Abstract

Recent work has shown impressive transform-invariant modeling and clustering for sets of images of objects with similar appearance. We seek to expand these capabilities to sets of images of an object class that show considerable variation across individual instances (e.g., pedestrian images) using a representation based on pixel-wise similarities, similarity templates. Because of its invariance to the colors of particular components of an object, this representation enables detection of instances of an object class and enables alignment of those instances. Further, this model implicitly represents the regions of color regularity in the class-specific image set enabling a decomposition of that object class into component regions.

1 Introduction

Images of a class of objects are often not effectively characterized by a Gaussian distribution or even a mixture of Gaussians. In particular, we are interested in modeling classes of objects that are characterized by similarities and differences between image pixels rather than by the values of those pixels. For instance, images of pedestrians (at a certain scale and pose) can be characterized by a few regions of regularity (RORs) such as shirt, pants, background, and head, that have fixed properties such as constant color or constant texture within the region, but tend to be different from each other. The particular color (or texture) of those regions is largely irrelevant. We shall refer to sets of images that fit this general description as images characterized by regions of regularity, or ICRORs.

Jojic and Frey [1] and others [2] have investigated transform-invariant modeling and clustering for images of a particular object (e.g., an individual’s face). Their method can simultaneously converge on a model and align the data to that model. This method has shown positive results for many types of objects that are effectively modeled by a Gaussian or a mixture of Gaussians. Their work with transformed component analysis (TCA) shows promise for handling considerable variation within the images resulting from lighting or slight misalignments. However, because these models rely on an image set with a fixed mean or mixture of means, they are not directly applicable to ICRORs.
We would also like to address transform-invariant modeling, but use a model which is invariant to the particular color of component regions. One simple way to achieve this is to use edge templates to model local differences in image color. In contrast, we have chosen to model global similarities in color using a similarity template (ST).

While representations of pixel similarity have previously been exploited for segmentation of single images [3, 4], we have chosen to use them for aggregate modeling of image sets. Similarity templates enable alignment of image sets and decomposition of images into class-specific pixel regions. We note also that registration of two ICRORs can be accomplished by minimizing the mutual information between corresponding pixels [5]. But, there is no obvious way of extending this method to large sets of images without a combinatorial explosion.

Section 2 briefly introduces similarity templates. We investigate their uses for modeling and detection. Section 3 discusses dataset alignment. Section 4 covers their application to decomposing a class-specific set of images into component regions. Future avenues of research and conclusions are discussed Section 5.

2 Similarity templates

This section begins with a brief explanation of the similarity template followed by the mechanics of computing and comparing similarity templates. A similarity template $S$ for an $N$-pixel image is an $N \times N$ matrix. The element $S_{i,j}$ represents the probability that pixel locations $p_i$ and $p_j$ would result from choosing a region and drawing (iid) two samples (pixel locations) from it. More formally,

$$S_{i,j} = \sum_r p(r)p(p_i|r)p(p_j|r),$$

where $p(r)$ is the probability of choosing region $r$ and $p(p_i|r)$ is the probability of choosing pixel location $p_i$ from region $r$.

2.1 The “ideal” similarity template

Consider sampling pixel pairs as described above from an $N$-pixel image of a particular object (e.g., a pedestrian) segmented by an oracle into disjoint regions (e.g., shirt, pants, head, feet, background). Assuming each region is equally likely to be sampled and that the pixels in the region are selected with uniform probability, then

$$S_{i,j} = \begin{cases} \left(\frac{1}{R}\right)(\frac{1}{S_r})^2 & \text{if } r_i = r_j \\ 0 & \text{otherwise,} \end{cases}$$

where $R$ is the number of regions, $S_r$ is the number of pixels in region $r$, and $r_i$ is the region label of $p_i$. If two pixels are from the same region, the corresponding value is the product of the probability $\frac{1}{R}$ of choosing a particular region and the probability $\left(\frac{1}{S_r}\right)^2$ of drawing that pixel pair. This can be interpreted as a block diagonal co-occurrence matrix of sampled pixel pairs.

In this ideal case, two images of different pedestrians with the same body size and shape would result in the same similarity template regardless of the colors of their clothes, since the ST is a function only of the segmentation. An ST of an image without a pedestrian would exhibit different statistics. Note that even the ST of an image of a blank wall (segmented as a single region) would be different because pixels that are in different regions under the ideal pedestrian ST would be in the same region.
Unfortunately, images do not typically come with labeled regions, and so computation of a similarity template is impossible. However, in this paper, we take advantage of the observation that properties within a region, such as color, are often approximately constant. Using this observation, we can approximate true similarity templates from unsegmented images.

### 2.2 Computing similarity templates

For the purposes of this paper, our model for similarity is based solely on color. Since there is a correlation between color similarity and two pixels being in the same region, we approximate the corresponding value $\tilde{S}_{i,j}$ with a measure of color similarity:

$$\tilde{S}_{i,j} = \frac{1}{NZ_i} \exp \left( -\frac{||I_i - I_j||^2}{\sigma_i^2} \right),$$

where $I_i$ and $I_j$ are pixel color values, $\sigma_i^2$ is a parameter that adjusts the color similarity measure as a function of the pixel color distribution in the image, and $Z_i$ is the sum of the $i^{th}$ row. This normalization is required because large regions have a disproportionate effect on the ST estimate. The choice of $\sigma_i^2$ had little effect on the resulting ST.

If each latent region had a constant but unique color and the regions were of equal size, then as $\sigma_i^2$ approaches zero this process reconstructs the “ideal” similarity template defined in Equation 1. Although region colors are neither constant nor unique, this approximation has proven to work well in practice.

It is possible to add a spatial prior based on the relative pixel location to model the fact that similarities tend to local, but we will rely on the statistics of the images in our data set to determine whether (and to what extent) this is the case. Also, it may be possible to achieve better results using a more complex color model (e.g., hsv with full covariance) or broadening the measure of similarity to include other modalities (e.g., texture, motion, depth, etc.).

Figure 1 shows two views of the same similarity template. The first view represents each pixel’s similarity to every other pixel. The second view contains a sub-image for each pixel which highlights the pixels that are most likely produced by the same region. Pixels in the shirt tend to highlight the entire shirt and the pants (to a lesser amount). Pixels in the background tend to be very dissimilar to all pixels in the foreground.

### 2.3 Aggregate similarity templates (AST)

We assume each estimated ST is a noisy measurement of the true underlying joint distribution. Hence we compute an aggregate similarity template (AST) as the mean $\bar{S}$ of the ST estimates over an entire class-specific set of $K$ images:

$$\bar{S}_{i,j} = \frac{1}{K} \sum_{k=1}^{K} \tilde{S}_{i,j}^k.$$  

For this quantity to be meaningful, the RORs must be in at least partial correspondence across the training set. Note that this is a less restrictive assumption than assuming edges of regions are in correspondence across an image set, since regions have greater support. Being the mean of a set of probability distributions, the AST is also a valid joint probability distribution.