A Survey on Context-Based Computer Vision Systems

Renqiang Min

Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, ON M5S3G4, Canada

Abstract. It is widely known that contextual information plays a very important role in object recognition and detection. In this paper, several context-based computer vision systems, which include shape contexts based systems along with some correponding matching and indexing methods, context sets based systems and many probabilistic systems modeling context, will be reviewed and compared first. Then some challenging problems still existing in the context-based computer vision research and potential extensions to the current context-based computer vision systems will be discussed.

Key words

Context-based Vision, Object Recognition, Object Detection, Image Segmentation

1 Introduction

In the early stage of computer vision research, researchers often assume that objects to be recognized can be represented by a small number of shape models and must have distinguishable features. However, these assumptions don't always hold in all kinds of object recognition and detection tasks, especially for recognizing and detecting objects from natural scenes. Based on such kind of observations, Strat and Fischler [1] proposed a complete and successful system for recognizing pictures of outdoor scenes based on context. Since then, context-based systems has appealed more and more attention in the computer vision research community. And a lot of contex-based vision systems have been presented and have demonstrated great success for handling complicated recognition tasks compared to the early systems.

Before the review of some representative context-based systems, some background knowledge about context-based vision will be presented first. In Barll and Ullman's paper [2], several psychological experiments show that: proper spatial relations between objects that often co-occur in the same scene help the identification of individual obejcts; the presence of objects having unique interpretations helps the identification of related ambiguous objects. Oliva et al [3] used the saliency of local features and the global scene context information

to predict the location of people in natural scene images. Psychological experiments on human eye movements showed the predicted regions by the model agreed with the regions that drew the most attention of human observers when they were told to search people from the given images. The above experimental findings clearly show that context information helps object identification. For e. g., when we see some vehicle-like stuff on the water, it is more likely that we believe it is a boat rather than a car based on the object-region context. In fact, humans often use contextual information to faciliate recognizing and detecting objects not only in the setting of natural scenes but also in the setting of general recognition tasks. For e. g., after we see the handle of a cup, we can easily infer the orientation of the cup; after we see two wheels and part of the contour of a car-like object, we tend to believe it is a car and we have no difficulty in infering its orientation. Theorectical study underlying those examples can be found in [4]. Thus, the clue information provided by the identification of some parts of an object also belong to context information.

Here we will discuss the types of context exploited in the computer vision research. According to the levels of modeling scenes, context can be classified as scene context and spatial context. Scene context represents the gist of scene and often refers to the types of the whole scene. For e. g., we say this represents an indoor office scene and that represents an outdoor rural scene. This kind of context stands at the highest level and can be used to give priors about the configurations of the scene. Spatial context refers to the relations among different regions in a scene. According to the relations among regions in different scales, spatial context can be further classified as local-range context, long-range context, and global context. Spatial context can also be classifed as intra-object context, which models relations within an object, and inter-object context, which models relations among different objects. Of course, the classification given here is not very strict. The scale for distinguishing local-range context from long-range context cannot be defined quantitatively. Besides, there is no scene context for an image only containing one object used by shape context [5], and now the global range only corresponds to one object. Because shape context models the context within an object, it is classified as intra-oject context. And because the shape context at each point holds the information about all the other points, shape context is also classified as global spatial context.

The paper will be organized as follows: In section 2, we will describe the privious work about shape context along with some fast mactching and indexing methods. In section 2.1, systems based on shape contexts and generalized shape contexts will be reviewed. In section 2.2, a non-rigid point matching method and its modified version used for matching shape contexts will be discussed. In section 2.3, fast pruning approaches to speeding shape retrieval based on generalized shape contexts will be described. In section 2.4, many-to-many feature matching and hierarchical indexing using graph spectra will be briefly introduced. In section 3, we will review an early sucessful rule-based system which models context using context sets. In section 4, several representative probabilistic models based on context, which include parts-based systems modeling

intra-object context, systems modeling all levels of spatial context, and systems modeling scene contexts and global spatial context, will be reviewed. In section 5, I will give comparisons and discussions about the above context-based models. I will discuss the existing challenges in the context-based computer vision research and propose potential extensions to the current systems. In section 6, we will conclude by summarizing the paper.

2 Shape contexts based systems and fast matching and indexing methods

2.1 Shape context and generalized shape context

Motivated by producing a rich and robust descriptor for shapes to reduce the ambiguity in matching, Belongie et al [5] proposed shape context to do shape matching and object recognition. Shape Context can be generated as follows: each object can be viewed as a point set, and we sample n points from its internal and external contours. Then the shape of the object can be represented by a point set $P = \{p_1, \ldots, p_n\}$. At each point p_i, we calculate the set of vectors originating from this point to all the other points on the shape. The distribution of those vectors can be compactly represented by a histogram over a log-polar space.

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\}. \tag{1}$$

h_i is defined as the shape context at point p_i. Since h_i measures the distribution and the orientation of all the other points relative to the point p_i, shape context can be viewed as a global intra-object context. And due to the property of log-polar space, the shape context at each point gives more precise descriptions about nearby points and gives less precise descriptions about points far away.

Generalized shape context (GSC) [**?**] is an extension of shape context. It gives richer discriptions than shape context does. For each point q_j in a shape, a unit tangent vector t_j reprenting the direction of the edge at that point is associated with the point. Instead of just counting the number of corresponding points falling into each bin in the log-polar space, GSC sums the tangent vectors for all points falling in the bin. Now the descriptor for a point p_i becomes:

$$\hat{h}_i(k) = \sum_{q_j \in Q} t_j, where \;\; Q = \{q_j \neq p_i, (q_j - p_i) \in bin(k)\}. \tag{2}$$

The generalized shape context $\hat{h}_i$ at point p_i carries more information than the original shape context h_i. It is also an intra-object global context.

2.2 Non-rigid point matching based on shape context

Chui and Rangarajan [6] proposed a non-rigid point matching algorithm based on thin-plate-spline (TPS) and softassign by minimizing an objective function which is a combination of several energy terms. The objective function is composed of

a bending energy term associated with TPS, a penalty term penalizing null matches, another penalty term penalizing unphysical reflection mappings, and an entropy barrier term to ensure the positivity of softassign coefficients. A linear annealing schedule is used to control the degree of non-rigid warping in TPS and the degree of unphysical reflection mappings. An alternating update strategy is used to minimize the objective function to learn the parameters.

Based on the similarity matrix, in which similarity between any two points on two different shapes is defined as the χ^2 distance between shape contexts, a bipartite graph matching problem is solved to establish the initial correspondences between the two point sets (shapes) by minimizing matching cost. Then Belongie et al [5] borrowed the above point matching idea to calculate the best transformation between the two point sets, but used hard matching instead of softassign given the correspondences. The above two procedures of establishing correspondences and estimating a transformation are iterated several iterations to produce refined matchings. The distance between two shapes is computed as the sum of the matching cost, the bending energy, and image appearance distance. The tasks of recognition and retrieval is achieved through k Nearest Neighbor (k-NN).

2.3 Fast pruning approaches to speeding shape retrieval

When an unknow shape is given, shape retrieval returns a set of similar shapes to the given shape from a stored shape database. Although the retrieval approach based on shape context discussed in section 2.2 has demonstrated success on several image databases, it is very computationally expensitive. When the shape database is large, the computations needed are prohibitively daunting. Therefore, fast matching, fast indexing, or fast pruning becomes dominant for handling these situations. Mori et al [?] used fast pruning to handle the problem with a large shape database. After pruning, only a small set of potential candidate shapes are chosen from a very large shape database, then expensive and accurate matching procedures can be applied to this small set to produce the final set of similar shapes to the query shape.

Two approaches to pruning are presented. One is based on representative shape contexts (RSC). The motivation to RSC is that we can avoid matching a pair of shapes if they are obviously very different. The detailed matching process using RSC is as follows: we precompute a large number s (about 100) of shape contexts for each known shape S_i, and for each query shape Q, we only compute a small number r (5 to 10) of shape contexts. These shape contexts are randomly sampled over the entire shapes. We find the best matches in each known shape S_i for each of the r RSCs, where the distance is calclulated based on the Euclidean distance between GSCs. The cost of each match is normalized by dividing the dicriminative significance of the corresponding GSC in Q. The average of these match costs is defined as the distance between S_i and Q. By sorting these distances, we can determine a small number of candidates for the query shape Q.

The other approach to pruning is called shapemes using vector quantization on the shape contexts. Clustering all the shape context (or GSC) vectors into several clusters, and the clusters are called shapemes. Then each shape context can be represented by the index of the shapemes it is in. Each shape will be represented by a histogram of shapeme frequencies. Thereby, it's very fast to get a small set of candidate shapes for a query shape Q by sorting the distances between Q and known shapes in the space of the histograms.

2.4 Many-to-many feature matching and fast hierarchical indexing

Except shape contexts, intra-object contexts can also have natural representations as graphs. Graph-based representations can easily show the relations between different parts of objects. As discussed earlier in section 1, the identification of some parts can provide clue information for the identification of some other parts. Therefore, we believe that graph can be used to model the intra-object contextual information effectively. In this section, we will review a many-to-many feature matching method [7] and a fast indexing method [8] for objects represented by graphs.

Many-to-many feature matching draws researchers' attention because sometimes segmentation errors, articulation, scale difference, and within-class deformation makes one-to-one feature matching impossible but many-to-many feature matching natural. In this matching framework, each object is represented by a vertex-labeled graph in which nodes represent image features and edges represent relations. Matching two graphs means establishing correspondeces between their nodes, and the quality of a match is measured by the overall distance which depends on both node and edge similarity.

The matching method [7] works as follows: embed two graphs to be matched into a d-dimensional vector space by respectively following the caterpillar decompositions of the metric trees of the graphs. d is specified by the user so that the embedding preserves pairwise distances between the nodes with some appropriate degree of distortion. Then compute Earth Mover's Distances between the embeddings by applying the FT iteration to get the optimal tranformation T. Then the many-to-many vertex matching is trivially obtained from the resulting optimal flow.

The above algorithm gives an effective approach to calculating many-to-many feature mappings between objects represented by vertex-labeled graphs. However, it requires the edge weights in the original graph be accurately measured, and it also requires the attributes of each node in a metric tree reflect meaningful many-to-many correspondences between the nodes in the original graph. Besides, it is not invariant to large occlusion.

Shokoufandeh et al [8] proposed a fast indexing method using graph spectra to index the hierarchical structures of directed acylic graphs (DAG). In the method, a topological signature vector (TSV) reflecting diverging and subtree structures is associated with each nonterminal node. A small number of candidate models similar to the given query are chosen by k-NN searching and calculating votes. Robust evidence accumulation based on Mutiple One-to-One

Vote Correspondence (MOOVC) algorithm is used to sum the votes, so that one-to-one vote matching is assured between the query and each candidate model. Detailed descriptions about the method can be found in [8].

The above method is very efficient in indexing hierarchical structures, and has great potential applications in computational biology due to its robustness to accomodate structure noise (node split/merge). It's also robust in accomodating large-scale occlusion, and it scales well with increasing database size making handling large dataset possible.

The limitations of the two methods discussed above are also obvious: they both require graph-based object representations. However, deriving accurate graph representations for objects is still a challenging problem.

3 Context Sets Based Systems

The context sets based systems introduced by Strat and Fischler [1] is one of the earliest systems that abandoned the geometric shape models and turned to using contextual information to recognize objects in natural scenes. Contexts used in the system include spatial contexts and some other contextual information for generating the natural scene images. Each context in the system is represented by a context set which is a set of context elements defining the conditions associated with the context. Each context set is embedded into a rule denoted by the name associated with the class of the context set and the context set followed by an action. If all the conditions in the context set are satisfied, the action will be triggered. For e. g., a rule can be: SKY: {image-is-color, camera-is-horizontal, sky-is-clear, time-is-daytime} $\Rightarrow$ BLUE-SKY.

All the rules containing the contextual knowledge that drives the recognition is encoded in a core knowledge structure (CKS). Recognition involves four processes: candidate generation (hypothesis generation), candidate comparison (hypothesis evaluation), clique formation (grouping mutually consistent hypotheses), and clique selection (selection of a best "descritption"). Three different types of context sets are respectively used for the first three stages of recognition process. It is argued in the paper that, when the knowledge base is constructed, rules designed for each recognition stage can be imperfect, and the reliable recognition results can be achieved through the use of large numbers of redundant operators in each recognition stage.

When generating candidates, a large number of simple procedures, in which each individual one handles a specific context, are used to collectively predict a hypothesis in a wide range of contexts. In the candidate comparison stage, several evaluators are used to evaluate each candidate. Partial-order relations are established between the candidates, but a preference is established only if one candidate is clearly better than the other. In the clique formation stage, a set of hyptheses that are mutually consistent and together explain larger portion of the image are grouped together. A best-first strategy is used. The best candidates of each class are first chosen to build cliques. In the clique selection stage, the

best set of candidates that explains the largest portion of the image is chosen as the final interpretation of the scene.

In [9], the above context-based system is adapted to a semiautomated system. In this system, only context sets for generating candidates are used to trigger different labeling algorithm. In each labelling algorithm, a mathematical objective function is minimized to enforce the contextual constraints. The quality of the potential labelings is evaluated by humans. Compared to the above pure rule-based system, this system has the trend to approaching the framework of probabilistic models modeling contexts.

4 Probabilistic Context-Based Systems

In this section, I will review several representative systemes based on probabilistic models modeling contexts, which are parts-based models modeling intra-object context (relations), systems modeling all levels of spatial contexts, and systems modeling scene context and global spatial context.

4.1 Parts-based models

In section 1 and 2, I have mentioned that the intra-object relations and the (partial) identification of some parts can act as contextual information to help recognition. Parts-based models represent the intra-object context implicitly and handle the intra-object parts interactions directly. These models calculate the probability of the object category based on the configurations of different parts. In [10], a hierarchical parts-based model is described for detecting obejcts from cluttered natural scenes. A shared set of feature patterns are obtained by clustering all the SIFT descriptors [11] from training images into several discrete bins. A shared set of parts are assumed, and each part is represented as a multinomial distribution over the feature patterns, and each object is represented as a multinomial distribution over the sharted set of parts. In [12], a similar hierarchical parts-based system is described to model intra-object contextual information (part relations). Unlike [10], no feature detector is used and the parts are learned from the data directly.

4.2 Probabilistic systems modeling all levels of spatial contexts

Singhal et al [13] introduced a probabilistic spatial context system for scene content understanding. Given an input natural scene image, the system will segment the image into several regions having semantic labels from a predefined list such as sky, grass, foliage, water, and snow etc, which is a standard image labelling task. In the system, a number of individual material detectors are used to generate raw labelings of the input image first. Then, for each segmented region, a trained two layer bayesian network is used to integrate the output labelings of all the material detectors for that region given the observed rough

location (top, middle or bottom) of the region. Then a probabilistic model based on spatial contexts is used to refine the fused labeling.

Spatial contexts modeling is done as follows: a set of spatial relationships is defined beforehand, that is, {above, far above, below, far below, beside, enclosed, and enclosing}. Given a set of labeled training images, for each relationship, the probability between pairwise labeled regions is calculated simply by counting and normalizing. After the fused labelings are obtained, the regions are ranked according to the associated confidence factor of their labels, say, they are arranged as $\#1, \#2, \ldots, \#n$, where n is the total number of the segmented regions. Then a greedy approach is used to build n Bayesian networks to approximate the spatial contexts among regions. The first network is rooted the region #1, its posterial probability of its labeling doesn't change. The i-th network makes the first i-1 regions as leafs, given the spatial relationships between the first i-1 regions and the i-th region, the posteria probability of the labelings of the first i regions are reestimated. At last, for each region, the label with the highest probability is picked.

Instead of just approximating the spatial contexts, He et al [14] used multiscale conditional random fields (CRF) to model the local-range context, long-range (regional) context, and global range context in order to solve the image labeling problem. In the system, different classifiers focus on different range context at different scales. The whole consistant labeling is achieved by multiplying these classifiers together to get a product-of-experts model. A neural network based classifer looks at the patch centered at each pixel to predict the label of that pixel. Given the labels predicted by the classifier, many Restribed Boltzmann Machines (RBM) are used to model the label patterns in many long-range regions. And a RBM is used to model the global-range label patterns in the whole image. Based on the same idea as He et al's, the system introduced in [15] used boosted CRF (BRF) to model the spatial context at different scales to do image labeling. In Kumar and Hebert [16], two-layer hierarchical CRF is used to do image labeling. In each layer, unary potential is simply modeled as softmax logistic regression.

4.3 Probabilistic systems modeling global scene contexts

Torralba et al [17] built a system that models scene contexts which is the "gist" of the scene. When the system moves through the world, it can tell where it is and what is is looking at by analyzing the global scene context. Texture features of an image are obtained using a wavelet image decompostion, and the global scene context of the image is a low-dimensional representation of the texture features reduced by PCA. An HMM is learned with the place being the state and the global context being the output vector. The place is predicted by calcuating the posteria probability of the state given the observed global scene context vectors. The probability of an object appearing in the current scene given the current place and the observed global scene context vetors so far can be easily calculated using Bayes rule.

Oliva et al [3] combined the global scene context and the saliency of local features to detect the locations of humans from natural scene images. The experimental results showed that the top-down control provided by the global scene contextual information is essential for efficient object recognition.

5 Comparisons and Discussions

In this paper, several systems for modeling several types of different contexts have been reviewed. Different models make different assumptions when emphasizing different contextual information. Shape context is only good at capturing intra-object context. But shape context itself ignores detailed geometry in the shape, so it is not invariant to articulation. And because it is a global shape context, it is sensitive to occlusion and local distortion. Generalized shape context still has these problems. Besides, it's hard to get good shape contexts for objects with cluttered background. Despite these disadvantages, shape context provides very good representations images of generic objects taken under restricted situations.

Context sets based system is an engieering system. It not only models contexts in images, but also models contexts under which the images are taken. It is only suited for analyzing simple natural scene images. Although the system is only a rule-based knowledge-base driven system and building a good knowledge base is very difficult, the ideas used in the system are very significant. For e. g., using redundant operators, establishing partial-order in a conserved way, and hierarchical processing are all important ideas. In some sense, Singhal et al's system [13] can be viewed as a probabilistic translation of the context sets based system.

Current parts-based models only models configurations of parts. The intra-object context here is not as strong as that given by shape context, since current parts-based models often model object category independent of part positions or ignoring modeling positions. Future extensions can be made to the current parts-based models by incorporating relative postions and orientations of parts to model the object category. This will make the parameter estimation slightly harder.

The success of Singhal et al's system [13] is highly dependent on the performance of individual material detectors. It only considers spatial contexts but igores global scene contexts. In [14], [15] and [16], scene contexts are not considered, and they cannot handle a large training set with a big set of labels. In [3] and [17], only global scene contexts are considered.

From the above discussions, we can find that, the systems mentioned either cannot handle complicated scene understanding task, or are not capable of modeling all types of context, or cannot handle real large dataset. Because learning all the context information from scratch is hard, we can turn to robust individual filters to sequentially process images as in [13]. Then we can use CRF or BRF to learn all levels of spatial contexts and use global contexts to define priors. And

we might use graphs to represent all types of context information, then we can do computations based on the graphs.

6 Conclusions

In the paper, I reviewed several context-based computer vision systems, which include shape context based systems along with some fast pruning, fast matching and fast indexing algorithms that scale well with large datasets, context-sets based systems, and several representative probabilistic models. My motivation for doing this is to hope that the context-based models with some fast matching methods in computer vision can help my research in motif discovery. I believe that the fast indexing and matching algorithm based on graphs are very useful for structure motif retrieval. And the models reviewed here also illustrate what all the types of contexts are and how they can be modeled.

References

1. T. Strat and M. Fischler: Context-based vision: recognizing objects using information from both 2D and 3D imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 13, Issue 10, Oct. 1991 Page(s):1050 - 1065.
2. M. Barll and S. Ullman: Spatial context in recognition. Perception. 1996. Volume 25, pages 343-352
3. Aude Oliva, Antonio Torralba, Monica S. Castelhano, John M. Henderson: Top-Down Control of Visual Attention in Object Detection. Journal of Vision. 2003
4. I. Biederman: Recognition-by-Components: A Theory of Human Image Understanding. Psychological Review, 94, 115-147.
5. S. Belongie, J. Malik and J. Puzicha: Shape Matching and Object Recognition Using Shape Contexts, PAMI, 24(4):509-522, April 2002.
6. G. Mori, S. Belongie, and J. Malik: Efficient Shape Matching Using Shape Contexts. IEEE PAMI. November 2005 (Vol. 27, No. 11), pp. 1832-1837.
7. H. Chui and A. Rangarajan: A New Algorithm for Non-Rigid Point Matching. CVPR, vol. 2, June 2000, pp. 44-51.
8. F. Demirci, A. Shokoufandeh, Y. Keselman, L. Bretzner, and S. Dickinson: Object Recognition as Many-to-Many Feature Matching. International Journal of Computer Vision, 2006, to appear
9. A. Shokoufandeh, D. Macrini, S. Dickinson, K. Siddiqi, and S. Zucker: Indexing Hierarchical Structures using Graph Spectra IEEE Transactions on Pattern Analysis and Machine Intelligence, special issue on Syntactic and Structural pattern Recognition, Volume 27, Number 7, July 2005.
10. T. M. Strat: Employing Contextual Information in Computer Vision DARPA93, 1993
11. E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky: Learning hierarchical models of scenes,objects and parts. ICCV. 2005.
12. D. Lowe: Object recognition from local scale invariant features, ICCV, 1999.
13. D. A. Ross and R. S. Zemel: Learning parts-based representations of data. Journal of Machine Learning Research. 2006. To Appear.

14. A. Singhal, J. Luo, and W. Zhu: Probabilistic spatial context models for scene content understanding. CVPR, 2003.
15. X. He, R. S. Zemel, and M. A. Carreira-Perpina: Multiscale conditional random fields for image labeling. CVPR. 2004.
16. A. Torralba, K. Murphy and W. Freeman: Contextual Models for Object Detection using Boosted Random Fields. NIPS'04.
17. S. Kumar and M. Hebert: A Hierarchical Field Framework for Unified Context-Based Classification. Proceedings, IEEE International Conference on Computer Vision (ICCV), 2005.
18. A. Torralba, K. P. Murphy, W. T. Freeman and M.Rubin: Context-based vision system for place and object recognition. ICCV. 2003.

This article was processed using the LaTeX macro package with LLNCS style