PAPER

A Novel Framework for Extracting Visual Feature-Based Keyword Relationships from an Image Database

Marie KATSURAI[†a)], ***Student Member***, **Takahiro OGAWA**[†b)], ***and*** **Miki HASEYAMA**[†c)], ***Members***

SUMMARY In this paper, a novel framework for extracting visual feature-based keyword relationships from an image database is proposed. From the characteristic that a set of relevant keywords tends to have common visual features, the keyword relationships in a target image database are extracted by using the following two steps. First, the relationship between each keyword and its corresponding visual features is modeled by using a classifier. This step enables detection of visual features related to each keyword. In the second step, the keyword relationships are extracted from the obtained results. Specifically, in order to measure the relevance between two keywords, the proposed method removes visual features related to one keyword from training images and monitors the performance of the classifier obtained for the other keyword. This measurement is the biggest difference from other conventional methods that focus on only keyword co-occurrences or visual similarities. Results of experiments conducted using an image database showed the effectiveness of the proposed method.

key words: *keyword relationship extraction, semantic analysis, inter-keyword relation, image annotation*

1. Introduction

Keyword relationship extraction from an image database has attracted much attention as a means to facilitate image annotation and retrieval methods [1]–[10]. Some conventional methods regard each training image in the target database as a document containing keywords and use the co-occurrence frequency of two keywords to measure their relevance [1]–[3]. This scheme is independent of visual features and there is therefore no guarantee that the relevant keywords always co-occur in the same image. In [4], keyword relevance is measured by calculating the similarity between two probabilistic models that are trained by using visual features. This method effectively uses visual features but is independent of context information from keyword co-occurrences. In [7], Ngo et al. focus on two similarities in the target database by using the co-occurrence frequency of keywords and visual similarities obtained by comparing outputs of classifiers to construct a keyword relationship graph for image annotation. However, the extracted relationships are not always true due to the separate use of the visual similarities and keyword co-occurrences. When considering application to automatic image annotation, if we focus on visual feature-based keyword relationships, a set of relevant keywords can accurately enhance each other. In order to extract such relationships, we need to know which visual features these keywords share in images.

In this paper, a novel framework for extracting visual feature-based keyword relationships from an image database is proposed. In order to extract visual feature-based relationships between keywords, we use the following two steps: (i) modeling of the relationship between each keyword and visual features and (ii) calculation of the relevance between keywords based on performance of a classifier. In the first step, we detect which visual features are related to each keyword. In the second step, we obtain an indication representing performance through training. To satisfy these requirements, we use logistic regression [11], which is a one of the well-known discriminative classifiers used for image annotation [12], [13] and feature selection [14]. The main contributions of this paper are twofold:

1. The proposed method focuses on the visual feature-based relationships between keywords. In order to extract such relationships, the proposed method adopts a step that detects which features two keywords share in images.
2. Compared to the conventional methods that use only single similarity or that simply combine two similarities, the proposed method can adaptively extract relationships from the target image database by finding visual feature-based relevance in images.

This paper is organized as follows. In Sect. 2, related works in the area of keyword relationship extraction are presented. In Sect. 3, extraction of the relationship between each keyword and visual features is described. A novel framework for extraction of the relationship between keywords by focusing on visual feature-based relevance is proposed in Sect. 4. Section 5 shows experimental results to verify the effectiveness of the proposed method. Finally, in Sect. 6, we summarize our method and suggest possible directions for future works.

2. Related Work

In this section, related works in the area of keyword relationship extraction are presented. There exist some methods that use only keyword co-occurrences to measure semantic relevance [1]–[3]. On the other hand, visual similarities are

Manuscript received August 31, 2011.
Manuscript revised December 28, 2011.

[†]The authors are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.
a) E-mail: katsurai@lmd.ist.hokudai.ac.jp
b) E-mail: ogawa@lmd.ist.hokudai.ac.jp
c) E-mail: miki@ist.hokudai.ac.jp

DOI: 10.1587/transfun.E95.A.927

important cues in the image database [4], [15]. For example, in [15], several kinds of visual features are used to show visually similar keywords. Recently, both of keyword co-occurrences and visual similarities have been used [7], [10]. In [7], context is modeled using a semantic graph that is weighted by keyword co-occurrences, and then outputs of classifiers trained by using visual features are fused. Such a fusion approach is also used in [16] for annotation. In [10], two types of probabilistic models are trained by using visual features and keyword co-occurrences respectively, and their outputs are combined for measuring keyword relevance. The conventional methods extract keyword relationships by focusing on each of keyword co-occurrences and visual similarities or by linearly combining them. Our novelty is in detecting visual feature-based relationships from the target image database.

Keyword relationship extraction methods also have been developed in the research field of multimedia ontology, which contains inter-concept relationships and their hierarchical structure. Representative examples of existing concept ontologies are WordNet [17] and Large-Scale Concept Ontology for Multimedia (LSCOM) [18], which are utilized as semantic networks. For example, some methods extract inter-concept relationships by measuring the path length of two concepts in WordNet [19]–[23]. Since semantics obtained from WordNet reflect only linguistic common sense, not visual properties, the linguistic similarities are linearly combined with visual similarities to measure relevance among the keywords [23]. Meanwhile, LSCOM concepts have mainly been used for video annotation and retrieval. In [24], multimedia ontology based on LSCOM concepts is used to improve the performance of video retrieval. In [25], a clustering-based method is used for measuring the overlap among concepts of LSCOM on a feature space. Our work in this paper focuses on measuring visual feature-based relationships from the image database, not organizing the relationships like ontologies.

3. Extraction of the Relationship Between Each Keyword and Visual Features

In this section, the extraction of the relationship between each keyword and visual features is described. We use logistic regression, which is used in a wide range of applications [26]–[28], for modeling the input-output relationship. Given a database of annotated images I_i ($i = 1, 2, \cdots, N$, where N is the number of images), we first extract their visual feature vectors $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,D}]^T$. Let w_k ($k = 1, 2, \cdots, K$, where K is the number of keywords) be a keyword in the database. We set the binary variable $t_i^k = 1$ if image I_i is annotated with keyword w_k and $t_i^k = 0$ otherwise. Then the set of training data is represented by $S = \{\boldsymbol{x}_i, t_i^k | \, i = 1, 2, \cdots, N, k = 1, 2, \cdots, K\}$. By training the logistic regression from the data set S, the probability of image I_i containing keyword w_k is calculated as follows:

$$y^k = p(t^k = 1|\boldsymbol{x}; \boldsymbol{\beta}_k) = \frac{1}{1 + \exp(-\boldsymbol{\beta}_k^T \boldsymbol{x})}, \tag{1}$$

where parameter vector $\boldsymbol{\beta}_k = [\beta_{k,1}, \beta_{k,2}, \cdots, \beta_{k,D}]^T$ is usually determined by minimizing the following negative log-likelihood function:

$$\begin{aligned} -\log P(S|\boldsymbol{\beta}_k) &= -\sum_{i=1}^{N} \log P(\boldsymbol{x}_i, t_i^k) \\ &= -\sum_{i=1}^{N} \left[t_i^k \log y_i^k + (1-t_i^k) \log(1-y_i^k) \right]. \end{aligned} \tag{2}$$

Since there exists no closed form solution for Eq. (2), the parameter vector is generally estimated by using the Newton-Raphson update or stochastic gradient descent (SGD) algorithm. In this paper, we utilize the SGD algorithm because this method performs efficiently even when the number of data points is very large [29].

Logistic regression is able to not only classify the images but also detect the relevant features for each keyword. This is because the logistic regression is a linear discriminative model, and Eq. (1) has a decision boundary in which $\boldsymbol{\beta}_k^T \boldsymbol{x}_i$ is zero. In the case in which an input image I_i has d-th visual feature $x_{i,d}$ which is related to the target keyword w_k, the corresponding parameter $\beta_{k,d}$ should be positive in order to let the visual feature $x_{i,d}$ contribute to the annotation of keyword w_k. Thus, in this paper, we assume that the d-th feature corresponding to the parameter such that $\beta_{k,d} > 0$ is a relevant feature with keyword w_k[†]. This characteristic is used in the keyword relationship extraction step, which is presented in the following section.

4. Extraction of Visual Feature-Based Relationships Between Keywords

A novel method for extraction of visual feature-based relationships between keywords is proposed in this section. If keyword w_j ($j = 1, 2, \cdots, K$) has visual feature-based relevance with keyword w_k ($k \neq j$), it is clear that they have the same visual features of the images in the database. Since it is difficult to directly find such keywords that have the same visual features, a novel approach is used to measure the relevance between the keywords. Below, this approach is divided into three steps and performed sequentially.

1. **Removal of relevant features**
   In the proposed method, we remove the visual features related to keyword w_k from all of the images containing keyword w_k in the database and monitor how difficult it becomes to estimate keyword w_j from these images (See 4.1).
2. **Determination of relevance presence between keywords**
   If the performance of keyword w_j becomes worse by removing the features related to keyword w_k, this means keyword w_j has a visual feature-based relevance with keyword w_k (See 4.2).

[†] In this paper, each element of the visual feature vector is positive.

3. Measurement of relevance between keywords
We regard change in keyword w_j performance as a strength of the visual feature-based relevance from keyword w_k to keyword w_j. This is a novel idea for measuring keyword relevance (See 4.3).

In the following subsections, we describe the details of the above three steps. Furthermore, in Sect. 4.4, we discuss classifiers for the proposed framework.

4.1 Removal of Relevant Features

After training the logistic regression for keyword w_k, we can detect the visual features that are relevant to keyword w_k using the obtained parameter as described in Sect. 3. We first remove the visual features relevant to keyword w_k from the training image I_i containing keyword w_k by using the obtained parameter vector $\boldsymbol{\beta}_k$ and generate new visual feature vector $\boldsymbol{x}_i^k = [x_{i,1}^k, x_{i,2}^k, \cdots, x_{i,D}^k]^T$ as follows:

$$x_{i,j}^k = m_j^k x_{i,j}, \qquad (3)$$
$$m_j^k = \begin{cases} 0 & \beta_{k,j} > 0 \\ 1 & \text{otherwise.} \end{cases}$$

This equation provides a new training set, $S^k = \{\boldsymbol{x}_i^k, t_i^k | i = 1, 2, \cdots, N, k = 1, 2, \cdots, K\}$, whose visual features related to keyword w_k are removed. The proposed method removes visual features related to a target keyword from images annotated with that keyword. This corresponds to using the keyword co-occurrence information because the removal should affect performance of keywords that co-occur with the target keyword. If we remove features related to the target keyword from all images, it affects estimation of keywords even if they do not co-occur with the target keyword, which means that only the visual similarities are computed for a pair of keywords. For an explanation of this removal process, some examples are shown in Fig. 1, where the target keyword is "*bloom*". Figure 1 shows that the visual features representing keyword "*bloom*" are effectively removed by using Eq. (3). Also, we can see that removal of visual features related to keyword "*bloom*" affects the performance of keyword "*flower*" that co-occurs with "*bloom*" in the image, which is equivalent to detecting the visual feature-based relationship.

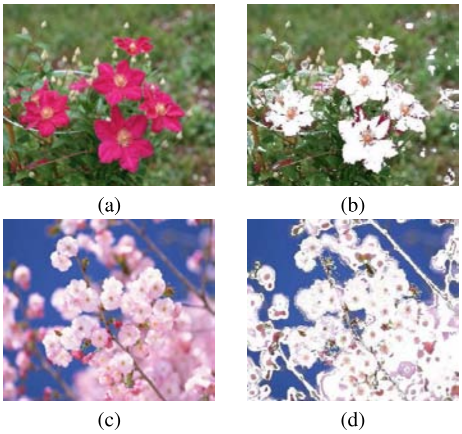

(a) (b) (c) (d)

Fig. 1 Examples of removing visual features relevant to keyword *"bloom"*: (a) Original image, (b) Image in which visual features relevant to keyword *"bloom"* are removed by using Eq. (3) from (a), (c) Another original image, (d) Image obtained from (c) in the same way as (b). White areas correspond to removed visual features.

4.2 Determination of Relevance Presence Between Keywords

Based on the new training set S^k, we train the other logistic regression for keyword w_j $(j \neq k)$. This training provides a new parameter vector $\boldsymbol{\beta}_j^k$. From the obtained results, we monitor the performance of keyword w_j. Specifically, we utilize

$$\begin{aligned} d_k(j) &= -\log P(S^k|\boldsymbol{\beta}_j^k) - [-\log P(S|\boldsymbol{\beta}_j)] \\ &= -\sum_{i=1}^{N} \Big[t_i^j \left\{ \log z_i^{k,j} - \log y_i^j \right\} \\ &\quad + (1 - t_i^j) \left\{ \log(1 - z_i^{k,j}) - \log(1 - y_i^j) \right\} \Big], \end{aligned} \qquad (4)$$

as the criterion for the relevance from keyword w_k to keyword w_j, where

$$z^{k,j} = \frac{1}{1 + \exp(-\boldsymbol{\beta}_j^{k^T} \boldsymbol{x})}.$$

The above criterion $d_k(j)$ represents the difference in performance for keyword w_j between the two logistic regressions whose parameter vectors are $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_j^k$, respectively. Since the logistic regression is trained to minimize the estimation error as shown in Eq. (2), we compare the performance of the two models using this criterion. If $d_k(j) > 0$, we can see that the performance for keyword w_j becomes worse due to the removal of visual features relevant to keyword w_k. This means that the removed visual features are necessary for the estimation of keyword w_j. On the other hand, if $d_k(j) < 0$, the performance for keyword w_j becomes better due to the removal of noise. We denote R_k as a set of keywords w_j satisfying $d_k(j) > 0$, i.e., relating to keyword w_k based on visual features.

4.3 Measurement of Relevance Between Keywords

We measure the degree of visual feature-based relevance from keyword w_k to keyword w_j using the following equation:

$$D(w_j|w_k) = -\sum_{i=1}^{N} \left\{ y_i^j \log z_i^{k,j} + (1 - y_i^j) \log(1 - z_i^{k,j}) \right\}. \qquad (5)$$

Using the above equation is equivalent to calculating the degree of change in performance by the removal of visual features that are related to keyword w_k. If the calculated degree $D(w_j|w_k)$ is a large value, this means that the keyword

w_j shares most of the visual features with keyword w_k. We normalize the degree for each keyword w_k and calculate a relevance score between 0 and 1 as follows:

$$r(w_j|w_k) = \frac{D(w_j|w_k)}{D(w_k|w_k)}. \tag{6}$$

This conversion is useful for comparing strengths of relations among keyword pairs and using the extracted relationships in applications such as automatic image annotation. The score $r(w_j|w_k)$ approaches one when the relevance is strong, where this behavior is the same with $D(w_j|w_k)$. Thus, visual feature-based keyword relationship extraction is realized by using the proposed method.

4.4 Discussion of Classifiers

In this subsection, we discuss classifiers used in the proposed framework. Our approach requires a classifier that enables the following two steps:

(i) Detection of visual features relevant to each keyword,
(ii) Evaluation of performance in the learning stage to measure the degree of relevance.

Since logistic regression can satisfy these requirements in a simple way, we apply it to the proposed framework in this paper. If other classifiers can satisfy these requirements, e.g., to accurately classify images or to effectively detect visual features better than the logistic regression, it is expected that the performance of keyword relationship extraction will be improved. For example, some non-linear methods such as SVM and kernel logistic regression have been widely used for many applications. These can be used for our framework by introducing some procedures. The issue of which classifier is suitable for the proposed framework will be discussed in our future work.

5. Experimental Results

In this section, experimental results to verify the effectiveness of the proposed method are presented. We use the following two different databases to show whether the proposed method can extract keyword relationships adaptively from a target image database.

IAPR-TC12. IAPR-TC12 was originally used in ImageCLEF [30] and has been developed for cross-lingual retrieval [31]. This set of 19,627 color images is a standard benchmark for automatic image annotation. Each image is annotated with about 6 words from the 291 candidate noun words in the database [12]. We use 17,665 images for training and 1,962 images for testing.

Hokkaido landscape database. This database has 22,000 color images of Hokkaido landscape, with each image being manually annotated with 10–20 keywords as ground truth. There are 230 keywords in the database. We use 20,000 images for training and 2,000 images for testing.

For feature extraction, each image in the database is divided into about 20 regions by using a multiresolution implementation of the well-known recursive shortest spanning tree (RSST) method [32]. Then 54-dimensional features representing color and textures [33] are extracted from each region and used to form a 200-dimensional feature vector $\boldsymbol{x}_i$ by means of a bag-of-words approach.

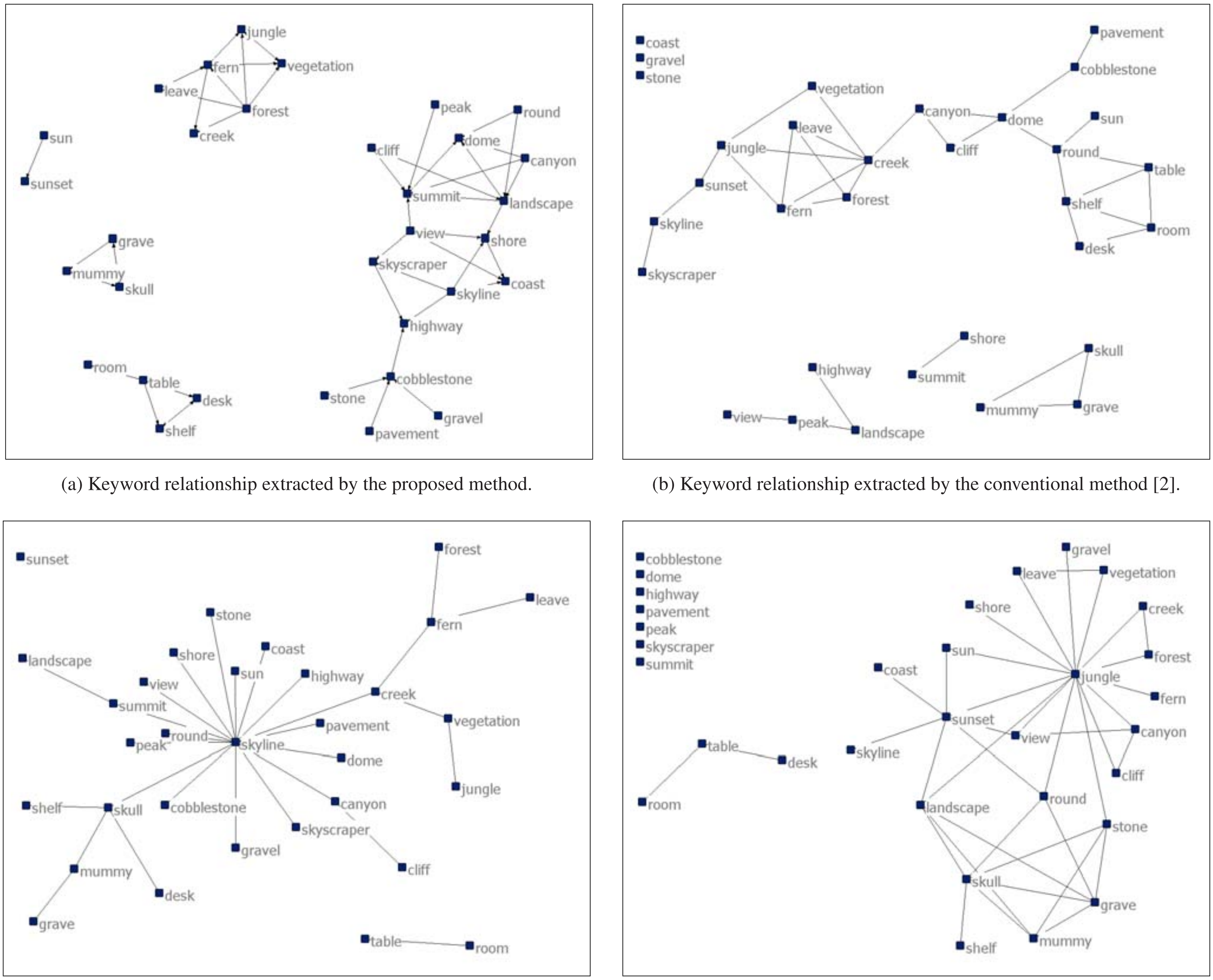
Section 5.1 shows the results of keyword relationship extraction from each of the target databases. Furthermore, in Sect. 5.2, we apply the extracted relationships to automatic image annotation on testing images and evaluate the performance of the proposed method.

5.1 Keyword Relationship Extraction

In this subsection, we utilize the proposed method to extract keyword relationships from each of the target databases and present the extracted relationships. We compare our keyword relationship extraction method with the conventional methods [2], [4], [7]. The proposed method and the conventional methods are applied to the same image database, and the extracted keyword relationships are depicted as networks by using NetDraw [34]. Since it is difficult to show relationships of all of the keywords in the database, we randomly select 32 keywords from each database to draw the relationships respectively. Moreover, strong relationships are chosen by thresholding for easier viewing. The drawn relationships are shown in Figs. 2 and 3, where each node represents a keyword and each edge between keywords represents a keyword relevance. As shown in Figs. 2 and 3, although the two networks are constructed from the same database, their keyword relationships are different. We can find the following from Figs. 2 and 3:

- The proposed method finds the visual feature-based relationships that are derived from the same source origin, not reflecting just keyword co-occurrences.
- The method in [2] extracts the keyword pairs that tend to co-occur in the image database, not reflecting visual contents in images.
- The method in [4] finds visually similar keyword pairs. Since this method does not have a scheme that detects the features relevant to each keyword, the similarities might be affected by some visual features that are derived from other keywords in images.
- The method in [7] finds relevant keyword pairs by combining keyword co-occurrences and visual similarities. This method cannot extract the visual feature-based relationships but can represent co-occurrence probability or visual similarity.

On the other hand, we can find that the proposed method has the following weakness: From Fig. 3, the proposed method fails to describe the relationship around the keyword "*round*". In this experiment, the features used do not include shape information. Thus, if the features cannot model each



(a) Keyword relationship extracted by the proposed method.

(b) Keyword relationship extracted by the conventional method [2].

(c) Keyword relationship extracted by the conventional method [4].

(d) Keyword relationship extracted by the conventional method [7].

Fig. 2 Keyword relationships extracted from the IAPR-TC12 dataset by the proposed method and conventional methods [2], [4], [7].

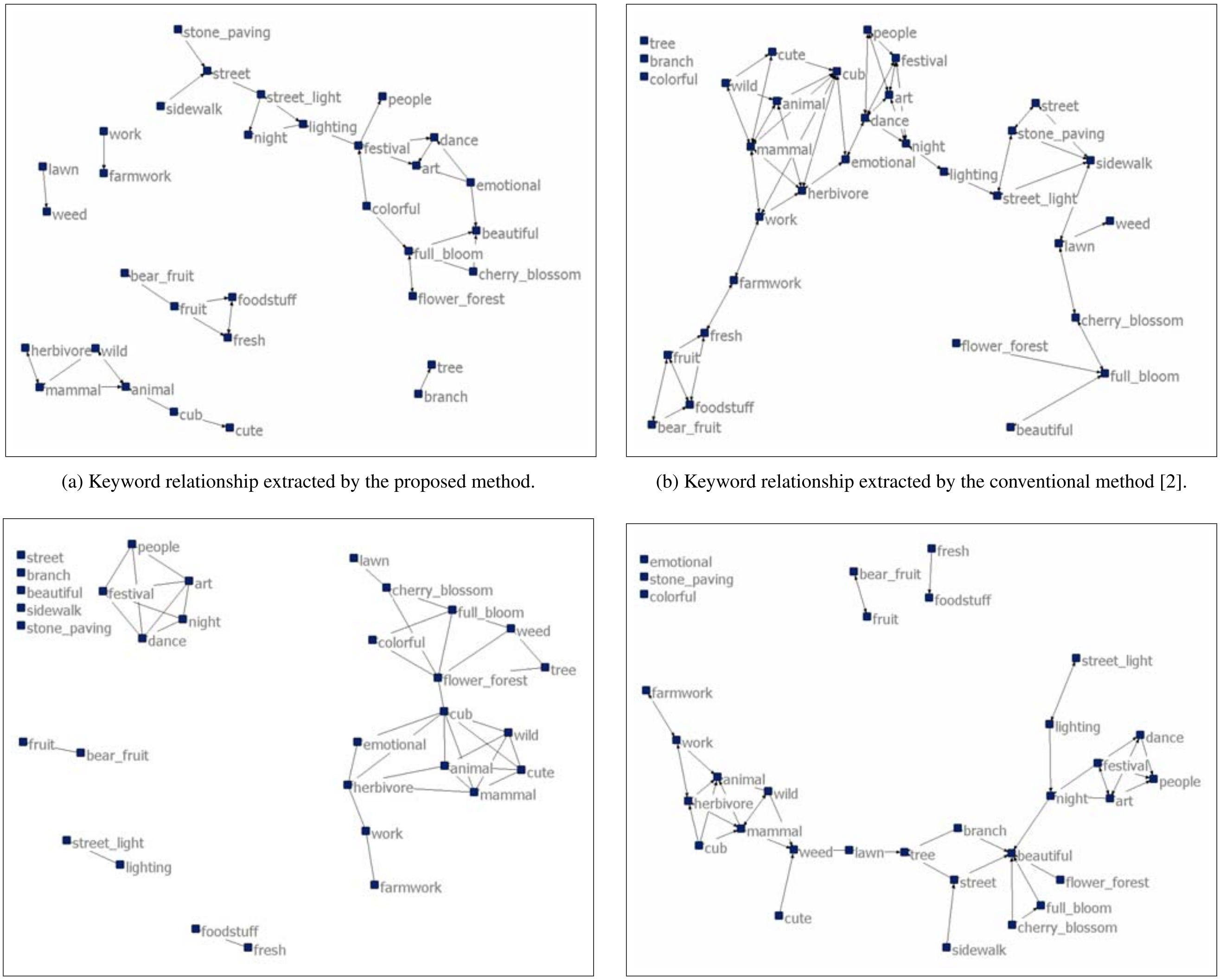
keyword well, the relationships around the keywords cannot be extracted. This weakness is common to other conventional methods that use visual similarities. For these keywords, the method using only keyword co-occurrences [2] can extract the relationships better than the other methods can. In future works, we should investigate various kinds of visual features to overcome this problem in the experiments.

The extracted strong relationships are also manually evaluated by users. We present 11 users a list of 55 keyword pairs with the highest relevance from each of the databases. For each keyword pair, the user is required to give a score ranging from 1 to 5. Note that if the score becomes higher, this means the relationship of the two keywords also becomes stronger. By averaging the scores from all of the users, we obtain the final score for each keyword pair, which are shown in Tables 1 and 2. In the tables, only some examples of keyword pairs and their scores are shown due to space limitation, and the bottom scores are average scores of 55 keyword pairs. From these results, the proposed method extracts the relationships that are always true better than the conventional method does [7]. This is due to the effective extraction of visual feature-based relationships.

In the following subsection, we apply the keyword relationships to automatic image annotation to quantitatively evaluate the keyword relationship extraction methods.

5.2 Automatic Image Annotation

In this subsection, we apply the keyword relationships to automatic image annotation in order to quantitatively evaluate the effectiveness of the proposed method. For introducing the semantic relationships between keywords into image annotation, the Dual Cross-Media Relevance Model (DCMRM) [35] has been proposed. The DCMRM calcu-



(a) Keyword relationship extracted by the proposed method.

(b) Keyword relationship extracted by the conventional method [2].

(c) Keyword relationship extracted by the conventional method [4].

(d) Keyword relationship extracted by the conventional method [7].

Fig. 3 Keyword relationships extracted from the Hokkaido landscape database by the proposed method and conventional methods [2], [4], [7].

Table 1 User assessments of the IAPR-TC12 dataset.

Proposed method		Conventional method [7]	
Example of keyword pair	Score	Example of keyword pair	Score
cyclist - bicycle	4.81	lake - cloud	2.63
wave - sea	4.91	waterfall - rock	3.82
dune - beach	2.64	table - wall	3.45
desk - shelf	4.10	portrait - man	3.18
fern - jungle	3.55	kid - classroom	3.10
gravel - cobblestone	4.37	horizon - cloud	3.54
hill - ground	3.91	stadium - shirt	2.09
cloth - cape	2.64	jungle - stone	2.00
highway - skyscraper	3.18	wood - wall	2.10
shirt - pullover	4.10	lagoon - cloud	2.27
Average	3.17	Average	2.87

Table 2 User assessments of the Hokkaido landscape database.

Proposed method		Conventional method [7]	
Example of keyword pair	Score	Example of keyword pair	Score
waterside - river	4.64	herbivore - work	1.82
maple - autumn	3.64	crop - agriculture	4.88
sea - horizon	3.73	horse - lawn	3.55
tree - branch	4.91	lawn - tree	2.73
building - exterior	3.82	crop - bloom	2.45
crop - agriculture	4.88	snow - mountain	3.34
water flow - liquid	4.73	mountain - sunset	3.55
people - clothes	3.82	reflection - beautiful	3.00
lawn - weed	3.82	night - festival	3.55
stone - solid	4.18	active - morning	2.27
Average	4.24	Average	3.53

lates an annotation score for each keyword, and keywords that have top m values of Eq. (A·2) are provided as annotations of the testing image (See Appendix). Figures 4 and 5 show annotation results of testing images obtained by the proposed method and the conventional method [2], in which $m = 7$. We can find that the keywords provided by the proposed method are semantically correlated since keywords that have visual feature-based relevance enhance each other.

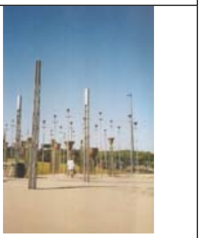
<table>
<tr><td colspan="5">Testing images.</td></tr>
<tr><td> </td><td> </td><td> </td><td> </td><td></td></tr>
<tr><td colspan="5">The top seven keywords provided by the proposed method.</td></tr>
<tr><td>tee-shirt, man, court, player, tennis, woman, stadium</td><td>flower, lamp, people, bench, tree, palm, building</td><td>gate, corridor, wall, brick, flower, pot, window</td><td>port, sky, cloud, boat, building, ship, view</td><td>cobblestone, dune, square, range, plane, lookout, bell, highway</td></tr>
<tr><td colspan="5">The top seven keywords provided by a conventional method [2].</td></tr>
<tr><td>woman, court, fence, people, meadow, tennis, flag</td><td>people, man, woman, tree, flower, lamp, sky</td><td>window, gate, sea, building, flag, fence, rock</td><td>airport, sky, cloud, house, man, car, sea</td><td>cobblestone, dune, rock, hill, mountain, road, car</td></tr>
</table>

Fig. 4 Image annotation results of the proposed method and conventional methods [2] for the IAPR-TC12 dataset. The top seven keywords are provided.

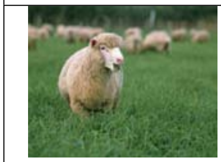
<table>
<tr><td colspan="5">Testing images.</td></tr>
<tr><td></td><td> </td><td> </td><td> </td><td> </td></tr>
<tr><td colspan="5">The top seven keywords provided by the proposed method.</td></tr>
<tr><td>farm, animal, work, pasture, herbivore, grazing, crowd</td><td>people, art, cloth, dance, team, building, festival</td><td>water, flow, liquid, river, people, animal, waterside</td><td>morning, sun, morning glow, silhouette, romantic, backlighting, cloud</td><td>sea, illumination, night, mountain, city, romantic, building</td></tr>
<tr><td colspan="5">The top seven keywords provided by a conventional method [2].</td></tr>
<tr><td>mammal, animal, fence, work, herbivore, wall, emotional</td><td>life, evening, people, reflection, dark, summer, shade</td><td>river, water, liquid, flow, solid, waterside, rock</td><td>evening, silhouette, reflection, bright, dazzling, morning, bright</td><td>romantic, night, reflection, tower, building, harbor, dark</td></tr>
<tr><td colspan="5">Testing images.</td></tr>
<tr><td> </td><td> </td><td> </td><td> </td><td> </td></tr>
<tr><td colspan="5">The top seven keywords provided by the proposed method.</td></tr>
<tr><td>food, seafood, leaf, maple, fresh, waterside, animal</td><td>fallen leaves, maple, sun, leaf, autumn, street, evening</td><td>farm, flower forest, agriculture, full bloom, flower, gradation, forest</td><td>lighting, art, night, people, building, flickering, street</td><td>food, fruit, forest, animal, fresh, bloom, wild</td></tr>
<tr><td colspan="5">The top seven keywords provided by a conventional method [2].</td></tr>
<tr><td>morning, people, autumn, solid, gradation, picnic, maple</td><td>evening, reflection, leaf, gradation, shadow, dark, tree</td><td>blue sky, agriculture, cloud, mountain, hill, forest, tree</td><td>evening, people, reflection, tree, bright, building, dark</td><td>land, summer, weed, tree, grass, agriculture, leaf</td></tr>
</table>

Fig. 5 Image annotation results of the proposed method and conventional methods [2] for the Hokkaido landscape database. The top seven keywords are provided.

However, we can find a weakness of the proposed method from Fig. 4. In the right-hand result of Fig. 4, keywords such as "*square*" and "*plane*" are incorrectly provided. In fact, the proposed method incorrectly extracted the relationships of the shape-based keyword "*square*", which was connected with the keywords "*cobblestone*" and "*dune*". As mentioned in the previous subsection, the proposed method cannot describe the relationships around the keywords such as "round" and "square". This weakness causes the poor annotation results, and we should therefore use more effective

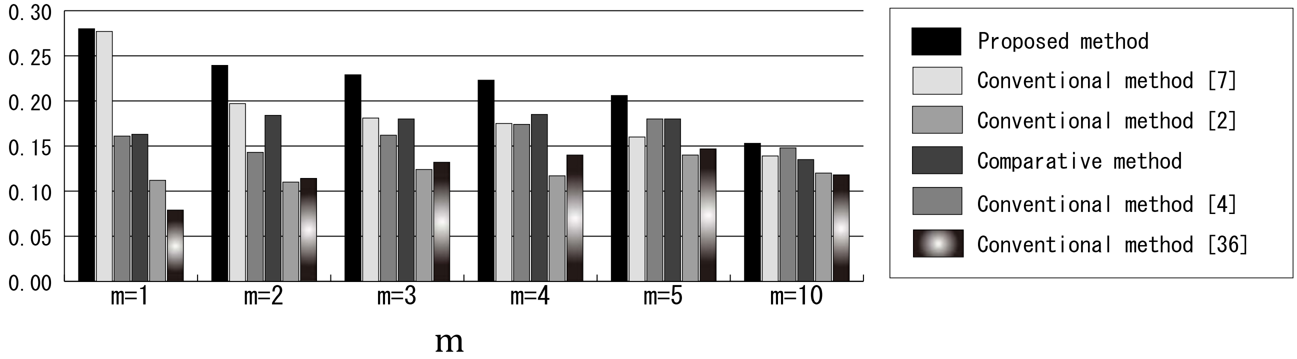


Fig. 6 Average precision ($m = 1, 2, 3, 4, 5$, and 10) on testing images of the IAPR-TC12 dataset.

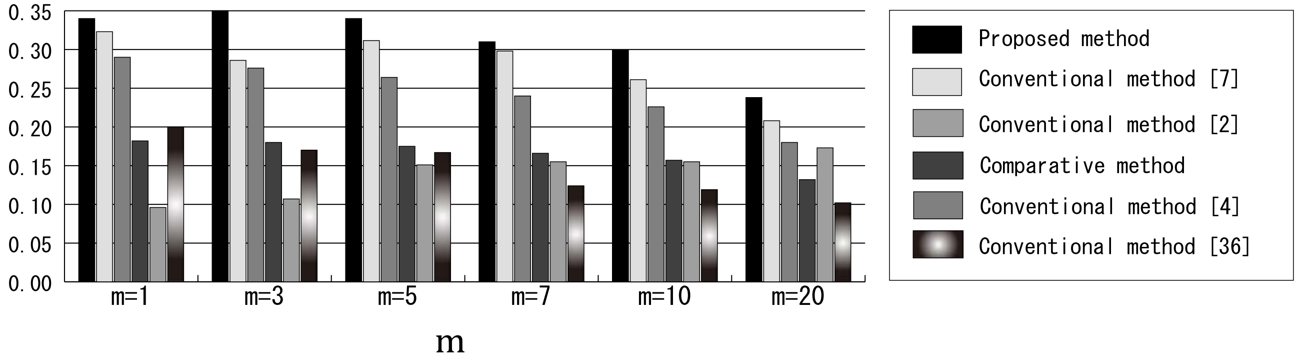


Fig. 7 Average precision ($m = 1, 3, 5, 7, 10$, and 20) on testing images of the Hokkaido landscape database.

Table 3 Precision, recall, and F-measure of the conventional methods [2], [4], [7], [36] and the proposed method for the IAPR-TC12 dataset, in which $m = 5$.

Method	Precision	Recall	F-measure
Proposed method	**0.206**	**0.183**	**0.194**
Conventional method [7]	0.160	0.162	0.161
Conventional method [2]	0.180	0.159	0.169
Conventional method [4]	0.140	0.158	0.148
Conventional method [36]	0.147	0.146	0.146
Comparative method	0.180	0.177	0.179

Table 4 Precision, recall, and F-measure of the conventional methods [2], [4], [7], [36] and the proposed method for the Hokkaido landscape database, in which $m = 12$.

Method	Precision	Recall	F-measure
Proposed method	**0.289**	**0.258**	**0.273**
Conventional method [7]	0.245	0.242	0.243
Conventional method [2]	0.206	0.179	0.192
Conventional method [4]	0.169	0.107	0.131
Conventional method [36]	0.112	0.143	0.126
Comparative method	0.151	0.215	0.177

features to describe each keyword.

Furthermore, the quality of automatic image annotation is measured through the process of retrieving testing images with a single keyword. For each keyword w_k, the number of correctly annotated images is denoted as C_k, the number of retrieved images is denoted as S_k, and the number of truly related images in the testing set is denoted as R_k. The precision, recall and F-measure are computed as follows:

$$precision(w_k) = \frac{C_k}{S_k}, \quad recall(w_k) = \frac{C_k}{R_k}$$
$$F(w_k) = \frac{2 \times precision(w_k) \times recall(w_k)}{precision(w_k) + recall(w_k)}, \tag{7}$$

We calculate the average precision, recall and F-measure over the words which exist in the testing images to evaluate the performance. Performance comparisons are shown in Tables 3 and 4, in which $m = 5$ and $m = 12$, respectively. In the tables, in order to show the effectiveness of the keyword relationship, we compare the proposed method with the Cross-Media Relevance Model (CMRM) [36] and non-linear SVM with RBF kernel (denoted by Comparative method). Tables 3 and 4 show that image annotation with our keyword relationship extraction method performs better than the conventional methods [2], [4], [7]. Compared to the method [7] that separately extracts context information from keyword co-occurrences and visual similarities and combines them, our method shows a superiority due to simultaneous use of context information and visual similarities. Also, compared with the conventional method [36], it becomes clear that keyword relationships are effective for image annotation. The precisions with different values of m are shown in Figs. 6 and 7. These figures show that precision of image annotation by the proposed method outperforms the precisions of the conventional methods with respect to each variable m. From these experimental results, we can conclude that the proposed method can extract suitable keyword relationships for image annotation by focusing on visual feature-based relevance. Even if we use logistic regression, which is a simple discriminative classifier, the proposed framework can effectively extract keyword relationships. It is expected that if non-linear classifiers including SVM are used for the proposed framework, we will be able to improve performance of keyword relationship extraction. We should discuss the possibility of improvement by comparing classifiers, which is our future work.

6. Conclusions and Future Works

This paper presents a novel framework for extracting visual feature-based keyword relationships from an image database. In the proposed approach, we first detect visual features related to each keyword. Then, in order to measure the keyword relevance, we remove the detected features from training images and monitor the performance of the other keywords. This is the biggest difference from the conventional approaches. The proposed framework needs a classifier that can perform the following steps: (i) detection of visual features relevant to each keyword and (ii) monitoring performance after training. To satisfy these requirements, we use logistic regression in this work. Results of experiments using image databases show that our method can analyze keyword relationships by using visual features. We applied extracted keyword relationships to automatic image annotation in the experiments. From the obtained results, we can see that our method has better image annotation performance compared to the conventional methods. We also find that more effective features are needed to detect features related to each keyword. We will investigate various kinds of feature extraction methods for the proposed approach to improve performance.

In addition, two kinds of verifications are needed in future works. First, we should consider the relationship between the performance of the proposed method and the number of training images. Apart from experiments in Sect. 5, we experimentally reduced the number of training samples and evaluated performance in image annotation. In the IAPR-TC12 dataset, when reducing the number of training samples from 17665 to 10000, the F-measure at $m = 5$ decreased to 0.181. In the Hokkaido landscape database, when reducing the number of training samples from 20000 to 10000, the F-measure at $m = 12$ decreased to 0.258. It is considered that such low-performance is due to insufficient training of the logistic regression. These results suggest that we should investigate how the number of training images affects the performance of the method and consider which classifiers are suitable for the proposed framework in future works. Second, we should investigate keywords that the proposed method do not work well with due to their visual diversity [23]. For this, we experimentally adapted the method proposed in [37] to compute the degree of visual diversity for each of keywords in each database. In the IAPR-TC12 dataset, the most visually diverse keywords were "*area*", "*bit*" and "*lot*", and the most visually representative keywords were "*sky*", "*tree*", and "*mountain*". In the Hokkaido landscape database, the most visually diverse keywords were "*picnic*", "*curve*" and "*transparency*", and the most visually representative keywords were "*cloud*", "*mountain*" and "*forest*". In fact, annotation performances of the visually diverse keywords tend to be lower than the performances of the visually representative keywords, even if the proposed method is applied. It can be said that visually diverse keywords should be separated from other keywords, and then some special procedures should be prepared to extract relationships for those keywords. We will conduct these experiments in detail and improve the proposed framework in future works.

Acknowledgements

This work was partly supported by SCOPE (Strategic Information and Communication R&D Promotion Programme) of the Japanese Ministry of Internal Affairs and Communications. In our experiments, we used an image database provided by Datacraft Co., Ltd (http://www.datacraft.co.jp).

References

[1] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Image annotation refinement using random walk with restarts," Proc. ACM Int'l Conf. on Multimedia, pp.647–650, 2006.

[2] R.L. Cilibrasi and P.M. B. Vitanyi, "The google similarity distance," IEEE Trans. Knowl. Data Eng., vol.19, no.3, pp.370–383, 2007.

[3] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma, "Image annotation via graph learning," Pattern Recognit., vol.42, no.2, pp.218–228, 2009.

[4] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr distance," Proc. ACM Int'l Conf. on Multimedia, pp.31–40, 2008.

[5] Z.J. Zha, T. Mei, Z. Wang, and X.S. Hua, "Building a comprehensive ontology to refine video concept detection," Proc. Int'l Workshop on Multimedia Information Retrieval, pp.227–236, 2007.

[6] Y. Wang and S. Gong, "Refining image annotation using contextual relations between words," Proc. ACM Int'l Conf. on Image and Video Retrieval, pp.425–432, 2007.

[7] C.W. Ngo, Y.G. Jiang, X.Y. Wei, W. Zhao, J Y. Liu, S. Zhu Wang, and S.F. Chang, "VIREO/DVMM at TRECVID 2009: High-level feature extraction, automatic video search, and content-based copy detection," Proc. TRECVID 2009 Workshop, pp.415–432, 2009.

[8] A. Llorente, E. Motta, and S. Rüger, "Image annotation refinement using web-based keyword correlation," in Semantic Multimedia, vol.5887 of Lect. Notes Comput. Sci., pp.188–191, Springer Berlin, Heidelberg, 2009.

[9] H. Kawakubo, Y. Akima, and K. Yanai, "Automatic construction of a folksonomy-based visual ontology," Proc. IEEE Int'l Symp. on Multimedia, pp.330–335, 2010.

[10] Y. Akima, H. Kawakubo, and K. Yanai, "Folksonomy-based automatic construction of visual ontology using visual features and tag co-occurrence," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J94-D, no.8, pp.1248–1259, Aug. 2011.

[11] D.W. Hosmer and S. Lemeshow, Applied logistic regression (Wiley Series in probability and statistics), 2nd ed., Wiley-Interscience Publication, 2000.

[12] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," Proc. European Conf. on Computer Vision, pp.316–329, 2008.

[13] L. Cao, J. Yu, J. Luo, and T.S. Huang, "Enhancing semantic and

geographic annotation of web images via logistic canonical correlation regression," Proc. ACM Int'l Conf. on Multimedia, pp.125–134, 2009.

[14] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns," NeuroImage, vol.42, no.4, pp.1414–1429, 2008.

[15] C. Yang, X. Feng, J. Peng, and J. Fan, "Efficient large-scale image data set exploration: Visual concept network and image summarization," in Advances in Multimedia Modeling, vol.6524 of LNCS, pp.111–121, Springer Berlin, Heidelberg, 2011.

[16] C. Diou, G. Stephanopoulos, P. Panagiotopoulos, C. Papachristou, N. Dimitriou, and A. Delopoulos, "Large-scale concept detection in multimedia data using small training sets and cross-domain concept fusion," IEEE Trans. Circuits Syst. Video Technol., vol.20, no.12, pp.1808–1821, 2010.

[17] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[18] M. Naphade, J.R. Smith, J. Tesic, S.F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," IEEE Multimedia, vol.13, no.3, pp.86 –91, 2006.

[19] J.J. Jiang and D.W. Conath, "Semantic similarity based on corpus statistics and lexical taxonomy," Int'l. Conf. on Research and Computational Linguistics, 1997.

[20] C. Breen, L. Khan, and A. Ponnusamy, "Image classification using neural networks and ontologies," Proc. Int'l Workshop on Database and Expert Systems Applications, p.98, 2002.

[21] Y. Wu, B.L. Tseng, and J.R. Smith, "Ontology-based multi-classification learning for video concept detection," Proc. IEEE Int'l Conf. on Multimedia and Expo, vol.2, pp.1003–1006, 2004.

[22] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan, "Exploiting ontologies for automatic image annotation," Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp.552–558, 2005.

[23] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," IEEE Trans. Image Process., vol.17, no.3, pp.407–426, 2008.

[24] K. Shirahama and K. Uehara, "Effectiveness of video ontology in query by example approach," Active Media Technology, vol.6890 of Lect. Notes Comput. Sci., pp.49–58, Springer Berlin, Heidelberg, 2011.

[25] M. Koskela, A.F. Smeaton, and J. Laaksonen, "Measuring concept similarities in multimedia ontologies: Analysis and evaluations," IEEE Trans. Multimedia, vol.9, no.5, pp.912–922, 2007.

[26] D.S. Rosario, "Highly effective logistic regression model for signal (anomaly) detection," Proc. IEEE Int'l Conf. on Acoustics, Speech, Signal Process., vol.5, pp.V–817–20, 2004.

[27] O. Birkenes, T. Matsui, K. Tanabe, S.M. Siniscalchi, T.A. Myrvoll, and M.H. Johnsen, "Penalized logistic regression with HMM log-likelihood regressors for speech recognition," IEEE Trans. Audio Speech Language Process., vol.18, no.6, pp.1440–1454, 2010.

[28] T. Gilliam, R.C. Wilson, and J.A. Clark, "Scribe identification in medieval English manuscripts," Proc. Int'l Conf. on Pattern Recognition, pp.1880–1883, 2010.

[29] Tong Z., "Solving large scale linear prediction problems using stochastic gradient descent algorithms," Proc. Int'l Conf. on Machine learning, pp.116–123, 2004.

[30] "ImageCLEF," http://www.imageclef.org/

[31] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR-TC12 benchmark: A new evaluation resource for visual information systems," Proc. Int'l Conf. on Language Resources and Evaluation, 2006.

[32] Y.S. Avrithis, A.D. Doulamis, N.D. Doulamis, and S.D. Kollias, "A stochastic framework for optimal key frame extraction from mpeg video databases," Comput. Vis. Image Understanding, vol.75, no.1-2, pp.3–24, 1999.

[33] X. Li, C.G.M. Snoek, and M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," Proc. ACM Int'l Conf. on Multimedia Information Retrieval, pp.180–187, 2008.

[34] S. Borgatti, "Netdraw network visualization," http://www.analytictech.com/netdraw/netdraw.htm

[35] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma, "Dual cross-media relevance model for image annotation," Proc. ACM Int'l Conf. on Multimedia, pp.605–614, 2007.

[36] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," Proc. ACM SIGIR Conf. on Research and Development in Informaion Retrieval, pp.119–126, 2003.

[37] A. Sun and S.S. Bhowmick, "Quantifying tag representativeness of visual content of social images," Proc. ACM Int'l Conf. on Multimedia, pp.471–480, 2010.

Appendix

In this appendix, the dual cross-media relevance model (DCMRM) presented in [35] which we employed as an image annotation model in the experiments is described. In the following explanations, we use the same notation as those in [35]. The goal of automatic image annotation is to predict the joint probability of a testing image I_q and a keyword w_k as follows:

$$w^* = \arg\max_{w_k \in V} p(w_k, I_q), \tag{A$\cdot$1}$$

where V is a set of keywords and w^* denotes an optimal keyword for annotation. The DCMRM assumes that the probability of observing both the annotation keyword w_k and the image I_q are mutually independent given a keyword w_j, so that the relevance model is represented as follows:

$$w^* = \arg\max_{w_k \in V} \sum_{w_j \in V} P(I_q|w_j)P(w_k|w_j)P(w_j), \tag{A$\cdot$2}$$

where $P(w_k|w_j)$ denotes the probability of a keyword w_k given a keyword w_j (See (1)) and $P(I_q|w_j)$ denotes the probability of image I_q given a keyword w_j (See (2)).

(1) Calculation of $P(w_k|w_j)$

In the method presented in [35], they calculate the keyword relevance as distance $dist(w_k|w_j)$ by using the conventional method [2] and calculate the conditional probability $P(w_k|w_j)$ as $P(w_k|w_j) = \exp[-\gamma \cdot dist(w_k|w_j)]$, where γ is an adjustable parameter. This conversion is not a calculation of strict probability but is effective in a relevance model such as DCMRM. We introduce the keyword distance from the conventional methods [2], [4], [7] and set the inverse of the average distance of each method to γ. Similarly, in our implementation, the proposed method introduces $r(w_k|w_j)$ in Eq. (6) to $P(w_k|w_j)$.

(2) Calculation of $P(I_q|w_j)$

In a way similar to that in [35], we first calculate image-based similarity as follows:

$$S(I_q, R_k) = \sum_{\boldsymbol{I}_i \in R_k} \alpha_i \exp\left\{-\frac{d(I_i, I_q)}{\sigma_l}\right\} \qquad (\text{A}\cdot 3)$$

where R_k is a set of images annotated with keyword w_k. α_i is an adjustable parameter which aims to support similar image-pairs and penalize dissimilar image-pairs, and $d(I_i, I_q)$ is certain distance metric between image I_i and query image I_q, which is L_1-distance in our experiments. We set the average distance to parameter σ_l. The final word-to-image relation can be approximated by this measure, which is given as:

$$S_{WIR}(I_q, w_k) = [S(I_q, R_k)]^{\eta} \qquad (\text{A}\cdot 4)$$

where η is a parameter. In the experiments, we set $\eta = 1.0$, and the above score is set to $P(I_q|w_k)$ in Eq. (A·2).

Marie Katsurai received her B.S. degree in Electronics and Information Engineering from Hokkaido University, Sapporo, Japan in 2010. She is currently pursuing an M.S. degree at the Graduate School of Information Science and Technology, Hokkaido University. Her research interests include digital image processing. She is a student member of the IEEE.

Takahiro Ogawa received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He is currently an assistant professor in the Graduate School of Information Science and Technology, Hokkaido University. His research interests are digital image processing and its applications. He is a member of the IEEE and Institute of Image Information and Television Engineers (ITE).

Miki Haseyama received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 2005 to 2006. She is currently a professor in the Graduate School of Information Science and Technology, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a member of the IEEE, Institute of Image Information and Television Engineers (ITE) and Acoustical Society of Japan (ASJ).