# Extracting Spatial Templates for Image Indexing

Huizhen Yu & W. Eric L. Grimson

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

**The Problem:** We propose a graph-based method for automatically extracting common spatial templates from query images for image indexing. These templates, describing common patterns by their features (e.g., color/texture) as well as their spatial arrangements, can be used in an inference system, along with the statistical features of the query examples, to retrieve more relevant images.

**Motivation:** Image indexing using information from images alone is a difficult task, given our current level of object recognition and scene interpretation. To ease this difficulty, one should exploit as much as possible both the prior knowledge about the database and the power in different representations or different retrieval methods. It is in the spirit of the latter, which is proposed as "a society of models" by Minka and Picard [1], that we believe constructing spatial templates for indexing is a meaningful goal to pursue. This is because, in the form of an attributed graph [2], spatial templates can contribute valuable contextual information, that may not be captured by only the statistical measurements of the query images.

**Previous Work:** Lipson [3] manually constructed spatial templates for several scene classes and demonstrated their promising performance in image retrieval. Lakshmi Ratan further formalized it as a multiple instance learning problem [4] and Diverse Density, an algorithm proposed by Maron, was applied to learn the template from a set of example images automatically [5]. One difficulty in their approach arises from the fact that each possible candidate is expressed explicitly in the space where the learning algorithm is seeking the optimal template, and the number of such candidates increases combinatorially as the size of the template increases, thus in practice restricting the complexity of the templates under consideration.

Among graph-based methods, our approach bears similarity to that of Das [6] in forming the template as an attributed graph, whereas in their context they know the target pattern a priori and build the template by constructing an adjacency graph for significant color patches in that pattern.

**Approach:** Decision graphs and maximum clique methods have been used to efficiently generate matching hypotheses for model based matching of rigid planar objects [7]. A decision graph is constructed such that each vertex represents an assignment of one image-component to one model-component, and each edge indicates consistency between two assignments under some constraints, such as the one-to-one matching and the rigidity constraint. Therefore, any set of vertices forming a completely connected subgraph, in particular a clique, generates a valid hypothesis of a possible matching.

We apply an extended decision graph approach to the problem of extracting a common template from two segmented images. Our approach differs from the standard one, as described above, in two aspects. First, each vertex in the decision graph represents a pair of component-assignments. Second, each vertex is associated with a height label indicating the goodness in its representing assignments. The reasons for these differences are briefly stated below.

Without the rigidity constraint under our problem setting, it is not necessary that all pairs of components in the matching assignment satisfy the spatial constraints. As a consequence, when directly using the standard approach, a set of consistent assignments may not form a completely connected subgraph. The solution we choose, essentially, is to let the assignments represented by each vertex satisfy those one and two-variable constraints, and to let the three-variable constraints be enforced by edges between vertices.

Introducing height labels brings flexibility into the matching scheme. As cliques are greedily grown in induced graphs at a gradually lowered height level, in the image plane, the corresponded templates are "grown" from where the matching is more reliable.

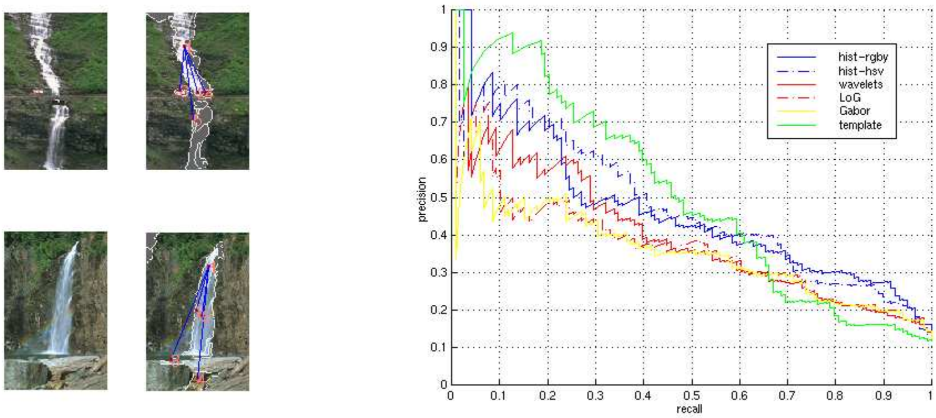


Figure 1: Left: An extracted template is superimposed on two images. Right: Retrieval results on waterfall images. The template approach uses the template shown at left, and the other methods use statistical measurements from one of the left images.

More details of our matching method is in [8]. Figure 1 shows one extracted template and its retrieval performance, compared to other methods using statistical measurements.

**Difficulty:** The computational cost is one of the major concerns in all template based approaches. As subgraph isomorphism is NP complete, we have to rely more on the nature of our problem to reduce the space and computational complexity. The method proposed above makes a reasonable compromise between the simplification of image contents and the computational demand, and is efficient in handling images with about 20 regions after segmentation.

The impreciseness in segmentation, reflecting the current gap between machine and human vision, affects the relevance of the extracted templates to the query concept. As segmentation errors are inevitable in an automatic system, the uncertainty information obtained at the matching stage needs to be preserved and utilized at the following inference stage. In our method, this uncertainty is indicated by the height labels associated with matching assignments and the lowest level where the clique is finally derived.

**Impact:** With an algorithm able to quickly generate hypotheses of relevant templates, it is then possible to combine the structural information in spatial templates with statistical features in image indexing.

**Future Work:** We are currently working on an inference framework that utilizes both the spatial templates and statistical measurements. With extremely few example images, how to handle correlations between different features is the most critical part.

**Research Support:** This work is supported by the Industrial Technology Research Institute.

**References:**

[1] T. P. Minka and R. W. Picard. Interactive learning using a ”society of models”. *IEEE Conference on Computer Vision and Pattern Recognition*, 1996.

[2] R. Haralick, and L. Shapiro. *Computer and Robot Vision*, Addison-Wesley, 1992.

[3] P. Lipson. Context and Configuration Based Scene Classification. *Ph.D. Thesis*, MIT, 1996.

[4] A. Lakshmi Ratan, and E. L. Grimson. Training Templates for Scene Classification using a Few Examples. *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997.

[5] O. Maron, and A. Lakshmi Ratan. Multiple Instance Learning for Natural Scene Classification. *Machine Learning: Proc. the 11th International Conference*, 1998.

[6] M. Das, E. Riseman, and B. Draper. FOCUS: Searching for Multi-colored Objects in a Diverse Image Database. *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[7] R. C. Bolles and R. A. Cain. Recognizing and Locating Partially Visible Objects: The Local-Feature-Focus Method. *The International Journal of Robotics Research*, Vol. 1, Num. 3, 1982.

[8] H. Yu, and E. L. Grimson. Extracting Spatial Templates from Query Images for Image Indexing. *Unpublished*, 1999.