7-2009

A Bayesian approach integrating regional and global features for image semantic learning

Luong-Dong NGUYEN

Ghim-Eng YAP

Ying LIU

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Liang-Tien CHIA

See next page for additional authors

Author
Luong-Dong NGUYEN, Ghim-Eng YAP, Ying LIU, Ah-hwee TAN, Liang-Tien CHIA, and Joo-Hwee LIM

A BAYESIAN APPROACH INTEGRATING REGIONAL AND GLOBAL FEATURES FOR IMAGE SEMANTIC LEARNING

Luong-Dong Nguyen[1], *Ghim-Eng Yap*[2], *Ying Liu*[1], *Ah-Hwee Tan*[1], *Liang-Tien Chia*[1], *Joo-Hwee Lim*[2]

[1]School of Computer Engineering, Nanyang Technological University
[2]Institute for Infocomm Research (I^2R)
Singapore

ABSTRACT

In content-based image retrieval, the "semantic gap" between visual image features and user semantics makes it hard to predict abstract image categories from low-level features. We present a hybrid system that integrates global features (G-features) and region features (R-features) for predicting image semantics. As an intermediary between image features and categories, we introduce the notion of mid-level concepts, which enables us to predict an image's category in three steps. First, a G-prediction system uses G-features to predict the probability of each category for an image. Simultaneously, a R-prediction system analyzes R-features to identify the probabilities of mid-level concepts in that image. Finally, our hybrid H-prediction system based on a Bayesian network reconciles the predictions from both R-prediction and G-prediction to produce the final classifications. Results of experimental validations show that this hybrid system outperforms both G-prediction and R-prediction significantly.

1. INTRODUCTION

Content-based image retrieval (CBIR) has been actively studied since the early 1980's. Eakins and Graham [1] defined three different levels of queries. *Level 1*: Retrieval using low-level features such as color, texture, shape or spatial location of images. *Level 2*: Retrieval of objects of a given type identified by derived features, e.g. "find the images with flowers". Lastly, *Level 3*: Retrieval by abstract attributes describing the theme of scenes, e.g. "find images depicting a joyful crowd".

The traditional CBIR systems support Level 1 queries and index the images by their low-level features [2]. However, prior experiments have shown that low-level features often fail to aptly describe high-level semantics in users' mind [3]. To perform satisfactorily, a CBIR system should thus strive to narrow down this 'semantic gap'.

Image retrieval at either Level 2 or Level 3 are referred to as *semantic image retrieval*. Retrieval at Level 3 is difficult as it involves a significant amount of high-level reasoning about the purpose of the objects or scenes depicted [1]. Most existing systems operate at Level 2 [4, 5, 6], and usually define a set of concepts describing concrete objects such as flower and grass. Depending on whether salient regions are available, existing systems rely on either features from image regions (R-features) [5, 6], or global features describing the entire image (G-features) [4, 7].

The proposed system in our paper is different from previous works in the following ways:

- We advocate the use of mid-level concept. As category labels are often abstract, it is difficult for existing machine learning techniques to learn the relationships between low-level image features and image categories. Hence, we introduce mid-level concepts as an intermediary between the low-level features and the high-level categories. The mid-level concepts can be related to the image categories in an ontology.

- We make use of both R-features and G-features for classification. These features describe an image from two different perspectives. To get the final classification, we learn a Bayesian network to reconcile the outputs derived from the G-features and R-features.

Although there exist past works that have similarly proposed the usage of ontology-based Bayesian networks to enable probabilistic category predictions [8], they mainly model the dependencies among the image categories and specific salient objects, like *a flower*. To the best of our knowledge, there has not been any previous attempt to predict simultaneously both the mid-level concepts and the image category and then improve on their individual performance through Bayesian hybrid reconciliation.

The rest of this paper is organized as follows. In Section 2, we first explain each component of the proposed system in details. Experimental results are then presented in Section 3. Finally, Section 4 concludes this paper and highlights the differences between our hybrid system and the existing systems.

The work reported in this paper was supported in part by the I^2R-SCE Joint Lab, a collaboration between A*Star Institute for Infocomm Research and School of Computer Engineering, Nanyang Technological University.

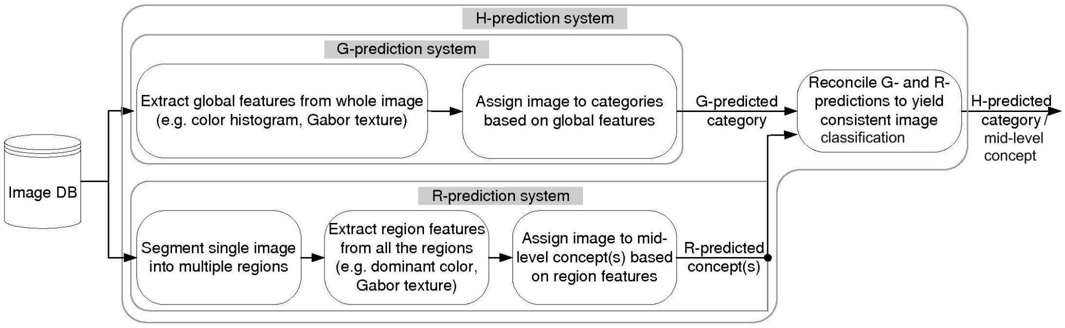

Fig. 1. Schematic diagram of the proposed hybrid system.

2. SYSTEM DESCRIPTION

Prior works on image semantic learning have generally used independent methods to learn the region and global aspects of images, which neglects the dependencies of semantics between the two levels. To capture these dependencies, so as to allow the region-based concepts and the global-feature-based categories to reconcile, we use a Bayesian network (BN) [9] based on an ontology describing the relations between the mid-level concepts and category labels.

We propose to analyze the semantics of an image at its region- and global-levels, before integrating these information to improve its classification. Figure 1 presents the schematic diagram of our proposed hybrid system, which comprises of three components: R-prediction, G-prediction and H-prediction. R-prediction learns the mid-level concepts within an image based on its R-features, and G-prediction learns the image's category based on its G-features. Our hybrid H-prediction system then uses an ontology-based Bayesian network to reconcile the outputs from the R- and G-prediction systems and yield a final prediction of the category.

At the global level, an image is described by features from the entire image, that include HSV domain color histogram and Gabor texture feature. Hue, Saturation and Value are uniformly quantized into 18*4*4 bins, resulting in a global color feature vector of 288 dimensions. For global texture feature, Gabor filters with four scales and six orientations are applied on the entire image to generate a feature vector of 48 dimensions.

At the regional level, each image is described by features from each of its segments. The features includes HSV domain color histogram , Gabor texture, and SIFT features. In contrast to global level, segment color feature is of three dimensions only, as each region is almost color-homogeneous and the dominant color is sufficient to represent the segment color.

Each global and regional prediction systems employ a relevant learning algorithm to best classify an image's category and its mid-level concepts. The R-prediction and G-prediction systems are flexible in that any existing machine learning method can be used as a part of a system. We will demonstrate two set of different learning algorithms for our experiment in two different domains.

With reference to Figure 1, our H-prediction system learns a BN that relates the predicted image categories from the G-prediction system to the mid-level concepts from R-prediction. The domain ontology built from the training data set is used as the background knowledge for the popular K2 structure learning algorithm proposed previously by Cooper and Herskovits [10]. The learned BN contains new relations and dependencies in comparison with the background network. The conditional probabilities that relate the nodes in the BN are learned automatically from the regional and global predictions. The network structure and the encoded joint probability distribution concisely capture the dependencies between the results of R-prediction and G-prediction systems for reconciliation.

For inconsistencies to be automatically corrected in a BN, the concept values and the category values have to be entered as inputs with allowances for uncertainty rather than as specific values. Their uncertainty values can be derived based on the confidence associated with each input. Probabilities of the BN nodes are then updated by applying the Bayes' Theorem:

$$Bel(X = x_i) = P(X = x_i|e) = \frac{P(e|X = x_i)P(X = x_i)}{P(e)} \quad (1)$$

where x_i is the i^{th} state of node X, e is the available evidence (the concept and category input values), and $Bel(X = x_i)$ is the probability of node X being in state x_i given evidence e.

The learned BN updates the probabilities in each node to be consistent with the rest of the network based on its encoded dependencies and conditional probabilities. In the end, the nodes with the highest scores (updated probability) is chosen as the predicted results. As such, our BN-based hybrid system ensures that the predictions from the region and global levels are reconciled for better classification.

3. EXPERIMENTAL VALIDATION

In this section, we present the experimental evaluation on two real-world problem domains. We investigate the image classification performances of the proposed hybrid system on our own Terrorist data set and the popular LabelMe data set.

3.1. Terrorist data set

The Terrorist data set comprises of 875 terrorism-related images that are downloaded from the Internet. The data set is divided into seven categories (key themes): *attack-scene*, *anti-terror*, *ceremony*, *rescue*, *government response*, *suspect*, and *victim*. Figure 2 shows some example images.

Fig. 2. Example images taken from the Terrorist data set.

3.1.1. Global and regional learning

At the global level, an image is described by features from the entire image, that include HSV domain color histogram and Gabor texture feature. A decision tree [11] is then used to predict the category of a given image.

On the other hand, each image is segmented into different regions. The Terrorist data set uses the JSEG algorithm [12]. The above mentioned decision tree is used to learn the concept for each region. Our method differs from the popular C4.5 decision tree in that we exploit the semantic templates of concepts in discretizing image features.

3.1.2. Hybrid learning

Describing the relations between the mid-level concepts and categories, an ontology in Figure 3 captures dependency between two layers of meanings.

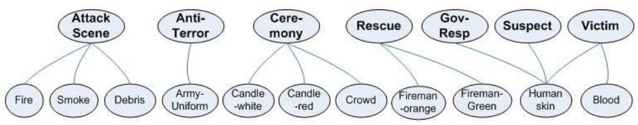

Fig. 3. Ontology of categories (above) and mid-level concepts (below) built from the Terrorist data set.

A BN model is learned for the H-prediction system based on this ontology. Regional and global predictions are used as the inputs to the BN model learned for the H-prediction system. The probabilities or the *beliefs* of the network nodes are then updated as described earlier in Section 2.

Figure 4 compares the performance of the global and hybrid prediction systems on the Terrorist data set. The H-prediction system outperforms the G-prediction in categorizing images for most of the categories. The average performance is 43% accuracy in G-prediction and up to 50.6% in H-prediction.

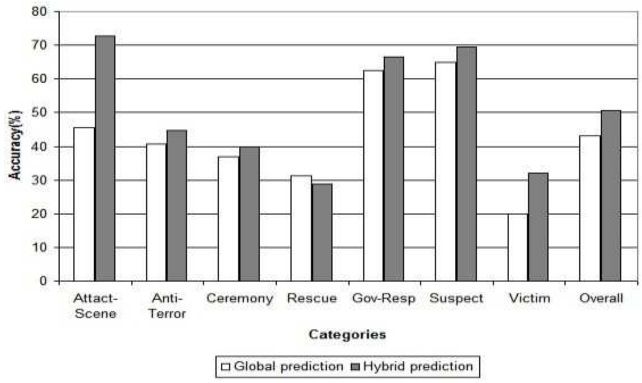

Fig. 4. Categorization Performance on the Terrorist data set.

3.2. LabelMe data set

Besides the Terrorist data set, we also use a part of the popular LabelMe data set in our experiment to demonstrate the ability of the hybrid system. The part of LabelMe data set we use includes 526 indoor images, 242 landscape ones, and 178 street ones. The three categories are: *indoor*, *street*, and *landscape*

3.2.1. Global and regional learning

The same kind of global features used in Terrorist data set is applied for LabelMe images. For learning algorithms, we choose support vector machine to predict the image category as SVM provides good results in this case.

Different from Terrorist data set, every image in LabelMe is provided with image segments and segment's labels. We extract features like shape, texture, color, and location from each segment. Applying a method proposed by [13], we learn separate distance function for each image segment or exemplar. Then the learned distance function combined with the nearest neighbor algorithm are used for classification of each testing image segment.

3.2.2. Hybrid learning

We have twenty eight mid-level concepts in relation with the three categories *indoor*, *landscape*, and *street*. The ontology in Figure 5 describes the relations between the mid-level concepts and categories.

This ontology is used as background knowledge to build a BN for reconciling results from global and regional prediction systems as described earlier in Section 2.

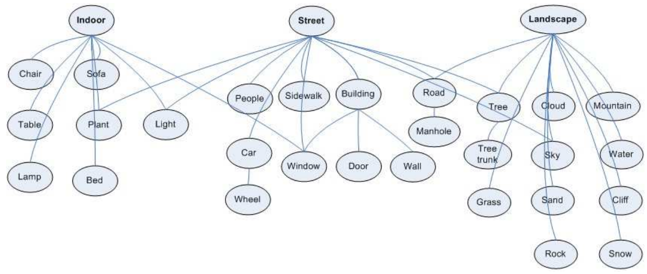

Fig. 5. Ontology of categories (above) and mid-level concepts (below) built from LabelMe data set.

Figure 6 compares the performance of the H-prediction system with the R-prediction's one. In this case G-prediction over categories is already very good, which helps H-prediction to improve mid-level concept classification but also makes little space to improve categorization. Some mid-level concept accuracies are higher considerably while few are lower. Overall we increase the accuracy from 31.5% in R-prediction up to 40.9% in H-prediction. Compared with result from [8] which also used a part of LabelMe data set, certain common concept like *car* performs much better.

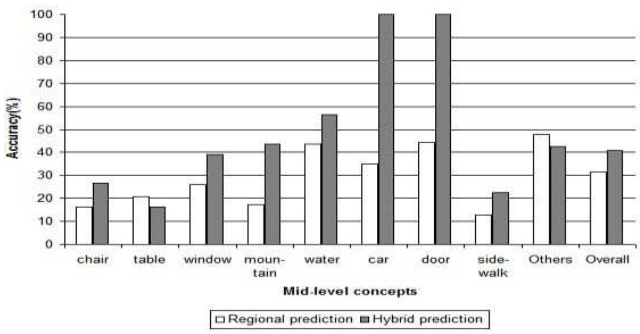

Fig. 6. Mid-level Concepts Classification Performance on the LabelMe data set.

4. CONCLUSION

We have proposed a hybrid ontology-based approach for image semantic learning. The two key features of this approach are as follows:

- It introduces mid-level concepts that bridge the gap between low-level image features and high-level abstract image semantics.
- It integrates the region features with global features through a Bayesian network built on ontologies which relate mid-level concepts and image categories.

Experimental evaluations on two real-life domains demonstrate that our hybrid approach improves significantly pure region- or global-feature-based methods.

5. REFERENCES

[1] J. Eakins and M. Graham, "Content-based image retrieval," Tech. Rep., University of Northumbria, 1999.

[2] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *J. of Intelligent Information Systems*, vol. 3, no. 3-4, pp. 231–262, 1994.

[3] X. S. Zhou and T. S. Huang, "CBIR: From low-level features to high-level semantics," in *Proc. of SPIE Image & Video Communic. and Processing*, 2000, pp. 426–431.

[4] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. of CVPR*, 2004, pp. 178–185.

[5] F. Jing, M. Li, L. Zhang, H. Zhang, and B. Zhang, "Learning in region-based image retrieval," in *Proc. of Inter. Conf. on Image & Video Retr.*, 2003, pp. 206–215.

[6] C. P. Town and D. Sinclair, "Content-based image retrieval using semantic visual categories," Tech. Rep. MV01-211, Society for Manufacturing Engineers, 2001.

[7] I. K. Sethi, I. L. Coman, and D. Stan, "Mining association rules between low-level image features and high-level concepts," in *Proc. of SPIE DMKD: Theory, Tools, and Technology III*, Mar. 2001, vol. 4384, pp. 279–290.

[8] Y. Gao and J. Fan, "Incorporating concept ontology to enable probabilistic concept reasoning for multi-level image annotation," in *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, New York, NY, USA, 2006, pp. 79–88, ACM.

[9] J. Pearl, *Probabilistic Reasoning in Intell. Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

[10] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[11] Y. Liu, D. S. Zhang, G. Lu, and W.-Y. Ma, "Deriving high-level concepts using fuzzy ID3 decision tree for image retrieval," in *Proc. of Inter. Conf. on Acoustics, Speech and Signal Processing*, Mar. 2005, pp. 501–504.

[12] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE TPAMI*, vol. 23, no. 8, pp. 800–810, Aug. 2001.

[13] T. Malisiewicz and A. A. Efros, "Recognition by association via learning per-exemplar distances," in *CVPR*, June 2008.