A Strongly Coupled Architecture for Contextual Object and Scene Identification

Tina Ehtiati, James J. Clark
Centre for Intelligent Machines, McGill University
Montréal, Québec, Canada
E-mail: {tina, clark}@cim.mcgill.ca

Abstract

The context-centered approach to object detection and recognition is based on the intuition that the contextual information of real-world scenes provides relevant information for these tasks. This intuition is supported by psychophysical experiments in human scene perception and visual search, which provide evidence that the human visual system uses the relationship between the environment and the objects to facilitate object recognition. Here we use a probabilistic model to investigate the possible interactions between object class hypotheses and scene class hypotheses in a visual system. The architecture of the model is based on separate modules interacting with each other via feedforward and feedback connections. A competitive-priors structure is used to implement the feedback connections.

1. Introduction: context based object and scene identification

Psychophysical studies in scene perception have provided evidence which shows object identification in human visual system does not operate exclusively on a bottom-up basis, but rather the conceptual meaning of the scene influences object identification [1][2]. This scheme can be useful in artificial visual systems, but the computational implementation of such a scheme requires dealing with several central questions. Is it possible to encode the context of a scene using the global scene information before any local object identification is performed? What kind of information could be used to identify a scene as a certain scene type? Is the scene identity inferred from the identity of the individual objects present in the scene? How could the scene and the object identification processes interact?

Schyns and Oliva [3] have demonstrated that contexts of scenes can be identified from the low spatial frequency images that preserve the spatial relationships between large-scale structures in the scene. Oliva and Torralba [4] have shown that it is possible to construct definitions of scene context which are not dependent on identification of individual objects in the scene. In their research the scenes are represented globally, based on their second order statistical regularities. Torralba and Sinha [5] have proposed a representation of the context based on the statistics of the low level features of the scenes, encoding spatially localized structural information using Gabor filters. They have furthermore shown that the contextual information can be useful for the object detection task. Their approach is based on conditioning the statistics of the contextual features of the scene according to the presence or absence of object categories [6].

In this paper a computational model is presented of how hypotheses could be formed simultaneously about both the objects present in the scene and the conceptual meaning of the scene. Inferences are made about objects present in the natural scene images based on their low-level context features. As a higher level of conceptual abstraction of the scene, an estimate of the likelihood of the scene class is then computed based on the object-level likelihood estimates. Joint probabilities of the presence of different object categories influence the estimate of the likelihood of each scene category. A strongly-coupled structure is proposed to implement the interactions between the object and scene hypotheses.

2. A strongly-coupled Bayesian model for object and scene identification

In this model we are interested in deriving two types of inferences from the images. First, we want to make a hypothesis about possible object classes present in the scene. Second, we want to make a hypothesis about the abstract classes of scenes. The computations for assigning probabilities for each object class and each scene class, for any given image, is performed in two separate

IEEE COMPUTER SOCIETY

modules, which are referred to as the object module and the scene module.

Assuming n object classes $O_1, O_2, \ldots, O_n$ and m scene classes $S_1, S_2, \ldots, S_m$, and the low-level context features V_C, as defined by Torralba and Sinha [5], as the input to the Bayesian system, the problem of inferring the presence of a certain object class O_i from the input data is formulated as the following

$$P(O_i|V_C) = \frac{P(V_C|O_i)P(O_i)}{P(V_C)} \qquad (1)$$

where $P(V_C|O_i)$ is a model of the context feature values given the condition of the presence of object class O_i in the scene. This probability density is estimated using the expectation-maximization (EM) algorithm and the V_C values computed from the set of images in the image database which contain object class O_i [6]. $P(O_i)$ is the *a priori* expectation which is determined by measurement on the image database. $P(V_C)$ is a normalization factor, which can be computed from $P(V_C|O_i)$ and $P(O_i)$. The probabilities $P_i = P(O_i|V_C)$ thus computed for each object class O_i are then combined into a single vector $\hat{P} = \{P_1, P_2, \cdots, P_n\}$ and is used for computation of the scene class probabilities. The Bayesian formulation of the estimation of the probability of the image with the corresponding vector $\hat{P}$ belonging to scene class S_j given $\hat{P}$ is as follows

$$P(S_j|\hat{P}) = \frac{P(\hat{P}|S_j)P(S_j)}{P(\hat{P})} \qquad (2)$$

where the probability density $P(\hat{P}|S_j)$ is estimated using the EM algorithm based on the values of $\hat{P}$ computed for the set of images in the data set which belong to the scene class S_j. The computations in equations (1) and (2) represent the functioning of the object and the scene hypotheses generating modules. The two modules can be connected simply through a feedforward connection as implied by these two equations or the equations can be revised in order to embed a feedback interaction between the two modules.

In order to make feedback connections possible between the two modules, the *a priori* terms in equations (1) and (2) are expanded as following

$$P(O_i) = \sum_j P(O_i, S_j) = \sum_j P(O_i|S_j)P(S_j) \qquad (3)$$

and

$$P(S_j) = \sum_i P(S_j, O_i) = \sum_i P(S_j|O_i)P(O_i) \qquad (4)$$

In the Bayesian formulation the constraining assumptions derived from specific experimental domains are incorporated as *a priori* terms. The *a priori* constraints make the ill-posed problem of making interpretations from the information derived from the input image possible, by providing supplementary previously acquired knowledge of the world. Expanding the *a priori* terms as in equations (3) and (4) provides a way to feed back the output of the scene module to the object module in order to update the *a priori* values. The term $P(S_j)$ in equation (3) is revised based on $P(S_j|\hat{P})$ and the term $P(O_j)$ in equation (4) is revised based on the new values of $P(O_j|V_C)$. This feedback is designed to use the new information inferred by the system from the input image, to determine a more accurate *a priori* model for the system. We can perceive that the *a priori* terms of the scene module could also be altered using the new values of the object module. This alteration of the *a priori* modules distinguishes the proposed system as a strongly-coupled architecture [7]. This architecture is visualized in detail in figure (1).

3. Experimental results

A database of 400 images from natural scenes has been gathered for the purpose of experimentation with the proposed architecture. The database consists of three sets of natural scene images from parks, streets, and indoor. Three object classes are chosen to be cars, trees, and people. Examples of images are shown in figure (4).

The likelihood terms in equations (1) and (2) are estimated from the images in the database using the EM algorithm. The *a priori* terms are also initialized using the statistics of the image database. Given a new image, not belonging to the training data set, the object module estimates the probability of the image containing cars, trees, and peoples.

The computed values form the three-dimensional vector $\hat{P} = \{P_1, P_2, P_3\}$, which is the input to the scene module. The scene module estimates the probability of the scene belonging to the three scene classes, parks, streets, and indoors, based on the previously estimated joint distribution $P(\hat{p}\,|S_j)$ and the *a priori* values.

Tables (1) and (2) illustrate the results of experimentation with a training set of 300 images. The classification results are shown for a set of 20 images in each scene class, not contained in the training set.

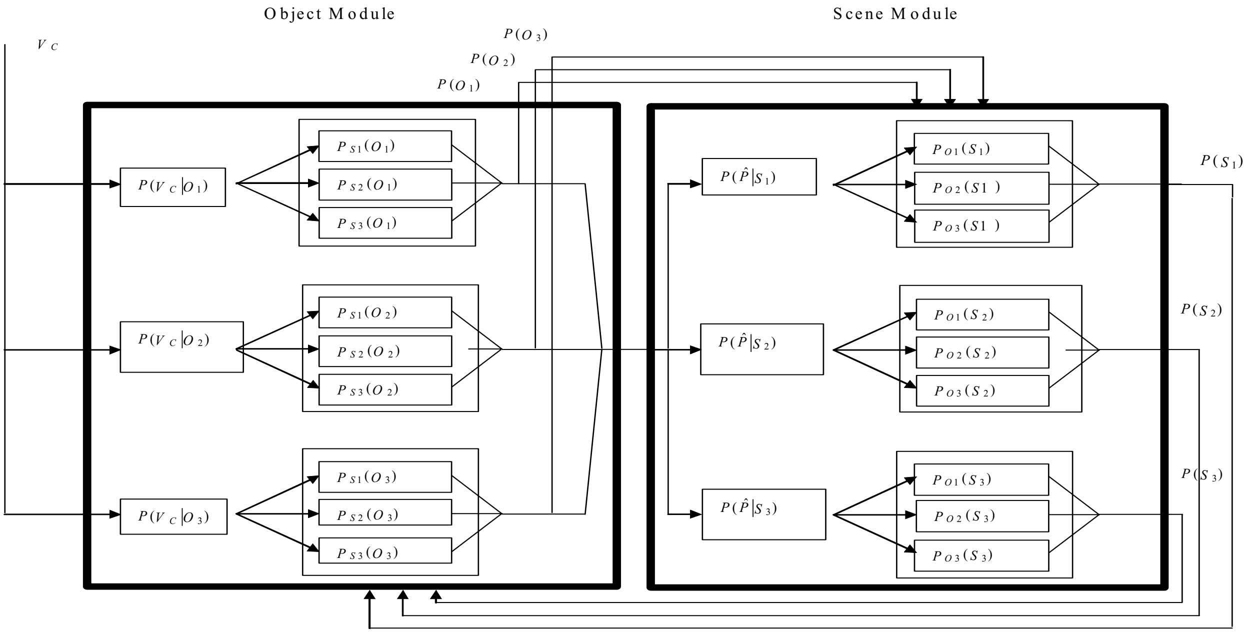

Figure 1. The strongly coupled model consists of two modules, object module, and scene module, which compute the hypothesis of the model about the object classes present in the scene, and the scene class, given a new test image. The *a priori* models in each module are revised through the feedback from the other module. In this figure it is assumed that $m = n = 3$ for the purpose of simplicity.

Table 1. The initial classification results for 20 test images in each scene class, before the starting of the feedback process.

Initial results	Correctly classified	Mis-classified	Unclassified
Park scenes	13	3	4
Street scenes	9	5	6
Indoor scenes	16	0	4

Table 2. The classification results of the same image set as in table 1, after 35 iterations of the model.

After 35 iterations	Correctly classified	Mis-classified	Unclassified
Park scenes	15	4	1
Street scenes	12	4	4
Indoor scenes	18	0	2

Table (1) shows the initial results of the classification of the scenes, and table (2) shows the results after 35 iterations of the model. The comparison of the two tables indicates that the number of unclassified images has decreased for this set of test images after 35 iterations, while the number of correctly classified images has increased. It is possible to have an initially unclassified image misclassified after a number of iterations.

The study of the dynamics of the system shows that for the correctly classified images the system is able to reach a stable decision, while for the unclassified images the system is not able to stabilize. Figures (2) and (3) illustrate the behavior of the system for reaching a stable hypothesis about the identity of the test images. In each figure the three curves illustrate the probability of the image being a park scene, a street scene, or an indoor scene as a function of the number of iterations. In figure (2) the system behavior is averaged for the 15 park scenes which are correctly classified. Figure (3) illustrates the scene probabilities averaged over the 18 indoor scenes which are correctly classified. It is interesting to note that while the system has produced a correct hypothesis about the identity of the scene, the system does not stabilize for scene classes with small probabilities.

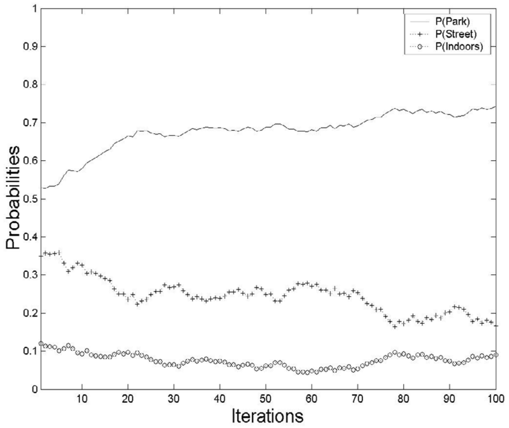

Figure 2. The probability of the image being a park, indoor, or street scene, for the first 100 iterations, averaged over the 15 park images, correctly classified in the experiment.

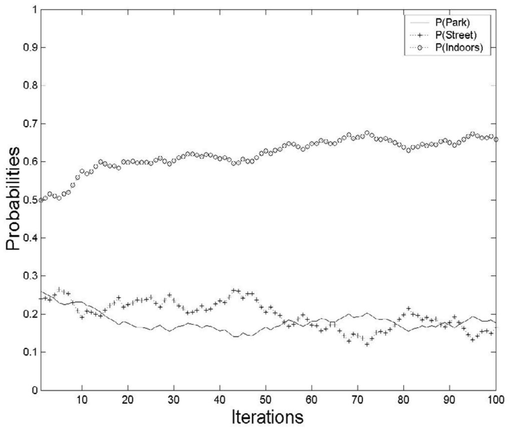

Figure 3. The probability of the image being a park, indoor, or street scene, for the first 100 iterations, averaged over the 18 indoor images correctly classified in the experiment.

4. Conclusions

A strongly-coupled Bayesian model is presented for modeling the influences between object and scene hypothesis inferred from images. The system uses the dependencies of the prior probability models of objects and scenes in order to incorporate feedback connections between the two object and scene modules. We have demonstrated the capability of the model for achieving better results for identifying scenes, compared to the simple feedforward probabilistic solution.

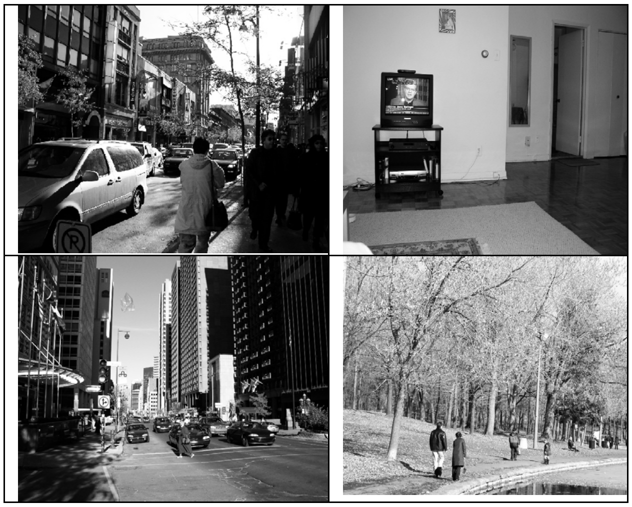

Figure 4. Sample images of the data base from the parks, streets, and indoors scene classes.

References

[1] S. M. Kosslyn, *Image and Brain*, MIT Press, Cambridge, MA, 1994.

[2] I. Biederman, R. J. Mezzanotte, J. C. Rabinowitz, “Scene perception: detecting and judging objects undergoing relational violations”, *Cognitive Psychol*ogy, 1982, 14:143-77.

[3] P. G. Schyns, A. Oliva, “From blobs to boundary edges: evidence for time and spatial scale dependent scene recognition”, *Psychological Science*, 1994,5:195-200

[4] A. Oliva, A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope”, *International Journal of Computer Vision,* 2001, 42: 145-75.

[5] A. Torralba, P. Sinha, “Statistical Context Priming for Object Detection”, *Proceedings of the International Conference on Computer Vision*, ICCV, BC, Canada, 2001, pp. 763-770.

[6] A. Torralba, “Contextual Priming for Object Detection”, *International Journal of Computer Vision*, 2003, 53(2):153-167.

[7] J. J. Clark, A. L. Yuille, *Data Fusion for Sensory Information Processing Systems*, Kluwer Academic Publishers, USA, I990.