Regular Paper

A Robust Clustering Method for Missing Metadata in Image Search Results

Masaharu Hirota[1,†1,a)] Naoki Fukuta[2] Shohei Yokoyama[2]
Hiroshi Ishikawa[2]

Received: September 16, 2011, Accepted: February 3, 2012

Abstract: Although metadata are useful to obtain better clustering results on image clustering, some images do not have social tags or metadata about photo-taking conditions. In this paper, we propose an image clustering method that is robust for those missing metadata of photo images that appear in search results on the Web. The method has an integrated estimation mechanism for missing social tags or photo-taking conditions from other images in the image search result. An advantage of our method is that our approach does not require another training set that is constructed from other images that are not included in the search result. We demonstrate that the proposed method can effectively cluster images which have some missing metadata by showing the performance of on-demand clustering on a photo sharing site.

Keywords: image search, image clustering, metadata estimation

1. Introduction

Image data clustering is an approach for effectively browsing a lot of images without any prior supervision of partitioning [1], [2]. There is strong demand for searching and browsing a large amount of image data via social media sites such as Flickr [3]. Therefore, a better clustering method that enables users to effectively search and browse images is required [1]. There are a lot of keyword-based image search services, and in most cases, the image search results are displayed as a ranked list structure. Keyword-based search has been established as the predominant method for discovering information on the Web [4]. This ranking functionality reflects the similarities of the metadata and the query, according to text-based retrieval models. However, it is difficult to preserve visual diversity when such models are used [1], [3]. The ranked list only considers the relevance to the query words, and therefore it may induce similar near the top rank of it. The image search results to a user should satisfy both diversity and high precision [5]. To overcome this problem, a clustering-based approach is used for presenting diverse results in searching images by keywords.

The necessity of a clustering-based approach is, for example, that it may be difficult for a user to prepare appropriate queries for keyword-based image search engines [6]. The user-generated query may include several aspects to be matched to diverse images [1]. Most users don't know how to express those query terms explicitly and in more general terms that match a number of images [5]. Indeed, the query length is typically very short. In Ref. [7], it was reported that approximate 90% of queries their length $l(q) \leq 4$. As a result, the query is often ambiguous [8]. Therefore, the result may contain a lot of different images that the user does not expect [9]. As a result, it is hard for the user to browse them.

On clustering a lot of images, we may have to consider the problem of a semantic *semantic gap* in content-based image retrieval (CBIR). Smeulders et al. defined the semantic gap as follows: "The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation the same data have for a user in a given situation" [10]. They also argued that an important issue in CBIR is filling this semantic gap. For example, a red flower may be regarded as the same as a rising sun, and a fish the same as an airplane [11].

The first issue is that we should consider the case that some images in the image in image search results do not have metadata. In social media site like Flickr, users can freely add tags to images. Therefore, uses can also leave the images untagged, so some images do not have tags. Also, some metadata are added by a digital camera, but these but their values are varied of them are varied, depending on the model of devices. In such case, it is difficult to calculate the similarities between images based on metadata. For example, **Fig. 1** shows the image search result about "polar bear." In this search result, some images do not have tags (In this figure, we describe "no tags" for the images that do not have tags). This search results has two different kinds of photos, "stuffed doll" and "real animal." In such case, the search result might be clustered by images which have metadata and the images which don't. Furthermore, as shown here, it is difficult to distinguish images based

[1] Graduate School of Informatics, Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
[2] Faculty of Informatics, Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
[†1] Presently with Graduate School of Science and Technology, Shizuoka University
[a)] dgs11538@s.inf.shizuoka.ac.jp

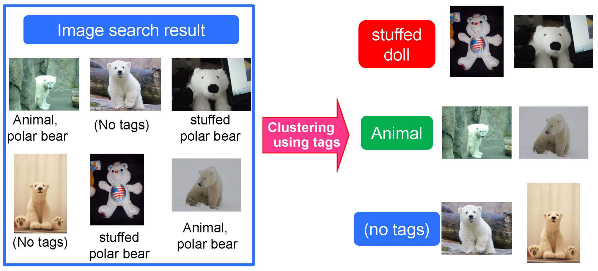

Fig. 1 Example of clustering the image search result that contains the images that do not have tags.

on their meanings when such key metadata has been lost. As a result, the quality of image clustering results might be degraded and it might provide a poor visual diversity. To solve these issues, an approach to estimate the missing metadata is necessary for the purpose of better image search result clustering.

The other issue is that the tags of images in the image search results are often images in image search results often imprecise and/or incomplete [12]. Also, the use of user-provided tags added to images has some drawbacks [6]. Lexical variability of a concept (e.g., "Louvre," "Le Louvre," "Louvre Museum," etc.), and the large amount of highly specific tags are among the most important disadvantages. The others are irrelevant tags to images, the difference of the language, and emerging words that that were not used before [13].

As a result, since some images in image search results do not have tags, or sufficient metadata, the quality of clustering has degraded and is poor visual diversity. Therefore, we estimate the missing metadata and aim to present the better image search results that have sufficient visual diversity to users.

When estimating tags of image search result that are obtained from a social media site, preparing the training data set has some limitations. A limitation is that preparing an appropriate training data set that can be applied to the images that has the emerging tags about new events is difficult. Another limitation is that it is difficult to prepare an appropriate training data set that can cover all queries. The training data set are cases due to the model drivenness [14]. This is caused by the user-provided queries that often are very diverse and ambiguous. Therefore, it is demanded to realize an estimation method that does not require any training data set in advance for estimating of image tags.

2. Preliminary

2.1 Ranked List

In many cases, a query for image search is represented as one or several keywords related to the search result, and its result is often presented as a ranked list of the resulting images. Many modern IR systems employ their own ranking approaches. These approaches mainly consider the relevance of each item to the query but ignore the content of other items ranked in the search results. The IR systems implementing this approach are mostly appropriate when the relevant documents are very few and high-recall is required. However, this approach has some issues to be resolved (e.g., Refs. [15], [16]).

One major issue is that the presentation method may drop good visual diversity in the result. For instance, when a specific type or brand of car is requested by the query, it may very well be that the top n of the ranking was shared by the same image that was released by a single source (e.g., the marketing division of the company). Furthermore, it may be difficult for a user to type in the exact query that is most relevant for the retrieval. Improving the ranked list will not be the solution for this problem. Therefore, cluster-based approaches have been proposed for presenting a lot of relevant images obtained by the query [17].

2.2 Diversity in Image Search

Some methods have been proposed to produce visually diversified image search results.

Zwol et al. [18] proposed a retrieval model which provides diverse results as a property of the model itself, rather than in a post-retrieval step. The models operating only on tags offers the highest level of diversity with no significant decrease in precision.

Leuken et al. [1] proposed clustering methods that are called Folding, Maxmin, and reciprocal election. These methods are for presenting a clustering result that maintains image diversity and representative images for each cluster. However, these methods only use low-level image features so they don't consider the meanings of images that can be figured out with supplemental metadata.

To overcome this issue, Yang et al. [19] proposed both supervised concept-based search and unsupervised search reranking to satisfy both conceptual and visual diversity of image search results. However, this approach suffers from scalability issues in a concept-based search mechanism. In this paper, we adopt a cluster-based method to satisfy the visual and semantic diversity of image search results with high scalability.

We propose a constrained agglomerative clustering method with must-link constraints for better clustering results using a multiple similarity metric [20]. Our approach tries to overcome these drawbacks by using constrained agglomerative clustering with must-link constraints.

Two measurements similar in image appearance, tags, and photo-taking conditions (e.g., tags and image appearances) are used for preparing must-link constraints. The other similarity measurement (e.g., photo-taking conditions) is used for clustering. The must-link constraint is calculated by:

$$sim(a, b) = T(a, b) + I(a, b) \tag{1}$$

where $T(a, b)$ and $I(a, b)$ are the tags and image similarities between two images a and b. Must-link constraints $(M_1, M_2, ..., M_r)$ are made based on these values in ascending order. In constrained agglomerative clustering with must-link constraints, any two images that have must-link constraint must belong to the same cluster. These clusters are agglomerated. The method uses group average method computing for the distance between each cluster. The detail of this method was reported in Ref. [17]. However, the method implicitly assumes that all images have sufficient metadata.

2.3 Metadata Estimation

Although metadata are useful to obtain better clustering results on image clustering, some images do not have social tags or metadata about photo-taking conditions. Therefore, on applying clustering methods to image search result, keeping the quality of the image clustering is an important issue. To this end, it is necessary to estimate or refine the tags of images in social media sites.

Some methods have been proposed to estimate metadata using low-level image features. Zhu et al. [12] proposed the metadata refinement method on the images in social media sites, which utilizes an approach where the tag refinement problem is formulated the tag refinement problem is formulated as a decomposition of the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix. Then, the optimality is measured by four aspects: low-rank, content consistency, tag correlation and error sparsity.

Lee et al. [21] proposed the metadata estimation method, which uses the two categories based on folksonomy. This method considers the aspects of tags that are objective tags and subjective tags.

Yang et al. [14] proposed the metadata estimation method using a weighted association rule and near-duplicate clusters. This approach firstly initializes the candidate tag set from its near-duplicate cluster's document. Then, the candidate tag set is expanded by considering the implicit multi-tag associations mined from all the clusters' images, where each cluster's document is regarded as a transaction. To further reduce noisy tags, a visual similarity is also computed for each candidate tag to the test image based on new tags. Tags with very low scores can be removed from the final tag set.

These approaches can be applied to the large image data set. However, in our case, there is fewer number of image search results than the image data set for training data set. Therefore, it is difficult to apply these approaches to image search results.

2.4 Metadata for Photo-taking Condition

The Exif (Exchangeable image file format) is a commonly used metadata format for representing photo-taking conditions. The metadata is automatically generated when the picture is taken by using a typical digital camera. This metadata contains ISO speed, Aperture, Time stamp, etc.

Boutell et al. proposed a method that uses Exif metadata for sunset detection and indoor-outdoor classification [22]. The method uses a Bayesian network based on low-level image similarity and Exif. However, this method may be applied for a specific classification problem but not for a generic clustering problem.

Sinha et al. [23] proposed deriving useful semantics about the digital photo. Also, they compare its results with classical relevance models used for automatic photo annotation.

When some images do not have tags or the tags are incorrect, the estimating of the metadata based on the tag co-occurrence or distribution may be difficult. Since Exif is another metadata that is annotated by different aspects from tags, Exif could be helpful to estimate features.

3. Metadata Estimation Using Similar Images in the Search Result

We describe the proposed method about metadata estimation for missing metadata of images to be used in the clustering process proposed in this section. Since our proposed method should be able to be applied to queries in any domains, the proposed metadata estimation our approach does not require another training set that is constructed from other images that are not included in the search result. Instead, our approach uses similar images in the image search result. We estimate social tags or photo-taking conditions from other images in the image search result. We applied the proposed constrained agglomerative clustering method to obtain better clustering results using multiple similarities based on estimated metadata [17]. We demonstrate that the proposed method effectively estimates missing metadata by showing the performance of on-demand clustering on a photo sharing site.

3.1 Our Proposed Metadata Estimation Approach

Our approach estimates the metadata of an image p_i in an image search result $P = \{p_0, p_1, ..., p_n\}$ using similarities. In this paper, we estimate two kinds of metadata: tag t_i, Exif e_i that the image p_i has.

For example, our approach is applied to an image p_i that does not have tags $|t_i| = 0$), to estimate the correct tags t_i' using images p_x that are similar to image p_i. This approach annotates all tags t_x and exif data e_x to an image p_i as estimated tags and exif. Potentially might wrongly add metadata that affects the performance of clustering. Therefore, we also prepared an algorithm to try to remove such noisy tags. However, in the later section, we will present that our algorithm does not need such *noisy tag removal* process.

Figure 2 shows the algorithm to estimate metadata. The details of our proposed algorithm for metadata estimation are as follows. To estimate the appropriate metadata using similar images, we apply k-nearest neighbor method to the image p_x in an image search result P that have missing metadata. For example, when an image p_x does not have tags ($|t_x| = 0$), we estimate appropriate tags using similar image. Therefore, we use an image p_x as a test image and apply k-nearest neighbor method using an image search result P without p_x to estimate metadata. The image search result that contains the metadata estimated images is denoted by P'.

Figure 3 shows a possible algorithm for removing noisy metadata. In the algorithm, it clusters the images that have each tag $T = \{w_0, w_1, ..., w_l\}$ in image search result P' and denotes the clustering results as $C = \{C_0, C_1, ..., C_w\}$, respectively. Here, since

```
Algorithm 1 Algorithm for Metadata estimation
INPUT: Image search result $P$
for all $p_i \in P$ do
    if $|t_i| = 0$ then
        $t_i \leftarrow kNN(p_i, P)$
    end if
    if $|e_i| = 0$ then
        $e_i \leftarrow kNN(p_i, P)$
    end if
end for
```

Fig. 2 Algorithm for metadata estimation.

Algorithm 2 Algorithm for removing noisy metadata
INPUT: Image search result P, thresholds α, β

```
for all $Image\ sub\ set\ P'\ has\ w_i \in T$ do
    $Clustering\ result\ C \leftarrow Clustering(P')$
    for all $Cluster\ c_i \in C$ do
        if $|c_i| < \alpha * |P'|$ then
            for all $p_i \in c_i$ do
                $delete\ w_i\ in\ p_i$
            end for
        end if
        if $|c_i| < \beta * |P'|$ then
            for all $p_i \in c_i$ do
                if $e_i\ is\ null$ then
                    $e_i \leftarrow ave(e_i \in c)$
                end if
            end for
        end if
    end for
end for
```

Fig. 3 Algorithm for removing potentially noisy metadata.

$T = \{w_0, w_1, ..., w_l\}$ is all tags w_a that appear in an image search result P', $|T|$ is the number of tags in image search result P'. Here, the algorithm uses the clustering algorithm for image search results which do not need to specify the number of clusters, known as Maxmin [1]. When the number of clusters $|c_i|$ in a clustering result C_y about the tag w_a is less than the thresholds α, then the algorithm removes the tag w_a from the images in a cluster c_i. This algorithm tries to remove the incorrect tags or unneeded tags that are wrongly added by the proposed estimation method from images. In similar way, when the number of cluster $|c_i|$ in a clustering result C_y about the tag w_a less than the thresholds β, then the algorithm tries to refine exif value e_a by using the average of exif value in cluster c_i. This approach can add the appropriate exif value to images based on the similar images. In this paper, the threshold α is 0.25 and β is 0.5.

Here, we briefly describe the clustering algorithm Maxmin [1]. Maxmin tries to get visually diverse representatives as much as possible. To achieve this, it uses a maxmin heuristic on the distances between cluster representatives. First, the first representative R_1 image is randomly chosen in the image search result. The second representative image R_2 is the image with the largest distance from the R_1. For each following representative, the image is selected that has the largest minimum distance to all the other selected representatives. This process is continued until this maximum minimal distance is smaller than a threshold. The threshold is defined as the mean distance of all images from the average image. Each image is assigned to the closest representative.

3.2 Image Similarity

We calculate the similarity between two images to consider image appearances, using the same method appearing in Leuken et al. [1]. We use six low-level image features: Color histogram [24], Color layout [25], Scalable color [25], CEDD [26], Edge histogram [25], and Tamura [27].

The image similarity (I) between two images a and b is calculated by:

$$I(a, b) = \frac{1}{f} \sum_{i=1}^{f} \frac{1}{\sigma_i^2} d_i(a, b) \tag{2}$$

where f is the total number of features, $d_i(a, b)$ is the similarity between a and b in terms of the i-th feature and σ_i^2 is the variance of all image similarities according to the i-th feature within this set of image search results.

3.3 Photo-taking Condition Similarity

We calculate the similarity between two images to consider photo-taking conditions defined in Exif based on Ref. [23]. Sinha et al. [23] proposed a metric which quantifies the ambient light in an image. We use four Exif metadata: ISO speed, Exposure Time, Aperture, and Focal Length. ISO speed is the sensitivity of a film recording light. When the value is higher, the more sensitive picture elements are in the picture. Exposure Time is the time that a film is exposed. The higher this value is, the slower the shutter speed was. Aperture is the amount of light that passes through the camera lens. The higher this value is, the lower the amount of light to the camera lens was. Focal length is the distance between lens and picture elements. The higher this value is, the longer the focal point was. The photo-taking condition feature is calculated using LogLight Metric [23] by:

$$LogLightMetric = lg(K * ET * A * ISO/FL^2) \tag{3}$$

where K is the proportionality constant, ET is the Exposure time, A is the Aperture, ISO is ISO speed rating and FL is the Focal length. LogLight Metric will have a small value when the ambient light is high (the camera will have a low exposure time, small aperture and low ISO). Similarly it will have a large value if the outdoor light is small.

The photo-taking condition similarity is calculated using L1-norm.

$$exif(a, b) = |LLM_a - LLM_b| \tag{4}$$

where LLM_a and LLM_b are the LogLight Metric of image a and b.

3.4 Tag Similarity

The method calculates the similarity between two images to consider the image semantics by using tags. To consider the tag's significance, we calculate the idf_i of each tag by:

$$idf_i = \log \frac{N}{n_i} \tag{5}$$

where N is the total number of images and n_i is the number of images that have i-th tag in all images. We calculate the tag vector $\vec{a}$ of an image a using this idf value for each elements. The tag vector $\vec{a} = \{idf_{1a}, idf_{2a}, ..., idf_{na}\}$ of an image a is calculated by:

$$idf_{ia} = \begin{cases} idf_i & w_i \in t_a \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

We calculate a cosine similarity for tag similarity (Tag) by:

$$Tag(a, b) = \frac{\vec{a} \cdot \vec{b}}{|a||b|} \tag{7}$$

where $\vec{a} = \{idf_{1a}, idf_{2a}, ..., idf_{na}\}$ and $\vec{b} = \{idf_{1b}, idf_{2b}, ..., idf_{nb}\}$ are the tag vectors of the two images of two images a and b.

4. Evaluation

In this section, we evaluate the performance of clustering image search results using our proposed estimation method. We use the ground truth by assessors, for evaluating our proposed method. In this experiment, we use two evaluation criteria: Fowlkes-Mallows index [28] and variation of information [29].

4.1 Evaluation Criteria

Comparing two clustering results on the same data set is an important research issue itself, thus many different measures have been proposed. We adopt two clustering comparison measures that reflect different properties. We use two evaluation criteria, Fowlkes-Mallows index and variation of information. We describe them briefly below.

The Fowlkes-Mallows index [28] is a measurement based on counting pairs. Given a result set I and two clustering C and C', all possible image pairs based on I are divided over the **Table 1**. This comparison can be seen as the precision and recall in clustering. A high score on the Fowlkes-Mallows index indicates that the two clustering are similar. The precision and recall for using the Fowlkes-Mallows index are calculated by:

$$W_I(C, C') = \frac{N_{11}}{N_{11} + N_{01}} \tag{8}$$

$$W_{II}(C, C') = \frac{N_{11}}{N_{11} + N_{10}} \tag{9}$$

The Fowlkes-Mallows index is the geometric mean of these two, making it a symmetric criterion by:

$$FM(C, C') = \sqrt{W_I(C, C')W_{II}(C, C')} \tag{10}$$

We use variation of information (VI), which is a theoretically based an information theoretic measure [29]. This is calculated using mutual information and entropy. For calculating the entropy, we calculate the probability that an image belongs to cluster k by:

$$P(k) = \frac{n_k}{n} \tag{11}$$

where n_k is the total number of images contained in the clustering result C. We calculate the entropy $H(C)$ about a clustering result C by:

$$H(C) = -\sum_{k=1}^{K} P(k) log P(k) \tag{12}$$

Next, we calculate the mutual information between two clustering results C and C'. Therefore, we calculate the probability that a randomly selected image belongs to cluster k in a clustering result C and cluster k' in a clustering result C' by:

$$P(k, k') = \frac{|C_k \cap C'_k|}{n} \tag{13}$$

Table 1 Classes of image pairs in Fowlkes-Mallows index.

	Image a belongs to C	Image a belongs to C'
Image b belongs to C	N_{11}	N_{01}
Image b belongs to C'	N_{10}	N_{00}

Then, the mutual information $I(C, C')$ is defined by:

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P'(k')} \tag{14}$$

Variation of information $VI(C, C')$ is calculated based on these expressions by:

$$VI(C, C') = [H(C) - I(C, C')] + [H(C') - I(C, C')] \tag{15}$$

The variation of information coefficient focuses on the relationship between a point and its cluster. It measures the difference in this relationship between the two clusterings and averaged overall points. Hence, a low variation of information score indicates that two clusterings are similar.

4.2 Evaluation of Our Proposed Method

In this experiment, there are three reasons why we evaluate performance using clustering results from estimated metadata rather than using the accuracy of metadata estimation directly. First, in this paper, the goal of our research is presenting better image search results to users. Therefore, we should confirm that metadata estimation using our proposed method contributes the quality improvement on clustering image search results. Second, when our proposed method is used for estimating exif values, since exif is linear value, it is difficult to make a meaningful evaluation to the estimated exif values directly. Therefore, we evaluate the performance of clustering image search result based on estimated exif data using both our proposed metadata estimation method and clustering method. Finally, increasing the accuracy of tag estimation does not always contribute to the quality of clustering results. Image sets in this experiment contain 5771 kinds of tags and the frequency of appearance about 70 percent in these tags is once. **Figure 4** shows the distribution of the tags frequency in the images. It is difficult to estimate these tags without using a training data set. Also, if an estimation method using training data set to estimate these tags is applied, it may be difficult to estimate the tags whose frequency of appearances are extremely low. However, the tags that the appearance in image search result is once are often about camera maker, time, the words by language other than English, and the other noisy tags. Therefore, the estimation of the tags whose frequency of appearances is extremely low does not contribute to improving the quality of clustering results.

To prepare a ground truth of clustering, we asked assessors to make 30 queries and to prepare clustered images to fit their natural feelings. **Figure 5** is 30 ambiguous search queries for obtaining the image search results from Flickr. In the experiments,

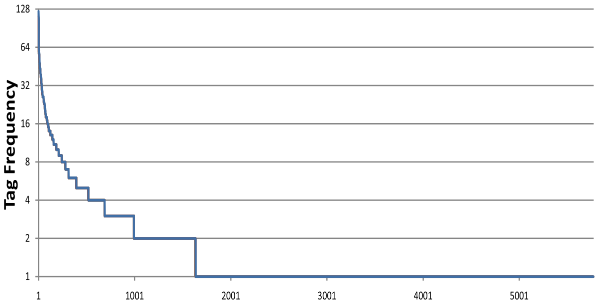

Fig. 4 Tag frequency in the image search result for this experiment.

apple
arashi
bicycle
building
cardgame
Christmas
computer
disaster
earthquake
eel

final fantasy
Gundam
Hamamatsu
hamster
hydrangea
Jellyfish
kyoto
landscape
Mercedes
Mount Fuji

Mozart
Muffler
nuclear plant
Sagrada Familia
sandwich
stationery
tea
tiger
Ultraman
windows

Fig. 5 Query terms used in experiments.

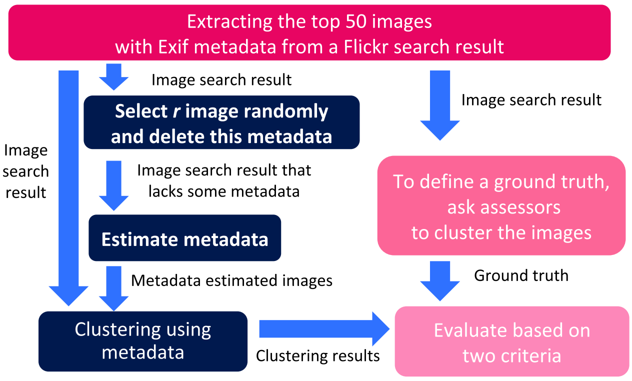

Fig. 6 Experimental procedure on the evaluation of our metadata estimation method.

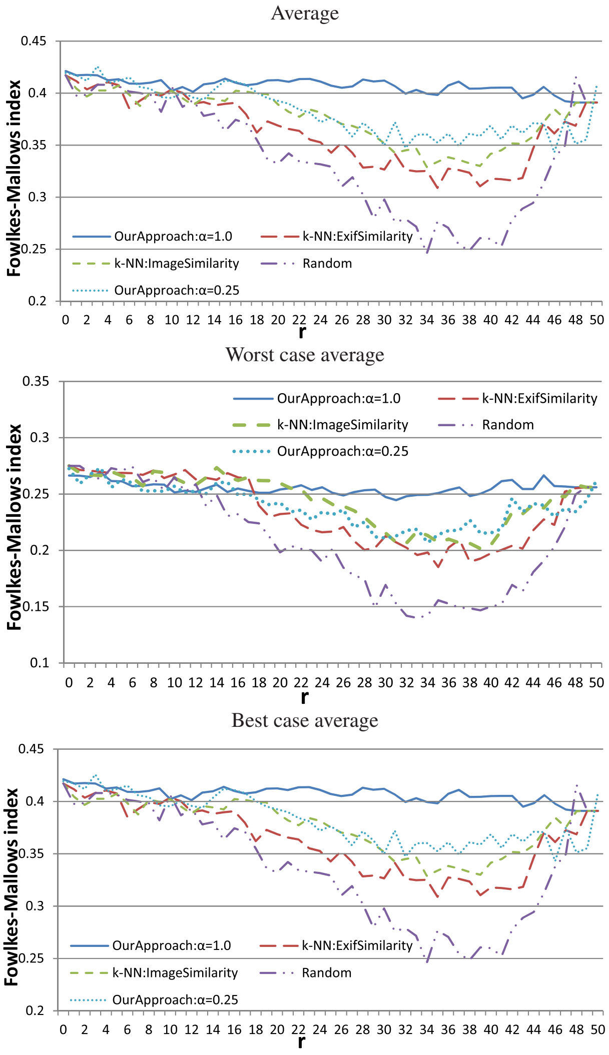

Fig. 7 Performance: Tag (Fowlkes-Mallows index).

2 assessors make each ground truth on each query. Then, we get the ranked list of an image search result and choose the top 50 images that have non empty data for all metadata of user-provided tags and 4 metadata of exif: ISO speed, Exposure Time, Aperture, and Focal Length from the list as experimental data. As a result, we use 30 image search results for the evaluation of our proposed method. Therefore, we obtain 60 ground truths for the evaluation of clustering results. We evaluate the clustering results by two evaluation criteria.

In this experiment, we use the all images in image search results that have the metadata of tags and exif as the data for the experiment. Also, the *k* for k-nearest neighbor method is 5 and the image similarity are used for metadata estimation. For evaluation of robustness for missing metadata, we remove the existing metadata on some images in the data and our proposed method is then applied to these data for estimation of removed metadata. **Figure 6** shows this experimental procedure.

The target metadata to be estimated are missing tags and four exif metadata: ISO speed, Exposure Time, Aperture and Focal Length. The number of the metadata to be estimated r is $r = 0, 1, 2...50$.

In this experiment, after the metadata estimation, we apply our constrained agglomerative clustering method proposed in Ref. [17]. The preparation of must-link constraints for that clustering method is based on two similarities in image, tag, and photo-taking condition similarities. Then the clustering method based on the other similarity is applied to images containing constrained images. In this experiment, the number of must-link constraint is 160, which is the same value used in Ref. [17]. Also, in the clustering, the used similarities for making constraints are image and tag similarities. These parameters are decided by the preliminary experiments in Ref. [17]. Then, we compare the clustering result to one of the k-nearest neighbor method as base line to evaluate the performance of our proposed method. To consider the differences among assessors' natural feeling, we have prepared two ground truths for each clustering problem and we will also show the best and worst case results to ground truths.

4.3 Performance on Data with Missing Tags

Figures 7 and **8** show the results of clustering performance with missing tags on Fowlkes-Mallows index and variation of information. In these figures, the average value and the range of performance in two ground truths (i.e., the top value shows the best case and the bottom value shows the worst case) are shown. Note that, in Fowlkes-Mallows index, a higher value is better. However, in variation of information, the lower value is better in the performance. The $r = 0$ means that the value of the clustering result is based on the using original tag metadata (i.e., using the tags that are not removed or estimated). These values present the average value of the clustering result to image search results using estimated tags. These tags are estimated by five approaches. The first approach is k-nearest neighbor method by exif similarity. The second approach is k-nearest neighbor method

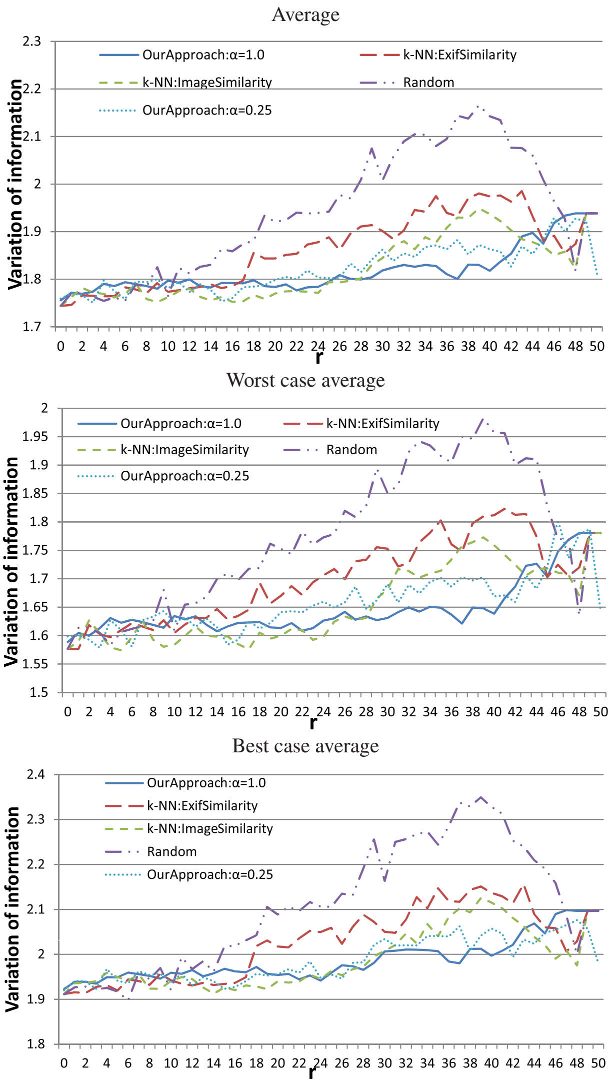

Fig. 8 Performance: Tag (variation of information).

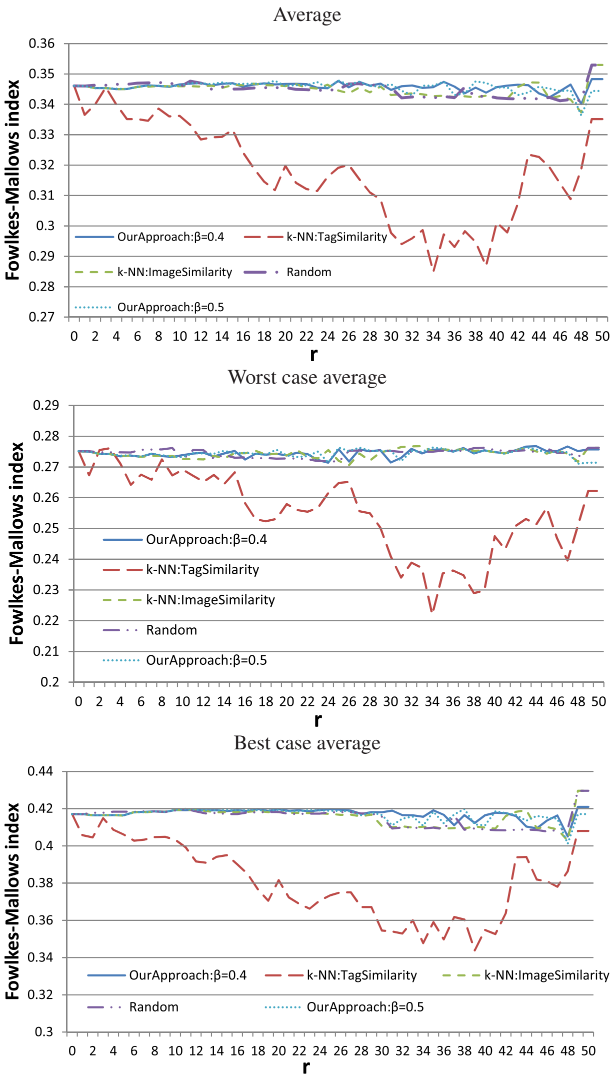

Fig. 9 Performance: Exif (Fowlkes-Mallows index).

by image similarity. The third approach is our approach that is k-nearest neighbor by image similarity with noisy tag removal ($\alpha = 1.0$). This *alpha* is the best average score that is ranked by Fowlkes-Mallows index and variation of information in our preliminarily experiments. The fourth approach is our approach that is k-nearest neighbor by image similarity with noisy tag removal ($\alpha = 0.25$). The other is the missing tag that is replaced by the values that are chosen from other images randomly.

On estimating tags, our proposed estimation method which uses image similarity with noisy tag removal shows the best performance in almost all cases. However, the method that removed some metadata in noisy tag removal produced slightly worse performance in this case. This means that our approach might wrongly remove some necessary tags by the noisy metadata removal. We can say our proposed method totally estimated the appropriate tags and it contributes the improvement of the clustering result in most cases.

Note that the best result is obtained in the case when the number of images that their metadata should be estimated is 0. This shows the estimation approach is helpful but it does not provide a better result than the case that all correct metadata are available.

In Figs. 7 and 8, when the number of the metadata to be estimated is low (e.g., r is about 0...10), our approach is worse than the others. In this experiment, we used the two α values, 0.25, and 1.0, the first one is the value we predicted best before the experiment, and the other one is the best value for obtaining the highest average value through the experiment in the range of $0 \leq r < 50$. The best value of α for each query depends on the resulting images.

Also, when we estimate tags using our approach, the accuracy of estimated tags to original tags is not so large. However, the rate of estimated wrong tags is about 2 percent or less. Therefore, our approach can estimate the tags that are useful for clustering.

4.4 Performance on Data with Missing Exif

Figures 9 and **10** show the results of clustering performance with missing exif on Fowlkes-Mallows index and variation of information. In these figures, the average value and the range of performance in two ground truth (i.e., the top value shows the best case and the bottom value shows the worst case) are shown. As mentioned in the previous part, in Fowlkes-Mallows index, a higher value is better. However, in variation of information, the lower value is better in the performance. The $r = 0$ means the value of the clustering result using true exif metadata (i.e., using

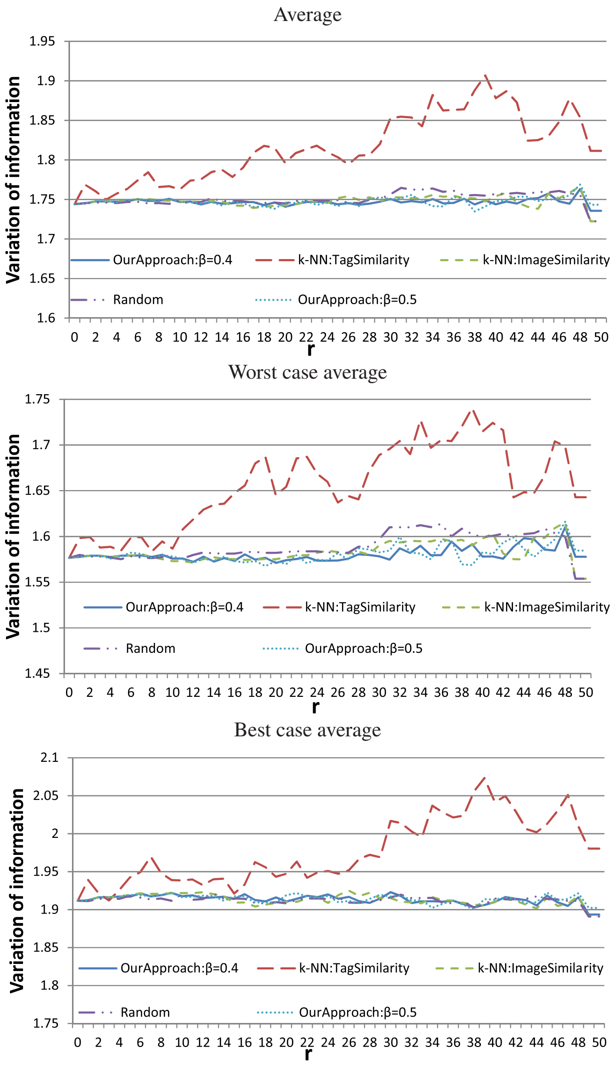

Fig. 10 Performance: Exif (variation of information).

the exif that are not removed and estimated). These values present the average value of the clustering result to image search results using estimated exif. These tags are estimated five approaches. The first approach is k-nearest neighbor method by tag similarity. The second approach is k-nearest neighbor method by image similarity. The third approach is our approach that is k-nearest neighbor by image similarity with noisy exif removal ($\beta = 0.4$). This β is the best average score that is ranked by Fowlkes-Mallows index and variation of information in the preliminarily experiments. The fourth approach is our approach that is k-nearest neighbor by image similarity with noisy exif removal ($\beta = 0.5$). The other is the missing exif that is replaced by the values that are chosen from other images randomly.

On estimating exif, our proposed estimation method outperformed other methods in almost cases. Especially, in $r = 30$ or more, our proposed estimation method outperformed other methods. This means that when a lot of images without exif metadata are clustered, our proposed method works better than the other approaches.

Also, in Figs. 9 and 10, the difference between our approach and the others are very small. Furthermore, the values are not deeply depended on the value of β. The accuracy with exif estimation using our approach depends on an image data set, since

Table 2 Ranking: Exif (Fowlkes-Mallows index).

Approach	Average Ranking (FM)
OurApproach: $\beta = 0.4$	1.922
OurApproach: $\beta = 0.5$	2.039
k-NN: TagSimilarity	4.863
k-NN: ImageSimilarity	2.902
Random	2.804

Table 3 Ranking: Exif (variation of information).

Approach	Average Ranking (VI)
OurApproach: $\beta = 0.4$	1.961
OurApproach: $\beta = 0.5$	2.176
k-NN: TagSimilarity	4.922
k-NN: ImageSimilarity	2.490
Random	2.980

exif similarities are not always helpful for better clustering. Also, our clustering algorithm should have appropriate parameters to obtain the best result in each case. The reason why the differences of the results in our approach and in "random" were small could be derived from the nature of exif metadata. The metadata used in this paper consists of a few attributes. These values are continuous values (e.g., focal length) or discrete values within limited ranges (e.g., ISO speed). When the missing exif data were just filled by the values that are chosen from other randomly selected images, a close value from the original one could often be chosen as the pseudo estimated value. Therefore, sometimes the results on randomly estimated values may perform well. In comparison on clustering methods, the results on constraint agglomerative clustering outperformed the ones on ordinary agglomerative clustering on both estimation cases. Here, **Tables 2** and **3** show the average ranking value on Fowlkes-Mallows index and variation of information. The ranking value is the value of rank within five methods, e.g., when the value was best in the methods, the value was 1, regardless of the difference to others. The average ranking value is the average of the ranking value in the whole range of possible r. In those tables, our proposed method is best. Therefore, on average, our approach outperforms the others.

Also, in some cases, the clustering results using estimated metadata provide better performance than true exif metadata. This shows that the estimated metadata is more useful for clustering than the true exif metadata. In this case, the noisy tag removal algorithm recalculates estimated exif data to refine the estimated exif data. The effect of containing missing exif data has been reduced by using our proposed method considering multiple similarities.

4.5 Effects of the Choice of Clustering Algorithm and Metadata Estimation

In the experiments we presented in Sections 4.3 and 4.4, our proposed constrained clustering approach has been used in all cases. There we can see differences in performance among the methods. In this section, we demonstrate how the metadata estimation performance affects the overall performance when different clustering methods have been used. In **Figs. 11**, **12**, **13**, and **14**, we show the performance of the number of estimated tags or exif based on Fowlkes-Mallows index and variation of infor-

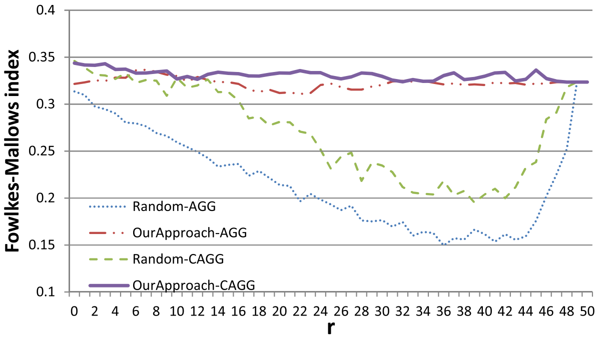

Fig. 11 Performance: Tag (Fowlkes-Mallows index).

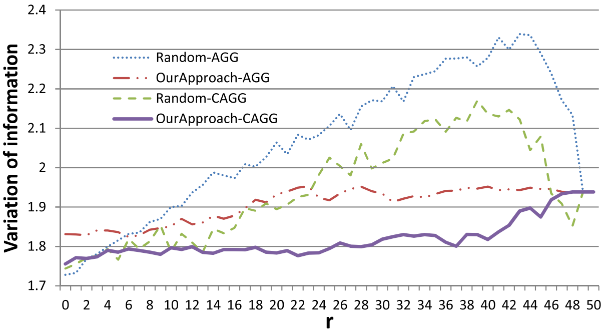

Fig. 12 Performance: Tag (variation of information).

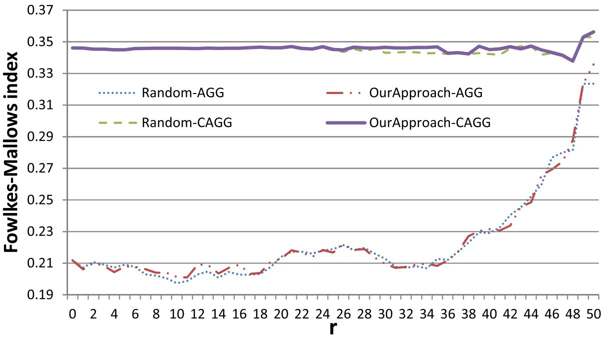

Fig. 13 Performance: Exif (Fowlkes-Mallows index).

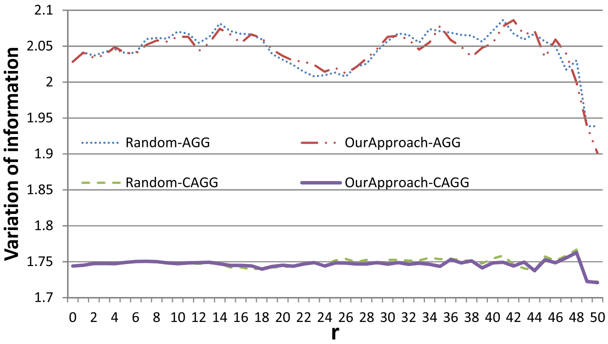

Fig. 14 Performance: Exif (variation of information).

mation. Here, we denote each result by combining its metadata estimation method and its clustering algorithm. On the part of metadata estimation method, "Random" means the result without metadata estimation. Instead, on "Random", the missing metadata are replaced by the values that are chosen from other images randomly. "OurApproach" means the result with metadata estimation using our proposed estimation method (α is 1.0 and β is 0.4). On the clustering algorithm part, "AGG" means the result on ordinary agglomerative clustering, and "CAGG" means the result on constrained agglomerative clustering with must-link constraints. In Figs. 11 and 12, our proposed estimation method outperformed the clustering results in most cases. In the metadata method, our proposed estimation method outperformed the clustering results using both clustering methods. This means that the use of our proposed estimation method is useful for clustering image search result on both clustering algorithm. Also, in the clustering method, the constrained agglomerative clustering outperformed the clustering result using both estimation methods. This means that the use of constraints is useful for clustering image search result regardless of the use of tag estimation, at least on this experimental setting. In Figs. 13 and 14, the results on our proposed estimation method are sometimes slightly better. This might be caused by the nature of exif metadata. These values are continuous values (e.g., focal length) or discrete values within limited ranges (e.g., ISO speed). When the missing exif data were just filled by the values that are chosen from other randomly selected images, a close value from the original one could often be chosen as the pseudo estimated value. Therefore, sometimes the results on randomly estimated values may perform well. In the clustering method, the constraints agglomerative clustering outperformed the clustering result using both estimation methods.

5. Conclusion

We proposed the metadata estimation method for clustering image search results to provide users with better search results. Our proposed method estimates the metadata in image search results that do not have tags, without using a training data set to cover diverse and ambiguous queries. Rather, we use the similar image to estimate metadata. We demonstrate the performance of our proposed method comparing to base line methods. We evaluated our proposed method based on Fowlkes-Mallows index and variation of information. On estimating tags, in many cases, our approach showed better performance in both the Fowlkes-Mallows index and variation of information. In contrast, the observed performance differences on exif estimation were very small. We left to improve the performance in exif estimation as a future work.

In this paper, our proposed method uses low-level image, tag and photo-taking condition features. We can use the spatio-temporal information for estimating metadata such as the GIS and time stamp. Especially, using GIS information will help the algorithm to find the correlation between them and the usage of tags or deciding the place where the user took the photo. As a result, we can estimate the missing metadata and refine the incorrect metadata based on more rich information. For example, Hays et al. [30] proposed the method for estimating of geographic information from a single image using similar images. Also, Crandall et al. [31] proposed the method for mapping a large collection of geotagged photos to a world map. We can use these data as training data set for estimation in where the photo was taken. Extending the approach to effective use those data is a topic for future work.

References

[1] van Leuken, R.H., Pueyo, L.G., Olivares, X. and van Zwol, R.: Visual diversification of image search results, *International World Wide Web Conference*, pp.341–350 (2009).

[2] Datta, R., Li, J. and Wang, J.: Content-based image retrieval: Approaches and trends of the new age, *Proc. 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, November*, pp.10–11 (2005).

[3] van Zwol, R. and Sigurbjornsson, B.: Faceted exploration of image search results, *Proc. 19th International Conference on World Wide Web*, pp.961–970 (2010).

[4] Skoutas, D., Minack, E. and Nejdl, W.: Increasing Diversity in Web Search Results, *Web Science Conference 2010* (2010).

[5] Song, K., Tian, Y., Gao, W. and Huang, T.: Diversifying the image retrieval results, *Proc. 14th Annual ACM International Conference on Multimedia*, pp.707–710 (2006).

[6] Moëllic, P.-A., Haugeard, J.-E. and Pitel, G.: Image clustering based on a shared nearest neighbors approach for tagged collections, *ACM International Conference on Image and Video Retrieval*, pp.269–278 (2008).

[7] Bendersky, M. and Croft, W.: Analysis of long queries in a large scale search log, *Proc. 2009 Workshop on Web Search Click Data*, pp.8–14 (2009).

[8] Agrawal, R., Gollapudi, S., Halverson, A. and Ieong, S.: Diversifying search results, *Proc. 2nd ACM International Conference on Web Search and Data Mining*, pp.5–14 (2009).

[9] Dou, Z., Hu, S., Chen, K., Song, R. and Wen, J.: Multi-dimensional search result diversification, *Proc. 4th ACM International Conference on Web Search and Data Mining*, pp.475–484, ACM (2011).

[10] Smeulders, A., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.12, pp.1349–1380 (2000).

[11] Hou, J., Zhang, D., Chen, Z., Jiang, L., Zhang, H. and Qin, X.: Web Image Search by Automatic Image Annotation and Translation, *Proc. 17th International Conference on Systems, Signals and Image Processing* (*IWSSIP'10*), pp.105–108 (2010).

[12] Zhu, G., Yan, S. and Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity, *ACM International Conference on Multimedia*, pp.461–470 (2010).

[13] Hirota, M., Fukuta, N., Yokoyama, S. and Ishikawa, H.: Implementing Constraint-based Clustering for a Photo Search System Using Estimated Metadata, *2nd International Symposium on Applied Informatics* (2011).

[14] Yang, Y., Huang, Z., Shen, H. and Zhou, X.: Mining multi-tag association for image tagging, *World Wide Web*, pp.1–24 (2011).

[15] Leelanupab, T., Halvey, M. and Jose, J.: Application and evaluation of multi-dimensional diversity, *ImageClef 2009 Workshop, Corfu, Greece* (2009).

[16] Arni, T., Tang, J., Sanderson, M. and Clough, P.: Creating a test collection to evaluate diversity in image retrieval, *Proc. beyond Binary Relevance: Preferences, Diversity and Set-Level Judgments, a workshop in SIGIR 2008* (2008).

[17] Hirota, M., Yokoyama, S., Fukuta, N. and Ishikawa, H.: Constraint-Based Clustering of Image Search Results Using Photo Metadata and Low-Level Image Features, *Computer and Information Science*, pp.165–178 (2010).

[18] Van Zwol, R., Murdock, V., Garcia Pueyo, L. and Ramirez, G.: Diversifying image search with user generated content, *Proc. 1st ACM International Conference on Multimedia Information Retrieval*, pp.67–74 (2008).

[19] Yang, L. and Hanjalic, A.: Supervised reranking for web image search, *Proc. International Conference on Multimedia*, pp.183–192 (2010).

[20] Davidson, I. and Ravi, S.S.: Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp.59–70 (2005).

[21] Lee, S., De Neve, W. and Ro, Y.M.: Tag refinement in an image folksonomy using visual similarity and tag co-occurrence statistics, *Image Commun.*, Vol.25, pp.761–773 (2010).

[22] Boutell, M.R. and Luo, J.: Bayesian Fusion of Camera Metadata Cues in Semantic Scene Classification, *Computer Vision and Pattern Recognition*, pp.623–630 (2004).

[23] Sinha, P. and Jain, R.: Classification and annotation of digital photos using optical context data, *Proc. 2008 International Conference on Content-based Image and Video Retrieval* (*CIVR'08*), pp.309–318, ACM, New York, NY, USA (2008).

[24] Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions, *Bull. Calcutta Math. Soc.*, Vol.35, No.99-109, p.4 (1943).

[25] Salembier, P. and Sikora, T.: *Introduction to MPEG-7: Multimedia Content Description Interface*, Hohn Wiley & Sons, Inc, New York, USA (2002).

[26] Chatzichristofis, S.A. and Boutalis, Y.S.: CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval, *International Conference on Computer Vision Systems*, pp.312–322 (2008).

[27] Tamura, H., Mori, S. and Yamawaki, T.: Textural features corresponding to visual perception, *IEEE Trans. Systems, Man and Cybernetics*, Vol.8, No.6, pp.460–473 (1978).

[28] Fowlkes, E. and Mallows, C.: A method for comparing two hierarchical clusterings, *J. Am. Stat. Assoc.*, pp.553–569 (1983).

[29] Meilă, M.: Comparing clusterings: an information based distance, *Journal of Multivariate Analysis* (2007).

[30] Hays, J. and Efros, A.: IM2GPS: Estimating geographic information from a single image, *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2008*), pp.1–8, IEEE (2008).

[31] Crandall, D., Backstrom, L., Huttenlocher, D. and Kleinberg, J.: Mapping the world's photos, *Proc. 18th International Conference on World Wide Web*, pp.761–770, ACM (2009).

Masaharu Hirota is a graduate student of Shizuoka University. He received his Bachelor of Informatics degree from Shizuoka University in 2010. His research interests include Web mining and Data engineering. He is a student member of ACM, IPSJ and IEICE.

Naoki Fukuta received his B.E. and M.E. from Nagoya Institute of Technology in 1997 and 1999 respectively. He received his Doctor of Engineering from Nagoya Institute of Technology in 2002. Since April 2002, he has been working as a research associate at Shizuoka University. Since April 2007, he has been working as an assistant professor. In 2012, he received the IPSJ Yamashita SIG Research Award. His main research interests include Mobile Agents, SemanticWeb, Konwledge-based Software Engineering, Logic Programming, Applications of Auction Mechanisms, and WWW-based Intelligent Systems. He is a member of ACM, IEEE-CS, JSAI (Japanese Society for Artificial Intelligence), IPSJ, IEICE, JSSST (Japan Society of Software Science and Technology), and ISSJ (Information Systems Society of Japan).

Shohei Yokoyama received his B.E., M.E., and Ph.D. degrees in Computer Science from Tokyo Metropolitan University. After working for National Institute of Advanced Industrial Science and Technology, he is now an assistant professor of Shizuoka University. His research interests include Web engineering and Data engineering. He is a member of IPSJ and IEICE.

Hiroshi Ishikawa received his B.S. and Ph.D. degrees in Computer Science from the University of Tokyo. After working for Fujitsu Laboratories and being a professor of Tokyo Metropolitan University, he is now a professor of Shizuoka University. His research interests include database and Web mining. He has published actively in international, refereed journals and conferences, such as ACM TODS, IEEE TKDE, VLDB, IEEE ICDE. He authored some books, which include books entitled Object-Oriented Database System (Springer-Verlag), Next-Generation of Databases and Data Mining (CQ Publishing) and Databases (Morikita Publishing). He received the Sakai Memorial Distinguished Award from Information Processing Society of Japan (IPSJ) and the Director General Award from Science and Technology Agency of Japan. He was an invited professor at the Polytechnic School of the University of Nantes, France. He was a trustee board member of the Database Society of Japan, an editorial board member of VLDB Journal, the chairman of the SIG on Database Systems of IPSJ, and an editor-in-chief of IPSJ Transaction on Databases. He is a fellow of IPSJ and IEICE and a member of ACM and IEEE.