Hindawi Publishing Corporation
Advances in Multimedia
Volume 2011, Article ID 786427, 10 pages
doi:10.1155/2011/786427

Research Article

Representing Images' Meanings by Associative Values with Given Lexicons Considering the Semantic Tolerance Relation

Ying Dai

Faculty of Software and Information Science, Iwate Prefectual University, Sugo 152-52, Iwate, 020-0193 Takizawa, Japan

Correspondence should be addressed to Ying Dai, dai@iwate-pu.ac.jp

Received 28 January 2011; Revised 11 May 2011; Accepted 30 May 2011

Academic Editor: Nicolas Tsapatsoulis

An approach of representing meanings of images based on associative values with lexicons is proposed. For this, the semantic tolerance relation model (STRM) that reflects the tolerance degree between defined lexicons is generated, and two factors of semantic relevance (SR) and visual similarity (VS) are involved in generating associative values. Furthermore, the algorithm of calculating associative values using pixel-based bidirectional associative memories (BAMs) in combination with the STRM, which is easy in implementation, is depicted. The experiment results of multilexicons-based retrieval by individuals show the effectiveness and efficiency of our proposed method in finding the expected images and the improvement in retrieving accuracy because of incorporating SR with VS in representing meanings of images.

1. Introduction

With the technological advances in digital imaging, networking, and data storage, more and more people communicate with one another and express themselves by sharing images, videos, and other forms of media on line. However, it is difficult to fully utilize the semantic messages that the image/video carries, because the nature of the concepts regarding images in many domains are imprecise, and the interpretation of finding similar images/videos is also ambiguous and subjective on the level of human perception. Accordingly, there are some techniques that focus on annotating the images by folksonomy (Flickr, del.icio.us). But the deliberately idiosyncratic annotation induced by folksonomies has a risk to decrease the systems' performance in formation retrieval utility. In Xie et al. [1], by examining the effects of different choices of lexicons and input detectors, such a conclusion is reached that more concepts than necessary can hurt performance.

Thus, much research aims to the image/video automatic annotation or presentation based on the semantic messages that the images carry. In order to avoid the expense and limitations of text annotations on images, there is considerable interest in efficient database access by perceptual and other automatically extractable attributes of images. However, most current retrieval systems only rely on low-level image features such as color and texture, whereas human users think in terms of concepts [2–5]. Usually relevance feedback is the only attempt to close the semantic gap between user and system.

Recently, there is much research to reduce the semantic gap between users and retrieval systems with the different levels of abstraction employed by human and machine. In Rogowitz [6], how human observers judge image similarity was analyzed to reach a conclusion that the human observers are very systematic to judge image similarity, following semantics, color, and structural characteristics. Following this conclusion, in Mojsilović et al. [7], the extraction of color features and the interpretation of these features based on five image similarity criteria were proposed. However, it was shown that color could not be used as a single measure to capture the semantics of images. In Vogel and Schiele [8], a concept called "vocabulary-supported image retrieval" was proposed, which allowed the system to translate the user query into an internal query. However, the user query as "find images with 10–30% of sky" is not a natural way to present the semantics of the images. In Mojsilovic et al. [9], a semantic-friendly query language for searching diverse collections of images was proposed. However, same

as Vogel and Schiele [8], the query language such as (nature <10 and contrast >800) is not easy to utilize for modeling the categories. In [10], the proposed system was aimed to place an image into a category to help user to navigate retrieval results more efficiently. However, the definition of categories was a mixture of semantic, syntactic, and statistical approaches, which seemed not to be the genuine semantic categories. In Dai and Cai [11], a scheme of image retrieval system, which followed the human perceptual similarity criteria described in Rogowitz [6], was proposed. However, the semantic categorization of images was made manually. In [12–14], the automatic semantic-based image categorization methods using the probabilistic approach or the subspace discovery were proposed. However, only the binary classes (indoor-outdoor, manmade-natural, sunset-nonsunset) were handled. In Shen et al. [15], a framework to handle the classification problem of overlapped classes was presented and was applied to the problem of multiscene classification. However, the extension to the other concepts' classification was not described. In Dai [16], a method of semantic tolerance relation-based image presentation and classification was proposed, and, as the demonstration, the semantics of each image regarding nature versus manmade domain was represented by assigning images to 7 categories based on the Bayesian classifier. Obviously, it is not enough for 7 classes to embody the whole semantics of images. In Carneiro et al. [17], 350 visual concepts were learned under the minimum probability of error retrieval framework, and each image was annotated with the concepts of largest probability. However, with the retrieval precision of about 30%, the proposed method cannot truly resolve the problem of semantic gap between the human and machine, because in fact, it is merely equated with the classification based on the low level visional features. On the other hand, for [17] and [16], the selection of training images set for generating the probability models of new concepts and the regeneration of the probability models for the new concepts are not so easy in practice, if they are required to add. In Xie [1], the statistical test results showed that the concept detection performance is better than baseline if the number of defined lexicons is in the range of (40, 240). However, these defined lexicons are only visual lexicons, and the higher-order concept dependency is still not considered in multiconcept learning. In Li and Wang [18], 2D MHMMs stochastic processes are utilized to implement the system of automatic linguistic indexing of pictures. For the future work, this paper indicates that besides assigning words to an image, weights can be given to the words in the meantime to show the believed extent of description appropriateness. However, how to generate the values of weights is still an open problem.

Because most images/videos have multiple semantic interpretations and people judge similar images with different criterion, in this paper, an approach of representing meanings of images based on associative values with lexicons is proposed. As for this approach, the meanings of images or keyframes (images are called afterwards) are described by the associative values with defined lexicons regarding different domains, while the tolerance degree between these lexicons is embodied by the semantic tolerance relation model (STRM).

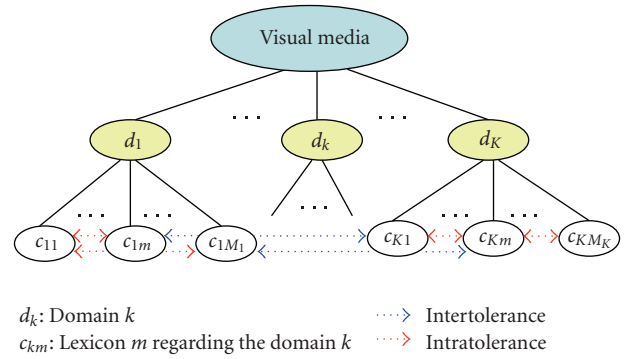


Figure 1: Form of STRM.

Furthermore, based on the experiment result depicted in Rogowitz [6] that people judge image similarity by semantic relevance (SR), following visual similarity (VS), the factor of SR is incorporated with the factor of VS in generating above associative values. Moreover, how to calculate associative values with defined lexicons by bidirectional associative memories (BAMs) is introduced. On the basis of generated associative values, a scheme of retrieving images according to multilexicons queries is proposed not only in the case of single domain but also in the case of cross domains. The influence of factors of SR and VS on the accuracy of retrieving images is analyzed, and the utility of the proposed approach in finding individual's expected images is investigated. The results show that combining SR with VS in generating the associative values improves the accuracy of retrieving images. Moreover, with the 40 generally defined lexicons, 82% target images are retrieved by multilexicons queries with 1.4 lexicons and 2.2 query times on average requested by 5 subjects.

2. Semantic Tolerance Relation Model

"A picture is worth of thousands of words." Indeed, the meanings of an image are diverse and ambiguous. In order to systematically describe the general meanings of images, we propose a semantic tolerance relation model (STRM) that reflects the tolerance degree between lexicons. The form of STRM is illustrated in Figure 1.

The meanings of images are described by many domains. However, it seems better to define the core domains for describing images' semantics. In views of the point that depicting news should include components of 5W1H (who, where, when, what, way, how), we define the following domains as the core domains. They are Nature versus manmade domain which represents concepts from nature to manmade regarding "what"; human versus nonhuman domain which represents the concepts from portrait, small face to nonface regarding "who"; temporal domain representing the time information regarding "when"; spatial domain representing the location information regarding "where"; action domain presenting the doing information regarding "way"; impression domain reflecting the impression information regarding "how." Of course, other new domains can be added as supplement domains.

For a certain domain d_k, concepts are depicted by some lexicons. Lexicon i regarding d_k is denoted as c_{ki}. The number of the lexicons in d_k is denoted as M_k. Similar to domains, we define general lexicons for a certain domain. It is taken into account that the words having the higher observed frequency counts within a very large text corpus are regarded as general lexicons. For example, these lexicons (landscape, tree, flower, beach, lake, mountain, sunset, building, building parts, clothing, furniture, kitchen items, tools, vegetables, vehicles) can be chosen as general lexicons regarding nature versus manmade domain. 3 lexicons (portrait, face, nonface) are selected as core lexicons regarding human versus nonhuman domain. However, it is obvious that the meanings of some defined lexicons are mutually tolerated, such as furniture and kitchen items. We define such overlapping of the meanings of two lexicons in the same domain as intratolerance, denoted as (c_{ki}, c_{kj}), and overlapping of two lexicons in the different domain as intertolerance, denoted as $(c_{ki}, c_{lj})(k \neq l)$. The rate of c_{ki} overlapped by c_{lj} is defined as the tolerance degree of c_{ki} to c_{lj}, denoted as $td(c_{ki}, c_{lj})$. It is assumed that the $td(c_{ki}, c_{lj})$ is represented by the co-occurrence count $cu(c_{ki}, c_{lj})$ of c_{ki} with c_{lj} within the very large text corpus, while it is in accordance with the counts giving the number of times of c_{ki} with c_{lj} as 2 grams appeared in a large corpus containing over a trillion total tokens. In particular, we can use the text corpus data provided by Google Inc. [19] to acquire such cooccurrence count. Let $cu\max$ give the maximal value of all cooccurrence counts regarding all defined lexicons, $td(c_{ki}, c_{lj})$ is calculated by (1), to make the values of tolerance degree be within the range of (1, 0):

$$td\left(c_{ki}, c_{lj}\right) = \frac{cu\left(c_{ki}, c_{lj}\right)}{cu\max}. \tag{1}$$

Therefore, STRM regarding the lexicons is expressed in a matrix TR^{kl}. When $k = l$, TR^{kk} expresses the intratolerance relation model of the lexicons regarding domain k. Otherwise, TR^{kl} expresses the intertolerance relation model of the lexicons regarding domains k and l. The entry at row i, column j in the matrix is the value of tolerance degree of lexicon c_{ki} to c_{lj}:

$$\begin{aligned} TR^{kl} &= \left(\mathrm{tr}_{ij}^{kl}\right) \\ &= \left[td\left(c_{ki}, c_{lj}\right) : i \in [1, M_k], cj \in [1, M_l]\right]. \end{aligned} \tag{2}$$

To add a new lexicon t in domain k, it must be registered firstly. However, the re-alignment of entries in the matrix TR^{kl} is accomplished by determining values of $(td(c_{ki}, c_{lt}) : i \in [1, M_k + 1]; t = M_k + 1; k, l \in [1, K])$, and $(td(c_{kt}, c_{li}) : i \in [1, M_k + 1]; t = M_k + 1; k, l \in [1, K])$. It means that all old entries are not needed to recompute in the realignment process, and only the new entries of tth row and tth column of domain k, which are the values of tolerance degree of the new lexicon t regarding the old lexicons, are required to determine. Accordingly, the realignment of entries in the matrix TR^{kl} caused by adding the new lexicon will not result in the expensive computation.

3. Representing Images' Meanings

3.1. Associative Values with Given Lexicons. Because most images have the multiple and ambiguous semantic interpretations and the human's criteria of judging images with the similar meanings are multiple, the meanings of each image is represented by a vector $\mathrm{img}_n = [A_1^n, \ldots, A_k^n, \ldots, A_K^n]$ of associative values with given lexicons, while $A_k^n = [a_{k1}^n, \ldots, a_{ki}^n, \ldots, a_{k,M_k}^n]$, which is a subvector of associative values with the lexicons regarding domain k. M_k denotes the number of given lexicons regarding domain k, and a_{ki}^n means the associative degree of the image n with the lexicon c_{ki}.

Based on the experiment described in [6], we see that people judge image similarity by semantic relevance (SR), following visual similarity (VS). Particularly, besides the images which are either semantically or visually similar, there are images that are semantically tolerant, but not similar visually, such as building and building parts. Also, there are images that are similar in shape, but not semantically tolerant, such as melon and ball. Accordingly, the associative values of image with the lexicons are affected by two facts: SR and VS. Let the value of SR regarding an image n to lexicon c_{ki} is denoted as sr_{ki}^n, and the value of VS is denoted as vs_{ki}^n. sr_{ki}^n embodies the degree of an image having the meaning of c_{ki}, and vs_{ki}^n reflects the degree of an image looking like c_{ki}. Therefore, the value of a_{ki}^n is generated by the weighted sum of sr_{ki}^n and a_{ki}^n, which is expressed by

$$a_{ki}^n = w_s sr_{ki}^n + w_v vs_{ki}^n, \tag{3}$$

where w_s is the weight of SR in generating a_{ki}^n, and w_v is the weight of VS.

Accordingly, if there are K domains, and M_k lexicons are given, the meanings of each image is represented by a vector

$$\begin{aligned} \mathrm{img}_n &= \left[A_1^n, \ldots, A_k^n, \ldots, A_K^n\right], \\ \text{while } A_k^n &= \left[a_{k,1}^n, \ldots, a_{k,i}^n, \ldots, a_{k,M_k}^n\right]. \end{aligned} \tag{4}$$

3.2. Calculating Associative Values. To fully specify the associative value within this computational modeling framework of (3), it is necessary to derive the values of sr_{ki}^n and the value of vs_{ki}^n. For this, the sufficient features are needed to extract from the images, and the detector of lexicon c_{ki} is needed to design. In fact, different detectors based on the different features perform similarly in terms of mean infAP (inferred average precision) on average [1]. So, considering the easily implementation, the pixel-based bidirectional associative memories (BAMs) [20] are used as the detector of lexicons to generating the values of sr_{ki}^n and vs_{ki}^n in this paper, and the performance of these generated values in representing the meanings of images is analyzed afterwards.

BAM is a two-layer network structure that maps specific input representations to specific output representations, but the connections between two layers are bi directional. It is a system that "associates" two patterns (X, Y) such that when one is encountered, the other can be recalled. Typically, X and Y are the vectors with the length of m and n, respectively. The structure of BAM utilized in our case is shown in Figure 2.

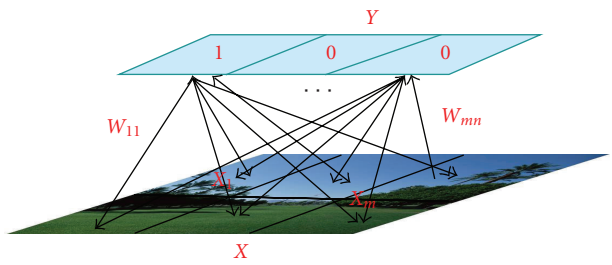


Figure 2: Structure of BAM.

The pattern X is a vector, the entries of which correspond to values of pixels of a learned image; the pattern Y is a vector regarding a code of lexicon representing the learned pattern X. In general, for a domain k, the lexicon c_{ki} is encoded as $Y_i = [0 \cdots 0 \underset{i}{1} 0 \cdots 0]_{M_k}$. Pattern X and pattern Y combine a pattern pair which is associated by building the connection weight matrix. The weight matrix for simultaneously storing several associated pattern pairs is calculated by

$$W = \alpha \sum_{k=1}^{K} X_k^T. \tag{5}$$

With continuously updating the outputs of X layer and Y layer for an input image n, the network will eventually converge to an energy local minimum. Then, the output values of units in the Y layer embody the recalling degree of the input image n to the learned image patterns. Let the output units of Y layer, that is, the recalling values, are expressed as

$$y_n = \left[r_1^n, \ldots r_i^n, \ldots r_{M_k}^n\right]. \tag{6}$$

Here, it assumes that the input image n is fully belonged to c_{ki}, if $r_i^n = \max\{r_m^n, m \in [1, M_k]\} \& r_i^n > T_r$ (T_r is a threshold of recalling values). Accordingly, sr_{ki}^n is set as 1 in this case. However, it is obvious that sr_{kj}^n will not be 0, if c_{ki} is relevant to c_{kj}. The value of sr_{kj}^n is affected by the tolerance relation of c_{ki} with c_{kj}. Therefore, the value of sr_{ki}^n is determined according to the following rules ($i, j \in [1, M_k]$):

$$\begin{aligned} &\text{if } \text{img}_n \in c_{ki}, \quad sr_{ki}^n = 1, \\ &\text{if } \text{img}_n \in c_{ki}, \quad sr_{kj}^n = tr_{ij}^{kk}, \\ &\text{if } \text{img}_n \notin c_{ki}\ (\forall i), \quad sr_{ki}^n\ (\forall i) = 0. \end{aligned} \tag{7}$$

On the other hand, for the pixel-based BAM, the value of r_i^n particularly reflects the visual similarity degree of the input image n to the learned pattern image i, which represents the concept of lexicon c_{ki}, and vs_{ki}^n is given by

$$vs_{ki}^n = r_i^n. \tag{8}$$

As a whole, (2) generating associative values of an image with given lexicons is converted as

$$a_{ki}^n = \begin{cases} w_s + w_v r_i^n, & \text{if } \text{img}_n \in c_{ki}, \\ w_s \text{tr}_{ji}^{kk} + w_v r_i^n, & \text{if } \text{img}_n \in c_{kj},\ j \neq i, \\ w_v r_i^n, & \text{if } \text{img}_n \notin c_{ki}\ (\forall i), \end{cases} \tag{9}$$

Some learned pattern images which represent the meanings of given lexicons regarding the domain of nature versus manmade are shown in Figure 3. The rule of selecting such images follows that visual features of the image representing one lexicon perfectly embody the general features of this lexicon. For some concepts, more than one pattern images which are obviously different in visual sensation are learned for one lexicon. For example, four pattern images are learned for lexicon tree. The resolution of them is set as 96×64. So, the relative capacity of BAM recalling pattern pairs is $96 \times 64 \times 0.1998 = 1227$ pairs [20]. On the other hand, according to the statement in [1] that about 40 to 240 lexicons are needed to define for perfectly detecting the concepts carried by images, the capacity of the constructed pixel-based BAM capacity is pretty sufficient in calculating the recalling values r_i^n, which are used to generate the associative values with given lexicons by (9), if the number of given lexicons for a domain is in the range of (40, 240).

3.3. Scheme of Retrieving Images. The scheme of image retrieval based on the vectors of associative values of images with given lexicons is considered as the following.

3.3.1. Retrieval Regarding Single Domain. If the lexicons regarding the other domain are strongly associated with the current one, considering the association between these lexicons will improve the performance of retrieval. Therefore, the retrieving scheme based on the given lexicons is expressed as

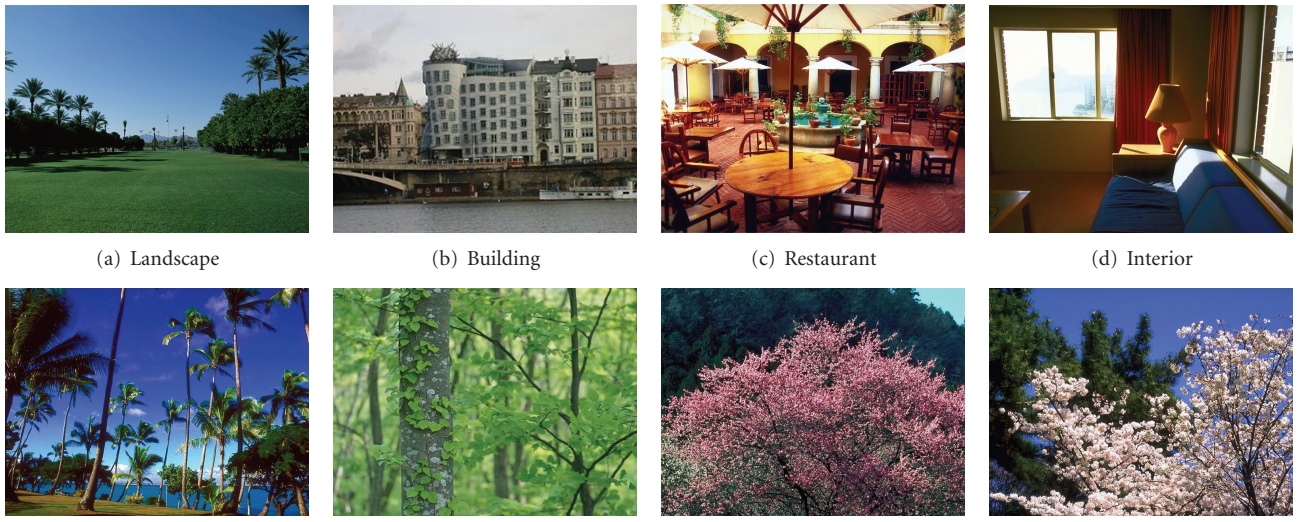
$$\begin{aligned} &\text{if } \text{tr}_{ij}^{kl} > T_{ij}^{kl} (\forall k, l \in [1, K], k \neq l) \\ &\text{IMG}_{ki} = \left\{\text{img}_n \in c_{ki}\right\} \\ &\qquad = \left\{\text{img}_n, a_{ki}^n \geq \sigma_{ki} \ \&\ a_{lj}^n \geq \sigma_{lj}\right\} \\ &\text{else } \text{IMG}_{ki} = \left\{\text{img}_n \in c_{ki}\right\} \\ &\qquad = \left\{\text{img}_n, a_{ki}^n \geq \sigma_{ki}\right\}, \end{aligned} \tag{10}$$

where T_{ij}^{kl} is called the tolerance value that is threshold of tolerance degree between two lexicons i and j regarding domains k and l. σ_{ki} or σ_{lj} are called the retrieval values that are the threshold of extracting image belonging to the lexicon i regarding k, or lexicon j regarding l, respectively.

Moreover, if a query cannot be depicted explicitly by the given lexicons, the user can specify the query to multiple related lexicons. Let the query be denoted as q_m:

$$\begin{aligned} &\text{if } q_m \in c_{ki}, \quad q_m \in c_{kj}, \\ &\text{IMG}_{q_m} = \left(\left\{\text{img}_n \in c_{ki}\right\} \bigcap \left\{\text{img}_n \in c_{kj}\right\}\right). \end{aligned} \tag{11}$$

3.3.2. Retrieval Regarding Cross-Domains. Obviously, the image retrieval regarding cross-domains is the intersection of extracted images regarding the different lexicons with the different domains. Therefore, the scheme of retrieving image



(a) Landscape (b) Building (c) Restaurant (d) Interior

(e) Tree (f) Tree (g) Tree (h) Tree

Figure 3: Some learned pattern images.

(a) (b) (c)

Figure 4: Images retrieved by furniture with face.

that carries the meanings of lexicon i regarding domain k and lexicon j regarding domain l is given by

$$\text{if } \mathrm{tr}_{ij}^{kl} \neq 0, \quad \mathrm{tr}_{ji}^{lk} \neq 0,$$
$$\mathrm{IMG}_{ki,lj} = \left\{\mathrm{img}_n, \mathrm{img}_n \in c_{ki}\right\} \bigcap \left\{\mathrm{img}_n, \mathrm{img}_n i \in c_{lj}\right\}. \tag{12}$$

Some images retrieved by the query of furniture with face are shown in Figure 4.

3.4. Implementation. For implementation, three domains together with corresponding 40 lexicons as general lexicons are defined. They are listed in Table 1. These lexicons are selected according to the observation that the frequency counts of these lexicons appeared in the large text corpus [19] are higher. For each domain, pixel-based BAM is separately constructed with its given lexicons as Y patterns and the corresponding representative images as X pattern pairs. Using (9), associative values with these lexicons are calculated and stored as index for retrieval. The users select one or more lexicons representing the semantics which they want to express to retrieve their expected images or keyframes of videos.

Figure 5 is a screenshot of system interface. At the top, the pull-down menus are provided for the users to select related domains and lexicons embodying the meanings of images which are expected to search. The sliding is used to input the threshold of associative values regarding a given lexicon for retrieving. It would be operated easily because the user can set the sliding position optionally and adjust that to left or right if the number of retrieved images is too few or too many. The middle space shows the inputted queries which can be edited by the buttons of Add, Delete, and Reset. The thumbnails of retrieved images or keyframes are listed at the bottom, from which users can find their expected images or videos easily.

4. Experiments and Analysis

400 images randomly selected from the personal albums and Sozaijiten Image Book 1 [21] and 1220 keyframes extracted from Video Traxx 1 [22] are prepared for testing the system performance. However, the images relevant to the action,

Table 1: Defined domains and lexicons.

Domains	Lexicons
Nature/manmade	Ground, landscape, flower, tree, light, building, sunset, restaurant, texture_brick, texture_wood, texture_paper, texture_cloth, food_materials, food, mall, beach, snow_viewing, mountain_lake, flower_arrangement, interior, street, waterfall, mountain, pool, guard, ship, board, rocket, brriage, factory
Face/nonface	Portrait, face, non_face
Impression	Active, cool, calm, soft, hard, grayish, clear

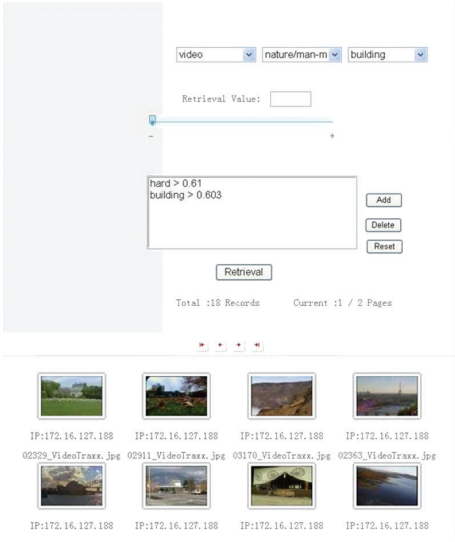


Figure 5: System interface.

such as sports, football, and camping, are not included because the action domain relevant to these lexicons is not defined for implementation.

4.1. Per-Lexicon Retrieval Accuracy. The per-lexicon retrieval accuracy is defined as the following expression (# stands for "number of")

$$\text{accuracy} = \frac{\#\text{true positive}}{\#\text{top ranking}}, \tag{13}$$

where # top ranking denotes the number of top ranking images according to their associative values with a certain lexicon. # true positive denotes the number of images which are positive to the relevant lexicon among the top ranking images.

Figures 6(a), 6(b), and 6(c) show the influence on the accuracy regarding the lexicons building, furniture, and landscape with varying the weights of SR and VS, while the number of top ranking images is varied at the range of (12, 120).

The case of $w_s = 0, w_v = 1$ indicates that the factor of VS is merely considered in generating the associative values of images with defined lexicons; the case of $w_s = 1, w_v = 0$ indicates that the factor of SR is only taken into account; the case of $w_s = 0.25, w_v = 0.75$ indicates that both of SR and VS are used in calculating associative values. From Figures 6(a), 6(b), and 6(c), it is observed that although the accuracy is different for the different lexicons the case of $w_s = 0.25, w_v = 0.75$, that is, the case of considering both of SR and VS in generating associated values with the lexicons, has the best retrieval performance as a whole. The accuracy of it is often highest among the above three cases, and the varying tendency of accuracy with changing the number of top ranking images is more gentle than those of the other cases. In the case of $w_s = 0, w_v = 1$, that is, the case of only VS being considered, the performance of retrieval accuracy drops a little as a whole. However, in the case of $w_s = 1, w_v = 0$, that is, the case of only SR being considered, the performance of retrieval accuracy becomes bad obviously (the reason for the number of top ranking does not getting to 120 is that the number of images having the nonzero associative values with the queried lexicon does not reach 120). From the results, we can see that the performance of retrieval accuracy is improved if both of factors SR and VS are taken into account in generating the associative values with defined lexicons. On the other hand, the case of only one factor VS is considered does not influence the retrieval accuracy so much. However, that only one factor SR is processed in generating the associative values with defined lexicons is not a good means for representing the meanings of images.

On the other hand, Table 2 shows the accuracy of the top 36 retrieved images regarding some implemented lexicons while $w_s = 0, w_v = 1$,$w_s = 0.25, w_v = 0.75$, or $w_s = 1.0, w_v = 0$.

From Table 2, we can see that although the accuracy to the different lexicons is varied, it is obvious that the accuracy is highest for all the lexicons when $w_s = 0.25, w_v = 0.75$. That is, the method of considering both of SR and VS in generating the associative values to represent the meanings of images is effective and improves the retrieving accuracy of images well.

On the other hand, besides using BAM to generate the values of SR and VS of images, there are other methods to use in the literature for this purpose, especially, the statistical modeling approach proposed in [18]. However, the focus of this paper is to verify that considering the semantic tolerance relation in representing the meanings of images will improve the performance of image retrieval while that is not proposed in the literature yet and not to verify that using BAM is better than using other methods in image retrieval. So the BAM for which it is easy to implement is utilized, and the performance comparison to the other methods in the literature is not done in this paper.

4.2. Effectiveness of Multilexicons Queries for Retrieval. Five students as subjects attend the experiment, and each

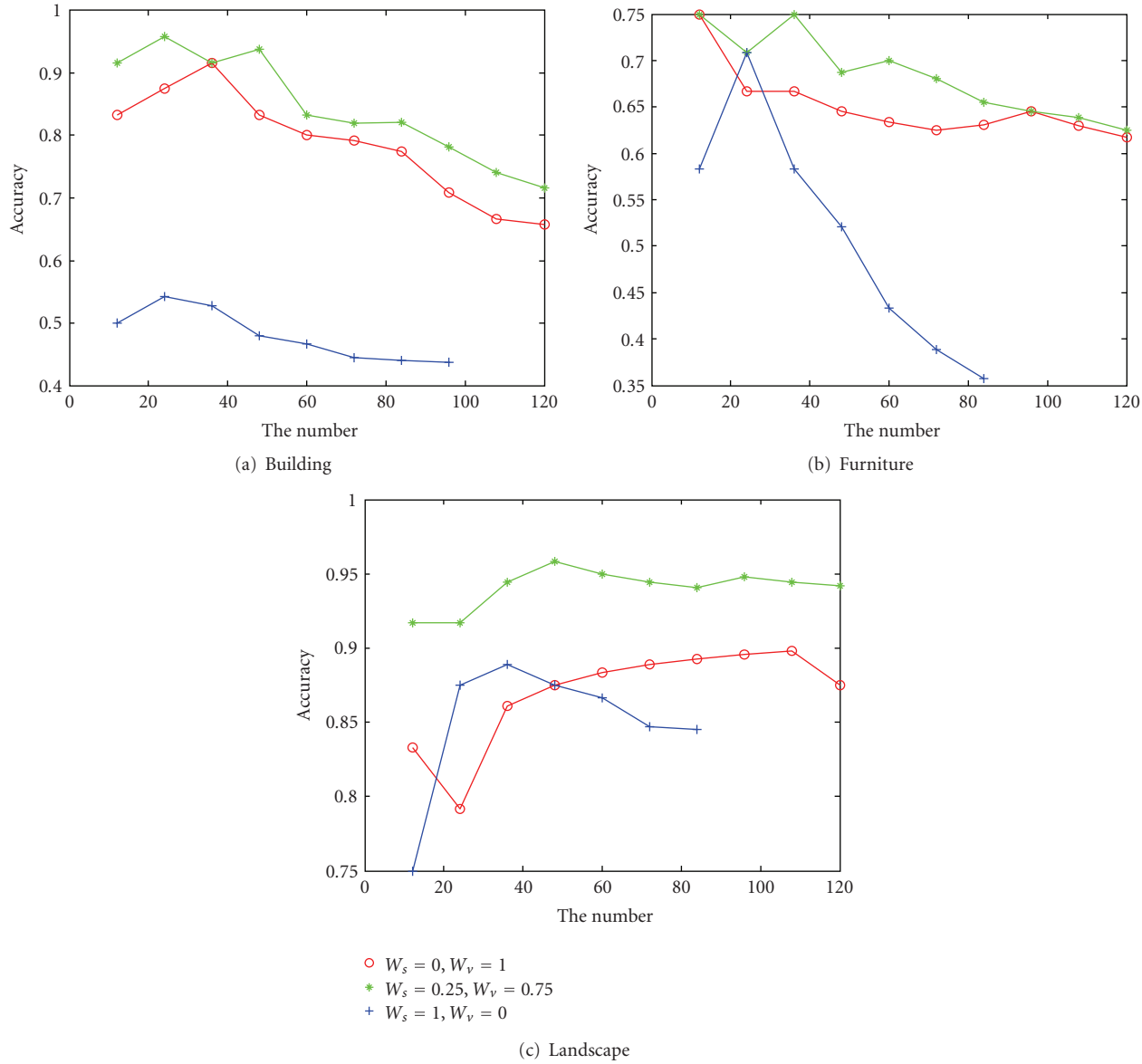


Figure 6: Accuracy with varying weights of SR and VS.

Table 2: Accuracy of retrieved images.

W_s	Landscape	Building	Furniture	Flower	Tree	Sunset	Restaurant	Food	Snow	Beach	Mountain	Street	Face
0	0.86	0.9	0.66	0.64	0.72	0.53	0.72	0.58	0.78	0.83	0.81	0.66	0.81
0.25	0.95	0.9	0.75	0.78	0.83	0.66	0.86	0.66	0.83	0.92	0.89	0.72	0.89
1.0	0.88	0.53	0.58	0.58	0.61	0.5	0.64	0.56	0.72	0.75	0.75	0.69	0.78

experimenter is asked to randomly target some images from the test sets for retrieving. For the experiments, the experimenters specify domains and lexicons according to meanings of a target image and investigate whether this image is retrieved by querying such domains/lexicons. The threshold of associative values for retrieving a lexicon (called retrieval values afterwards) can be adjusted in the query process. The whole testing process is showed in the following. QT_i denotes the number of query times for retrieving the target image i.

(1) For a target image i, selecting a lexicon form the given domains/lexicons, which is mostly considered to represent the meanings of the image, and setting the retrieval value for this lexicon.

(2) Retrieving and investigating whether the target image or images which are similar to the target image in semantics and visual sensation are retrieved. If there are no corresponding images in the top 24 listed images,

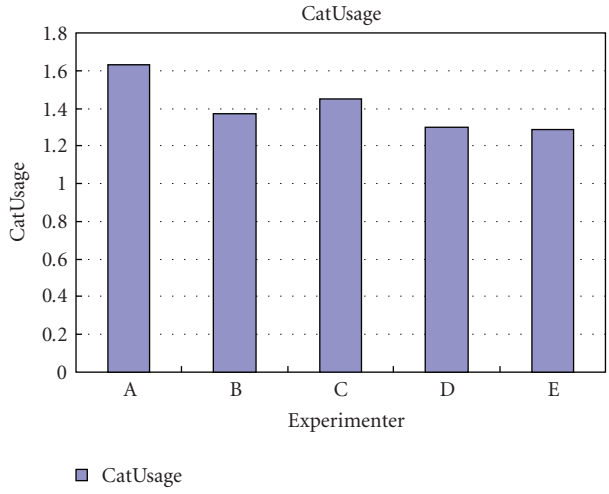


Figure 7: CatUsage of five experimenters.

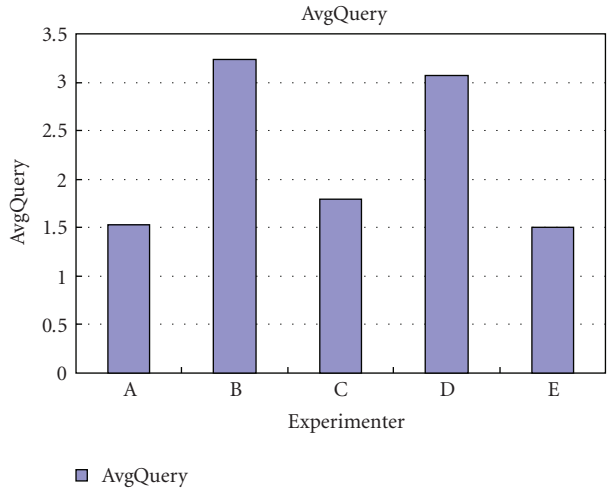


Figure 8: AvgQuery of five experimenters.

(i) picking up a new lexicon from the given domains/lexicons to add to the query condition, setting the retrieving value for this lexicon, and going to step (2), if the amount of listed images is larger than 24. Here, let $QT_i = QT_i + 1$,

(ii) deleting the latest added lexicon, and going to step (2), if the amount of listed images is smaller than 24. Here, let $QT_i = QT_i + 1$.

(3) Ending the retrieving process, if the target image is found on the top 24 listed images or the QT_i reach 5. The experimenter records the status whether the target image is found (successful retrieval (SRL)), the number of queried lexicons for retrieving this target image (AQC_i) and QT_i.

In this test, each experimenter is required to target 30 images for retrieving and investigate whether these images can be detected in the above retrieving process.

Three measurements are gotten from the above process. They are the average value of utilized lexicons in the image retrieval process to each experimenter (CatUsage), average query times for retrieving target image to each experimenter (AvgQuery), and the detection rate of target images to each experimenter (Detection), which are expressed in the following:

$$\begin{aligned} \text{CatUsage} &= \frac{\sum_{i=1}^{M} AQC_i}{\text{amount of SRL}}, \\ \text{AvgQuery} &= \frac{\sum_{i=1}^{M} QT_i}{M}, \\ \text{Detection} &= \frac{\text{amount of SRL}}{M}, \end{aligned} \tag{14}$$

where M denotes the number of target images. Here, $M = 30$. If ith image is not detected, AQC_i is assigned to 0.

The CatUsage embodies the utilized frequency counts of lexicons in the query process. The lower value the CatUsage has, the more general the defined lexicons are. The AvgQuery embodies the average query times of finding target images in retrieving. The smaller value means that the images can be searched more easily. The Detection embodies the successful rate of finding target images. The higher value means that the retrieving performance is good.

Figure 7 shows the values of CatUsage regarding 5 experimenters. The average value is 1.41, and the standard deviation is 0.126. From the results, we observed that the average amount of utilized lexicons was less than 2 in the query process for finding the target images, while the deviation of the amount of utilized lexicons is small. It means that the above given lexicons are efficient and effective for retrieving images with general meanings.

Figure 8 shows the average query times of retrieving target images to five experimenters. The values vary in the range of (1.5, 3), and the average value is 2.23. These values show that almost target images can be found by querying lexicons twice or thrice. The query times are not so large that it can satisfy users' retrieving request. It also shows that the defined lexicons are efficient and effective for retrieving images with general meanings.

Figure 9 shows the detection rate of target images in the image retrieval experiments to 5 experimenters. The values vary from 63% to 97%, and its average value is 82%, which means that more than 80% of target images are successfully retrieved with less than 5 query times on average, although the subjective behaviors and criteria of individuals for retrieving somewhat effect the retrieval results.

As a whole, the above experiment results show that the proposed method of representing the images' meanings by the associative values with the given lexicons for retrieval is with generality and being able to absorb the effect of individuals' criteria in annotating the meanings of images. The proposed method is efficient and effective for retrieving images with general meanings.

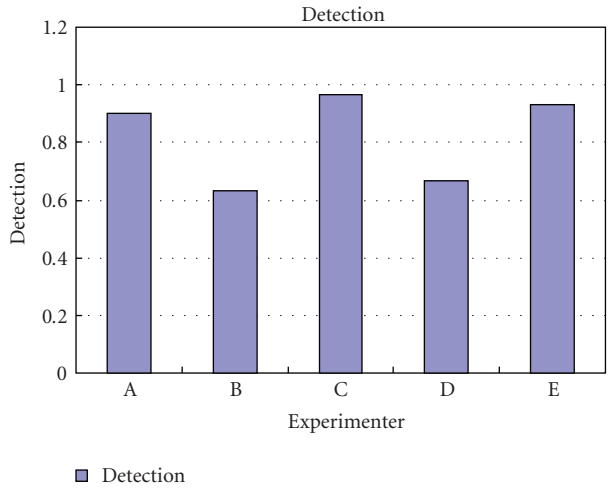


Figure 9: Detection of five experimenters.

5. Conclusion

In this paper, in order to systematically describe the meanings of images, semantic tolerance relation model (STRM) that reflects the tolerance degree between lexicons regarding different domains is proposed and specified by using the observed frequency counts of 2 grams in the large text corpus. Considering the influence of semantic relevance (SR) and visual similarity (VS) on interpreting the meanings of images, the images are represented by the associative values with the given lexicons, while such associative values are affected by the factors of SR and VS. Moreover, the associative values are calculated by using pixel-based bidirectional associative memories (BAMs) and incorporating with STRM, which are easy in implementation compared with the methods such as [17] based on the Bayesian statistical model. Using the associative values, the scheme for retrieving images by lexicons was proposed. Finally, the influence of involving the factors of SR and VS in generating associative values was tested, and the effectiveness of proposed method is examined by retrieving target images queried by 5 subjects. The accuracy analysis of retrieving images showed that considering the factors of SR and VS in generating the associative values improves the accuracy of retrieving images. On the other hand, 82% target images are retrieved by multilexicons queries with 1.4 lexicons and 2.2 query times on average by 5 subjects.

For the future work, the influence of learning pattern images' selection for representing the meanings of given lexicons on retrieval performance is needed to be analyzed. On the other hand, how to regulate the number of given lexicons in constructing BAM according to the context of actual images set is required to be explored . Further, the pixel values of images were used as learned patterns in X layer of BAM in this paper. However, the influence of other types of learned patterns, such as the feature vector of color, shape, and texture, on the accuracy of retrieving images is needed to be analyzed .

On the other hand, although the experiment results regarding users' usability in Section 4.2 somewhat show the efficiency and effectiveness of the proposed approach, the system interface of querying images may give rise to the inconvenience for the user to determine the concepts from the given lexicons when the number of them increases. Accordingly, how to design the more convenient interface for retrieving images based on the proposed mechanism in this paper is also required to consider in the future.

Acknowledgments

This work is supported by JSPS KAKENHI, Grant-in-Aid for Scientific Research, 19500175, and the research funds of Iwate Pref. University. The author would like to thank Professor Shaozi Li and Feng Guo for their cooperation in system implementation and experiments.

References

[1] L. Xie, R. Yan, and J. Yang, "Multi-concept learning with large-scale multimedia lexicons," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '08)*, pp. 2148–2151, October 2008.

[2] T. Gevers, "Color constant ratio gradients for image segmentation and similarity of texture objects," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. I, pp. 8–25, Hawaii, USA, December 2001.

[3] W. Y. Ma and H. J. Zhang, "Content-based image indexing and retrieval," in *Handbook of Multimedia Computing*, CRC Press, New York, NY, USA, 1999.

[4] Y. Rui and T. Huang, "Optimizing learning in image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, pp. 236–243, June 2000.

[5] X. S. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using BiasMap," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, pp. I11–I17, Urbana, Ill, USA, December 2001.

[6] B. E. Rogowitz, "Perceptual image similarity experiments," in *Human Vision and Electronic Imaging III*, vol. 3299 of *Proceedings of SPIE*, San Jose, Calif, USA, January 1998.

[7] A. Mojsilović, J. Hu, and E. Soljanin, "Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis," *IEEE Transactions on Image Processing*, vol. 11, no. 11, pp. 1238–1248, 2002.

[8] J. Vogel and B. Schiele, "Performance prediction for vocabulary-supported image retrieval," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '01)*, October 2001.

[9] A. Mojsilović, J. Gomes, and B. Rogowitz, "Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues," *International Journal of Computer Vision*, vol. 56, no. 1-2, pp. 79–107, 2004.

[10] A. Wardhani and T. Thomson, "Content based image retrieval using category-based indexing," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '04)*, pp. 783–786, Taipei, Taiwan, June 2004.

[11] Y. Dai and D. Cai, "Imagery-based digital collection retrieval on web using compact perception features," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '05)*, pp. 572–576, September 2005.

[12] M. Boutell and J. Luo, "Beyond pixels: exploiting camera metadata for photo classification," *Pattern Recognition*, vol. 38, no. 6, pp. 935–946, 2005.
[13] B. Bradshaw, "Semantic based image retrieval: a probabilistic approach," in *Proceedings of the 8th ACM International Conference on Multimedia*, pp. 167–176, New York, NY, USA, November 2000.
[14] J. Yu and Q. Tian, "Toward intelligent use of semantic information on subspace discovery for image retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '06)*, pp. 293–296, Toronto, Canada, July 2006.
[15] X. Shen, M. Boutell, J. Luo, and C. Brown, "Multi-label machine learning and its application to semantic scene classification," in *Storage and Retrieval Methods and Applications for Multimedia*, Proceedings of SPIE, January 2004.
[16] Y. Dai, "Class-based image representation for Kansei retrieval considering semantic tolerance relation," *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, vol. 21, no. 2, pp. 184–193, 2009.
[17] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.
[18] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," in *Proceedings of the IEEE Transactions on Pattern Recognation and Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, September 2003.
[19] B. Thorsten and A. Franz, "Web 1T 5-gram. Linguistic data consortium," 2006, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T13.
[20] B. Kosko, "Bidirectional associative memories," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, no. 1, pp. 49–60, 1988.
[21] "Sozaijiten image book 1," Datacraft Co.,Ltd.
[22] Video Traxx 1, "Film & video library," Digital Juice.