MIL at ImageCLEF 2013: Personal Photo Retrieval

Masaru Mizuochi, Takayuki Higuchi,
Chie Kamada, and Tatsuya Harada

Machine Intelligence Laboratory, The University of Tokyo
{mizuochi,higuchi,kamada,
harada}@mi.t.u-tokyo.ac.jp
http://www.mi.t.u-tokyo.ac.jp

Abstract. In this paper, we describe our methods for ImageCLEF 2013 Personal Photo Retrieval Task. We devote our attention to making our system efficient in retrieving documents which have the similar topic with few query data. We train a ranking function using rankSVM. We extract Fisher Vectors (FVs) from several local descriptors as visual features, and use Bag-of-Words (BoW) as metadata features. The final similarity score is the weighted average of the visual similarities and the metadata similarities, where the weights are determined by relevance feedback. Results have shown that our system achieves the best performance in the Personal Photo Retrieval Task.

1 Introduction

This paper describes our methods for the ImageCLEF 2013 Personal Photo Retrieval Task. Our objective is to investigate efficient methods for retrieving documents that have the similar topic with few query data. The dataset is obtained such that it imitates actual browsing data. It contains query by example(QBE) and browsing data, though some of them are unavailable. We compare three methods to measure a visual similarity between a document and a topic: SVM, rankSVM, and a similarity between a query and the nearest sample. We use the distance metric like radial basis function (RBF) kernel for a metadata similarity. We extract Fisher Vectors (FVs) as visual features, and use Bag-of-Words (BoW) as metadata features. The final similarity score is the weighted average of the visual similarities from several local descriptors and the metadata similarity, where the weights are determined by relevance feedback.

2 Feature Extraction

In this section, we describe the features we use in the task. We extract FVs as visual features and BoW of EXIF tags as metadata features.

2.1 Visual Features

Fisher Vectors. As a visual feature, we use Fisher Vectors (FVs) computed from four kinds of local descriptors: SIFT, C-SIFT, LBP, and GIST. Actually, GIST is usually used to describe a whole image, but we use it as a local descriptor. All local descriptors are reduced to 64 dimensions using Principal Component Analysis (PCA). Local descriptors are densely extracted from five scales of patches on a regular grid every six pixels and learn a Gaussian Mixture Model (GMM) with 256 components, which have a diagonal matrix as its covariance matrix. To use spatial information, we divide images into 1×1, 2×2, and 3×1 cells. Then FVs are calculated over each region as follows.

Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ be a set of N local descriptors extracted from an image, and w_i, $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ be the mixture weight, mean vector, covariance matrix of the i-th Gaussian, respectively. Then we difine,

$$\boldsymbol{u}_i = \frac{1}{N\sqrt{w_i}} \sum_{n=1}^{N} \gamma_n(i) \boldsymbol{\Sigma}_i^{-\frac{1}{2}} (\boldsymbol{x}_n - \boldsymbol{\mu}_i),$$

$$\boldsymbol{v}_i = \frac{1}{N\sqrt{2w_i}} \sum_{n=1}^{N} \gamma_n(i) \left[\boldsymbol{\Sigma}_i^{-1} \mathrm{diag}((\boldsymbol{x}_n - \boldsymbol{\mu}_i)(\boldsymbol{x}_n - \boldsymbol{\mu}_i)^T) - \mathbf{1}\right],$$

where $\mathbf{1}$ is a column vector whose components are all 1 and diag(X) for matrix X is a column vector which is composed of diagonal components of X. $\gamma_n(i)$ is the soft assignment of $\boldsymbol{x}_n$ to i-th Gaussian as

$$\gamma_n(i) = \frac{w_i u_i(x_n)}{\Sigma_{j=1}^{K} w_j u_j(x_n)},$$

where u_i is the i-th Gaussian, and it is also known as the posterior probability. The FV representation is therefore given as

$$\mathcal{G} = \left[\mathbf{u}_1^T \ \mathbf{v}_1^T \ldots \mathbf{u}_K^T \ \mathbf{v}_K^T\right]^T,$$

where K is the number of GMM components.

Following [2], we apply power normalization and L2 normalization to each of the extracted FVs. Power normalization is done by applying the function,

$$g(z) = sign(z)|z|^a,$$

to each component of FVs, where a is a parameter and is set to 1/2 in this work. After normalization, we concatenate them into a single vector. The dimension of our FVs is 262144. In the following sections, we represent visual features as $\boldsymbol{x}^v$.

2.2 Metadata Features

As metadata features, we use Bag-of-Words (BoW) representation, which is based on the idea that each word in a text appears independently. BoW is obtained by counting the occurrence frequency of words in a text. In our method, each EXIF datum is represented as a component of the histogram separately. We use 10 Exif data in xml files as metadata features. Each feature and its dimension are shown in Table 1. We represent metadata features as $\boldsymbol{x}^m$.

Table 1. Metadata Features

EXIF data name	dimension
Make	20
Model	38
Flash	13
SceneCaptureType	4
DateTime	10
GPS Altitude	41
GPS Latitude Ref	2
GPS Latitude	143
GPS Longitude Ref	4
GPS Longitude	151

3 Retrieving Methods

This section describes retrieving methods that we applied in the task. The dataset I for the task is composed of 5,555 images with metadata. Therefore, each image $i \in I$ has visual information and meta information. For a topic t, we represent the query by example(QBE) as q_t, and browsing data as $B_t = \{b_1, b_2, ...\} \in I$ in which b_i is browsed later than b_j $(i > j)$. QBE is an image which is admitted as the similar image for the topic. Browsing data is obtained while seeking QBE. A n-dimensional feature vector of image i is represented as $\boldsymbol{x}_i^v \in \mathbb{R}^n$.

In our approach, we use rankSVM to ascertain the similarity between visual features, and use the distance metric as the similarity between metadata features. Finally, we obtain a similarity score of the document by summing the weighted similarity scores (rankSVM score for visual features + similarity score for metadata features). We use relevance feedback to obtain the weights for each score. Relevance is based on the variance between QBE and browsing data. For a comparison, we also adopt Nearest Neighbor and Support Vector Machine for visual features.

3.1 Nearest Neighbor

We first adopt simple Nearest Neighbor method. For a topic t, $s_{t,i}^v$is a visual similarity score of an image i, and calculated as:

$$s_{t,i}^v = \frac{1}{1 + \min_{\boldsymbol{x}^v \in \{q_t\} \cup B_t} d(\boldsymbol{x}^v, \boldsymbol{x}_i^v)},$$

where $d(\boldsymbol{x}^v, \boldsymbol{x}_i^v)$ is a distance between $\boldsymbol{x}^v$ and $\boldsymbol{x}_i^v$, e.g. a Euclidean distance.

3.2 Support Vector Machine

Support Vector Machine (SVM) is a popular learning method that applies a maximum margin manner for binary classification. $s_{t,i}^v$ is a similarity score of an image i for topic t, and calculated as:

$$s_{t,i}^v = \boldsymbol{w}_t \cdot {\boldsymbol{x}^v}_i - h_t.$$

For topic t, to train a linear model $(\boldsymbol{w}_t, h_t)$, SVM needs training data $D_t = \{(\boldsymbol{x}_i^v, y_i^{v,t})\}_{i=1}^{|I|}$. We choose a label of image i for topic t as:

$$y_i^{v,t} = \begin{cases} 1 & (i \in \{q_t\} \cup B_t) \\ -1 & (\text{otherwise}). \end{cases}$$

Here, SVM is formulated as following optimization problem:

$$\begin{aligned} & \min_{\boldsymbol{w}_t, \{\xi_i\}_{i=1}^{|D_t|}} \frac{1}{2}||\boldsymbol{w}_t||^2 + C\sum_{i=1}^{|D_t|}\xi_i, \\ \text{s.t. } & y_i^{v,t}(\boldsymbol{w}_t \cdot \boldsymbol{x}_i^v - h_t) \geq 1 - \xi_i \quad (i = 1, \cdots, |D_t|), \\ & \xi_i \geq 0. \end{aligned}$$

3.3 RankSVM

To learn a model for each topic from the QBE, browsing data, and other images, we applied rankSVM proposed by T. Joachims[1] for search engine optimization. $s_{t,i}^v$ is a similarity score of an image i for topic t, and calculate as:

$$s_{t,i}^v = \boldsymbol{w}_t \cdot \boldsymbol{x}_i^v.$$

RankSVM is a pairwise learning method that applies a large margin principle used in SVM for ranking learning. The QBE has the strongest presence for the topic in the given image set. The browsing data have stronger presence than images except the QBE and other browsing data. Later browsed images are regarded as higher ranked than earlier ones. P_t is a pair set of topic t, and defined as:

$$\begin{aligned} P_t &= O_1^t \cup O_2^t \cup O_3^t, \\ O_1^t &= \{(i,j) | i = q_t, j \in I - \{q_t\}\}, \\ O_2^t &= \{(i,j) | i \in B_t, j \in I - B_t - \{q_t\}\}, \\ O_3^t &= \{(b_i, b_j) | b_i, b_j \in B_t, i > j\}. \end{aligned}$$

Here, rankSVM is formulated as following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{w}_t, \{\xi_i\}_{i=1}^{|P_t|}} \frac{1}{2}||\boldsymbol{w}_t||^2 + C\sum_{i=1}^{|P_t|}\xi_i, \\ \text{s.t. } \boldsymbol{w}_t \cdot (\boldsymbol{x}_j^v - \boldsymbol{x}_k^v) \geq 1 - \xi_j \quad (\forall (j,k) \in P_t), \\ \xi_i \geq 0. \end{aligned}$$

3.4 Distance Metric between Metadata Features

Similarity between metadata features is calculated by the distance metric of BoW feature vectors as:

$$\begin{aligned} d(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m) &= 1 - e^{-\tau||\boldsymbol{x}_i^m - \boldsymbol{x}_j^m||^2}, \\ c_{i,j} &= \frac{1}{1 + d(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)}, \\ s_{t,i}^m &= \sum_{k \in B_t \cup q_t} c_{k,i}, \end{aligned}$$

where $\boldsymbol{x}_i^m$ and $\boldsymbol{x}_j^m$ are feature vectors $i, j \in I$. $d(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)$ is the distance between two metadata features. In this method, we use 1 as τ. Here, $c_{i,j}$ is a similarity between $\boldsymbol{x}_i^m$ and $\boldsymbol{x}_j^m$. $s_{t,i}^m$ is a similarity score of metadata of image i for topic t, and is obtained by summing the scores with all of the query browsed images. Then we normalize the scores to zero mean and unit variance.

3.5 Relevance Feedback

Relevance feedback is a technique of information retrieval that improves a user's query and facilitates retrieval of information that a user needs. In this task, which feature is important for retrieval varies by topics. Therefore, the system must recognize the criteria for retrieval automatically from the query such as QBE document and browsing data. We calculate the weights of each feature. The final similarity score is the weighted average of the visual similarities and the metadata similarity.

In this method, the weights are calculated utilizing the process of the creation of the query data. Each query image is browsed sequentially. The weights of features are updated by chosen images. For example, the feature of which variance in query images is small should be important. We expand the weight formula as:

$$\begin{aligned} \omega_{l,t}^{new} &= \frac{\sigma_l^I}{\sigma_l^{B_t}}, \\ \omega_{l,t} &= \alpha \times \omega_{l,t}^{new} + (1 - \alpha) \times \omega_{l,t}^{old}, \end{aligned}$$

where α is the updating weight. $\omega_{l,t}$ is the weight used in summing all of the similarity scores. σ_l^I is the variance of images about feature l. $\sigma_l^{B_t}$ is the variance of images in B_t. An initial value of $\omega_{l,t}$ is 1, and as the element count of Q_t increases by one, $\omega_{l,t}$ is recalculated.

Finally, the score of image i is calculated by summing visual feature scores and metadata feature scores which are weighted with relevance feedback. The final score $s_{t,i}^{fin}$ is calculated as:

$$\begin{aligned} s_{t,i}^{fin} &= \sum_{l \in L^v} \gamma_l^v \, \omega_{l,t} \, s_{l,t,i}^v + \sum_{l \in L^m} \gamma_l^m \, \omega_{l,t} \, s_{l,t,i}^m, \\ \gamma_l^v &= \frac{\lambda_l^v}{N_l^v}, \\ \gamma_l^m &= \frac{\lambda_l^m}{N_l^m}, \end{aligned}$$

where $s_{l,t,i}^v$ and $s_{l,t,i}^m$ are the visual and metadata scores of the image i for topic t. L^v and L^m are all of the visual and metadata features respectively. N_l^v and N_l^m represent the number of combinations of the visual features and the metadata features respectively. In this paper, we set $N_l^m = 10$. N_l^v is 1, 2, 3, and 4. λ_l^v and λ_l^m are the parameters to change ratio of the weight between visual and metadata features.

4 Results

This section describes details of the comparison of learning methods and the result of visual feature's combinations.

4.1 Retrieving method comparison

In our experiment, four kinds of local descriptors were extracted from each image: SIFT, C-SIFT, LBP, and GIST. Each descriptor was sampled densely on a regular grids (every six pixels). The dimensionalities of SIFT, C-SIFT, LBP, and GIST were 128, 384, 1024, and 960 respectively. They were coded into two state-of-the-art global feature representations ($4 \times 2 = 8$ visual features in total). One is FV, as explained in the previous section. First, the mixture model of 256 Gaussians was trained using a standard EM-algorithm. To use spatial information, FVs were calculated respectively over 1×1, 2×2, and 3×1 cells. Thereby, we obtained FVs of which the dimensionality was $64 \times 256 \times 8 \times 2 = 262,144$. The other is Locality-constrained Linear Coding (LLC) [3], which describes each local descriptor as a linear weighted sum of a few nearest codewords. In our experiment, 1,024 codewords were generated with the k-means algorithm, and then each local descriptor was approximated using 5-NN of the descriptor. The images were divided into 1×1, 1×3 and 3×1 spatial grids differently from FV. Therefore, the dimensionality was $1024 \times 7 = 7168$.

In this experiment, we examined which learning method is effective to train a ranking function for the task using only visual features. We investigated which visual feature should be combined to achieve the best performance through the next experiment. We use NN, SVM, and rankSVM to train the ranking function. Therefore, we examine $8 \times 3 = 24$ experiments in total. We use trec_eval system published for evaluation. We use a ndcg_cut_100 value in the evaluation of average users. The ndcg_cut_100 value indicates the rate of compliance with the top 100 documents of correct data.

Table 2. Retrieval method comparison.

	NN	SVM	rankSVM
LLC SIFT	0.2946	0.3066	0.3308
LLC C-SIFT	0.2856	0.2967	0.3257
LLC LBP	0.3043	0.3199	0.3385
LLC GIST	0.2796	0.2943	0.3175
FV SIFT	0.3135	0.3278	0.3357
FV C-SIFT	0.3492	0.3486	0.3696
FV LBP	0.3636	0.3363	0.3861
FV GIST	0.3376	0.3145	0.3572

Table 2 shows the trec_eval result of 12 runs. Results show that rankSVM achieves superior performance among the learning methods we use. Therefore, we apply only rankSVM for next experiment of the feature combination.

4.2 Feature combination

This experiment addresses a combination of visual features and metadata features. First, we evaluated only metadata features for retrieval, then we got a score of 0.6228. With only visual features, the best score is 0.3861 using FV, LBP, and rankSVM.

The combination formula is given as:

$$\begin{aligned} s_{t,i}^{fin} &= \sum_{l \in L^v} \gamma_l^v \, \omega_{l,t} \, s_{l,t,i}^v + \sum_{l \in L^m} \gamma_l^m \, \omega_{l,t} \, s_{l,t,i}^m, \\ \gamma_l^v &= \frac{\lambda_l^v}{N_l^v}, \\ \gamma_l^m &= \frac{\lambda_l^m}{N_l^m}, \end{aligned}$$

where λ_l^v and λ_l^m are the parameters to change ratio of the weight between visual and metadata features. We use $\lambda_l^v = 3$ and $\lambda_l^m = 1$.

Finally, we present the top six combinations of four visual features in two circumstances where 10 metadata features are used and not used respectively. We

Table 3. Top combinations of visual features only.

FV SIFT	-	✓	-	✓	-	✓
FV C-SIFT	✓	✓	✓	✓	-	✓
FV LBP	✓	✓	✓	✓	✓	-
FV GIST	✓	✓	-	-	✓	✓
10 Metadata	-	-	-	-	-	-
ndcg_cut_100	0.4236	0.4186	0.4186	0.4118	0.4058	0.4008

Table 4. Top combinations of visual features and metadata features.

FV SIFT	✓	-	-	✓	✓	✓
FV C-SIFT	✓	✓	✓	✓	-	✓
FV LBP	✓	✓	✓	✓	✓	-
FV GIST	✓	✓	-	-	✓	✓
10 Metadata	✓	✓	✓	✓	✓	✓
ndcg_cut_100	0.6998	0.6986	0.6985	0.6983	0.6982	0.6967

obtained the scores shown in Table 3 and Table 4. Without metadata features, the combination of three descriptors (C-SIFT, LBP, and GIST) got the best score among them. Additionally when metadata features were combined, the combination of all of the features (SIFT, C-SIFT, LBP, GIST, and Metadata) got the best score. We could not use this evaluation system before the submission. Therefore, we submitted the score of the experiment which uses four features (C-SIFT, LBP, GIST, and Metadata) we considered the best.

5 Conclusions

This paper explained our approach for ImageCLEF 2013 Personal Photo Retrieval Task. Results of comparative experiments show FVs as useful for visual representation and BoWs as metadata features. Retrieval is performed using rankSVM with visual features and the distance metric between metadata features. Using combinations of FVs from various local descriptors (C-SIFT, LBP, and GIST) and metadata features, we achieved the top score among all teams.

References

1. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM SIGKDD. pp. 133–142 (2002)
2. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV, pp. 143–156 (2010)
3. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR. pp. 3360–3367 (2010)