A KNOWLEDGE STRUCTURING TECHNIQUE FOR IMAGE CLASSIFICATION

Le Dong, Student Member, IEEE, and Ebroul Izquierdo, Senior Member, IEEE

Department of Electronic Engineering, Queen Mary, University of London,
London E1 4NS, U.K.
{le.dong, ebroul.izquierdo}@elec.qmul.ac.uk

ABSTRACT

A system for image analysis and classification based on a knowledge structuring technique is presented. The knowledge structuring technique automatically creates a relevance map from salient areas of natural images. It also derives a set of well-structured representations from low-level description to drive the final classification. The backbone of the knowledge structuring technique is a distribution mapping strategy involving two basic modules: structured low-level feature extraction using convolution neural network and a topology representation module based on a growing cell structure network. Classification is achieved by simulating high-level top-down visual information perception and classifying using an incremental Bayesian parameter estimation method. The proposed modular system architecture offers straightforward expansion to include user relevance feedback, contextual input, and multimodal information if available.

***Index Terms*—** Image classification, knowledge structuring, topology representation

1. INTRODUCTION

To build the next generation of intelligent retrieval hinges on solving tasks such as indexing, classification, and relevance feedback. The specialized systems mostly based on the analysis of low-level image primitives have been powerful approaches to classification [1], [2]. Relying on low-level features only, it is possible to automatically extract important relationships between images. However, such an approach lacks potential to achieve accurate image classification for generic automatic retrieval. A significant number of semantic-based approaches address this fundamental problem by utilizing automatic generation of links between low- and high-level features. For instance, Dorado et al. introduced in [3] a system that exploits the ability of support vector classifiers to learn from relatively small number of patterns. Based on a better understanding of visual information elements and their role in synthesis and manipulation of their content, an approach called "computational media aesthetics" studies the dynamic nature of the narrative via analysis of the integration and sequencing of audio and video [4]. Semantic extraction using fuzzy inference rules has been used in [5]. These approaches are based on the premise that the rules needed to infer a set of high-level concepts from low-level descriptors can not be defined a priori. Rather, knowledge embedded in the database and interaction with an expert user is exploited to enable learning.

Closer to the models described in this paper, knowledge and feature based classification as well as topology representation is important aspect that can be used to improve classification performance. The proposed system uses a knowledge structuring technique to approximate human-like inference. The system consists of two main parts: knowledge structuring and classification. In this paper a knowledge structuring technique is exploited to build a system for image analysis and classification following human perception and interpretation of natural images. The proposed approach aims at, to some extent, mimicking the human knowledge structuring system and to use it to achieve higher accuracy in image classification. A method to generate a topology representation based on the structured low-level features is developed. Using this method, the preservation of new objects from a previously perceived ontology in conjunction with the colour and texture perceptions can be processed autonomously and incrementally. The topology representation network structure consists of the posterior probability and the prior frequency distribution map of each image cluster conveying a given semantic concept.

Contrasting related works from the conventional literature, the proposed system exploits known fundamental properties of a suitable knowledge structuring technique to achieve classification of natural images. An important contribution of the presented work is the dynamic preservation of high-level representation of natural scenes. Another important feature of the proposed system is the constant evaluation of the involved confidence and support measures used in the image classification. As a result, continually changing associations for each class is achieved. These two main novel features of the system together with an open and modular system architecture, enable important system extensions to include user relevance feedback, contextual input, and multimodal information if available. These important features are the scope of ongoing implementations and system extensions targeting enhanced robustness and classification accuracy. The knowledge structuring technique is given in Section 2. A detailed description of the classification process is given in Section 3. The selected result and a comparative analysis of the proposed approach with other existing methods are given in Section 4. The paper closes with conclusions and an outline of ongoing extensions in Section 5.

2. KNOWLEDGE STRUCTURING

The knowledge structuring technique is described here. The proposed knowledge structuring technique automatically creates a relevance map from the salient image areas detected by previously proposed biologically inspired visual selective attention model [6]. It also derives a set of well-structured representations from low-level description to drive the classification. The backbone of this technique is a distribution mapping strategy involving two basic modules: structured low-level feature extraction using convolution neural network (CNN) and topology representation based on growing cell structure network (GCS).

2.1. Structured Low-level Feature Extraction Using Convolution Neural Network

The CNN architecture is capable to characterize and recognize variable object patterns directly from images free of pre-processing, by automatically synthesizing its own feature extractors from a large data set [7]. A framework is set up to extract and build structured low-level features of an object via CNN architecture in this paper.

As shown in Fig. 1, a CNN for structured low-level feature extraction consists of a subsequent processing such as convolutional operation, local sampling, further convolutional operation and subsampling. The input of the CNN is an essential area detected by aforementioned biologically inspired visual selective attention model [6]. Following that, colour and texture features extracted from the essential area are used as parallel detailed information. The convolutional operation is typically followed by the local sampling that makes normalization and sampling around the considered neighbourhood. Technically, further convolutional operation and subsampling are necessary for the well represented feature maps. Finally, the structured low-level feature can be extracted using such kind of CNN architecture.

The colour/texture feature for the detected essential area is represented by a certain dimensional vector including specific feature maps. In our implementation, each area is represented by a 168-dimensional vector containing colour and texture features, with each feature map size of 2x2. These kind of structured features outweigh simplex low-level features in representation and application for the further clustering.

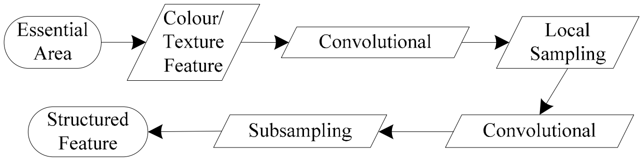

Fig. 1. The architecture of convolution neural network.

2.2 Topology Representation Based on Growing Cell Structure Network

In this paper topographic maps are used to reflect the result of the GCS [8]. A transformation of the input pattern space into the output feature space preserves the topology. A meaningful distribution preserving mapping coordinate system for different input features is created and spatial locations signify intrinsic statistical features of input patterns. The number of clusters and the connections among them are dynamically assigned during the training of the network. New cells can be inserted and existent cells can be removed in order to adapt the output map to the distribution of the input vectors. By growing the GCS in two dimensions, the appropriate topology of the network is found.

The topology representation based on the GCS algorithmic steps are summarized below:

1. Create initial network topology; initialize node parameters $Q(i)$, for $i = 1,\cdots,k+1$, with small values.
2. Repeat steps 3-5 for $t = 1,\cdots,L$, where t denotes the number of presentation epochs.
3. Randomly select input vector V and find the winner node z. Move the winner node towards the input $F_{Qz} = h_z(V - Q_z)$, where h_z denotes a learning rate of the winner node.
4. Increase the frequency counter of the winner node by a fixed value Δ, where Δ is set 1 in our experiments.
5. Correct parameters of its direct neighbours $F_{Qn} = h_n(V - Q_n)$.
6. After L epochs calculate renormalized frequency for each node.
7. Find node with highest frequency Q_1; find direct neighbour Q_2 with largest distance $\| Q_1 - Q_2 \|$; add new node $Q_n = (Q_1 + Q_2)/2$ in between, link Q_1, Q_2, and Q_n.
8. Stop if the epoch is too large; otherwise go to step 2.

The input of the algorithm is a set of extracted structured low-level features generated by CNN. GCS is also integrated into the mechanism of topology preserving to maintain the gracious network structure. Various topology maps subtly reflect the characteristics of distinct image groups which are closely related to the order of the forthcoming visual information. Furthermore, the extracted information from perceptions in colour and texture domains can also be used to represent objects [6].

3. CLASSIFICATION

Using the output generated by the knowledge structuring technique, high-level classification is achieved. The proposed high-level classification approach follows a high-level perception and classification model that mimics the top-down attention mechanism in primates' brain. The attentional area is generated using the task independent representation and detection model based on a maximum likelihood approach. A high-level perception and classification model employs a generative mode based on a dynamical Bayesian parameter estimation method. The structured low-level features generated by CNN architecture are used as the input information of the specific represented object.

On independently learning the conditional density of the pattern with all other existing patterns, the novel pattern might be added dynamically into the current pattern setting. Considering *n* training data samples from a pattern ω , with each pattern featured by f $(f < n)$ codebook vectors, learning is progressing with updating the corresponding codebook vectors whenever a novel data vector u is enrolled. The prior probabilities $p(\omega)$ and the conditional densities $p(u \mid \omega)$ of the pattern can be learned independently by generative approach. Furthermore, the posterior probabilities are obtained using the Bayes' theorem:

$$p(\omega \mid u) = \frac{p(u \mid \omega)p(\omega)}{p(u)} = \frac{p(u \mid \omega)p(\omega)}{\sum_{j} p(u \mid j)p(j)}$$

Based on [9], a vector quantizer is used to extract codebook vectors from training samples in order to estimate the conditional density of the feature vector u given the pattern ω . The conditional densities of the pattern are approximated using a mixture of Gaussians, assuming identity covariance matrices, with each centred at a codebook vector. Finally, the conditional densities of the pattern can be represented as [9]:

$$p_U(u \mid \omega) \propto \sum_{j=1}^{f} m_j * \exp\left(- \| u - v_j \|^2 / 2\right),$$

where $v_j (1 \le j \le f)$ denotes the codebook vectors, m_j is the proportion of training samples assigned to v_j .

As indicated in [10], a task for target detection activates a non-specific representation model for a desired target area. The high-level perception and classification model can just compute the similarity of the statistical properties for candidate attended areas [10]. Finally, an integrated essential map for the specific target is generated. When human beings focus its attention in a given image area, the prefrontal cortex gives a competition bias related to the target object in the inferior temporal area [10]. Then, the inferior temporal area generates specific information and transmits it to the high-level attention generator which conducts a biased competition [10]. Therefore, the high-level perception and classification model can assign a specific pattern to a target area, which possesses the maximum likelihood.

Assuming the prior density is essentially uniform, the posterior probability can be estimated as [9], [11]:

$$\arg\max_{\omega \in \Omega} \left\{ p(\omega \mid u) \right\} = \arg\max_{\omega \in \Omega} \left\{ p_U(u \mid \omega) p(\omega) \right\},$$

where Ω is the set of patterns. Moreover, the high-level perception and classification model can generate a specific attention area based on the pattern classification. On the other hand, it might provide informative control signals to the internal effectors [10]. To some extent, this might be regarded as an incremental framework for knowledge structuring with human interaction.

4. EXPERIMENTAL EVALUATION

Given a collection of completely unlabelled images, the goal is to automatically discover the visual categories present in the data and localize them in the topology representation of the image. To this end, a set of quantitative experiments with progressively increasing level of topology representation complexity was conducted.

The Corel database was used, which was labelled manually with eight predefined concepts. The concepts are "building", "car", "autumn", "rural scenery", "cloud", "elephant", "lion", and "tiger". However, other images not representing any of these concepts were also considered in the dataset for evaluation, thus containing total 7000 images. In order to assess the accuracy of the image classification, a performance evaluation based on the amount of missed detections (*MD*) and false alarms (*FA*) was conducted. The obtained results are given in Table I, where *D* is a sum of true memberships for the corresponding recognized class, *MD* is a sum of the complement of the full true memberships and *FA* is a sum of false memberships.

TABLE I
IMAGE CLASSIFICATION AND RETRIEVAL

Class	*D*	*MD*	*FA*	*Recall*	*Precision*
building	830	170	100	83%	89%
autumn	460	100	80	82%	85%
car	840	160	120	84%	88%
cloud	920	80	60	92%	94%
tiger	880	120	160	88%	85%
rural scenery	360	80	60	82%	86%
elephant	940	60	140	94%	87%
lion	850	150	120	85%	88%

The proposed technique was compared with an approach based on multi-objective optimization (MOO) [12] and another using Bayesian networks for concept propagation [13]. Table II shows a summary of results on some subsets of the image categories coming out from this comparative evaluation.

TABLE II
PRECISION COMPARISON WITH TWO OTHER APPROACHES

(%)	*Proposed*	*Bayesian*	*MOO*
building	89	72	70
cloud	94	84	79
lion	88	92	88
tiger	85	60	60

The selection of the subset depends on the common categories among comparable approaches. It can be observed that the proposed technique outperforms the other two approaches. Even though multi-objective optimization can be optimized for a given concept, the result of the proposed technique performs better in general. Except for the class lion, in which the Bayesian approach delivers the highest accuracy, the proposed technique performs substantially better in other cases. The exception of the lion is due to the interference from complex background environment, while there is more larruping colour and texture information in other categories. It could be compensated by the prior information between affiliated features and semantic meaning. This summary of results truly represents the observed outcomes with other classes and datasets used in the experimental evaluation and evidences our claim that the proposed technique has good discriminative power and it is suitable for retrieving natural images in large datasets.

We also compared the performance of the proposed approach with two binary image classifiers: one based on ant colony optimization (ACO/COP-K-Means) [14], and the other using particle swarm optimization and self organizing feature maps (PSO/SOFM) [15]. A summary of results on some subsets of the image categories is given in Table III. It can be concluded that the proposed technique also outperforms other classical approaches.

TABLE III
COMPARISON WITH TWO OTHER BINARY CLASSIFIERS

(%)	*ACO/COP-K-Means*		*PSO/SOFM*		*Proposed*	
	P	*R*	*P*	*R*	*P*	*R*
lion	55	62	48	69	88	85
elephant	71	71	74	65	87	94
tiger	63	58	68	64	85	88
cloud	62	57	69	63	94	92
car	65	56	70	64	88	84
building	65	62	51	74	89	83

5. CONCLUSION

A knowledge structuring technique for image classification is presented. By utilizing biologically inspired theory and knowledge structuring technique, the system simulates the human-like image classification and inference. Since the knowledge structuring base creation depends on information provided by expert users, the system can be easily extended to support intelligent retrieval wit enabled user relevance feedback. The whole system can automatically generate relevance maps from the visual information and classifying the visual information using learned information.

6. REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content based image retrieval at the end of the early years,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 12, pp.1349-1380, 2000.

[2] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7, Multimeida Content Description Interface*, John Wiley & Sons, 2003.

[3] A. Dorado, D. Djordjevic, W. Pedrycz, and E. Izquierdo, “Efficient image selection for concept learning,” *Proc. Vision, Image and Signal Processing*, vol. 153, no. 3, pp. 263-273, 2006.

[4] C. Dorai and S. Venkatesh, “Bridging the semantic gap with computational media aesthetics,” *IEEE Multimedia*, vol. 10, no. 2, pp. 15–17, 2003.

[5] A. Dorado, J. Calic, and E. Izquierdo, “A rule-based video annotation system,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no.5, pp. 622 – 633, 2004.

[6] L. Dong, S. W. Ban, I. Lee and M. Lee, “Incremental knowledge representation model based on visual selective attention”, *Neural Information Processing – Letters and Reviews*, vol. 10, no. 4-6, pp. 115-124, 2006.

[7] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, “Face recognition: A convolutional neural-network approach,” *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Feb. 1997.

[8] B. Fritzke, “Growing cell structures - a self-organizing network for unsupervised and supervised leaming,” *Neural Networks*, vol. 7, no. 9, pp. 1441-1460, 1994.

[9] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. J. Zhang, “Image classification for content-based indexing,” *IEEE Trans. Image Processing*, vol. 10, no. 1, pp. 117-130, 2001.

[10] L. J. Lanyon and S. L. Denham, “A model of active visual search with object-based attention guiding scan paths,” *Neural Networks*, vol. 17, no. 5-6, pp.873-897, 2004.

[11] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley & Sons, Inc. 2001.

[12] Q. Zhang, and E. Izquierdo, “A multi-feature optimization approach to object-based image classification,” *Proc. Int. Conf. Image and Video Retrieval*, pp. 310-319, 2006.

[13] F. F. Li, R. Fergus, and P. Perona, “A bayesian approach to unsupervised one-shot learning of object categories,” *Proc. IEEE Int. Conf. on Computer Vision*, vol. 2, pp. 1134-1141, 2003.

[14] S. Saatchi and Ch. Ch. Hung, “Hybridization of the ant colony optimization with the K-Means algorithm for clustering,” Springer-Verlag, Heidelberg, vol. 3540, pp. 511–520, 2005.

[15] K. Chandramouli and E. Izquierdo, “Image classification using chaotic particle swarm optimization,” *in Proc. IEEE Int. Conf. on Image Processing*, Atlanta, USA, pp. 3001-3004, Oct. 2006.