A Knowledge Synthesizing Approach for Classification of Visual Information

Le Dong and Ebroul Izquierdo

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

Abstract. An approach for visual information analysis and classification is presented. It is based on a knowledge synthesizing technique to automatically create a relevance map from essential areas in natural images. It also derives a set of well-structured representations from low-level description to drive the final classification. The backbone of this approach is a distribution mapping strategy involving a knowledge synthesizing module based on an intelligent growing when required network. Classification is achieved by simulating the high-level top-down visual information perception in primates followed by incremental Bayesian parameter estimation. The proposed modular system architecture offers straightforward expansion to include user relevance feedback, contextual input, and multimodal information if available.

Keywords: classification, essence map, knowledge synthesizing.

1 Introduction

Classification and retrieval of visual information is a critical task for high-level computer based understanding of visual information. Current systems for classification of visual information are mostly based on the analysis of low-level image primitives [1], [2]. Relying on low-level features only, it is possible to automatically extract important relationships between images. However, such approaches lack potential to achieve accurate 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.

G. Qiu et al. (Eds.): VISUAL 2007, LNCS 4781, pp. 17–25, 2007.

Closer to the models described in this paper, knowledge and feature based classification as well as topology preservation are important aspects that can be used to improve classification performance. The proposed system uses a knowledge synthesizing approach to approximate human-like inference. The system consists of two main parts: knowledge synthesizing and classification. In this paper a knowledge synthesizing approach is exploited to build a system for visual information analysis following human perception and interpretation of natural images. The proposed approach aims at, to some extent, mimicking the human knowledge synthesizing system and to use it to achieve higher accuracy in classification of visual information. A method to generate a knowledge synthesizing 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 knowledge synthesizing network consists of the posterior probability and the prior frequency distribution map of each visual information cluster conveying a given semantic concept.

Contrasting related works from the conventional literature, the proposed system exploits known fundamental properties of a suitable knowledge synthesizing model to achieve classification of natural images. An important contribution of the presented work is the dynamic preservation of high-level representation of natural scenes. As a result, continually changing associations for each class is achieved. This novel feature 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 essence map model for feature extraction is described in Section 2. The knowledge synthesizing approach is given in Section 3. A detailed description of the high-level classification is given in Section 4. The selected result and a comparative analysis of the proposed approach with other existing methods are given in Section 5. The paper closes with conclusions and an outline of ongoing extensions in section 6.

2 Essence Map Model

Five features of intensity (*I*), edge (*E*), colour (*C*), orientation (*O*), and symmetry (*S*) are used to model the human-like bottom-up visual attention mechanism [6], as shown in Fig. 1. The roles of retina cells and LGN are reflected in previously proposed attention models [7]. The feature maps are constructed by centre-surround difference and normalization (*CSD & N*) of the five bases. This mimics the on-centre and off-surround mechanism in the human brain. Subsequently, they are integrated using a conventional independent component analysis (*ICA*) algorithm [8]. The symmetry information is used as a joint basis to consider shape primitives in objects [9], which is obtained by the noise tolerant general symmetry transform (*NTGST*) method [7]. The ICA can be used for modelling the role of the primary visual cortex for the redundancy reduction according to Barlow's hypothesis and Sejnowski's results [8].

Barlow's hypothesis is that human visual cortical feature detectors might be the end result of a redundancy reduction process [10]. Sejnowski's result states that the ICA is the best way to reduce redundancy [8].

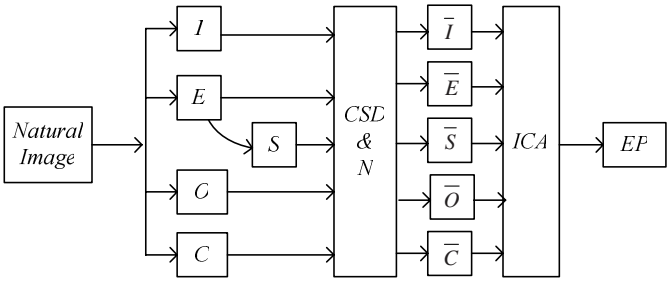

Fig. 1. The architecture of essence map model. $\bar{I}$: normalized intensity feature map, $\bar{E}$: normalized edge feature map, $\bar{S}$: normalized symmetry feature map, $\bar{O}$: normalized orientation feature map, $\bar{C}$: normalized colour feature map, *EP*: essence point.

Using a similar notation to that used in [11], and after the convolution between the channel of feature maps and filters obtained by ICA, the essence map is computed by the summation of all feature maps for every location [12]. In the course of preprocessing, a Gaussian pyramid with different scales from 0 to *n* level is used [7]. Each level is obtained by subsampling of 2^n, thus constructing five feature maps. Subsequently, the centre-surround mechanism is implemented in the model as the difference between the fine and coarse scales of Gaussian pyramid images [7]. Consequently, five feature maps are obtained by the following equations.

$$I(c,s) = |I(c) \bullet I(s)|, E(c,s) = |E(c) \bullet E(s)|, S(c,s) = |S(c) \bullet S(s)|. \quad (1)$$

$$O(c,s) = |O(c) \bullet O(s)|, C(c,s) = |C(c) \bullet C(s)|. \quad (2)$$

Here, $\bullet$ represents interpolation to the finer scale and point-by-point subtraction, *N* stands for the normalization operation, *c* and *s* are indexes of the finer scale and the coarse scale, respectively. Feature maps are combined into five characteristic maps.

$$\bar{I} = \bigoplus_{c,s} N(I(c,s)), \bar{E} = \bigoplus_{c,s} N(E(c,s)), \bar{S} = \bigoplus_{c,s} N(S(c,s)), \bar{O} = \bigoplus_{c,s} N(O(c,s)), \bar{C} = \bigoplus_{c,s} N(C(c,s)). \quad (3)$$

Here, $\bar{I}$, $\bar{E}$, $\bar{S}$, $\bar{O}$, and $\bar{C}$ are obtained through across-scale addition "$\oplus$". To obtain ICA filters, the five feature maps are used for input patches of the ICA. The basis functions are determined using the extended infomax algorithm [13]. Each row of the basis functions represents an independent filter and that is ordered according to the length of the filter vector. The resulting ICA filters are then applied to the five feature maps to obtain the essence map according to [7]:

$$E_{qi} = FM_q * ICs_{qi} \ for \quad i = 1,\cdots,M, \ q = 1,\cdots,5, \ EM(x,y) = \sum E_{qi}(x,y) \ for \ all \ i. \quad (4)$$

Here, M denotes the number of filters; FM_q denotes feature maps, ICs_{qi} denotes each independent component accounting for the number of filters and feature maps, $EM(x, y)$ denotes the essence map. The convolution result E_{qi} represents the influences of the five feature maps on each independent component and the most essential point is computed by maximum operator, then an appropriate essential area centred by the most essential location is masked off and the next essential location in the input visual information is calculated using the essence map model.

3 Knowledge Synthesizing

In this section the knowledge synthesizing approach is described. The proposed knowledge synthesizing approach automatically creates a relevance map from the essential areas detected by our proposed essence map model. It also derives a set of well-structured representations from low-level description to drive the high-level classification. The backbone of this technique is a distribution mapping strategy involving knowledge synthesizing based on growing when required network (GWR).

The precise steps of the GWR algorithm will now be detailed as follows [14].

Let A be the set of map nodes, and $C \subset A \times A$ be the set of connections between nodes in the map field. Let the input distribution be $p(\xi)$ for inputs ξ. Define w_n as the weight vector of node n.

Initialisation. Create two nodes for the set A, $A = \{n_1, n_2\}$, with n_1, n_2 initialised randomly from $p(\xi)$.

Define C, the connection set, to be the empty set $C = \varnothing$. Then, each iteration of the algorithm looks like this:

1. Generate a data sample ξ for input to the network.
2. For each node i in the network, calculate the distance from the input $\| \xi - w_i \|$.
3. Select the best matching node, and the second best, that is the nodes $s, t \in A$ such that $s = \arg\min_{n \in A} \| \xi - w_n \|$ and $t = \arg\min_{n \in A / \{s\}} \| \xi - w_n \|$, where w_n is the weight vector of node n.
4. If there is not a connection between s and t, create it $C = C \cup \{(s,t)\}$, otherwise, set the age of the connection to 0.
5. Calculate the activity of the best matching unit $a = \exp(-\| \xi - w_s \|)$.
6. If the activity $a <$ activity threshold a_T and firing counter $<$ firing threshold h_T then a new node should be added between the two best matching nodes (s and t)
   - Add the new node r, $A = A \cup \{r\}$.
   - Create the new weight vector, setting the weights to be the average of the weights for the best matching node and the input vector $w_r = (w_s + \xi)/2$.
   - Insert edges between r and s and between r and t, $C = C \cup \{(r,s),(r,t)\}$.
   - Remove the link between s and t, $C = C / \{(s,t)\}$.

7. If a new node is not added, adapt the positions of the winning node and its neighbours, i, that is the nodes to which it is connected, $\Delta w_s = \varepsilon_b \times h_s \times (\xi - w_s)$, $\Delta w_i = \varepsilon_n \times h_i \times (\xi - w_i)$,where $0 < \varepsilon_n < \varepsilon_b < 1$ and h_s is the value of the firing counter for node s.
8. Age edges with an end at s, $age_{(s,i)} = age_{(s,i)} + 1$.
9. Reduce the counter of how frequently the winning node s has fired according to $h_s(t) = h_0 - \frac{S(t)}{\alpha_b}(1 - e^{(-\alpha_b t / \tau_b)})$ and the counters of its neighbours, (i), $h_i(t) = h_0 - \frac{S(t)}{\alpha_n}(1 - e^{(-\alpha_n t / \tau_n)})$, where $h_i(t)$ is the size of the firing variable for node i, h_0 the initial strength, and $S(t)$ is the stimulus strength, usually 1. α_n, α_b and τ_n, τ_b are constants controlling the behaviour of the curve. The firing counter of the winner reduces faster than those of its neighbours.
10. Check if there are any nodes or edges to delete, i.e. if there are any nodes that no longer have any neighbours, or edges that are older than the greatest allowed age, in which case, delete them.
11. If further inputs are available, return to step (1) unless some stopping criterion has been reached.

The input of the algorithm is a set of extracted low-level features generated by essence map model. Various topology maps of the network subtly reflect the characteristics of distinct visual information 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.

4 Classification

Using the output generated by the knowledge synthesizing approach, 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. A proposed high-level perception and classification model uses a generative approach based on an incremental Bayesian parameter estimation method. The input features of this generative object representation are the low-level information generated by knowledge synthesizing module. A new class can be added incrementally by learning its class-conditional density independently of all the previous classes. In this paper n training data samples from a class ω are considered. Each class is represented by f $(f < n)$ codebook vectors. Learning is conducted incrementally by updating these codebook vectors whenever a new data vector u is entered. The used generative approach learns the class prior probabilities $p(\omega)$ and the class-conditional densities $p(u \mid \omega)$ separately. The required posterior probabilities area then obtained using the Bayes' theorem:

$$p(\omega \mid u) = \frac{p(u \mid \omega)\, p(\omega)}{p(u)} = \frac{p(u \mid \omega)\, p(\omega)}{\sum_{j} p(u \mid j)\, p(j)}. \tag{5}$$

In order to estimate the class-conditional density of the feature vector u given the class ω, a vector quantizer is used to extract codebook vectors from training samples. Following Vailaya *et al.* in [15], the class-conditional densities are approximated using a mixture of Gaussians (with identity covariance matrices), each centred at a codebook vector. Then, the class-conditional densities can be represented as,

$$p_U(u \mid \omega) \propto \sum_{j=1}^{f} m_j * \exp\left(-\| u - v_j \|^2 / 2\right). \tag{6}$$

where $v_j (1 \le j \le f)$ denotes the codebook vectors, m_j is the proportion of training samples assigned to v_j. When human beings focus its attention in a given area, the prefrontal cortex gives a competition bias related to the target object in the inferior temporal area [16]. Subsequently, the inferior temporal area generates specific information and transmits it to the high-level attention generator which conducts a biased competition [16]. Therefore, the high-level perception and classification model can assign a specific class to a target area, which gives the maximum likelihood. If the prior density is assumed essentially uniform, the posterior probability can be estimated as follows [15],

$$\arg\max_{\omega \in \Omega} \left\{ p(\omega \mid u) \right\} = \arg\max_{\omega \in \Omega} \left\{ p_U(u \mid \omega)\, p(\omega) \right\}. \tag{7}$$

where Ω is the set of pattern classes. In addition, the high-level perception and classification model can generate a specific attention based on the class detection ability. Moreover, it may provide informative control signals to the internal effectors [16]. This in turn can be seen as an incremental framework for knowledge synthesizing with human interaction.

5 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 preservation of the network. To this end, a set of quantitative experiments with progressively increasing level of topology representation complexity was conducted. The Corel database containing 700 images was used, which was labelled manually with eight predefined concepts. The concepts are "building", "car", "autumn", "rural scenery", "cloud", "elephant", "lion", and "tiger". In order to assess the accuracy of the classification, a performance evaluation based on the amount of missed detections (*MD*) and false alarms (*FA*) for each class from the large dataset of the Corel database was conducted. In this evaluation recall (*R*) and precision (*P*) values were estimated and used: $recall = \frac{D}{D + MD}$, $precision = \frac{D}{D + FA}$, 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. The obtained results are given in Table 1.

Table 1. Recall/Precision Results of Classification and Retrieval

Class	D	MD	FA	R (%)	P (%)
Building	84	16	12	84	88
Autumn	42	14	10	75	81
Car	89	11	8	89	92
Cloud	90	10	10	90	90
Tiger	87	13	10	87	90
Rural scenery	36	8	9	82	80
Elephant	93	7	14	93	87
Lion	88	12	10	88	90

The proposed technique was compared with an approach based on multi-objective optimization [17] and another using Bayesian networks for concept propagation [18]. Table 2 shows a summary of results on some subsets of the categories coming out from this comparative evaluation. 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 belief approach delivers the highest accuracy, the proposed technique performs substantially better in other cases. 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 classifiers: one based on ant colony optimization and constraints of points with K-Means approach (ACO/COP-K-Means) [19], and the other using particle swarm optimization and self organizing feature maps (PSO/SOFM) [20]. A summary of results on some subsets of the categories is given in Table 3.

According to these results, it can be concluded that the proposed technique also outperforms other classical approaches and works well in the case of multi-mode classification.

Table 2. Precision Results of the Proposed Technique Compared with Two Other Approaches

(%)	Proposed Technique	Bayesian Belief	Multi-Objective Optimization
Building	88	72	70
Cloud	90	84	79
Lion	90	92	88
Tiger	90	60	60

Table 3. Results of the Proposed Technique Compared with Two Other Binary Classifiers

(%)	ACO/ COP-K-Means		PSO/ SOFM		Proposed technique	
	P	R	P	R	P	R
Lion	55	62	48	69	90	88
Elephant	71	71	74	65	87	93
Tiger	63	58	68	64	90	87
Cloud	62	57	69	63	90	90
Car	65	56	70	64	92	89
Building	65	62	51	74	88	84

6 Conclusion

A knowledge synthesizing approach for classification of visual information is presented. By utilizing biologically inspired theory and knowledge synthesizing, this system simulates the human-like classification and inference. Since the knowledge synthesizing 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. Additional expansion capabilities include learning from semantics and annotation-based approach and the use of multimodal information.

References

1. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content Based Image Retrieval at the End of the Early Years. IEEE Trans. Patt. Anal. Mach. Intell. 22(12), 1349–1380 (2000)
2. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7. In: Multimeida Content Description Interface, John Wiley & Sons, West Sussex (2003)
3. Dorado, A., Djordjevic, D., Pedrycz, W., Izquierdo, E.: Efficient Image Selection for Concept Learning. IEE Proc. on Vision, Image and Signal Processing 153(3), 263–273 (2006)
4. Dorai, C., Venkatesh, S.: Bridging the Semantic Gap with Computational Media Aesthetics. IEEE Multimedia 10(2), 15–17 (2003)
5. Dorado, A., Calic, J., Izquierdo, E.: A Rule-based Video Annotation System. IEEE Trans. on Circuits and Systems for Video Technology 14(5), 622–633 (2004)
6. Goldstein, E.B.: Sensation and Perception, 4th edn. An international Thomson Publishing Company, USA (1996)
7. Park, S.J., An, K.H., Lee, M.: Saliency Map Model with Adaptive Masking Based on Independent Component Analysis. Neurocomputing 49, 417–422 (2002)
8. Bell, A.J., Sejnowski, T.J.: The Independent Components of Natural Scenes Are Edge Filters. Vision Research 37, 3327–3338 (1997)
9. Vetter, T., Poggio, T., Bülthoff, H.: The Importance of Symmetry and Virtual Views in Three-Dimensional Object Recognition. Current Biology 4, 18–23 (1994)
10. Barlow, H.B., Tolhust, D.J.: Why Do You Have Edge Detectors? Optical Society of America Technical Digest 23(172) (1992)

11. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. Patt. Anal. Mach. Intell. 20(11), 1254–1259 (1998)
12. Ratnaparkhi, A.: Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. Dissertation. Comp.and Inf. Science, Univ. of Pennsylvania, USA (1998)
13. Jasinschi, R.S., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J., Li, D., Louie, J.: A Probabilistic Layered Framework for Integrating Multimedia Content and Context Information. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 2, pp. 2057–2060. IEEE Computer Society Press, Los Alamitos (2002)
14. Marsland, S., Shapiro, J., Nehmzow, U.: A Self-organising Network That Grows When Required. Neural Networks 15, 1041–1058 (2002)
15. Vailaya, A., Figueiredo, M.A.T., Jain, A.K., Zhang, H.J.: Image Classification for Content-based Indexing. IEEE Trans. on Image Processing 10(1), 117–130 (2001)
16. Lanyon, L.J., Denham, S.L.: A Model of Active Visual Search with Object-based Attention Guiding Scan Paths. Neural Networks Special Issue: Vision & Brain 17(5-6), 873–897 (2004)
17. Zhang, Q., Izquierdo, E.: A Multi-feature Optimization Approach to Object-based Image Classification. In: Sundaram, H., Naphade, M., Smith, J.R., Rui, Y. (eds.) CIVR 2006. LNCS, vol. 4071, pp. 310–319. Springer, Heidelberg (2006)
18. Li, F.F., Fergus, R., Perona, P.: A Bayesian Approach to Unsupervised One-shot Learning of Object Categories. In: Proc. IEEE Int. Conf. on Computer Vision, vol. 2, pp. 1134–1141. IEEE Computer Society Press, Los Alamitos (2003)
19. Saatchi, S., Hung, Ch.: Hybridization of the Ant Colony Optimization with the K-Means Algorithm for Clustering. In: Kalviainen, H., Parkkinen, J., Kaarna, A. (eds.) SCIA 2005. LNCS, vol. 3540, pp. 511–520. Springer, Heidelberg (2005)
20. Chandramouli, K., Izquierdo, E.: Image Classification Using Chaotic Particle Swarm Optimization. In: Proc. IEEE Int. Conf. on Image Processing, Atlanta, USA, pp. 3001–3004. IEEE Computer Society Press, Los Alamitos (2006)