Similarity Templates for Detection and Recognition

Chris Stauffer & Eric Grimson

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: This work investigates applications of a new representation for images, the similarity template(ST). A similarity template is a probabilistic representation of the similarity of pixels in an image patch. It has application to detection of a class of objects, because it is reasonably invariant to the color of a particular object. Further, it enables the decomposition of a class of objects into component parts over which robust statistics of color can be approximated. These regions can be used to create a factored color model that is useful for recognition. Detection results are shown on a system that learns to detect a class of objects (pedestrians) in static scenes based on examples of the object provided automatically by a tracking system. Applications of the factored color model to image indexing and anomaly detection are pursued on a database of images of pedestrians.

Motivation: Detection and recognition in color images are often approached with completely different representations of images. For detection of a class of objects, a representation is sought that is invariant to the color of a particular object (e.g., edge templates, gray-scale Haar wavelets, etc.). In contrast, for recognition of a particular instance, often the colors of particular regions are extremely important in differentiating instances. This work develops a new representation that models the pairwise similarity between all pixels in an image patch.

Previous Work: Object detection refers to detecting an instance of a particular class of object. Some examples of detection tasks are face detection [2], pedestrian detection [1], and vehicle detection [1]. Edge templates are often used for class distinctions because of their invariance to scene lighting and object color. They have similar properties to similarity templates (STs), but they are based on a measure of local differences as opposed to global similarities.

Principal Component Analysis, Multi-scale Gabor filters, and Haar wavelet functions are examples of projections of images into a lower dimensional space to facilitate recognition. Generally the coefficients in these spaces show invariance to noise within regions. Unfortunately, using these to make a general detector usually involves a complex supervised training algorithm [1], which is often run on only gray-scale images. While neglecting color information entirely is arguably ill advised, many researchers have found that learning on a color image space requires much more complexity in the classifier and extremely large data sets to train.

Approach: For an N-pixel image the corresponding template, S, is an NxN matrix. The element, $S_{i,j}$, is an estimate of the probability of pixel j being drawn from a similar colored region as pixel i. Aggregate similarity templates estimate the same statistics over an entire class-specific data set.

To estimate a similarity template, S, from a single image, I, a distribution over pixels p_j is derived, conditioned on each pixel, p_i, as follows

$$S(p_i, p_j) = Pr(p_i)Pr(p_j|p_i) \tag{8}$$

where $Pr(p_i)$ is the prior probability of choosing any pixel in the image and $Pr(p_j|p_i)$ is an estimate of the probability that the j^{th} pixel was produced by a region whose mean color is the same as the i^{th} pixel's color. Using this formulation, we are able to decide which pixels contribute most to the representation by altering $Pr(p_i)$. The general form of the second element of Equation 8 is

$$Pr(p_j|p_i) = \frac{f(I(p_j), I(p_i))}{\sum_{j \in N} f(I(p_i), I(p_j)} \tag{9}$$

where f is any function that measures a probability of being from a similar region. Effectively the i^{th} row of the ST represents the conditional probability that each pixel belongs to the same region as the i^{th} pixel. Further details are available in [4].

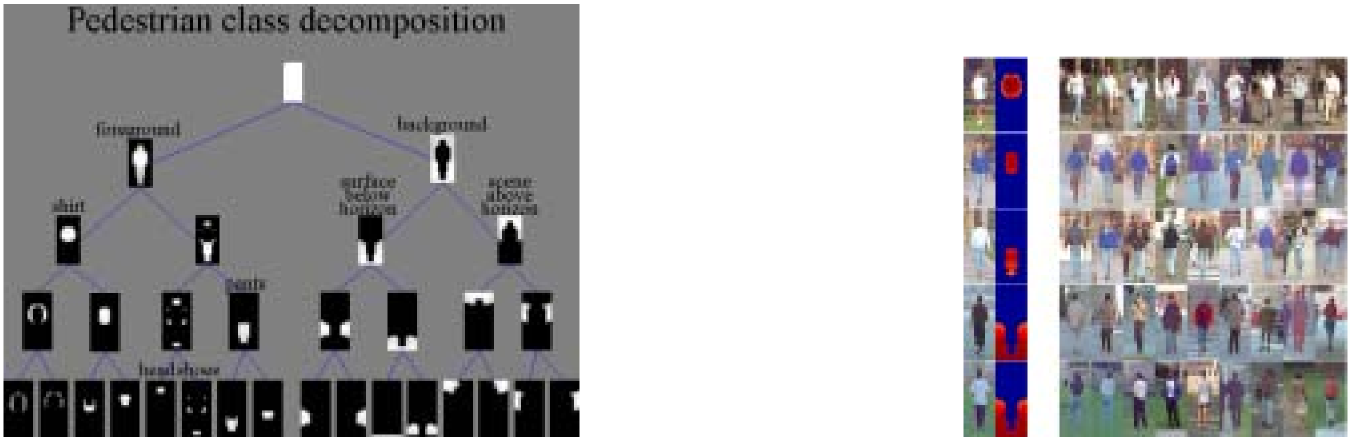

Figure 1: This figure shows the automatically generated binary decomposition of the image patch for the pedestrian data set. The root node represents every pixel in the image. The first branch represents foreground vs. background pixels. Further branches are discussed below. On the right are some simple example-based queries using this representation.

An aggregate similarity template is computed assuming that each similarity template in the training set, τ, is a noisy estimate of true class-specific distribution. Therefore, the aggregate similarity template is simply an average of all the training STs.

Using an existing tracking algorithm[3] in our laboratory environment, 32x32 patches centered on the centroid of pedestrians and scaled to include the entire person were automatically extracted from a live video source. The background images were extracted from the scenes randomly at approximately the same scale. We trained two agreggate STs for the pedestrians and non-pedestrians. Using a likelihood ratio, we were able to acheive a equal error rate of less than 2%.

The ST contains aggregated similarity statistics which can be exploited to determine a decomposition of the pixels in pedestrian images. Rather than performing a straight K-way clustering on the pixel's to obtain M pixel region models, we extracted a hierarchical segmentation in the form of a binary tree. This decomposition is extremely useful for image database queries (as shown in Figure 1) and data mining applications.

Impact: We have shown the application of similarity templates to both detection of a class of objects and recognition of a particular object within that class. We believe this representation offer a unified approach to both areas.

Future Work: Some future work will involve overcoming the computational and space complexity of this method by using multi-scale templates. There is ongoing work on simulataneously aligning a dataset and deriving the corresponding aggregate ST. If the alignment work matures, we will have the capability to detect instances of objects in a dataset, align them, derive informative features from them, and use that compact representation for content-based retrieval. Other future work will involve using this representation for other object classes, binding language to this factorable representation, incorporating other notions of similarity, and applications of this representation to object detection.

Research Support: This work has been funded by DARPA under contract number N00014-00-1-0907 administered by the Office of Naval Research.

References:

[1] M. Oren, C. Papageorgiou, E. Osuna P. Sinha, and T. Poggio. Pedestrian detection using wavelet templates. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 193–199, Puerto Rico, June 16–20 1997.

[2] T. Poggio and K.K. Sung. Example-based learning for view-based human face detection. In *Proc. of the ARPA Image Understanding Workshop*, volume II:843-850, 1994.

[3] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[4] C. Stauffer and W.E.L. Grimson. Similarity templates for detection and recognition. In *Proc. Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001.