# KIDS Lab at ImageCLEF 2012 Personal Photo Retrieval

Chia-Wei Ku, Been-Chian Chien, Guan-Bin Chen, Li-Ji Gaou, Rong-Sing Huang, and Siao-En Wang

Knowledge, Information, and Database System Laboratory
Department of Computer Science and Information Engineering
National University of Tainan, Tainan, Taiwan
cooperku@msn.com
bcchien@mail.nutn.edu.tw

**Abstract.** The personal photo retrieval task at ImageCLEF 2012 is a pilot task for testing QBE-based retrieval scenarios in the scope of personal information retrieval. This pilot task is organized as two subtasks: the visual concepts retrieval and the events retrieval. In this paper, we develop a framework of combining different visual features, EXIF data and similarity measures based on two clustering methods to retrieve the relevant images having similar visual concepts. We first analyze and select the effective visual features including color, shape, texture, and descriptor to be the basic elements of recognition. A flexible similarity measure is then given to achieve high precise image retrieval automatically. The experimental results show that the proposed framework can provide good effectiveness in distinct measures of evaluation.

**Keywords:** Image retrieval, Concept retrieval, Features clustering, Similarity measure

## 1 Introduction

The main aim of the ImageCLEF 2012 personal photo retrieval task is providing a test bed for image retrieval based on some given query images [1]. The task is further divided into two subtasks: the visual concepts retrieval and the events retrieval. Compare with traditional image retrieval, the topics of this task are more abstract or more general. It might cause image retrieval to be more difficult. The benchmark data set used in this task consists of 5,555 images downloaded from Flickr. Both the visual concepts retrieval and the events retrieval use the same dataset.

The visual concepts retrieval is a great challenge to the developers. Some of the concepts are abstract like the topic “sign,” and some of them are very subjective like “art object.” Even different people would draw different opinions on the same image. The events retrieval is to find the images with the same kinds of events. Some of the target topics like “fire” and “conference” are too general to define in visual concept. Parts of the events in this subtask connect with geographical topics. Thus, most of the topics are difficult to retrieve in visual. In such a case, EXIF features may support much more information about the event concept.

In our participation to the ImageCLEF 2012 personal photo retrieval task, we developed a framework for the visual concepts retrieval and the events retrieval. First, we selected 7 visual features from the given features set for the task. Each selected feature is used to cluster all the images into groups individually. We first define the similarity degree for visual features and EXIF's information. Then, the similarity measures for different image features are integrated to estimate the similarity scores between each image and the query image. The cluster of each feature is used to help weighting the image similarity. Finally, the framework combines and ranks the similarity degrees between an image and the different QBE images to retrieve the photos with the same concept.

The remainder of this paper is organized as follows. We describe the used features provided by the organizers in Section 2. Section 3 introduces the proposed similarity measures and retrieval methods. In Section 4, we present the experimental results of our proposed framework. Finally, we conclude the paper with a discussion and future work.

## 2 Process of Image Features

### 2.1 Visual Features

The original datasets in the personal photo retrieval task provided 19 extracted visual features. After our estimating test, 7 features were selected from the 19 features. They are AutoColorCorrelogram [2], BIC [3], CEDD [4], Color Structure, Edge Histogram, FCTH [5], and SURF [6]. The selected features cover different kinds of popular visual perception including color, shape, and texture. SURF is a robustly scale-invariant and rotation-invariant descriptor feature. The features are summarized in Table 1.

**Table 1.** The selected 7 features.

| Visual Features | Color | Shape | Texture | Descriptor |
|---|---|---|---|---|
| AutoColorCorrelogram | ○ | | | |
| BIC | | ○ | ○ | |
| CEDD | ○ | ○ | | |
| Color Structure | ○ | | | |
| Edge Histogram | | ○ | | |
| FCTH | ○ | ○ | | |
| SURF | | | | ○ |

**Visual Features Clustering.** We first cluster images by the individual visual features to find the groups of images with the similar visual features. Two different clustering methods are developed for the SURF descriptor and the others visual features, respectively. We depict the clustering algorithms in the following.

*SURF feature clustering*. The SURF descriptor is the feature with scale-invariant and rotation-invariant. In this paper we defined the *matching pair* to measure the similarity between two images. If the SURF descriptor $d_i$ for the image $I_i$ matches another descriptor $d_j$ in the image $I_j$ and vice versa, the descriptors $d_i$ and $d_j$ form a matching pair. The distance between two images $I_i$ and $I_j$ is defined as

$$dist_{SURF}(I_i, I_j) = \frac{1}{\sqrt{N_{mp}(I_i, I_j)}}, \tag{1}$$

where $N_{mp}(I_i, I_j)$ is the number of matching pairs between the two images $I_i$ and $I_j$. The larger $N_{mp}$ is, more similar two images are. Based on the measure of the matching pair, we propose the clustering algorithm for SURF descriptors, shown as Table 2. Before describing the detailed algorithm, we define two cluster distances: the *intra-cluster* $D_{intra}(C_k)$ and the *inter-cluster* $D_{inter}(C_k, C_l)$.

**Table 2.** The clustering algorithm for SURF feature.

**Algorithm:** *SurfCluster*

**Input**: the set of images **I**
**Output**: the clusters of images **C**
**C** = {};
**while** ( $\min\{dist_{SURF}(I_i, I_j)\} < \theta$ )  // $\theta$ is the threshold of SURF distance
  **Case 1:** $I_i \in C_k$ and $I_j \in C_l$ for $C_k, C_l \in$ **C** and $C_k \neq C_l$
    **if** ( $D_{intra}(C_k \cup C_l) \leqq \mu_1 \times \min\{D_{intra}(C_k), D_{intra}(C_l)\}$ ) // $\mu_1$ is a constant.
      **C** = **C** $\cup$ $\{C_k \cup C_l\}$ - $\{C_k\}$ - $\{C_l\}$;
    **else**
      *SurfCluster*($C_i \cup C_j$);
    **end if**
  **Case 2:** $I_i \in C_k$ and $I_j \notin C_k$ for $C_k \in$ **C**
    **if** ( $D_{inter}(I_j, C_k) \leqq \mu_2 \times D_{intra}(C_k)$ ) // $\mu_2$ is a constant.
      **C** = **C** $\cup$ $\{C_k \cup \{I_j\}\}$;
    **else**
      **C** = **C** $\cup$ $\{\{I_i, I_j\}\}$;
    **end if**
  **Case 3:** $I_i \notin C_k$ and $I_j \notin C_k$ for all $C_k \in$ **C**
    **C** = **C** $\cup$ $\{\{I_i, I_j\}\}$;
**end while**

**for** $C_k, C_l$ **in C**
  **if** ($C_k \cap C_l \neq \varnothing$)
    **if** ( $D_{intra}(C_k \cup C_l) \leqq \mu_1 \times \min\{D_{intra}(C_k), D_{intra}(C_l)\}$ )
      **C** = **C** $\cup$ $\{C_k \cup C_l\}$ - $\{C_k\}$ - $\{C_l\}$;
    **else**
      *SurfCluster*($C_i \cup C_j$);
    **end if**
  **end if**
**end for**

$$D_{intra}(C_k) = \frac{1}{|C_k|^2} \sum dist_{SURF}(I_i, I_j), \text{ for } I_i, I_j \in C_k; \tag{2}$$

$$D_{inter}(I_j, C_k) = \frac{1}{|C_k|} \sum dist_{SURF}(I_i, I_j), \text{ for } I_i \in C_k. \tag{3}$$

According to our observation, if the number of matching pairs is larger than four, the images look similar in visual. Hence, we define the similarity for SURF feature as

$$S_{SURF}(I_i, I_j) = \max\left\{1, \sqrt{\frac{N_{mp}(I_i, I_j) - 4}{4}}\right\}. \tag{4}$$

*Other Visual Features.* For other visual features, the clustering methods consider only the similarity between two images using the distance measures of Table 3. The detailed algorithm is list as Table 4.

**Table 3.** Features and their distance measures.

| Visual Feature | Distance Measure |
|---|---|
| AutoColorCorrelogram | $L_1$ measure |
| BIC | $L_1$ measure |
| CEDD | Tanimoto measure |
| Color Structure | $L_1$ measure |
| Edge Histogram | $L_1$ measure |
| FCTH | Tanimoto measure |

**Table** 4. The clustering algorithm for general visual features.

| **Algorithm**: *VisualCluster* |
|---|
| **Input:** the set of images **I** |
| **Output:** the cluster of images **C** |
| **C** = {}; |
| **for** $I_i \in$ **I** |
|     **C** = **C** $\cup$ {{$I_i$}}; |
| **end for** |
| **for** $C_k$, $C_l$ **in C** |
|     **if** ( min{$dist(C_k, C_l)$} < $\theta$ )   // $\theta$ is the threshold of the minimum distance. |
|         **C** = **C** $\cup$ {$C_k \cup C_l$} - {$C_k$} - {$C_l$}; |
|     **end if** |
| **end for** |

### 2.2 Textual Features

The textual features are mainly extracted from EXIFs. There are totally 63 features in EXIFs; for example, ApertureValue, BrightnessValue, ColorSpace, CompressedBitsPerPixel, Contrast, etc. However, only two features, the GPS and the time, were considered and used in our methods. The values of the GPS and the time are also clustered by the same clustering algorithm of general visual features shown in Table 4 using $L_1$ distance measure.

## 3 The Measure for Similarity Image Retrieval

### 3.1 Normalization of Visual Features

The ranges of feature distances are quite different for all visual features. Before combining all the features to measure the similarity of images, the normalization process is necessary. We use the approximation proposed by Abramowitz & Stegun [7] to approximate the values of normalization. The approximation step is very fast and accurate. Let $x$ be the similarity between two images of an image feature, the normalization was calculated by the following equation,

$$\Phi(x) = 1 - \phi(x)(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) + \varepsilon(x),\ t = \frac{1}{1 + b_0 x}, \tag{5}$$

where $\phi(x)$ is the normal probability density function of the similarity degrees among all images in the feature, $b_0$ to $b_5$ are constants, and the absolute error $|\varepsilon(x)|$ would be smaller than $7.5 \times 10^{-8}$.

### 3.2 Similarity Measures of Image Features

**The Similarity Measure of Visual Features.** Let $I_i$, $I_j$ denote two images. Then the visual similarity between the images $I_i$ and $I_j$, $S_V(I_i, I_j)$, is defined as

$$S_V(I_i, I_j) = S_{SURF}(I_i, I_j) + \sum_k w_k \Phi\left(S_{f_k}(I_i, I_j)\right), \tag{6}$$

where $S_{fk}(I_i, I_j)$ means the similarity between the images $I_i$ and $I_j$ of the $k$-th feature, $w_k$ is the weight of the $k$-th feature. Two weighting methods, the *cluster weighting* and the *non-cluster weighting*, are proposed as follows:

- *Cluster Weighting.* We use the clustering results of Section 2 to automatically weight the features. If a query image belongs to a cluster for a specific visual feature, the average similarity between the query image and each image in the cluster is computed as the weight of the specific visual feature.

- *Non-Cluster Weighting.* In this method, the weights $w_k$ are set to 1, except for the weights of AutoColorCorrelogram, Color Structure, and SURF features double other visual features.

**The Similarity Measure of the GPS feature.** Two distance similarity measures are proposed for the geographical distance:

- *Boolean measure.* The Boolean measure of the GPS feature is defined as

$$S_{G(\mathrm{B})}(I_i, I_j) = \begin{cases} 1 & \text{if } GPS(I_i) \text{ and } GPS(I_j) \text{ are in the same cluster,} \\ 0 & \text{otherwise;} \end{cases} \tag{7}$$

  where $GPS(I_i)$ and $GPS(I_j)$ denote the values of the GPS feature in EXIF for $I_i$, $I_j$.
- *Similarity measure.* The continuous similarity measure on geographical distance is defined as

$$S_{G(\mathrm{S})}(I_i, I_j) = 1 \Big/ 1 + e^{\frac{dist(GPS(I_i), GPS(I_j)) - radius}{\mu}}\ , \tag{8}$$

  where $\mu$ and *radius* are smoothing parameters; $dist(GPS(I_i), GPS(I_j))$ means the real geographical distance on earth between the two positions $GPS(I_i)$, $GPS(I_j)$.

**The Similarity Measure of the Time feature.** Two time similarity measures are proposed for time duration:

- *Boolean measure.* The Boolean measure of the time feature is defined as

$$S_{T(\mathrm{B})}(I_i, I_j) = \begin{cases} 1 & \text{if } T(I_i) \text{ and } T(I_j) \text{ are in the same cluster,} \\ 0 & \text{otherwise;} \end{cases} \tag{9}$$

  where $T(I_i)$ and $T(I_j)$ denote the time feature in EXIF of $I_i$ and $I_j$.
- *Similarity Measure.* The continuous similarity measure on time is defined as

$$S_{T(\mathrm{S})}(I_i, I_j) = 1 - \Phi\big(dist(T(I_t), T(I_j))\big). \tag{10}$$

  where $dist(T(I_i), T(I_j))$ denote the real time difference in second between two time-stamp $T(I_i)$ and $T(I_j)$.

### 3.3 The Ranking of Image Similarity

Finally, we define the similarity between two images $I_i$ and $I_j$ by integrate the features $S_V(I_i, I_j)$, $S_{G(\cdot)}(I_i, I_j)$, and $S_{T(\cdot)}(I_i, I_j)$ into a linear combination. The image similarity $Sim(I_i, I_j)$ is defined as

$$Sim(I_i, I_j) = w_V \times S_V(I_i, I_j) + w_G \times S_{G(\cdot)}(I_i, I_j) + w_T \times S_{T(\cdot)}(I_i, I_j)\ . \tag{11}$$

Given a set of query images $Q_j$, $1 \le j \le m$, the similarity of each query image $Q_j$ and the image $I_i$ in the image set is measured by $Sim(I_i, Q_j)$. The maximum similarity $Sim(I_i, Q_j)$ is the similarity degree of the image $I_i$ for the visual concept via the $m$ query images $Q_j$. It can be formally defined as

$$\max_{1 \le j \le m} \{Sim(I_i, Q_j)\} . \tag{12}$$

## 4 Experiments and Discussion

### 4.1 Experimental Environments

The system is implemented on a Microsoft Windows XP SP 3, 2.33 GHz PC with 3.00GB RAM. The developed software and related systems are written in Java language, so the system is cross-platform. The methods in five runs used different image features, which are shown in Table 5. The notations in the table are: V stands for the visual features; G denotes the GPS feature; T is the time feature. While the parameter C, N means the cluster weighting and the non-clustering weighting, respectively. Finally, the parameter B represents the Boolean measures and S is the similarity measures.

**Table 5.** Features we used in our methods.

| | Visual Features | GPS | Time |
|---|---|---|---|
| V | C | | |
| V + G | N | B | |
| V + T | N | | B |
| V + G + T | N | B | B |
| G + T | | S | S |
| T | | | S |

### 4.2 Results of Subtask 1: Retrieval of Visual Concepts

In this subtask, 24 visual concept queries were given to be evaluated from the totally 32 concepts. The retrieval results for the visual concepts are evaluated by three different measures: precision, NDCG (normalize discount cumulative gain) [8], and MAP (mean average precision). The experimental results are shown in Table 6.

As Table 6 shows, the Run 5 using all of the image features is the best one for all measures. The second place is the Run 2 which uses the time feature only. The Run 3 with the visual features and the GPS feature is the third place. The Run 1 and the Run 4 are worse than the above three runs.

The results show that the visual features are not useful for most of the visual concepts in the task. The reason is that most of the concept topics are semantically related to each other. There is not much common characteristic in visual features among QBE images. While combining the visual features with the EXIF features, the performance increases obviously. The GPS feature can help us to find the images photographed in the neighboring positions easily. The geographic-related topics like “Asian temple & palace” and “temple (ancient)” have good results. However, some topics are not expected to be good, like “animals” and “submarine scene,” which returned high precision. The main reason is that a photographer generally tries to take pictures with the similar topics at the same place. Although the GPS feature is precise for geographic-related topics, the missing values on the GPS feature will degrade the precision greatly. Some non-geographical topics have obviously bad results in the runs of using the GPS features, like “clouds.” The time feature is also an important factor for searching personal photos. Since the images photographed in short time are usually very similar or dependent in visual concept. As the above discussion, the Run 5 getting the best results shows that our image similarity measure method can combine the different image features effectively.

**Table 6.** Performance on retrieval of visual concepts.

| | Run 1 | Run 2 | Run 3 | | Run 4 | | Run 5 | | |
|---|---|---|---|---|---|---|---|---|---|
| Features | V | T | V + G | | G + T | | V + G + T | | |
| Weights | $w_V$ | $w_T$ | $w_V$ | $w_G$ | $w_G$ | $w_T$ | $w_V$ | $w_G$ | $w_T$ |
| | 1 | 1 | 0.45 | 0.17 | 0.975 | 0.025 | 0.45 | 0.18 | 0.22 |
| P@5 | 0.6750 | 0.8000 | 0.7667 | | 0.6500 | | 0.8333 | | |
| P@10 | 0.6125 | 0.7292 | 0.6583 | | 0.6500 | | 0.7833 | | |
| P@15 | 0.5778 | 0.6667 | 0.6222 | | 0.6083 | | 0.7222 | | |
| P@20 | 0.5354 | 0.6354 | 0.6104 | | 0.5771 | | 0.6896 | | |
| P@30 | 0.4486 | 0.6083 | 0.5639 | | 0.5611 | | **0.6347** | | |
| P@100 | 0.3054 | 0.4117 | 0.3925 | | 0.3925 | | **0.4379** | | |
| NDCG@5 | 0.5701 | 0.5858 | 0.5800 | | 0.4073 | | **0.6405** | | |
| NDCG@10 | 0.5062 | 0.5348 | 0.5184 | | 0.4268 | | **0.6017** | | |
| NDCG@15 | 0.4798 | 0.5028 | 0.4951 | | 0.4123 | | **0.5658** | | |
| NDCG@20 | 0.4545 | 0.4836 | 0.4872 | | 0.4066 | | **0.5459** | | |
| NDCG@30 | 0.4016 | 0.4728 | 0.4615 | | 0.4046 | | **0.5213** | | |
| NDCG@100 | 0.3303 | 0.4144 | 0.3979 | | 0.3717 | | **0.4436** | | |
| MAP@30 | 0.0632 | 0.0952 | 0.0906 | | 0.0854 | | **0.1026** | | |
| MAP@100 | 0.0930 | 0.1589 | 0.1558 | | 0.1518 | | **0.1777** | | |

### 4.3 Results of Subtask 2: Retrieval of Events

In the subtask, totally 15 different events queries are given to find the pictures with the same event. Each query contains three QBE images. The evaluations are done by precision, NDCG, and MAP as the subtask 1. The experimental results are shown in Table 7.

As Table 7 shows, the best results are the Run 2 and Run 5. The Run 1 using the visual features is still the worst as the subtask 1. The Run 3 using the visual and the GPS features is a little better than the Run 4 taking the visual and the time features.

Owing to the event queries usually describe the images with the properties of happening in specific time duration or location area, the time and the GPS features are relatively important here. For example, the topics “Australia,” “Bali,” and “Egypt” are related in geographical; the topics of activities like “conference,” “party,” and “rock concert” are temporal-related. Hence, the provided EXIFs of the images are very useful in this subtask of events retrieval. The Run 5 combining all features is not expected to be the best as the subtask 1. The reason might be that the event queries are not so related with the visual concept, but highly dependent on time and location. However, the proposed similarity measure method did not degrade the precision much.

**Table 7.** Performance on retrieval of events.

| | Run 1 | Run 2 | | Run 3 | | Run 4 | | Run 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Features | V | G + T | | V + G | | V + T | | V + G + T | | |
| Weights | $w_V$ | $w_G$ | $w_T$ | $w_V$ | $w_G$ | $w_G$ | $w_T$ | $w_V$ | $w_G$ | $w_T$ |
| | 1 | 0.975 | 0.025 | 0.45 | 0.17 | 0.45 | 0.22 | 0.45 | 0.18 | 0.22 |
| P@5 | 0.6533 | **1.0000** | | 0.9333 | | 0.9200 | | **1.0000** | | |
| P@10 | 0.5800 | **1.0000** | | 0.9000 | | 0.8733 | | **1.0000** | | |
| P@15 | 0.5156 | **0.9644** | | 0.8533 | | 0.8400 | | **0.9644** | | |
| P@20 | 0.4833 | **0.9333** | | 0.8100 | | 0.7867 | | 0.9267 | | |
| P@30 | 0.4467 | **0.8889** | | 0.7622 | | 0.6956 | | 0.8756 | | |
| P@100 | 0.2693 | **0.6787** | | 0.5740 | | 0.4613 | | 0.6307 | | |
| NDCG@5 | 0.6904 | **1.0000** | | 0.9417 | | 0.9201 | | **1.0000** | | |
| NDCG@10 | 0.6247 | **1.0000** | | 0.9153 | | 0.8877 | | **1.0000** | | |
| NDCG@15 | 0.5727 | 0.9837 | | 0.8884 | | 0.8681 | | **0.9841** | | |
| NDCG@20 | 0.5446 | **0.9697** | | 0.8636 | | 0.8357 | | 0.9655 | | |
| NDCG@30 | 0.5186 | **0.9586** | | 0.8458 | | 0.7854 | | 0.9489 | | |
| NDCG@100 | 0.4101 | **0.9126** | | 0.8042 | | 0.6638 | | 0.8601 | | |
| MAP@30 | 0.1100 | **0.3305** | | 0.2800 | | 0.2287 | | 0.3225 | | |
| MAP@100 | 0.1484 | **0.5533** | | 0.4282 | | 0.3179 | | 0.4947 | | |

## 5 Conclusion

In this paper we proposed a framework and similarity measure methods to combine different image features for retrieving images from a set of conceptual photos. The proposed method can handle the visual concepts retrieval subtask in part. However, the time and position information are more important than other visual features in the event retrieval subtask. Although the proposed method could adjust the weights to fit the requirements, it has still a lot of problems to be solved. The proposed framework retrieved the relevant images weighted by manual in most of the cases. As we know, the feature selection is important in retrieval individual concept. For example, the experimental results show that the GPS and the time features are very useful for retrieval in this dataset. However, it may be not so effective in other dataset. The problem of selecting and weighting the features automatically is a challenge in the task.

This pilot task is its first year announced at ImageCLEF. The dataset seems too small for evaluating modern applications. Further, the concept queries often contain some irrelevant images in visual. The procedure of determining concepts and their relevant images may need to be fixed for providing as a benchmark.

## References


1. Zellhöfer, D.: Overview of the Personal Photo Retrieval Pilot Task at ImageCLEF 2012. In: CLEF 2012 working notes, Rome, Italy (2012)
2. Huang, J., Kumar, S., Mitra, M., Zhu, W. J., Zabih, R.: Image Indexing Using Color Correlograms. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 762-768 (1997)
3. Stehling, R. O., Nascimento, M. A., Falcão, A. X.: A Compact and Efficient Image Retrieval Approach Based on Border/Interior Pixel Classification. In: Proceedings of the eleventh international conference on Information and knowledge management, pp. 102-109 (2002)
4. Chatzichristofis, S. A., Boutalis, Y. S.: CEDD: Color and Edge Directivity Descriptor. A Compact Descriptor for Image Indexing and Retrieval. In: Proceedings of the 6th International Conference on Computer Vision Systems, vol. 5008/2008, pp. 312-322, Springer (2008)
5. Chatzichristofis, S. A., Boutalis, Y. S.: FCTH: Fuzzy Color and Texture Histogram a Low Level Feature for Accurate Image Retrieval. In: Proceedings of the ninth International Workshop on Image Analysis for Multimedia Interactive Services, pp 191-196 (2008)
6. Bay, H., Tuytelaars, T., Gool, L. V.: SURF: Speeded-Up Robust Features. In: 9th European Conference on Computer Vision, vol. 3951/2006, pp. 404-417, Springer (2006)
7. Abramowitz, M. and Stegun, I. A.: Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables. ISBN: 0-486-61272-4. Dover Publications (1965)
8. Järvelin, K., Kekäläinen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems, vol. 20, no. 4, 422 - 446 (2002)