16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

HYPERSPHERE TOPOLOGY CREATION FOR IMAGE CLASSIFICATION

Le Dong and Ebroul Izquierdo

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

ABSTRACT

A kind of topology creation strategy for image analysis and classification is presented. The topology creation strategy automatically generates a relevance map from essential regions 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 topology creation strategy is a distribution mapping rule involving two basic modules: structured low-level feature extraction using convolution neural network and a topology creation module based on a hypersphere neural 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.

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 topology creation strategy to approximate human-like inference. The system consists of two main parts: topology creation and classification. In this paper a topology creation strategy 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 topology creation strategy for knowledge structuring 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. TOPOLOGY CREATION

The topology creation strategy for knowledge structuring is described here. The proposed topology creation strategy automatically generates a relevance map from the essential regions 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 rule involving two basic modules: structured low-level feature extraction using convolution neural network (CNN) and topology creation based on hypersphere neural network (HNN).

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 preprocessing, by automatically synthesizing its own feature extractors from a large data set [7]. Moreover, the use of receptive fields, shared weights, and spatial subsampling in such a neural model provides some degrees of partial invariance to translation, rotation, scale, and deformation. CNN has been applied to object detection and face recognition when sophisticated preprocessing is to be avoided and raw visual information are to be processed directly [7]. The CNN extracts successively large volume of features in a hierarchical set of layers. Furthermore, the convolution network topology is more similar to biological networks based on receptive fields and improves tolerance to local distortions. A framework is set up to extract and build structured low-level features of an object via CNN architecture in this paper. A sparse coding scheme is considered in order to extract and represent structured low-level features of an arbitrary object using CNN.

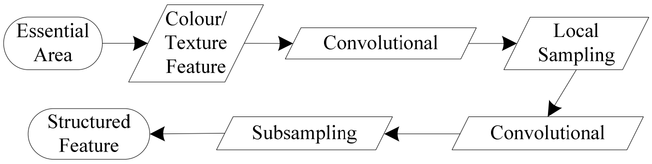

Figure 1 - The architecture of convolution neural network

As shown in Figure 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. Each processing contains one or more levels. Multi-levels are usually used in each processing in order to detect multi-features. The input of the CNN is essential areas detected by the aforementioned biologically inspired visual selective attention model [6]. In the illustrated CNN architecture shown in Figure 1, colour and texture features extracted from essential areas 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, the initial colour feature components contain hue, saturation and intensity. After the first convolutional operation, we have a 192-dimensional vector with feature map size of 8x8 for each component. Then a 48-dimensional vector is remained on local sampling, with feature map size of 4x4 for each component. Following the second convolutional operation, we get a 96-dimensional vector with feature map size of 4x4. The final structured colour feature vector is generated after the subsampling, 24-dimentional vector with feature map size of 2x2. The eight directional Gabor filter is used to generate the initial texture features. After the first convolutional operation, we have a 512-dimensional vector with feature map size of 8x8 for each direction. Then a 128-dimensional vector is remained on local sampling, with feature map size of 4x4 for each direction. Following the second convolutional operation, we get a 576-dimensional vector with feature map size of 4x4. The final structured texture feature vector is generated after the subsampling, 144-dimentional vector with feature map size of 2x2. Therefore, each area is totally 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.

2.2 Topology Creation Based on Hypersphere Neural Network

A formal definition for a topology defined in terms of set operations is given as follows:

A set $\boldsymbol{X}$ along with a collection $\boldsymbol{T}$ of subsets of it is said to be a topology if the subsets in $\boldsymbol{T}$ obey the following properties [8]:

- The subsets $\boldsymbol{X}$ and the empty set Φ are in $\boldsymbol{T}$
- Whenever sets $\boldsymbol{A}$ and $\boldsymbol{B}$ are in $\boldsymbol{T}$, then so is $A \bigcap B$
- Whenever two or more sets are in $\boldsymbol{T}$, then so is their union

In this paper hyper topographic maps are used to reflect the configuration of the HNN [9]. A transformation of the input pattern space into the output feature space preserves and develops 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 hyperspheres can be created and the original hypersphere can be spread in order to adapt the output map to the distribution of the input patterns. By developing the HNN in hypersphere dimensions, the appropriate topology of the input pattern is found.

The radius of the hypersphere is very important during the procedure of topology creation. Here, the maximum size of hypersphere is bounded by a user defined value between 0 and 1. The learning algorithm is composed of initialization and training stages [9].

Initialization: All topology structures are initialized by creating a hypersphere with a first pattern belonging to some class of the module.

Training: An input pattern of any class is applied to the corresponding topology only and membership of the input pattern with all the hyperspheres belonging to that topology is calculated. After this the input pattern is accepted either by spread of the original hypersphere or creation of the new hypersphere as described below.

- Acceptance by spread of the original hypersphere: Each hypersphere has maximum limit on its radius. The pattern is included in the existing hypersphere if radius of that hypersphere after increment is not larger than $\sqrt{\sum_{m=1}^{Num}\left(c_{lm}^{p} - tr_{gm}\right)^2}$, where $C_l^p = \{c_{l1}^p, \ldots, c_{lNum}^p\}$ denotes the centre points of hyperspheres, $TR_g = \{tr_{g1}, \ldots, tr_{gNum}\}$ represents the training set of pattern *g*. The hypersphere is spread to include the input pattern by modifying its radius if the aforementioned criterion is satisfied. The procedure is as follows:
  - Referring to the method proposed in [9] to determine whether the current input pattern is contained by any one of the existing hyperspheres.
  - If the input pattern is included then the remaining steps in the training process are skipped and the training continues with the next training pair.
  - If the current input pattern falls outside the hypersphere, then the hypersphere is spread to include pattern if the aforementioned criterion is met.
- Acceptance by creation of new hypersphere: If it fails to include the input pattern described above, then a new hypersphere is created.

The input of the algorithm is a set of extracted structured low-level features generated by CNN. HNN is also integrated into the mechanism of topology development 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 topology creation, 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 [10], 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 [10]:

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

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 [11], 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 [11]. 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 [11]. Then, the inferior temporal area generates specific information and transmits it to the high-level attention generator which conducts a biased competition [11]. 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 [10], [12]:

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

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 [11]. 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	845	155	102	85%	89%
autumn	490	70	68	88%	88%
car	878	122	83	88%	91%
cloud	926	74	57	93%	94%
tiger	891	109	127	89%	88%
rural scenery	392	48	56	89%	88%
elephant	920	80	109	92%	89%
lion	873	127	96	87%	90%

The proposed technique was compared with an approach based on multi-objective optimization (MOO) [13] and another using Bayesian networks for concept propagation [14]. 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	90	92	88
tiger	88	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 and balances the result on precision and recall measures. 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.

The proposed technique was also compared with the topology representation approaches based on Growing Neural Gas (GNG) [15] and another using Growing When Required (GWR) algorithm [16]. Table III and IV show the summary of results on precision and recall comparative evaluation, respectively.

TABLE III
PRECISION COMPARISON WITH OTHER TOPOLOGY REPRESENTATION APPROACHES

(%)	*Proposed*	*GNG*	*GWR*
building	89	88	88
autumn	88	85	81
car	91	90	92
cloud	94	92	90
tiger	88	86	90
rural scenery	88	83	80
elephant	89	88	87
lion	90	90	90

TABLE IV
RECALL COMPARISON WITH OTHER TOPOLOGY REPRESENTATION APPROACHES

(%)	*Proposed*	*GNG*	*GWR*
building	85	85	84
autumn	88	79	75
car	88	88	89
cloud	93	90	90
tiger	89	88	87
rural scenery	89	80	82
elephant	92	92	93
lion	87	87	88

It can be observed that the proposed technique using HNN outperforms the other two approaches. Even though the other two topology representation approaches can be optimised for a given concept, the result of the proposed technique performs better in general. It is convinced that these three kinds of neural network algorithms work for the topology representation in this case.

5. CONCLUSION

A topology creation strategy 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.

ACKNOWLEDGEMENT

The research work leading to this paper has been partially supported by the European Commission under the IST research network of MESH of the 6th Framework programme.

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] R. Bishop and S. Goldberg, *Tensor Analysis on Manifolds,* New York: Dover, 1980.

[9]U. V. Kulkarni, D.D Doye, and T. R. Sontakke, "Fuzzy hypersphere neural network for rotation invariant handwritten character recognition," in *Proc.of 10th int. IEEE Conference on Fuzzy Systems,* Melbourne, Australia, Dec. 2001.

[10] 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.

[11] 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.

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

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

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

[15] G. Tesauro, D. S. Touretzky, and T. K. Leen, *Advances in Neural Information Processing Systems*, MIT Press, Cambridge MA, pp. 625-632, 1995.

[16] S. Marsland, J. Shapiro, U. Nehmzow, “A self-organising network that grows when required,” *Neural Networks*, vol. 15, pp. 1041-1058, 2002.