Semantic Tolerance Relation-based Image Representation and Classification

Ying Dai
Iwate Pref. University
dai@iwate-pu.ac.jp

Abstract

The nature of the concepts regarding images in many domains are imprecise, and the interpretation of finding similar images is also ambiguous and subjective on the level of human perception. To solve these problems, in this paper, images’ semantic categories and the tolerance degree between them are defined systematically, and the approach of modeling tolerance relations between the semantic classes is proposed. Furthermore, for removing the induced false tolerance in the produce of using semantic tolerance relation model, the method of un-tolerating is introduced in image representation. We apply the proposed approach to the representations of images regarding the nature vs. man-made domain, human vs. non-human domain, and temporal domain, and compare the categorization results of them with the results not using semantic tolerance relation model. The results show the effectiveness of proposed method.

1. Introduction

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. Usually relevance feedback is the only attempt to close the semantic gap between user and system.

There has been some research which is to reduce the semantic gap between users and retrieval systems with the different levels of abstraction employed by human and machine. In [1], how human observers judge image similarity was analyzed to reach a conclusion that the human observers are very systematic to judge image similarity, following semantic, color, and structural characteristics. Following this conclusion, in [2], 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 [5], a semantic-friendly query language for and searching diverse collections of images was proposed. However, the query language such as «nature<10 && contrast>800» is not easy to utilize for modeling the categories. In [3], a scheme of image retrieval system, which followed the human perceptual similarity criteria described in [1], was proposed. However, the semantic categorization of images was made manually. In [6][7][9], 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-non-sunset) were handled. In [8], a framework to handle the classification problem of overlapped classes was presented and was applied to the problem of multi-scene classification. However, the extension to the other concepts’ classification was not described. In [10], image categorization regarding multi-
domain considering the tolerance between classes was proposed. However, how to remove
the induced false tolerance was not mentioned.

In this paper, in order to solve the problems that the concepts regarding images in
many domains are imprecise and can be overlapped, and also there is a gap between the
low-level features and the high level concepts of human perception, we give a
systematic definition of semantic categorization for images, and propose a method of
semantic tolerance relation model-based image representation and categorization. In
details, the tolerance relations between the semantic categories are modeled and the
tolerance degrees between them are used to represent each image in a weight vector
regarding the classes. How to gain the tolerance degree is also described by using
training images set. Furthermore, in order to eliminate the influence of false tolerance
induced by the training images set, the method of un-tolerating in the procedure of
representing images is introduced. According to the threshold of tolerance degree, set of
images sharing similar meanings are grouped together to assign to the specific
categories.

2. Semantic tolerance relation model

As we know, based on the perceptual image similarity experiments in [1], human observers
are very systematic to judge image similarity, following semantic, color, and structural
characteristics. Furthermore, the images can superimpose onto the space composed of two
axes, nature vs. man-made axis and human vs. non-human axis according to the semantics.
Accordingly, it is the first step for retrieving images to categorize images with the same or
similar semantics. As mentioned above, because the concepts of images in many domains are
imprecise, and the interpretation of finding similar images is ambiguous and subjective on the
level of human perception, we define the semantic categories of images together with the
tolerance degrees between them as shown in Fig. 1.

![Fig. 1 Semantic categorization](image-url)

Images are described by different domains. Nature vs. man-made axis or the human vs.
non-human axis can be considered as the one of the domains. For other domains, such
temporal domain, special domain, and impression domain are also included. Of course, a new
domain can be added. For a certain domain \( d_k \), concepts are divided into some classes, the
one of which is denoted as \( c_{km} \). The number of the classes in it is denoted as \( M_k \).

The class may be associated with the other in a domain \( d_k \). For example, the class of
landscape is strongly associated with the class of tree and plant; interior is done with the
pattern of wood. Such classes are called intra-associated classes of current class, denoted
as \((c_{ki}, c_{kj})\), where \(i\) denotes the current category, and \(j\) does its associated categories. Similarly, the class may be also associated with the other in the different domains. For example, the sunset is strongly associated with the time 5pm to 7 pm. Such classes are called inter-associated classes of the current class, denoted as \((c_{ki}, c_{lj})\), where \(k\) denotes the current domain, and \(l\) does its associated domain. The intra-associated or inter-associated classes are overlapped on portion. The overlapped images have both the meaning of \(i\) regarding domain \(k\) and the meaning of \(j\) regarding domain \(k\) or \(l\).

To measure the degree of tolerance regarding \((c_{ki}, c_{kj})\) or \((c_{ki}, c_{lj})\), we conduct the class co-occurrence matrix as
\[
CO^{kl} = [\text{co}_{(k,j)(i,j)}]_{M_k \times M_i} = [n(c_{ki}, c_{lj})]_{M_k \times M_i},
\]
where \(n(c_{ki}, c_{lj})\) is a function of the number of images in the collection while these images are both assigned to class \(i\) regarding domain \(k\) and class \(j\) regarding domain \(l\). \(k=l\) means the domain is same. In details, \(IMG_{ki} = \{img_{ki}^1, ..., img_{ki}^{N_k}\}\) be a set of training images of class \(i\) in domain \(k\). \(n(c_{ki}, c_{lj})\) can be obtained by counting the number of the images which are tolerant to belong to class \(j\) in domain \(l\).

Then, the tolerance degree \(td(\cdot, \cdot)\) between two classes, which can be in the same domain or not, is defined by the following expression (\# stand for number of).
\[
\text{td}(c_{ki}, c_{lj}) = \text{td}_{(k,j)(i,j)} = \begin{cases} 1, & \text{if } k = l, \text{and } i = j \\ \frac{\text{co}_{(k,j)(i,j)}}{\# IMG_{ki}}, & \text{if } i \neq j \end{cases}
\]

The larger the \(td(c_{ki}, c_{lj})\) or \(td(c_{kj}, c_{lj})\), the stronger the \(c_{ki}\) is associated with the \(c_{kj}\) or \(c_{lj}\).

Let tolerance relation \(R\) between the two classes be defined as \((k,i)R(l,j) \iff \text{td}_{(k,j)(i,j)}\). The tolerance relation model regarding the classes can be expressed in a matrix \(TR^{kl}\). when \(k=l\), \(TR^{kl}\) expresses the intra-tolerance relation model of the classes in domain \(k\). Otherwise, \(TR^{ki}\) expresses the inter-tolerance relation model of the classes regarding domains \(k\) and \(l\).
\[
TR^{kl} = [\text{td}_{(k,j)(i,j)}]: i \in [1, M_k], j \in [1, M_i]
\]

As demonstration, for the nature vs. man-made domain, 7 classes are prepared. They are patterns of fabrics & woods & papers (\(c_{11}\)), landscape (\(c_{12}\)), flower arrangement (\(c_{13}\)), tree & plant (\(c_{14}\)), sunset (\(c_{15}\)), building (\(c_{16}\)), and food & tableware (\(c_{17}\)). Of course, the categories can be re-defined, or the new categories can be added, if necessary. Except \(c_{16}\), 1300 training images for the other 6 classes are selected from the particular volume of the Sozaijiten Image Book 1, published in Japan. The 200 training images for \(c_{16}\) are selected from Web. Then, each image is assigned to the appropriated classes by the subjects, according to its semantics. More than one class is permitted to assign. Accordingly, on the basis of (1), the intra-tolerance relation model \(TR^{11}\) is gotten by collecting the means of \(\text{co}_{(1,0)(1,j)}\) regarding the above 7 classes. The result is represented in expression (3).
Similarly, for the human vs. non-human domain where have 3 classes: portrait, face, and non-face, the intra-tolerance relation model $TR^{22}$ is obtained as the following.

$$TR^{22} = \begin{bmatrix}
1.00 & 0.80 & 0.00 \\
0.10 & 1.00 & 0.00 \\
0.00 & 0.00 & 1.00
\end{bmatrix}$$

As for the inter-tolerance relation model, as an example, the $TR^{13}$ is presented here, which is the one of nature vs. man-made domain $d_1$ and temporal domain $d_3$. Because the images of sunset are almost taken during 5pm to 7 pm, the inter-tolerance relation model of $d_1$ and $d_3$ is $TR^{13}$ are written as

$$TR^{13} = \begin{cases}
0 & \text{if } 5 \leq t < 7, \text{ other } \text{ classes}
\end{cases}$$

Where, $M$ indicates the number of the classes in $d_1$, $N$ indicates the number of the classes in $d_3$, $\alpha$ denotes the duration of sunset.

To add a new class $t$ in domain $k$, it must be registered firstly. The re-alignment of tolerance relation model $TR^{kl}$ is accomplished by collecting a set of training images of the new class and calculating $\{t_{d_1(k,j)}i, j \in [1,M_k+1], t = M_k + 1, k, l \in [1,K] \}$, and $\{t_{d_3(k,j)}i, t \in [1,M_k+1], t = M_k + 1, k, l \in [1,K] \}$ . The classes are also scanned to determine if they need to be combined, separated, or removed.

3. Image representation and categorization

On the basis of semantic tolerance relation models, the semantics of each image is represented by a weight vector $W = [W_1, W_2, ..., W_K]$, while, $W_k = [w_{ij}]_{M_k}$. The element $w_{ij}$ is a function of $t_{d_1(k,j)}i$ regarding the $TR^{kl}$, which indicates the weight of the class $i$ in the domain $k$ to the image.

Furthermore, because the decision of training images for classes $c_k$ strongly affects the tolerance degrees, $t_{d_1(k,j)}i$ between the two classes, the un-tolerating is conducted in the image representation, in order to remove the induced false tolerance. As the whole,

$$w_{ij} = \begin{cases}
(1 - \lambda_{k,j})t_{d_1(k,j)}i, & \text{if } \text{img}_i \in c_k \\
\lambda_{k,j}t_{d_1(k,j)}i, & \text{if } \text{img}_i \in c_j, j \neq i
\end{cases}$$

where, $\lambda_{k,j} \in [0,1]$, and $\lambda_{k,j} \in [0,1]$ are the un-tolerance parameters regarding $t_{d_1(k,j)}i$, and $t_{d_1(k,j)}i$, which are used to eliminate the influence of the false tolerance
caused by the training image sets. The smaller \( \lambda_{(k,j)(k,l)} \) means that the association of the image that is assigned to the class \( i \) to the other class is smaller. The smaller \( \lambda_{(k,j)(k,l)} \) means that the association of the image that is assigned to the class \( j \) to the class \( i \) is smaller. How to determine the values of un-tolerance parameters is based on the performance of the utilized classifier, which will be introduced in section 5, automatic image categorization. That \( \{\lambda_{(k,j)(k,l)} = 0, \text{for all } k,l,i,j\} \) means the semantic tolerance relation model is withdrawn.

Accordingly, if there are \( K \) domains, each image is represented by the following expression:

\[
img_s = [W'_1, W'_2, \ldots, W'_k] \tag{7}
\]

Then, the semantic-based image categorization is considered as the following.

- For the image categorization regarding single domain
  
  If the classes in the other domain are strongly associated with the current class, considering the association between these classes will improve the performance of image categorization. For example, the images of sunset are mostly taken from 5pm to 7 pm. If considering this condition in categorizing images regarding sunset, the retrieval performance will be improved. Therefore, the categorizing algorithm is expressed as

\[
\text{if } d_{(k,j)(l,j)} > T_{uj} (\forall k,l \in [1, K], k \neq l)
\]

\[
IMG_{su} = \{\{img_s \in c_{u}\} \tag{8}
\]

\[
= \{\text{img}_s, \quad w'_i \geq \sigma^i_{s} \land w'_j \geq \sigma^j_{s}\}
\]

\[ \]

\[
\text{else}
\]

\[
IMG_{su} = \{\text{img}_s, \in c_{u}\}
\]

\[ 
= \{\text{img}_s, \quad w'_i \geq \sigma^i_{s}\}
\]

where, \( \sigma^k_i \) denotes the threshold of extracting images belonging to the class \( i \) in domain \( k \), and \( \sigma^l_j \) does the threshold of extracting images belonging to the class \( j \) in domain \( l \).

- For the image categorization regarding cross-domains
  
  Obviously, the image categorization of the cross-domains is the intersection of extracted images in the different groups which are corresponded to the classes in different domains. The intersection of images categorized to the different classes that cause the contradictory is not considered. Then, the retrieval algorithm is expressed as

\[
\text{IMG}_{su,y} = \{\text{img}_s, \text{img}_s \in c_{u}, d_{(k,j)(l,j)} \neq 0\}
\]

\[
\bigcap \{\text{img}_s \in c_{y}, d_{(l,j)(k,j)} \neq 0\} \tag{10}
\]
4. Automatic image representation

In order to realize the image automatic semantic representation by a weighted matrix of the classes based on the semantic tolerance relation model, it is necessary to extract the low-level features of images, design the classifiers, and classify each image using the designed classifier. Then, according to the classification result, each image is semantic-represented by a weight vector of the classes on the basis of (6).

However, for the different domains, the utilized methods of extracting low-level features and classifying images are always different. As demonstration, we mainly introduce how to represent images in nature vs. man-made domain $d_1$, human vs. non-human domain $d_2$, and the temporal domain $d_3$.

4.1. Image representation regarding $d_1$

Although any algorithms for the feature extraction and the classification can be used, we use the SGLD matrices of H, S, V components to extract features, and Bayesian classifier to classify each image. The detailed method was introduced in [10].

Here, how to classify images to the aforementioned 7 categories is simply explained as the following. $\text{img}_{f_1}$ indicates the image with the feature vector

$$f_1 = [h_1, h_2, p_{j_1}, p_{j_2}, d_1, d_2, t_1(10^\circ), t_1(1, 45^\circ), t_1(1, 90^\circ)] .$$

Let

$$\prod_{k \in \text{defined categories}} p(h / c_i) \cdot p(p_{j_1} / c_i) \cdot p(p_{j_2} / c_i) = p(f_1 / c_i) ,$$

if $t_1(1, 45^\circ) > 1.1$ and $t_1(1, 90^\circ) > 1.1$

then

$$p(f_1 / c_i) = \frac{P_k}{P_k} p(f_1 / c_i) ,$$

else

$$p(f_1 / c_i) = \frac{P_k}{P_k} p(f_1 / c_i)$$

if

$$\max p(f_1 / c_i) > p(f_1 / c_i) ,$$

then assign $\text{img}_{f_1}$ to $c_{i_k}$.

Then, according to (11) (12), let

$$\lambda_{(i,j)\in(1,1)} = \text{sub max} p(f_1 / c_{i_1}) / p(f_1 / c_{i_1}) ,$$

if $\text{img}_{f_1}$ is assigned to $c_{i_1}$, and let

$$\lambda_{(i,j)\in(1,1)} = p(f_1 / c_{i_1}) / p(f_1 / c_{i_1}) ,$$

if $\text{img}_{f_1}$ is assigned to $c_{i_1}$.

Furthermore, the $TR_{11}^{11}$ presented in (3) is utilized. Accordingly, for each image, a sub-vector $W_1$ of (7) can be obtained based on (6).

4.2. Image representation regarding $d_2$

Human vs. non-human domain ($d_2$) are divided into three classes, which are portrait $c_{21}$, face $c_{22}$, and non-face group $c_{23}$. The images in the portrait class are what the occupied rate of human face in the image is larger than a threshold. The images without humans belong to the non-face group. The rest are belonged to the small face group.
Generally, for the region of skin color $R_f$, the value of H components $h$ is almost equal to 1, and the value of S components $s$ is almost equal to 6 or 7. In expression, 

$$R_f : h = 1 \& (s = 6 \lor s = 7)$$

(13)

The following shows a procedure of representing the images in $d_2$. $I_w^H(i, j)$, $I_w^S(i, j)$, $I_w^I(i, j)$ indicates the intensities of pixel $(i, j)$ of an image $m$ regarding the HSV color space, respectively.

a) Extracting the pixels which belong to $R_f$,

$$\forall i, j, \text{ if } I_w^H(i, j) = 1 \& (I_w^S(i, j) = 6 \lor I_w^S(i, j) = 7)$$

$$\text{ then } I_w^I(i, j) = 0.$$  

(14)

b) Noise cleaning;

c) Calculating $SGLD, \Phi(2,0^o) = [\hat{f}(0,1\mid 2,0^o)]$ regarding $I_w^I(i, j)$;

d) Constructing intra-tolerance relation model $TR^{22} = \{td_{ij}^{22}, i, j \in [1,3]\}$.

$TR^{22}$ presented in (4) is utilized.

e) Representing an image by a sub-vector $W_2$ based on (6).

The un-tolerance parameters $\lambda_{i(2,2.2.2)}$ are set as the following,

$$\lambda_{i(2,2.2.2)} = \begin{cases} 
\frac{1}{k_1}, & \text{if } i = 1, j = 2 \\
\frac{1}{k_2}, & \text{if } i = 2, j = 1 \\
\frac{1}{k_3}, & \text{if } i = 2, j = 2 \\
\frac{1}{k_4}, & \text{if } i = 3, j = 3 \\
0, & \text{others}
\end{cases}$$

(15)

where, let $T^-$ a threshold assigning an image to the portrait, $T^+$ a threshold assigning an image to the non-face class, then, $k_1 = \frac{\hat{f}(1,1\mid 2,0^o)}{T}$, and $k_2 = \frac{T^-}{\hat{f}(1,1\mid 2,0^o)}$. Accordingly, On the basis of (4) and (6), each image is represented by the following algorithm.

$$\text{if } \hat{f}(1,1\mid 2,0^o) < T^- \text{ then } img = [0 \ 0 \ 1 - \frac{1}{k_2}];$$

$$\text{if } \hat{f}(1,1\mid 2,0^o) > T^+ \text{ then } img = [1 \ 0.8*\frac{1}{k_1} \ 0];$$

else $img = [0.1*k_1 \ 1 - k_2 \ 0]$.

The above procedure is easy to implement, however, if necessary, the method proposed in [4] or other facial detection methods are considered to utilize to reduce the false alarm of categorization.

4.3. Image representation regarding $d_1, d_2, d_3$

In default, the probabilities that the images of the classes in $d_1$ associated to the classes in $d_2$ are considered to be same. Same as the $d_2$ to $d_1$, $d_2$ to $d_3$, and $d_3$ to $d_2$. Accordingly, the inter-tolerance relation models regarding to these domains are not needed to utilize in image categorization. However, because the images of sunset are almost taken during 5pm to
7 pm, the inter-tolerance relation model of $d_1$ and $d_3$--$TR^{13}$, which is presented in (5), is needed to consider.

Accordingly, on the basis of (8), the image group categorizing to a class $i$ in a domain $d_4$ is gotten, and on the basis of (10), the image group categorizing to the cross-classes regarding cross-domains $c_{d_i}c_0$ is gotten.

Some images categorized to $c_{11}c_{22}$ are shown in Fig. 3.

![Fig. 3 Images in $D_{11}C_{12}$](image)

5. Experimental results and analysis

We evaluate the categorizing performance of images using the Recall and Precision. The recall and precision are defined in the following expression ($\#$ stands for “number of”),

\[
\text{recall} = \frac{\# \text{true positive}}{\# \text{relevant}} \quad (17)
\]

\[
\text{precision} = \frac{\# \text{true positive}}{\# \text{clustered}} \quad (18)
\]

The test data set has 1000 images randomly selected from the personal albums and painting collections, and includes images with humans or not. These images are previously assigned to the aforementioned mentioned 7 classes ($c_{11} \sim c_{17}$) regarding $d_1$, and to the 3 classes ($c_{21} \sim c_{23}$) regarding $d_2$. Here, more than one class can be assigned. Then, each is represented by a weight vector of classes using the aforementioned method.

![Fig. 4 Precision-recall regarding $d_1$](image)

Fig. 4 shows the precision-recall evaluation of image categorization regarding $c_{11}$ (patterns), and $c_{12}$ (landscape) on the basis of (8). The thresholds $\sigma_i$ presented
in (8) are dynamic adjusted. The precision-recall evaluations of the other classes are omitted because of the pages limited. “with TR” means that semantic tolerance relation model $TR^{11}$ is utilized. That is, on the basis of (6), $\lambda_{(i,j)\in[1,M_1]} \neq 0$ ($i, j \in [1,M_1]$). “without TR” means that $TR^{11}$ is withdrawn. That is, $\lambda_{(i,j)\in[1,M_1]} = 0$ ($\forall i, j \in [1,M_1]$).

From the results, we can see, compared with a single precision-recall when $TR^{11}$ is not used, $TR^{11}$’s conducting improves the performance of image categorization. Obviously, with the same recall, the precision is always larger than what $TR^{11}$ is not used. Moreover, $TR^{11}$’s conducting enables the dynamic association of precision and recall, which makes user’s requirements for image categorization accordant with the optimizing goals such as maximum recall, maximum precision, or the tradeoff between them by adjusting the weight thresholds of classes $\sigma_i$. That is, the precision and recall of image categorization can be controlled according to user’s requirements. As the whole, the image representation based on semantic tolerance relation model makes the image categorization more effective and flexible.

Furthermore, although the test images set included the images of humans with the background, which were not included in the training images set, for almost classes, the precision can reach more than 70%, while the recall is about 70%. Especially, for $c_{12}$ (landscape), the precision can reach 90%, while the recall is about 70%.

On the other hand, the images assigned to a certain class can be ranked according to the weight values of the class, which is very useful to the image searching.

The precision-recall evaluations of categorizing images regarding $d_j$, cross-domains, and that using $TR^{13}$ or not are omitted here because of the limited pages. However, the performance evaluations shows the similar conclusions that semantic tolerance relation model enables a good dynamic association of precision with recall, which improve the effectiveness and flexibility of image semantic categorization, compared with the performance while the semantic tolerance relation model is not used.

6. Conclusion

In this paper, for realizing the perceptual similarity-based image retrieval, we systematically defined semantic classes regarding multi-domains together with the tolerance degrees between the classes. Furthermore, the semantic tolerance relation models regarding the defined classes are proposed, and the matrix parameters regarding it were gotten by the classification results of the training images set to the defined classes. On the basis of the semantic tolerance relation model, each image was represented by the class-weighted vector according to its classification result to the defined classes, and the categorization algorithms of the vector-represented images were described. Finally, as demonstration, the image automatic representation and categorization methods regarding the nature/man-made domain, human/non-human domain, and temporal domain were implemented. The precision-recall evaluations regarding test image categorization to the defined classes showed the effectiveness of the proposed methods.
7. Reference


Author

Dr. Ying Dai received her M.Sc From Xian Jiaotong University, China in 1988. After some years in universities as a lecturer, she attended Department of Information Engineering, Shinshu University, Japan in 1992. She had a Dr. Eng from Shinshu University in 1996. She was granted JSPS Research Fellowships for Young Scientists from 1995 to 1997. She joined the Iwate Pref. University in 1998. She has been an associate professor in the Faculty of Software and Information Science, Iwate Pref. University since April, 2002. She is a member of IEEE.