Visual graph modeling for scene recognition and mobile robot localization

Trong-Ton Pham, Philippe Mulhem, Loïc Maisonnasse, Éric Gaussier, Joo-Hwee Lim

Visual graph modeling for scene recognition and mobile robot localization

Trong-Ton Pham · Philippe Mulhem · Loïc Maisonnasse · Eric Gaussier · Joo-Hwee Lim

Published online: 14 September 2010

Abstract Image retrieval and categorization may need to consider several types of visual features and spatial information between them (e.g., different point of views of an image). This paper presents a novel approach that exploits an extension of the language modeling approach from information retrieval to the problem of graph-based image retrieval and categorization. Such versatile graph model is needed to represent the multiple points of views of images. A language model is defined on such graphs to handle a fast graph matching. We present the experiments achieved with several instances of the proposed model on two collections of images: one composed of 3,849 touristic images and another composed of 3,633 images captured by a mobile robot. Experimental results show that using visual graph model (VGM) improves the

T.-T. Pham (✉)
Grenoble Institute of Technology—Laboratoire Informatique de Grenoble (LIG),
385 Av. de la Bibliothèque, 38400, Grenoble, France
e-mail: ttpham@imag.fr, trongtonfr@yahoo.fr

P. Mulhem · E. Gaussier
Multimedia Information Modeling and Retrieval—Laboratoire Informatique de Grenoble (LIG), 385 Av. de la Bibliothèque, 38400, Grenoble, France

P. Mulhem
e-mail: mulhem@imag.fr

E. Gaussier
e-mail: eric@imag.fr

L. Maisonnasse
R&D Department-TecKnowMetrix, 4 rue Léon Béridot, Voiron, France
e-mail: lm@tkm.fr

J.-H. Lim
Computer Vision and Image Understanding-Institute for Infocomm Research (I^2R),
1 Fusionpolis Way, #21-01, Connexis, 138632, Singapore
e-mail: joohwee@i2r.a-star.edu.sg

accuracies of the results of the standard language model (LM) and outperforms the Support Vector Machine (SVM) method.

Keywords Graph theory · Information retrieval · Language model · Scene Recognition · Robot localization

1 Introduction

Still image understanding and retrieval for computers are about combining multiple points of views. A broader perspective for multimedia document indexing and retrieval is given by R. Datta et al. in [4]:

> "*The future lies in harnessing as many channels of information as possible, and fusing them in smart, practical ways to solve real problems. Principled approaches to fusion, particularly probabilistic ones, can also help provide performance guarantees which in turn convert to quality standards for public-domain systems*."

This reflexion also holds in the specific context of image documents. The points of views of images rely on different regions extracted, different features generated and different ways to integrate these aspects in order to annotate or retrieve images based on their visual similarity. Let us present a short overview of the diversity of approaches encountered in the image annotation and retrieval domain. Image annotation and retrieval may use predefined segmentation in blocks [3], or try to consider segmentation techniques based on color/texture [6] or regions of interest like the well-known work of D. Lowe [13]. The feature considered are mostly represented using histograms of features (colors, textures or shapes) or of *bag-of-word (BOW)* proposed initially in [25]. Other approach may consider spatial relationships between regions as in [27]. When considering more complex representations, approaches may use graph representations like in [18].

Despite the fact that selecting relevant regions and extracting good features are *per se* very difficult tasks, we believe that the way we represent different points of views of the image (like several segmentations and/or several features for instance) also has a great impact on image annotation and image retrieval. Our interest in this paper is twofold. First, we focus on a representation of image content, more precisely graph-based representation, which is able to represent different points of views (namely several visual representations and spatial relationships between regions). Second, we define a language model on such graphs that tackles the problem of retrieval and classification of images. Considering a graph to represent the features intends to preserve the diversity of content when needed. In fact, such graphs are versatile, because they can handle early fusion-like approaches when considering several representations in an integrated matching process as well as late fusion-like approaches when considering matching on specific sub-graphs before fusion. The interest of considering language models for such graphs lies in the fact that it benefits from this successful research field of information retrieval since the end of the 90s and in particular the seminal work of Ponte and Croft in [24]. Such language models are well-defined theoretically, and also have shown interesting experimental results as synthesized in [16]. Therefore, our main focus in this paper is to propose an

extension of language models [21] in the context of graph-based representation for image content.

1.1 Related works

We consider here first the integration of relationships during the processing and/or the representation of images. When focusing on still images, the obvious kinds of relations are spatial relationships between image regions. For instance, image descriptions expressed by 2D-strings used in the Visualseek system [27] reflect the sequences of object occurrences along one or several reading directions. In this case an integration of visual feature of regions and spatial relationships is achieved. Retrieval on 2D-strings is complex: substring matching is a costly operation. However, heuristics to speed up the process do exist, as described in [2], which allows reduce the processing time by a factor of 10. Several works have considered visual features and relationships between image regions integrated into probabilistic models, e.g., through the use of 2D HMMs [11], through the use of normalized configuration vectors in [23], or graphical generative models and overlapping relationships in [8], but these works do not consider the relations during the retrieval process. Relationships between the elements of an image may also be expressed explicitly through naming conventions, as in [19] where the relations are used for indexing. Other attempts have integrated relationships in graphs like [17] for indexing and retrieval. Nevertheless, considering the explicit relationships may generate complex graphs representations, an the retrieval process is likely to suffer from the complexity of the graph matching [18]. One of our aims here is to be able to represent the different points of views of images using graphs, without suffering from the burden of computationally expensive matching procedures.

Language models for information retrieval exist since the end of the 90s [24]. In these generative models, the relevance status value of one document d for a query q is estimated using the probability $P(q|d)$ of document d to generate query q. In [24] the language model was based on a multiple Bernoulli distribution. Nevertheless, the predominant modeling assumption is now centered on multinomial distribution [28]. Such models can use unigram (terms taken one by one) or n-grams (sequences of n terms) [28]. However, these models lack an easy extension to integrate explicit relationships between terms. A core issue in language model is to estimate the probability when terms are absent in the documents due to the data sparseness. This may cause an inaccurate estimation of the overall probability (a.k.a. zero probability problems). To overcome this probability estimation problem, different smoothing methods [32] (such as Jelinek–Mercer, Dirichlet, etc.) have been proposed to adjust the maximum likelihood estimator of the language modeling. We will detail this technique in Section 3 using the Jelinek–Mercer smoothing technique.

Simple language models have been used for image indexing and retrieval [10] without incorporating relationships between image regions. Some works applied language models for image annotation incorporated bigrams [31] and trigrams [7] model, where these n-grams are built from spatially connected regions. Tirilly et al. in [29] proposed to generate "sentences" built from projecting visual words on one given axis (according to the idea similar to the 2D-strings described above), in a way to generate from unigram to 4-grams language models on these visual "sentences". In this case, the spatial relationships are the *precedence* along the projection axis.

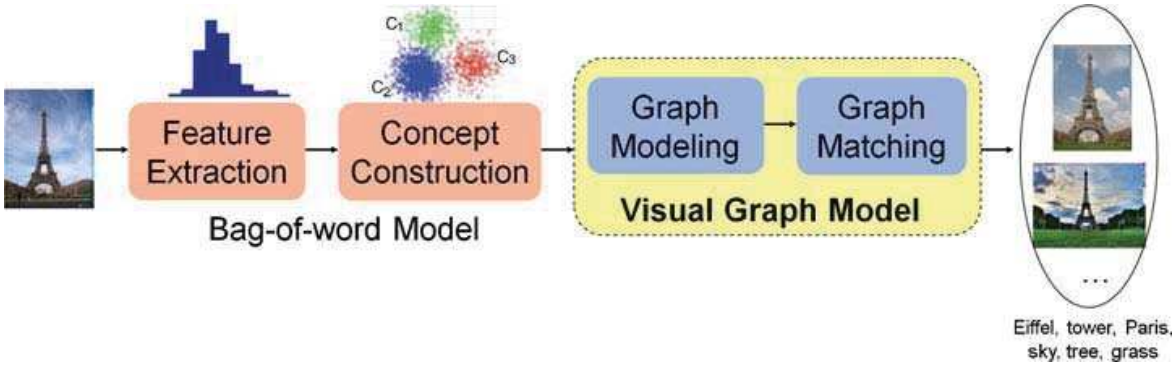

Fig. 1 System architecture of the graph-based model for scene retrieval/annotation

Named relationships with generative models have been used in [1] for image scenes, but the complexity of the learning of scene is prohibitive and experiments were achieved on small image sets.

Another work [14] has proposed to extend language models with relationships but focusing on texts. These approaches consider features and relationships, but do not consider several points of views according to several features extracted. Our work concentrates on modeling visual graphs composed of several types of concepts (i.e., visual words) and named relationships, which correspond to a different point of view of the images.

1.2 Our approach

We present in this paper the system architecture that consists of three main steps (see Fig. 1). The first step is to extract a set of visual concepts from each image feature considered, such as color, edge histogram or local features. The second step represents each image as a graph generated by a set of weighted concepts and a set of weighted relations. The third step is related to the fact that we want to retrieve relevant images to a given query. Therefore, we extend the work in [21, 22] by taking into account the different types of image representations and spatial relations during matching by computing likelihood of two graphs using a language model framework.

The contributions of this work are twofold. First, we present an unified graph-based framework for image representation which allows us to integrate different types of visual concepts and different spatial relations among them. Second, we extensively study the extension of language model for graph matching which allows a more reliable matching based on a well studied theory of information retrieval. The experimental results, performed on STOIC-101 and RobotVision '09 image collections, confirm the effectiveness of our visual graph model.

1.3 Outline of the paper

The remainder of this paper is structured as follows. We first present the visual graph model used to describe the image content in Section 2. Then, Section 3 discusses the matching of a trained graph given a query graph. Section 4 details our experimental results on two image collections. More precisely, the experiments focused on a relation between different types of concepts. Finally, we conclude the paper with a discussion in Section 5.

2 Visual graph modeling

2.1 Image modeling

Inspired by the bag-of-word model in [9], images are modeled as a set of *visual concept* (or *concept* in short) coming from different visual features. Our goal is to automatically induce, from a given image, a graph that represents the image content. Such a graph will contain concepts directly associated with the elements present in the image, as well as spatial relations which express how concepts are related in the image. To do so, our procedure is based on four main steps:

1. Identify regions within the image that will form the basic blocks for concept identification.
2. Index each region with a predefined set of visual features.
3. Cluster all the regions found in the collection into K classes, each class representing one concept.
4. Finally, extract relations between concepts.

As described before, several segmentation approaches already existed in the literature [26]. For the first step, we focus on three segmentation methods to define an image region: sampling pixel, grid partitioning and keypoint detection. The second step aims at representing each region as set of feature vectors for clustering purposes. We consider here several visual features (i.e., several points of views) extracted from one image region. For the pixel sampling method, each region is represented by its central pixel. The HSV color value of this pixel can be used as visual feature to represent images. We choose to focus on the HSV color space because of its robustness against illumination changes. In the case of grid partition, in our experiments, visual features extracted from the patches are color histograms and edge descriptors [30]. For the keypoint extraction, SIFT descriptors [13] are extracted within a given radius around the keypoint. The dimensionality for each feature is summarized in Table 1.

Next step consists of building visual concept vocabulary for each type of image representation (i.e., for each type of visual feature f) as follows: (1) Unsupervised learning with k-means clustering groups similar feature vectors into the same cluster (each cluster corresponds to a visual concept c). Clustering transforms the continuous feature space into a discrete number of clusters. (2) Image is then represented by the number of occurrences for each concept in this image. Each type of image representation for a specific visual feature corresponds to a concept $\mathcal{C}_f$.

Once these visual concepts are defined and characterized independently, the last step is to integrate the relationships between them. Existing work has proposed the use of topological relations between regions or between points in a 2D space [5].

Table 1 Visual features used for each type of image representation

Region type	Features	Dimensions
Pixel	(H,S,V) value	3
Patch	HSV histogram	64
Patch	Edge histogram (5 scales × 16 orientations)	80
Keypoint	SIFT descriptor	128

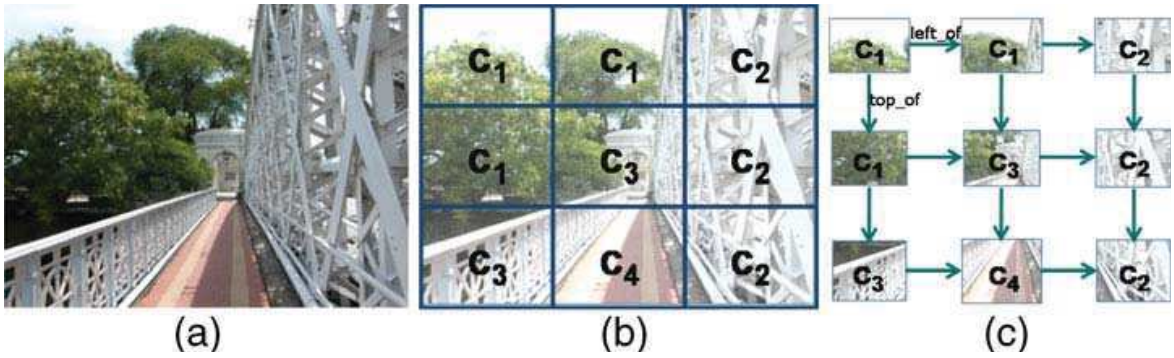

Fig. 2 Example of spatial relations extracted from image. **a** Scene of a bridge, **b** visual concept extraction, **c** relations *left_of* and *top_of* extracted from concepts

Based on this work, we will define the relationships between segmented regions. Figure 2 gives an example of spatial relations between visual concepts used in our experiment with STOIC collection. Relation sets $left\_of$ and top_of are extracted from the two connected concepts. These relations help to capture the spatial co-occurrence information of two visual concepts. For example, instances of the concept "sky" are usually on the top_of instances of the concept "tree", while instances of concept "tree" is more frequently on the $left\_of$ instances of concept "bridge". If the number of training image is large enough, the graph framework will capture the statistical consistency for this type of relation.

At the end of this procedure, we obtain a set of visual concepts $\mathcal{C}_f$ and a set of predefined relations E for each type of concept and relation. Each concept is associated with a weight that represents its number of occurrences in the image. Similarly, each relation is also given a weight corresponds to the number of times this relation has occured in the image. We will denote the weighted concepts set by WC_f and the weighted relations set by WE. As we may have several image representations (or point of views) and different kind of spatial relationships between them, in the end, we denote a set of weighted concept sets as $S^I_{WC_f} = \bigcup_f WC^I_f$ and a set of weighted relation sets as $S^I_{WE} = \bigcup WE^I$ for an image I.

Note that we tend to choose different types of visual features (i.e., color, edge, SIFT) which are visually independent [26] from each other to represent image content. Therefore, concept sets WC_f are disjoint (e.g., $WC_{\text{color}} \bigcup WC_{\text{edge}} \bigcup WC_{\text{sift}} = \emptyset$). From this observation, we could make an independent assumption based on the set of weighted concept set $S^I_{WC_f}$. The similar assumption is also applied for weighted relation sets S^I_{WE}.

2.2 Graph definition

Given a set of weighted concept sets $S^I_{WC_f}$ and a set of weighted relation sets S^I_{WE}, the visual graph representing an image I is defined by:

$$G^I =< S^I_{WC_f}, S^I_{WE} >$$

where each concept c of concept set WC^I_f corresponds to a visual concept used to represent the image according to the feature f associated to it. The weight of concept

captures the number of times of this concept appears in the image. Denoting $\mathcal{C}_f$ a set of concepts for one feature over the whole collection, WC_f^I is a set of pairs $(c, \#(c, I))$

$$WC_f^I = \{(c, \#(c, I))|c \in \mathcal{C}_f\}$$

where c is an element of $\mathcal{C}_f$ and $\#(c, I)$ is the number of times c occurs in the document image I.

Any labeled relation between any pair of concepts $(c, c') \in \mathcal{C}_f \times \mathcal{C}_{f'}$ is represented by a triple $((c, c'), l, \#(c, c', l, I))$, where l is an element of $\mathcal{L}$, the set of possible labels for the relation, and $\#(c, c', l, I)$ is the number of times c and c' are related with label l in image I. WE^I is then defined as:

$$WE^I = \{((c, c'), l, \#(c, c', l, I))|(c, c') \in \mathcal{C}_f \times \mathcal{C}_{f'}, l \in \mathcal{L}\}$$

If a pair of concepts (c, c') come from the same concept set (i.e., $\mathcal{C}_f = \mathcal{C}_{f'}$), we refer this relation as *intra-relation*. Otherwise, we refer it as *inter-relation*.

Figure 3 shows an example of our graph constructed from an image of a bridge scene. This example corresponds to a visual graph containing one visual concept set ($\mathcal{C}_{\text{color}}$) and two intra-relation sets (E_{left_of} and E_{top_of}). Each node corresponds to a concept and the number of time it occurs in the image. For example, concept c1 appeared two times in the image and is denoted by (c1,#2) in the figure. Likewise, the relation between a couple of concepts is also captured by the directed arcs in this graph. Here, the blue arcs express the relation $left\_of$ and the green arcs express the relation top_of of two connected concepts. For example, concept c1 is related to concept c2 with the relation top_of two times and is related to itself by the relation $left\_of$ two times. It is denoted by (c1, c2, top_of, #2) and (c1, c1, $left\_of$, #2).

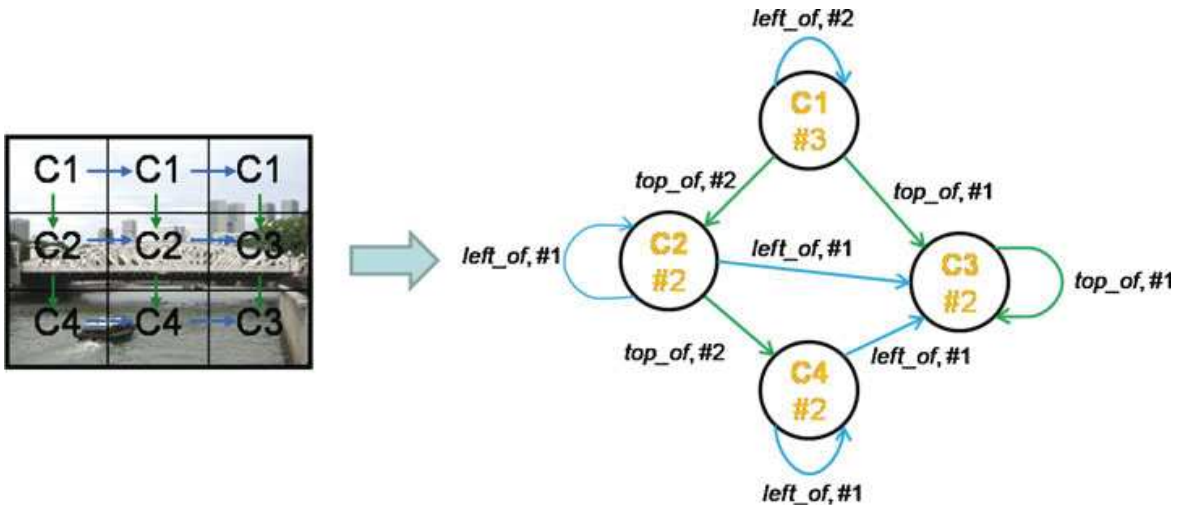

Fig. 3 Example of a visual graph extracted from an image. Concepts are represented by *nodes* and spatial relations are expressed by *directed arcs*. Nodes and links are weighted by the number of times they appear in the image

Finally, this representation of graph will be used for the graph matching in the next step.

3 Language model for graph matching

Based on the language model defined over the graph proposed in [15], we present in this paper an extension that handles set of concept sets and set of relation sets. The probability for a query image graph $G^{Iq} = < S^{Iq}_{WC_f}, S^{Iq}_{WE} >$ to be generated from one document image graph G^{Id} can be written as:

$$P(G^{Iq}|G^{Id}) = P(S^{Iq}_{WC_f}|G^{Id}) \times P(S^{Iq}_{WE}|S^{Iq}_{WC_f}, G^{Id}) \tag{1}$$

where $P(S^{Iq}_{WC_f}|G^{Id})$ is the probability of generating set of concept sets of the query graph from the document graph, and $P(S^{Iq}_{WE}|S^{Iq}_{WC_f}, G^{Id})$ is the probability of generating set of relation sets of the query graph from the document graph.

3.1 Concept set generation

For generating the probability of query concept sets from the document model $P(S^{q}_{WC_f}|G^{d})$, we assume a concept set independence hypothesis (see our explanation in Section 2.1). The probability can thus be written as:

$$P(S^{Iq}_{WC_f}|G^{Id}) = \prod_{WC_f{}^{Iq} \in S^{Iq}_{WC_f}} P(WC_f{}^{Iq}|G^{Id}) \tag{2}$$

Assuming concept independence, which is standard in information retrieval, the number of occurrences of the concepts (i.e., the weights considered previously) are integrated through the use of a multinomial model. We compute $P(WC_f{}^{Iq}|G^{Id})$ as follows:

$$P(WC_f{}^{Iq}|G^{Id}) \propto \prod_{c \in \mathcal{C}_f} P(c|G^{Id})^{\#(c,Iq)} \tag{3}$$

where $\#(c, Iq)$ denotes the number of times concept c occurs in the image query graph. This contribution corresponds to the concept probability as proposed in [15]. Similar to the previous work, the quantity $P(c|G^{Id})$ is estimated through maximum likelihood using Jelinek–Mercer smoothing:

$$P(c|G^{Id}) = (1 - \lambda_c)\frac{\#(c, Id)}{\#(*, Id)} + \lambda_c \frac{\#(c, D)}{\#(*, D)} \tag{4}$$

where λ_c is the smoothing parameter for each concept set $\mathcal{C}_f$. The quantity $\#(c, Id)$ represents the number of times c occurs in the document image Id and $\#(*, Id)$ is equal to $\sum_c \#(c, Id)$. The quantities $\#(c, D)$ and $\#(*, D)$ are similar, but defined over the whole collection D (i.e., over the union of all images in the collection).

3.2 Relation set generation

Assuming the relation set independence, as shown in previous section, we follow a similar process for generating the probability of the relation sets from document image graph, this leads to:

$$P(S_{WE}^{Iq}|S_{WC_f}^{Iq}, G^{Id}) = \prod_{W_E^{Iq} \in S_{WE}^{Iq}} P(W_E^{Iq}|S_{WC_f}^{Iq}, G^{Id}) \tag{5}$$

For the probability of generating query relation from the document, we assume that a relation depends only on the two linked sets. Assuming that the relations are independent and following a multinomial model, we compute:

$$P(W_E^{Iq}|S_{WC_f}^{Iq}, G^{Id}) \propto \prod_{(c,c',l) \in \mathcal{C}_f \times \mathcal{C}_{f'} \times \mathcal{L}} P(L(c,c') = l|WC_f^{Iq}, WC_{f'}^{Iq}, G^{Id})^{\#(c,c',l,Iq)}$$

where $c \in \mathcal{C}_f$, $c' \in \mathcal{C}_{f'}$ and $L(c,c')$ are variables which values in $\mathcal{L}$ and which reflect the possible relation labels between c and c', in this relation set. As before, the parameters of the model $P(L(c,c') = l|WC_f^{Iq}, WC_{f'}^{Iq}, G^{Id})$ are estimated by maximum likelihood with Jelinek–Mercer smoothing, giving:

$$P(L(c,c') = l|WC_f^{Iq}, WC_{f'}^{Iq}, G^{Id}) = (1-\lambda_l)\frac{\#(c,c',l,Id)}{\#(c,c',*,Id)} + \lambda_l \frac{\#(c,c',l,D)}{\#(c,c',*,D)} \tag{6}$$

where λ_l is the smoothing parameter for each relation set E. The quantity $\#(c,c',l,Id)$ represents the number of times concepts c and c' are linked with label l in the document image Id, and $\#(c,c',*,Id)$ is equal to $\sum_{l \in \mathcal{L}} \#(c,c',l,Id)$. By convention, when one of the two concepts does not appear in the image d, we set:

$$\frac{\#(c,c',l,Id)}{\#(c,c',*,Id)} = 0$$

Again, the quantities $\#(c,c',l,D)$ and $\#(c,c',*,D)$ are counted in a similar way but computed on the whole collection D (i.e., over the union of all the graphs from all the documents in the collection).

This graph model is a generalization of the model defined in [21] which corresponds to the special case where only one concept set and one relation set are used. In some special cases, our model corresponds to the standard language model (LM) used in [15, 20] where relations are not considered (i.e., documents and queries correspond to multiple bag-of-words model) .

In a more practical way, as done in [20], we compute the relevance status value (RSV) of a document image Id for query image Iq in the log-probability domain. Such domain, in the context of multinomial distributions, leads to the same ranking as the probability computed for G^{Id} and G^{Iq}. For image categorization, a query image Iq is then classified to a class of closest document image Id that estimated as follows:

$$class(Iq) = class(\arg\max_{Id \in D} RSV(G^{Iq}|G^{Id})) \tag{7}$$

4 Experiments

First, we describe the collections used to carry out our experimentation. Then, we present the results obtained with our model based on this collection. Our objective is to demonstrate that the visual graph model, as presented in previous section, is well adapted in representing image content. Furthermore, the integration of relationships to the graph helps to improve the image representation using only visual concepts. Finally, we give some discussions on our experimental results.

4.1 Scene recognition

4.1.1 STOIC-101 collection

The Singapore Tourist Object Identification Collection (STOIC) is a collection of 3,849 images containing 101 popular tourist landmarks (mainly outdoor). These images were taken, mostly with consumer digital cameras in a manner typical of a casual tourist, from three distances and four angles in natural light, with a mix of occlusions and cluttered background to ensure a minimum of 16 images per scene (see Fig. 4). Images in the collection are affected by different weather patterns and different image capturing styles. For experimental purposes, the STOIC-101 collection has been divided into a training set containing 3,189 images (82.8% of the collection) and a test set containing 660 images (17.15% of the collection). The average number of images per class for training is 31.7, and 6.53 for testing,

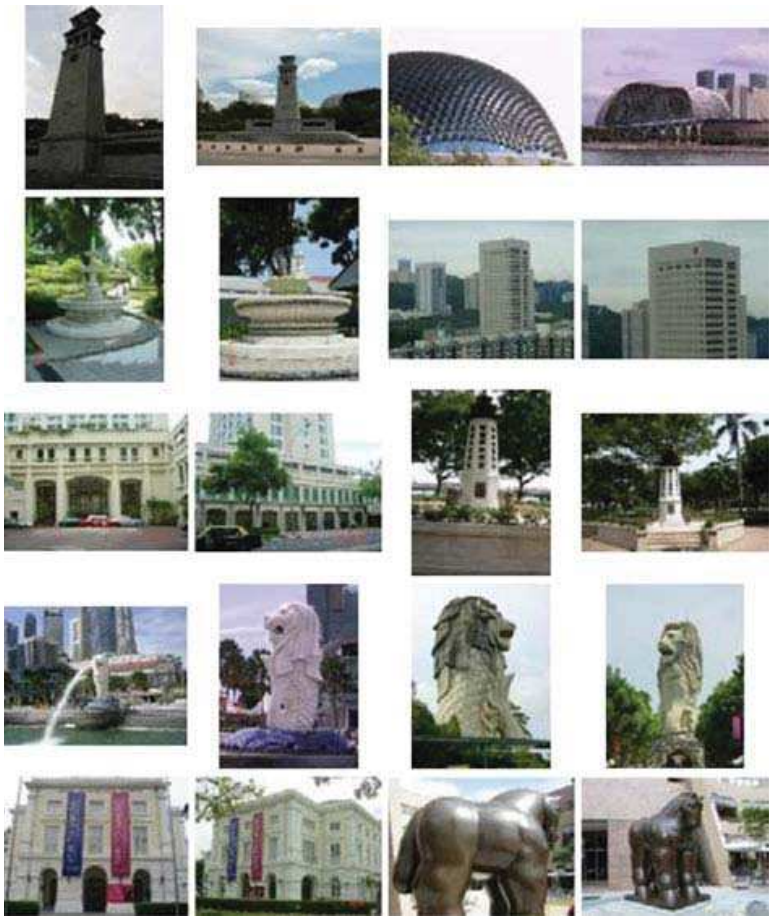

Fig. 4 Images of STOIC-101 collection are taken from different angles, viewpoints and weather conditions

Table 2 Summary of experiments on STOIC-101 collection

	Training by (I)	Training by (S)
Query by (I)	√	√
Query by (S)	√	√

respectively. In the test set, the minimum number of images per class is 1, and the maximum is 21.

The main application of STOIC collection is for mobile image search. A user can upload an image taken with a hand-phone and post it as a query to the system. On the server-side, the images from the 101 scenes of the STOIC collection are matched against the user query. The server-side of this search engine architecture is two-tiers: (a) the query processing server takes a query image as input and generates a query graph file and (b) the language model server receives the query graph and computes the matching function based on graphs built from training images. Finally, textual information related to the matched scenes will be sent back to the user.

As a user can take one or several images of the same scene and query the system accordingly, we have considered several usage scenarios. Table 2 summarizes these different scenarios (a scene (S) corresponds to a group of images and a single image (I)). Note that some images in the collection have been rotated into the correct orientation (for both portrait and landscape layouts).

4.1.2 Proposed models

Several studies on the STOIC collection have shown that color plays a dominant role, and should be preferred to other visual features such as edge or texture [12]. Furthermore, color histogram can be easily and efficiently extracted. For these reasons, we rely only on HSV color features in our experiments. In order to assess the validity of our methodology, we followed different ways to divide each image into regions and we retained:

1. A division of medium grain, where blocks of 10×10 pixels are used, the center pixel being considered as a representative for the region. We refer to this division as *mg*.
2. A patch division where the image is divided into 5×5 regions of equal size. We refer to this division as *gg*.

For *mg* divisions, we used the (H, S, V) values as a feature vector for each pixel. Similary, each patch in *gg* division is quantized by a HSV histogram (4 bins/channel) that yields a 64-dimension vector for each region. We then clustered the HSV feature vectors of all regions into $k = 500$ classes with *k-means* clustering algorithm. This results in a hard assignment of each region to one concept. The set of weighted concepts, W_C, is then obtained by counting how many times a given concept occurs in the image. The choice of $k = 500$ is motivated by the fact that we want a certain granularity in the number of concepts used to represent an image. With too few concepts, one is likely to miss important differences between images, whereas too many concepts will tend to make similar images look different. We will refer to the indexing obtained in this way as *mg-LM* and *gg-LM*, respectively for "division *mg* with automatically induced concepts" and "division *gg* with automatically induced concepts".

In addition, for the methods *mg-LM* and *gg-LM*, we extracted the spatial relations between concepts as mentioned previously: *left_of* and *top_of*, and counted how many times two given concepts are related through a particular relation in order to obtain the weights for our relations. This last step provides a complete graph representation for images. We will refer to these two complete methods as *mg-VGM* and *gg-VGM*. To summarize, we have constructed four models based on the visual concept sets and the relation sets:

1. *mg-LM*$=< \{W_{Cmg}\}, \emptyset >$, that used only *mg* division concepts.
2. *mg-VGM*$=< \{W_{Cmg}\}, \{W_{Eleft\_of}, W_{Etop\_of}\}$, that used *mg* division concepts and two intra-relation sets *left_of* and *top_of*.
3. *gg-LM*$=< \{W_{Cgg}\}, \emptyset >$, that used only *gg* concepts.
4. *gg-VGM*$=< \{W_{Cgg}\}, \{W_{Eleft\_of}, W_{Etop\_of}\}$, that used *gg* concepts and two intra-relation sets *left_of* and *top_of*.

Last but not least, to classify query images in the 101 scenes, we used the language model for visual graphs presented in (7). When there is no relation, as in the cases of *mg-LM* and *gg-LM*, the term $P(S^q_{WE}|S^q_{WC}, G^d) = 1$ so that only concepts are taken into account to compare images.

4.1.3 Results

The performance of the different methods was evaluated using the accuracy, per image and per scene. They are defined as the ratio of correctly classified images or scenes. More precisely:

$$\text{Image accuracy} = \frac{TP_i}{N_i}, \quad \text{Scene accuracy} = \frac{TP_s}{N_s}$$

where TP_i and TP_s represent the number of images and the number of scenes (respectively) correctly classified. N_i is the total number of test images (i.e., 660 images), and N_s the total number of scenes (i.e., 101 locations).

Table 3 shows the results we obtained when using automatically induced (through clustering) concepts. As one can see, automatically inducing concepts with a medium grain division of the image yields the best results (the difference with the patch division for the S–I scenario being marginal). Overall, the *mg* division outperforms the *gg* division in most of the cases. Especially in the S–S scenario, the *mg* models obtained the best performance. One possible reason is that in *mg* division the number of concepts is far more than the one in the *gg* division.

This being said, there is a difference between the I–S and S–I scenarios: The system is queried with more information in the I–S scenario than in the S–I scenario. This difference results in a performance which is, for all methods, worse for the S–I

Table 3 Impact of spatial relations on the performance (best results are in bold, relative improvement over the method without relations is in parentheses)

Training	Query	*mg-LM*	*mg-VGM*	*gg-LM*	*gg-VGM*
I	I	0.789	**0.794** (+0.6%)	0.484	0.551 (+13.8%)
I	S	0.822	**1.00** (+21.6%)	0.465	0.762 (+63.8%)
S	I	0.529	0.594 (+12.3%)	0.478	**0.603** (+26.1%)
S	S	**1.00**	**1.00**	0.891	0.920 (+3.2%)

scenario than for the other ones. We conjecture that this is why the results obtained for the *mg-VGM* method on S–I are not as good as the ones for I–I. There seems to be a plateau for this scenario around 0.6, a hypothesis we want to explore in future work.

We finally assessed the usefulness of spatial relationships by comparing the results obtained with the different methods that include or not such relations. These results are displayed in Table 3. As one can note, except for the S–S scenario with the *mg* division, the use of spatial relations always improves the accuracy of the classifier. This justifies the framework we developed in Section 3 of language model for visual graphs including automatically induced concepts and spatial relations among them.

4.2 Mobile robot localization

4.2.1 The RobotVision'09 collection

The image collection came from the RobotVision'09 competition[1] task (part of ImageCLEF campaign) aiming to address the problem of topological localization of a mobile robot using only visual information. The difficulty of this task is that the robot has to recognize the correct room in different illumination conditions and adapt as objects, new furniture, etc. are added over the time.

The RobotVision collection contains a sequence of 1,034 images for training and a sequence of 909 images for validation. Training and validation sets (see Fig. 5) were captured within an indoor laboratory environment consists of five rooms across a span of 6 months. Then, the official test has been carried out on a list of 1,690 images (recorded 20 months later). The collection comprises five annotated rooms (corridor-CR, printer area-PA, kitchen-KT, one-person office-BO, two-persons office-EO) and an *unknown* room from test set.

4.2.2 Proposed models

The system we used for the RobotVision competition was composed of two processes: a recognition step and a post-processing step. However, we describe and evaluate here only the recognition step, in such a way to assess the impact of the model proposed. The robot was trained with a sequence of images taken in the night condition. Then, we used a validation set captured in sunny condition to estimate the system parameters. The different concept sets and relation sets were extracted from the collection of images as follows:

1. Each image was divided into 5 × 5 patches. We extracted for each patch a HSV color histogram and an edge histogram as in Section 2.1. Then, the visual vocabulary of 500 visual concepts was constructed by using k-means clustering algorithm. From this vocabulary, we built the weighted concept set W_{Cpatch}.
2. Similar to the previous step except that the visual features were extracted from the local keypoints. To be more precise, we detected scale invariant keypoints using SIFT detector [13] for each images. Local features were then used to create the weighted concept set W_{Csift}.

[1] http://imageclef.org/2009/robot

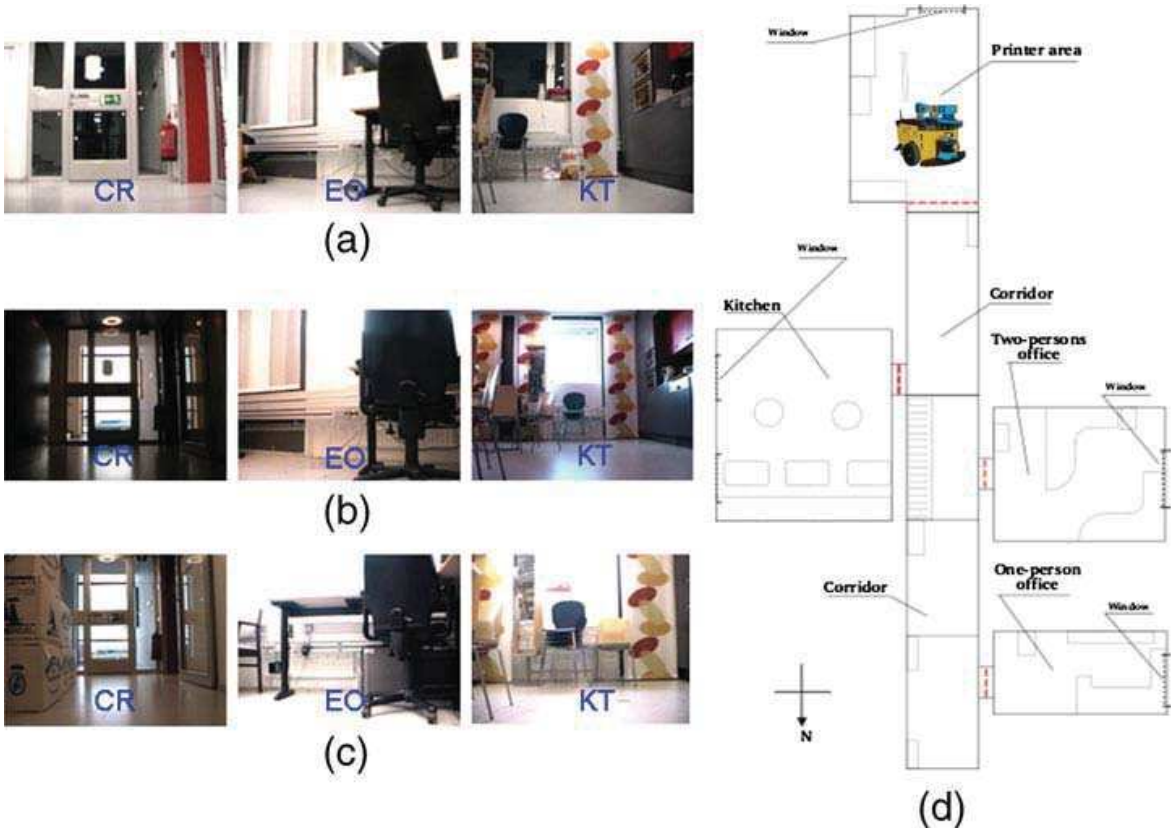

Fig. 5 Example images from RobotVision'09 collection: **a** training set in night condition, **b** validation set in sunny condition, **c** test set in unknown condition, **d** the local area map

3. Using the two previous features we defined an inter-relation set {*inside*} between patch concepts and SIFT concepts, denoted as W_{Einside}, if one key-point is located **inside** the area of a corresponding patch.

Similar to above, we referred to the model without relation as LM (simply the production of probability generated by different concept sets) and the graph model with the spatial relation as VGM (with the contributing of relation probability to graph model). Based on this definition, we have implemented several graphs to measure the performance of our proposed model:

1. $LM^P =< \{W_{\text{Cpatch}}\}, \emptyset >$, that used only patch concepts.
2. $LM^S =< \{W_{\text{Csift}}\}, \emptyset >$, that used only SIFT feature concepts.
3. $LM^{S.P} =< \{W_{\text{Csift}}, W_{\text{Cpatch}}\}, \emptyset >$, that used both patch and SIFT feature concepts.
4. $VGM^{S \to P} =< \{W_{\text{Csift}}, W_{\text{Cpatch}}\}, \{W_{\text{Einside}}\} >$, that used patch concepts, SIFT feature concepts and the *inside* relations between them.

Figure 6 gives an example of the graph extracted from the concept sets and relation sets defined above. In fact, the first three models were estimated following the equation presented in Section 3.1. The fourth model is the fusion graph combined with spatial relation. Its probability was computed according to the equation defined in Section 3.2.

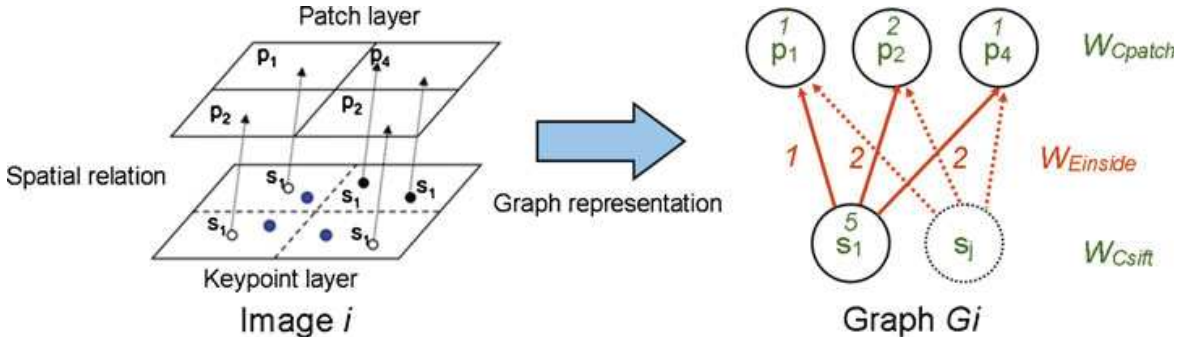

Fig. 6 Graph model constructed for RobotVision includes two type of image representation and one type of relation

4.2.3 Results

The evaluation measured the differences between the actual room id and the one classified by the systems. The following rules were used when calculating the official score for a test sequence: *+1.0 points* for each correctly classified image; *−0.5 points* for each misclassified image. So, higher score means higher accuracy.

Table 4 describes the results in terms of score value for each model. As expected, the two basic models LM^P and LM^S gave a good score for the validation set. However, the model LM^P did not perform well on the test set due to the introduction of new room and new arrangement of interior furniture. The simple fusion model $LM^{S.P}$ underperformed the best results of LM^P and LM^S. However, this result was more robust in the sense that it leveraged on the spurious effects of each visual feature (i.e., $LM^{S.P}$ outperformed the averaged result of LM^P and LM^S in both cases). Moreover, the introduction of *inside* relations to the completed graph $VGM^{S\rightarrow P}$ boosted its results respectively by 39.5 and 40.1% comparing to the fusion graph $LM^{S.P}$ for both validation set and test set. This fact confirmed that the integration of relations played a significant role to improve the results. In addition, it showed that the link between object details and its global presentation provides a better abstraction for image content.

We present in detail the classification accuracies for each class (represented by its room id) as categorized by our algorithms in Table 5. For each class, the accuracy is computed by the number of correctly labeled images divided by the total number of images belonging to this class. Here, we only consider the classification accuracies of five rooms as we did not treat the *unknown* room in the test sequence at this step. Due to the paper constrains, the reader may refer to [20] for more infomation on the post-processing step of the results.

Generally, the graph model for SIFT concepts LM^S performs better than the graph model for patch concepts LM^P. This leads us to a conclusion that the details of object are important clues for scene recognition. In addition, the simple fusion model $LM^{S.P}$ tried to leverage the effect on both model LM^S and LM^P and improved

Table 4 Results of different graph models

Graph model	LM^P	LM^S	$LM^{S.P}$	$VGM^{S\rightarrow P}$
Validation set	345	285	334.5	**466.5** (+39.5%)
Test set	80.5	263	209.5	**293.5** (+40.1%)

Table 5 Classification accuracies of graph models for each room

	BO	CR	EO	KT	PA	Mean
Validation set						
LM^P	0.257	0.779	0.524	0.450	0.434	0.489
LM^S	0.354	0.658	0.581	0.426	0.402	0.484
$LM^{S.P}$	0.398	0.679	0.613	0.519	0.426	0.527
$VGM^{S\rightarrow P}$	**0.416**	**0.829**	**0.702**	**0.550**	**0.492**	**0.598**
Test set						
LM^P	0.163	0.701	0.385	0.236	0.279	0.353
LM^S	0.331	0.721	0.478	0.509	**0.348**	0.477
$LM^{S.P}$	0.206	**0.756**	0.484	0.410	0.286	0.428
$VGM^{S\rightarrow P}$	**0.369**	0.736	**0.540**	**0.516**	0.344	**0.501**

the results only in the case of two-person office (EO). All four models gave good accuracies for the corridor (CR) regardless of brutal changes in light conditions. We also noted that the number of training images for corridor (CR) was the highest (483/1,034 images) comparing to other classes. It suggests that the higher the number of image samples, the more robust the performance is.

Overall, the fusion graph combined with spatial relations $VGM^{S\rightarrow P}$ gave better accuracies in the major cases except in the case of corridor (CR) for test set. However, the difference was not significant comparing to other models (only 2% less than the $LM^{S.P}$ graph model). Furthermore, the mean accuracy of model $VGM^{S\rightarrow P}$ achieved on the test set and the validation set were the best of four models, with more than 7% better than the simple fusion model $VGM^{S.P}$. This result confirms again the strength of spatial relationships contributed in our graph model.

4.3 Disscussions

4.3.1 Cross validation optimization with STOIC collection

The results presented above are optimized *a posteriori*, i.e., we exhaustively tested the parameters on the test set to get the best configuration. This approach overestimates the proposed algorithms, by giving an upper bound of the evaluation results and not a correct estimation. In a way to estimate more precisely the results, we optimized the smoothing parameters on a validation set for the *mg-LM* method, because this approach gives the best results. To achieve this optimization, a three-fold cross validation was performed. Once the parameters were optimized for each

Table 6 Comparison of the results *mg-LM-val* on three-fold cross validation, and percentage of difference in accuracy compared to the *a posteriori* optimization model *mg-LM*

Training	Query	*mg-LM*	*mg-LM-val*		Diff (%)
			Avg	Std-dev	
I	I	0.789	0.784	5.8×10^{-3}	−0.68
I	S	0.822	0.785	5.8×10^{-3}	−4.46
S	I	0.529	0.529	0.0	0
S	S	1.00	0.990	1.7×10^{-2}	−0.01

Table 7 Comparison of the results *mg-VGM-val* on three-fold cross validation, and percentage of difference in accuracy compared to the *a posteriori* optimization model *mg-VGM*

Training	Query	*mg-VGM*	*mg-VGM-val*		Diff (%)
			Avg	Std-dev	
I	I	0.794	0.788	6.4×10^{-3}	−2.64
I	S	1.00	0.939	5.3×10^{-2}	−6.07
S	I	0.594	0.594	0.0	0
S	S	1.00	0.990	1.7×10^{-2}	−0.01

of the three training/validation sets, we processed the test set using the whole training set. Table 6 shows the average (Avg) and standard deviation (Std-dev) of the three results obtained. The last column of Table 6 exhibits the difference in percentage for the evaluation measurement between the three-fold results and the *a posteriori* optimization.

As shown by Table 6, the results obtained by the cross validation and by a posteriori optimization are very similar. If we focus on the results of the I–I, S–I and S–S configurations, the differences are smaller than 1%, and for the configuration I–S the three-fold results are 4.46% lower. So, the optimization used on the validation sets provides satisfying results for a medium grain and for automatically defined visual concepts. We also tested three-fold cross validation with relationships, as presented in Table 7. Here again the results with the cross validations are very close to the *a posteriori* optimized results: the S–I and S–S results are almost equal.

Another conclusion drawn from Tables 6 and 7 is that, with a cross validation procedure, the usage of relationships still outperforms the results without relationships: +0.5% for the case I–I, +19.6% for I–S, and +12.3% for S–I. For the case S–S no improvement is achieved, which is also consistent with the *a posteriori* optimized results.

4.3.2 Comparing to SVM method

In order to assess the validity of our methods, we have compared the results with the state-of-the-art method in image categorization such as SVM classification method (implemented thanks to the *libsvm*[2]). We applied the same visual features used for graph model in our experiment. The input vector in SVM classifier is the early fusion of the multiple bag-of-word models. Then, each image class was trained with a corresponding SVM classifier using radial basis function (RBF). To optimize the kernel parametters, we train SVM classifiers with three-fold cross validation on the training set. Finally, these classifiers are used to classify the new query image.

Similar to above, we refer to the model with only the contribution of concept as LM and model with the spatial relation as VGM. For STOIC collection, we choose the *mg* division as a comparison model. Likewise for RobotVision collection, we choose the model $LM^{S.P}$ as LM and $VGM^{S\rightarrow P}$ as VGM.

[2] http://www.csie.ntu.edu.tw/cjlin/libsvm/

Table 8 Results on categorizing STOIC-101 and RobotVision'09 collections comparing to SVM method

	#class	*SVM*	*LM*	*VGM*
STOIC-101	101	0.744	0.789 (+6.0%)	**0.794** (+6.3%)
RobotVision'09				
Validation	5	0.535	0.579 (+8.2%)	**0.675** (+26.2%)
Test	6	0.439	0.416 (−5.2%)	**0.449** (+22.8%)

Table 8 summarizes the results obtained from both collection STOIC-101 and RobotVision'09. We can see that in all cases our VGMs outperform other methods. More precisely, with the integration of spatial relation into VGM helps improving the accuracy of classical approaches of LM by at least 2.5%. Especially with the RobotVision collection, VGMs increase roughly the accuracies of 22.8–26.2% comparing to SVM respectively for both test and validation set. Lastly, the VGMs retain medium to large improvements over the standard LMs in both image collections as confirmed in previous section.

4.3.3 Implementation

The system is implemented in C/C++ with the LTI-Lib[3] and compiled on a Linux platform. Image indexing and querying are performed on a 3.0 GHz quad-core computer with 8.0 Gb of memory. Training step takes about 2 h for the whole training images set from extracting visual features, clustering the concepts and modeling trained graphs. For the query step, it takes about 0.22 s on average (or 5 images/second) for computing the likelihood of graph query with all the graphs stored in database. However, the computation is highly parallelizable given graph models are stored and are processed independently. It shows that the graph matching step is very reliable for image matching comparing to classical graph matching algorithm.

5 Conclusion

We have introduced in this paper a graph-based model for representing image content. This graph captured the spatial relations among visual concepts associated with extracted regions of images. Theorically, our model fits within the language modeling approach for information retrieval, and extends previous proposals based on graph representation. On a more practical aspect, the consideration of regions and associated concepts allows us to gain generality in the description of images, a generality which may be beneficial when the usage of the system slightly differs from its training environment. This is likely to happen with image collections that, for example, use one or several images to represent a scene. On the other hand, querying

[3] http://ltilib.sourceforge.net/

a specific location with a group of images is very promising in future application (such as mobile localization) that allows higher accuracy rate with less computational effort comparing to video sequence. In addition, as demonstrated in the case of RobotVision, the way of combining of different image representations/features in the graph framework is more versatile comparing to other fusion approaches.

On the experimental side, we have conducted the test on two image collections (STOIC-101 and RobotVision'09). The experiments aim at assessing the validity of our approach in certain aspects. In particular, we showed that integrating spatial relations to represent images led to a significant improvement in the results. The model we have proposed is able to adequately match images and sets of images represented by graphs. As we conjectured, being able to abstract from a low level description allows robustness with respect to the usage scenarios. We also discussed on optimimizing the smoothing parameters of the language model with the cross validation technique based on training image set. We also demonstrated that our graph models outperformed the current state-of-the-art SVM method for image classification.

To summarize, the major contributions of our approach are: (1) a well-founded graph model for image indexing and retrieval, (2) with a smooth integration of spatial relations and visual concepts in the framework and (3) with a simpler and more effective graph matching process based on the language model.

In the future, many aspects can be considered to extend our graph model. First of all, as the language model is coming from textual domain, we could combine the graph representation of image with the graph representation of the annotated text as done in ImageCLEF photographic retrieval track. In our case, this could be integrated smoothly as they shared the same probabilistic framework. There is also the need to study different visual concepts and their spatial relations. This should be adapted following a specific image context or towards a typical scenario of image retrieval. Moreover, experiment on a large collection of images (such as ImageCLEF or VOC collection) could be interesting to test the scalability of our method. Last but not least, we also wish to investigate different possible couplings of the low level and high level representations, with the hope to come up with a single representation that could be used in a general case.

Acknowledgements This work was supported by the French National Agency of Research (ANR-06-MDCA-002). Pham Trong-Ton would like to thank Merlion programme of the French Embassy in Singapore for their supports during his Ph.D study.

References

1. Boutell MR, Luo J, Brown CM (2007) Scene parsing using region-based generative models. IEEE Trans Multimedia 9(1):136–146
2. Chang Y, Ann H, Yeh W (2000) A unique-id-based matrix strategy for efficient iconic indexing of symbolic pictures. Pattern Recogn 33(8):1263–1276
3. Chua TS, Tan KL, Ooi BC (1997) Fast signature-based color-spatial image retrieval. In: ICMCS 1997, pp 362–369
4. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2):1–60
5. Egenhofer M, Herring J (1991) Categorizing binary topological relationships between regions, lines and points in geographic databases. In: A framework for the definition of topological

relationships and an approach to spatial reasoning within this framework. Santa Barbara, CA
6. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. Int J Comput Vis 59(2):167–181
7. Gao S, Wang DH, Lee CH (2006) Automatic image annotation through multi-topic text categorization. In: Proc. of ICASSP 2006, pp 377–380
8. Han D, Li W, Li Z (2008) Semantic image classification using statistical local spatial relations model. Multimedia Tools and Applications 39(2):169–188
9. Hironobu YM, Takahashi H, Oka R (1999) Image-to-word transformation based on dividing and vector quantizing images with words. In: Neural networks, pp 405–409
10. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: SIGIR '03, pp 119–126
11. Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE PAMI 25(9):1075–1088
12. Lim J, Li Y, You Y, Chevallet J (2007) Scene recognition with camera phones for tourist information access. In: ICME'07
13. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2) 91–110
14. Maisonnasse L, Gaussier E, Chevallet J (2007) Revisiting the dependence language model for information retrieval. In: SIGIR '07
15. Maisonnasse L, Gaussier E, Chevalet J (2009) Model fusion in conceptual language modeling. In: ECIR '09, pp 240–251
16. Manning CD, Raghavan P, Schtze H (2009) Language models for information retrieval. In: An introduction to information retrieval. Cambridge University Press, pp 237–252
17. Mulhem P, Debanne E (2006) A framework for mixed symbolic-based and feature-based query by example image retrieval. Int J Inf Technol 12(1):74–98
18. Ounis I, Pasca M (1998) Relief: combining expressiveness and rapidity into a single system. In: SIGIR '98, pp 266–274
19. Papadopoulos G, Mezaris V, Kompatsiaris I, Strintzis MG (2007) Combining global and local information for knowledge-assisted image analysis and classification. EURASIP Journal on Advances in Signal Processing, Special Issue on Knowledge-Assisted Media Analysis for Interactive Multimedia Applications 2007
20. Pham TT, Maisonnasse L, Mulhem P (2009) Visual language modeling for mobile localization: Lig participation in Robotvision'09. In: CLEF working notes 2009. Corfu, Greece
21. Pham TT, Maisonnasse L, Mulhem P, Gaussier E (2010) Integration of spatial relationship in visual language model for scene retrieval. In: 8th IEEE int. workshop on content-based multimedia indexing
22. Pham TT, Mulhem P, Maisonnasse L (2010) Spatial relationships in visual graph modeling for image categorization. In: ACM SIGIR'10. Geneva, Switzerland
23. Pham TV, Smeulders AWM (2006) Learning spatial relations in object recognition. Pattern Recogn Lett 27(14):1673–1684
24. Ponte JM, Croft WB (1998) A language modeling approach to information retrieval. In: SIGIR '98
25. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proceedings of the international conference on computer vision, vol 2, pp 1470–1477
26. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content based image retrieval at the end of the early years. IEEE PAMI 22(12):1349–1380
27. Smith JR, Chang S-F (1996) Visualseek: a fully automated content-based image query system. In: Proceedings ACM MM, pp 87–98
28. Song F, Croft WB (1999) General language model for information retrieval. In: CIKM'99, pp 316–321
29. Tirilly P, Claveau V, Gros P (2008) Language modeling for bag-of-visual words image categorization. In: Proc. of CIVR 2008, pp 249–258
30. Won CS, Park DK, Park SJ (2002) Efficient use of mpeg-7 edge histogram descriptor. ETRI J 24(1)
31. Wu L, Li M, Li Z, Ma WY, Yu N (2007) Visual language modeling for image classification. In: MIR '07. ACM, New York, pp 115–124
32. Zhai C, Lafferty J (2001) A study of smoothing methods for language models applied to ad-hoc information retrieval. In: SIGIR '01, pp 334–342

Trong-Ton Pham received his B.Sc degree (with honor) in Information Technology from Vietnam National University (VNU) in 2004 and his M.Sc degree in Computer Science with speciality in Intelligent Artificial and Decision Making in 2006 from the University of Pièrre et Marie Curie (Paris 6), France. From 2007 to 2009, he was research officer with the Computer Vision and Image Understanding department at the Institute for Infocomm Research (I2R), Singapore. He is currently with the Computer Science Laboratory of Grenoble (LIG) at the Grenoble Institute of Technology (Grenoble-INP) to prepare his PhD thesis. His research interests include image processing/analysis, computer vision, information retrieval and image annotation.

Philippe Mulhem is currently a researcher in the Modeling and Multimedia Information Retrieval group (MRIM) of the Computer Science Laboratory of Grenoble (LIG), Grenoble, France. He was formerly director of the Image Processing and Applications Laboratory (IPAL) during five years in Singapore. His research interests include formalization and experimentation of image, video, and multimedia document indexing and retrieval. He received a PhD and Habilitation to Manage Research from the Joseph Fourier University, Grenoble. He is author and co-author of more than 80 papers in international and national journals, conference proceedings and book chapters.

Loic Maisonnasse is currently head of Research and Development at the TecKnowMetrix. He received his PhD degree in Computer Science in 2008 from the University Joseph Fourier (Grenoble I), France. From 2007 to 2008, he was research assistant with the DRIM team from the LIRIS Laboratory at INSA of Lyon, while he still connected to the MRIM team from the LIG Laboratory (Grenoble). His research interests include natural language processing, information retrieval, conceptual indexing, language modeling.

Eric Gaussier received his PhD in Computer Science, from University Paris 7, in 1995. After a year spent in the Linguistics Department of University Paris 7 as research assistant, he joined the Xerox Research Centre Europe (XRCE) in 1996, to work on textual indexing for information retrieval. He later became Area Manager of Learning and Content Analysis at XRCE, prior to joining the University Joseph Fourier and the Computer Science Laboratory of Grenoble as a professor in September 2006. He currently is a member of the Executive Board of the European Association for Computational Linguistics, a member of the Computer Science Panel of the European Research Council and a member of the Advisory Board of SIGDAT. His research focuses on probabilistic modeling of large document collections for information access, in particular on multilingual, multimedia collections, and applications as categorization, clustering and information retrieval.

Joo-Hwee Lim received his B.Sc (Hons I) and M.Sc (by research) degrees in Computer Science from the National University of Singapore and his Ph.D. degree in Computer Science & Engineering from the University of New South Wales. He is currently the Head of the Computer Vision and Image Understanding Department at the Institute for Infocomm Research (I2R), with staff strength of over sixty research scientists and engineers. He is the co-Director of IPAL (Image and Pervasive Access Laboratory), a French-Singapore Joint Lab (UMI 2955, January 2007–December 2010). He is bestowed the title of 'Chevallet dans l'ordre des Palmes Academiques' by the French Government in 2008. He is also the Director (Imaging) of a joint lab SAILOR (June 2009–June 2012) between I2R and Singapore Eye Research Institute where computer scientists and clinicians collaborate closely. Dr Lim's research experience includes connectionist expert systems, neural-fuzzy systems, handwritten character recognition, multi-agent systems, content-based image/video retrieval, scene/object recognition, and medical image analysis. Dr. Lim has published more than 150 international refereed journal and conference papers. He has also co-authored 15 patents (awarded and pending).