting efforts on artificial objects is that they are often more meaningful to human beings, and on the other hand, 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. SIFT-based feature points are further modeled and classified by the following processes, and the ones being declared as on natural scenes are put aside from the applications described in Section 4.

3.1 Point-Based Classification

SIFT-based description is based on orientation information of small patches in different resolutions, centered by the feature point. Therefore, the feature vector implicitly embeds local structure. Figure 3 shows SIFT-based description of feature points on artificial objects and natural scenes, respectively. These two figures are statistics of 1000 points on 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.

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. In this work, we respectively collect two types of feature points, and construct the mapping function by a binary SVM classifier. At the filtering stage, each feature point is evaluated by the classifier, and is then categorized into an artificial point or a natural point.

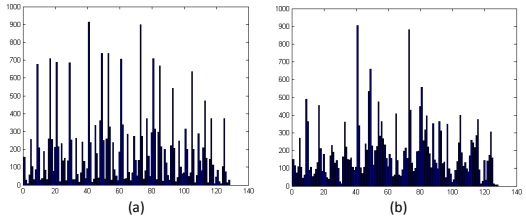

Figure 3. Orientation histograms of feature points on (a) artificial objects and (b) natural scenes, respectively.

3.2 Region-Based Classification

Although Figure 3 shows distinct characteristics on single SIFT descriptors, spatial correlation between feature points in neighborhood is not considered. Conceptually, examining a single feature point and classifying it into an element of artificial objects or natural objects may unavoidably suffer an issue that is 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,

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

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.

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 vectors of feature points in the same region. At the filtering stage, each region is evaluated by the SVM classifier, and is then categorized into an artificial region or a natural region. All feature points in an artificial region are then claimed as artificial points.

3.3 PLSA-Based Classification

Another approach to consider the context information between feature points was proposed in [4]. We modify their method 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 group them into a 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 all SIFT descriptors in this cluster, and is called as a *visual word* that represents this 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 words v_i^a and v_i^n in $\mathcal{V}^a$ and $\mathcal{V}^n$ by quantizing it into one of the pre-trained visual vocabularies. That is,

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

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

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 artificial objects is estimated based on the co-occurrence information between artificial feature points. We exploit probabilistic latent semantic analysis (pLSA) model to build a joint probability model over the image d_j and the visual word v_i^a:

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

where $z_\ell \in \mathcal{Z} = \{z_1, ..., z_{N_A}\}$ is a latent concept subtly embedded in the visual vocabulary $\mathcal{V}^a$. The pLSA model is defined by the conditional probability $P(v_i^a|z_\ell)$ that represents 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, using a set of training data that includes artificial points. 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 $z_{v_i^a}$ that are learned from artificial object training data. Based on the pLSA model, the most likely concept can be determined as follows:

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

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

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

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.

4. REPRESENTATIVE SELECTION

Given a set of photos $\mathrm{P} = \{p_1, p_2, ..., p_N\}$, we first filter out feature points that are claimed as natural points by the methods described above. Then, whether a pair of photos (p_i, p_j), $i \neq j$, $i, j \leq N$, is near-duplicate is determined by the method proposed in [3]. 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. Given this graph, we determine the most important node by checking the "centrality value" of each node. From the idea of social network modeling, the person who is "closest" to all others plays the most important role. Similarly, the photo that is mostly near-duplicate to others is the most representative one. We evaluate the centrality value of each node by the degree centrality [1]. The degree centrality of a node v_i is

$$\text{degree centrality}(v_i) = \frac{\sum_{k=1}^{n} a(v_i, v_k)}{n-1}, \quad (7)$$

where $a(v_i, v_k) = 1$ if v_i and v_k are connected, and otherwise $a(v_i, v_k) = 0$.

5. EXPERIMENTS

5.1 Evaluation Dataset

- Training for the point-based classification: There are totally 3483 artificial feature points and 6170 natural feature points for training, which are extracted from twelve photos. By labeling artificial points as positive samples and natural points as negative samples, we construct an SVM classifier [5] to determine whether a feature point is artificial or natural.
- Training for the region-based classification: 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, which are extracted from forty photos. Note that only the regions in which all feature points belong to artificial or natural points are selected as the training data.
- Training for the pLSA-based classification: Four hundred photos are used for training. There are totally 53655 artificial feature points, which are clustered into 600 visual words. Similarly, 63769 natural feature points are used to construct 600 visual words. In this work, we use the program provided in [6] to implement the proposed approach.

5.2 Performance of Feature Classification

Based on manually labeled ground truths in which ten photos include 3182 artificial feature points and 5173 natural feature points, we calculate precision rate for each classification method as $\text{Precision} = \frac{C_a + C_n}{N}$, where C_a is the number of artificial

feature points that are correctly classified, and C_n is the number of natural feature points that are correctly classified. The denominator N is always 3182+5173=8355. The precision rates for point-based, region-based, and pLSA-based methods are 0.81, 0.92, and 0.26, respectively. The point-based and region-based methods work much better than the pLSA-based approach.

Figure 4 gives some examples of classification results. From the third to the fifth 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 method, many feature points on trees are mis-classified, which cause many noises to near-duplicate detection.

The pLSA-like seems to be inappropriate in feature classification when only limited training data are available. Although promising performance has been reported in [4] and [7], large amounts of training data were needed, 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 [7]. A pLSA model that includes 60 concepts is then constructed based on 300 photos. On the other hand, the 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.

Figure 4. Sample results of different feature classification methods.

5.3 Performance of Representative Selection

To evaluate 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. A score ranges from one to five. A larger score is given if the observer thinks a photo better represents a scenic spot. 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 score.

We evaluate fifty-two photo sets, which include 1024 photos representing building, statute, and cityscape. We just show some results in Table 1 due to space limitation. The overall selection performances are 3.58, 3.61, 3.63, and 3.32 for photo sets (1) without feature filtering; (2) with the point-based filtering; (3) with the region-based filtering; and (4) with the pLSA-based filtering, respectively. We can clearly see that performances of the selections with point-based and region-based filtering are better than that without feature filtering. This confirms the idea we introduced in Section 1. The pLSA-based approach doesn't have good classification performance, and therefore leaves many noises to harm the selection module.

Table 1. Performance of representative selection.

Scenic spot	(1)	(2)	(3)	(4)
Athens	3.44	3	3.67	2.89
Back Bay	4.11	2.89	2.11	4.33
Baltimore	4.33	3.33	3.78	4.44
Basllique	3.89	4.78	3.89	3.78
Paris	3.22	3	4.78	3.66
…	…	…	…	…
Overall	3.58	3.61	3.63	3.32

6. CONCLUSION

We have presented that different feature points don't equally impact near-duplicate detection and related applications. Three feature classification methods are developed to discriminate each feature point as an artificial point or natural point. Artificial feature points provide more important clues in near-duplicate detection. With the results of near-duplicate detection, we describe relationships between photos as a graph and analyze its link structure to find the most representative photo. Experimental results show that the region-based classification method that considers locality of feature points has superior performance. Moreover, with appropriate feature classification and filtering, more satisfactory results can be obtained in representative selection. In the future, weighted SIFT matching based on feature classification results would be investigated, evaluation based on more extensive datasets will be conducted, and more promising applications will be developed.

7. ACKNOWLEDGEMENT

This work was partially supported by the National Science Council of the Republic of China under grants NSC 97-2221-E-194-050.

8. REFERENCES

[1] 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.

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

[3] 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.

[4] 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 ICCV.

[5] 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

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

[7] 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.