Translation Templates for Object Matching Across Predictable Pose Variation

Chris Stauffer and Matthew Antone
BAE Systems, Advanced Information Technologies Group
Burlington, MA USA
{chris.stauffer,matthew.antone}@baesystems.com

Abstract

Computer vision is the most promising technology for automated, passive tracking of multiple objects over large areas. Effective inter-camera and intra-camera visual tracking can enable information about a vehicle or a pedestrian to be integrated from various sources. Unfortunately, tracking objects across multiple non-overlapping cameras requires reliable comparison of the objects' appearance under widely-varying view angles and resolutions. Fortunately, in most cases, an object of a particular type entering a scene at a particular position and direction will tend to be in a very similar pose. This paper introduces Translation Templates (TTs). TTs exploit this regularity to learn a color-based matching metric for images from a pair of tracking source and sink points, without prior knowledge of object type or object pose. This model benefits from histogram-based aggregation while still preserving spatial relationships between the two images. The model can be learned directly from data and used to compare arbitrary types of objects observed from extremely different viewpoints, as long as the relationship between the viewpoints is preserved. This paper describes TTs, describes a method for efficient computation and for visualization of TTs, and presents experimental results from both an indoor pedestrian data set and an outdoor vehicle data set.

1 Introduction

Systems that track objects over a large area inevitably require multiple cameras and often include significant regions where objects are not visible. Even tracking within a single camera can involve matching the appearance of an object at two different locations in a scene, e.g., a car entering the east side of a parking structure and exiting the south side of the structure. Fortunately, in the majority of our experimentation a particular type of object appearing or disappearing at a particular location will present in a similar pose. In a five camera vehicle experiment, vehicles appeared within five degrees of the average direction in 99.19% of 862 vehicles. In a four camera indoor pedestrian experiment, pedestrians entered from ten different locations in four cameras in similar pose for 97.56% of 82 cases.

Figure 1 illustrates the uniformity within two automatically generated sets of images of pedestrians entering two different scenes at two different points. It also illustrates the

BMVC 2006 doi:10.5244/C.20.124

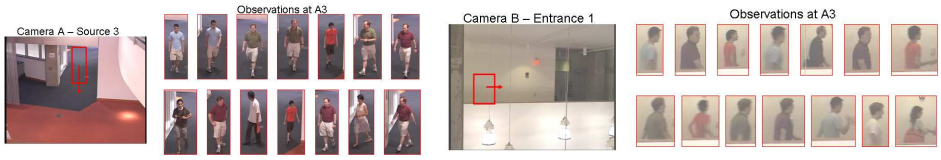

Figure 1: This figure shows examples of pedestrians entering one scene from the top and pedestrians entering a second scene from the left while suffering an unresolvable static occlusion that blocks the entire lower half of the pedestrians. Although the poses of the two sets are very different, within each of the sets, the poses are very similar.

non-trivial relationship between the two sets of images, one containing front-facing, full-view pedestrians and the other containing side-view pedestrians under significant (but predictable) occlusion. A template-based matching algorithm would require complex occlusion reasoning or hand-coding and would be fundamentally limited by observing different portions of the same object. A global histogram-based approach with no spatial reasoning will attempt to compare colors of an entire object with those of a small portion of an object. Clearly, neither approach will be particularly effective for this scenario.

If the object type and viewing angle were known, one could attempt to design specific detectors for informative color features and use robust aggregate estimates to compare the two objects. Taking vehicles as an example, such detectors would extract the color of the side panels, hood, hubcap, and roof to enable a robust comparison that is independent of viewing angle, but aware of when one or more of these features is unreliable (e.g. when the wheels are not visible). Unfortunately, this procedure would be difficult and time intensive and would require robust and accurate registration. This paper proposes an automated method for making a similar comparison and illustrates that it is capturing interesting relationships between the two images of the object.

1.1 Prior Work

This document is primarily concerned with color-based object matching. For this reason, we will not pursue a detailed discussion on the vast quantities of work in the areas of either object detection or object classification. We instead assume that object detection and localization has already been performed by a tracking system. We also assume that the general class of object, e.g., pedestrian or vehicle, is either known or can be easily determined using gross object characteristics.

Classical object matching involves either template matching or color histograms. Template matching requires spatial correspondence. Deformable templates [10] can be robust to more variation in pose. Histogram matching can be invariant to a much larger change in relative pose as long as similar colors are represented in similar amounts. A hybrid was introduced by Mittal and Davis[4] who built vertically binned histograms to match five different parts of pedestrians. The aggregation of pixels in each region resulted in more robust comparisons than direct pixel matching, but their approach only works for images of upright pedestrians in full view taken by a camera with an informative view of the pedestrians. Histograms of directed edge energies, Hessians, and other texture features can also be employed. Schettini et al. [6] produced a survey on color comparison methods for image database applications.

Hausdorff matching[5, 2], shape context[1], and alpha-edge-images[7] are examples of approaches that are invariant to lighting and slight misalignment. Unfortunately, none of these approaches are capable of dealing with gross misalignment. They are also all invariant to color, and thus show no hope of differentiating a red Toyota Camary from a blue Toyota Camary, and they show little promise in differentiating deformable objects, such as pedestrians. Recently, Shan et al. [7] built a system to learning embeddings of pair-wise similarity metrics to enable matching across gross changes. Though their system showed promise matching vehicles based on vehicle type, it is unclear whether a color-invariant, edge-based representation could differentiate pedestrians.

In this paper, we introduce Translation Templates (TTs). TTs leverage the advantages of both aggregation and spatial reasoning. TTs are capable of being applied to *both* pedestrians and vehicles. TTs can be trained automatically without specifying the object type, relative camera pose, relative object pose, or a detailed object model. By learning which regions are likely to contain the same colors (and which do not), a robust estimate of the similarity of the objects can be estimated even in the situation shown in Figure 1, where the "pants" region has no corresponding region in the second image.

Section 2 describes how TTs are estimated, how TTs can be decomposed and analyzed, and how they can be employed for robust matching. Section 3 explains how TTs subsume both simple template matching and simple global color histogram matching. Section 4 shows results from an indoor pedestrian matching experiment and an outdoor vehicle matching experiment. Section 5 discusses future work and conclusions.

2 Translation Templates (TTs)

Object matching involves estimating the likelihood that two images are observations of the same object as well as the likelihood that they are different objects. Using prior weights on each of these cases, the posterior probability that this pair of observations are observations of the same object can be estimated. This probability can be used together with linking constraints to stitch together tracks when an object passes between multiple cameras or suffers some other predictable occlusion.

2.1 Translation Template Overview

TTs are estimated directly from sets of image pairs of objects at two different locations. Given image pairs, $\{I_a, I_b\}$, taken at locations a and b, our goal is to define the match likelihood and non-match likelihood of a pair of images, or $p(I_a, I_b|m_{ab})$ and $p(I_a, I_b|\bar{m}_{ab})$. For simplicity, we will assume that all images from location a have been scaled to N_a pixels and all images from location b have been scaled to N_b pixels.

If we were presented with an infinite set of matching image pairs $\{I_a, I_b\}_i : i \in [1, \text{inf}]$ and we were given a ground truth segmentation for paired, similarly-colored regions that are contained in each pair of images, we could estimate the true likelihood that pixel p_i in the first image and a pixel p_j in the second image were drawn from corresponding regions in the two images data set, or

$$\mu^*_{ij} = p(p_i, p_j|r_i = r_j)p(r_i = r_j), \tag{1}$$

where r_i and r_j are the region labels of the respective pixels. The colors of pixel pairs for which the value of μ^*_{ij} is low will generally have no simple relationship, whereas the

colors of pixel pairs with a high value of μ^*_{ij} are more likely to have been drawn from the same color process. Intuitively, we would like our matching metric to exploit any and all pairs of pixels that are likely to be from similarly colored regions and to discount pairs of pixels that are likely to be from different regions. Thus, we will effectively be able to compare the color distribution in the shirt in one set of images to the color distribution of the shirt in the other set of images.

Unfortunately, producing ground truth segmentations of matched similarly-colored regions is extremely labor intensive, is not a well-defined task, and would involve characterizing aspects of appearance such as shadows and specularities that may change over time. By estimating this value directly from the data, a TT can be quickly approximated and can be adapted over time as conditions change.

2.2 Estimating TTs

In order to estimate the ideal measure μ^*_{ij}, we must estimate whether a pixel is "within the corresponding similarly-colored region". Given a set of M images, our estimate is

$$\mu_{ij} = p(p_i, p_j, r_i = r_j) = \frac{1}{M} \sum_{m=1}^{M} k(p_i, p_j), \tag{2}$$

where $k(p_i, p_j)$ is an estimate of the likelihood that two pixel colors are derived from the same region. In this work, we define the k-function as a Gaussian kernel. This function will be poor if there is significant color drift or gross lighting changes between the pairs of images. Alternatively, the k-function could be estimated from images as described in [8] and may require learning a color transfer function between pairs of cameras as described in [3]. The same kernel is used in the results section for the alternative approaches as the approach described here, enabling a consistent comparison. The k-function will approach a maximum value for pixels that are the same color and decrease for colors that are increasingly different. Unfortunately, this introduces two types of error to our estimation described in the previous section.

First, the k-function may be near zero even for pixels drawn from the same region as a result of specularities or shadows. If this is consistent across image pairs, our algorithm will develop a decreased dependence on the corresponding pixels. While this may not result in the "ideal" region-based representation as described above, the matching scores will be more invariant to these lighting effects as a result.

Second, two pixels may coincidentally be nearly the same color even though the regions they are drawn from are completely different, e.g., red pants and red shoes. In this work, we assume this type of error is uniformly distributed and no more common in the set of matching pairs of images than it is in the set of non-matching pairs of images. A major source of this type of error is common colors that occur in the background around the moving object. For this reason, we use silhouette information provided by our tracking system to mask estimation of Equation 1 to only pairs of pixels that are inside the moving objects in both images.

We estimate μ_{ij} for the set of matching image pairs as well as the non-matching image pairs. Similarity that is represented in both models will be discounted relative to similarity that is more common in the matching image pairs only. E.g., if both matching and non-matching image pairs tended to contain variable shadowed regions but only the *matching* image pairs contained shirt regions that were similar across image pairs, the shirt region will be more significant in the matching score.

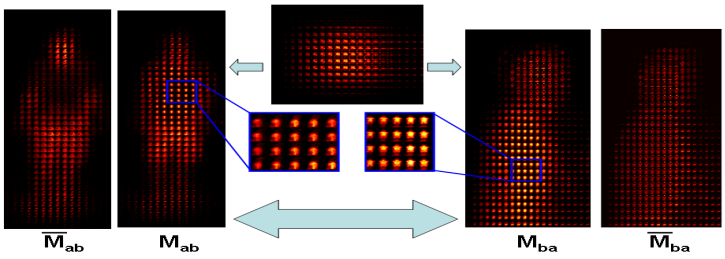

Figure 2: This figure shows an aggregate M-matrix derived from images pairs (top). Its transformed matching and non-matching counterparts (left and right) contain the same number of elements as the original matrix, but each row has been reconstructed into a template and placed in the corresponding location for the pixel that it represents. The enlarged portion in the upper-right of the left template illustrates that the center of the shirt region in I_a has a high affinity to the shoulders and chest regions of I_b. Similarly, the middle of the chest region in I_b has a high affinity to center of the chest in I_a.

2.3 Decomposing and Visualizing TTs

Figure 2 (top) shows an $N_a x N_b$ matrix M_{ab} containing the μ_{ij} values as calculated for matching pairs of images in the example in Figure 1. The $N_a x N_b$ matrix is difficult to evaluate and understand because each image is represented as a single vector. For this reason, we visualize TTs as shown in Figure 2 (left) and (right), see caption for description. The blurriness of the templates is the result of the rough, automated alignment and the variation in the over all shapes of the pedestrians. By comparing M_{ab} and $\overline{M}_{ab}$, it is apparent that there is significant regularity in the shirt regions of matching pedestrians that is not represented in the non-matching pedestrians, whereas there is significant similarity in the face and arm regions for both matching and non-matching pedestrians in our database. It is also evident that the shorts in the center of M_{ab} have regularity with the shorts at the very bottom of M_{ba}.

It is important to note that this model was derived automatically from matching and non-matching pairs of images with no a priori knowledge of object type. By using periods of time with very few pedestrians passing between cameras, it may be possible to bootstrap this model from automated correspondences, enabling completely automatic model building (without requiring manual ground-truth matching). It is also important to note that all images from one location could be translated, rotated, inverted, or distorted in arbitrary ways and a qualitatively similar TT would result.

It is also possible to estimate a set of K latent factors that approximate the entire $N_a x N_b$ maxtrix M_{ab} by minimizing the KL-divergence between our estimated Translation Template M_{ab} for matching object and the model

$$\hat{M}_{ij} = \sum_{k=1}^{K} p(p_i|k)p(p_j|k)p(k). \tag{3}$$

Figure 3: This figure shows three latent class decomposition of M_{ab} as approximated in Equation 3. Two example images with the corresponding class conditional distributions for each of three classes are shown. The first latent class primarily corresponds to the shirt region in both images. The second latent class primarily corresponds to the shorts in the first image and the very top of the shorts in the second image. The final latent class primarily corresponds to face, hair, arms, and legs in both images.

Beginning with random values for the latent class conditional distributions, $p(p_i|k)$ and $p(p_j|k)$, and uniform values for the latent class priors, $p(k)$, the KL-divergence can be minimized using the update equations

$$p'(p_i|k) \propto p(p_i|k) \sum_{p_j} p(k) p(p_j|k) \frac{\mu_{ij}}{\hat{\mu}_{ij}}, \tag{4}$$

$$p'(k) \propto p(k) \sum_{p_i} \sum_{p_j} p(p_i|k) p(p_j|k) \frac{\mu_{ij}}{\hat{\mu}_{ij}}. \tag{5}$$

Figure 3 shows a 3-way decomposition of the previous M_{ab} matrix. Not only is this helpful in understanding the underlying representation, but this can be used as an efficient approximation to the original similarity matrix, thereby avoiding the costly $O(N_a N_b)$ comparison operation described in the next section.

2.4 Matching with TTs

This section describes how a TT can be used to compare two images on a pixel-by-pixel basis with arbitrary transformations. To evaluate the likelihood of a particular image pair $\{I_a, I_b\}$, we estimate the likelihood of a matching pair as

$$p(I_a, I_b|match) = E_\mu[k(p_i, p_j)], \tag{6}$$

or the expectation of the likelihood of a pair of pixels as drawn from the distribution defined by the values μ_{ij}. This is the average likelihood of a pixel match given a pixel pair randomly drawn from the pixel pairs that are likely to contain similar pixel values for the particular views of the particular objects the Translation Template was trained on. The same value can be estimated for a the non-match Translation Template. Finally, the posterior of a match given the two pixels is

$$p(match|I_a, I_b) \quad = \quad \frac{p(I_a, I_b|match) p(match)}{p(I_a, I_b)}, \tag{7}$$

where $p(I_a, I_b)$ is the weighted sum of the likelihoods of a match and non-match. Given the sets of matching and non-matching pairs it is trivial to determine the values of $p(match)$ and $p(nonmatch)$ that minimize a particular cost function.

3 The extremes of TTs

This section describes how a template matching algorithm and a global color histogram matching algorithm are subsumed by TT matching. These models will be used to illustrate the need for our approach and will be a basis for comparison in the results section to follow.

3.1 A Template Matching Algorithm

Imagine that the pairs of images are of known objects in similar orientations containing the same number of pixels that are in rough correspondence. If this were the case, one could compare the two images directly as the mean of the likelihoods of the two colors at each image location. This is equivalent to having a TT with zeros everywhere except on pixels that are in correspondence. In this simple case, the TT would simply be an identity matrix. TTs can represent translation, rotation, scaling, or an arbitrary non-homeomorphic warping by simply changing the non-zero value for each row to the pixel in the second image that corresponds to the pixel represented by that row, or cases with occlusion, by removing the correspondence completely.

For arbitrarily textured objects for which the pixel-wise correspondence can be effectively predicted, this may be a reasonable approach. But, when the objects show regularity over regions and when pixel-wise correspondence is difficult to predict (such as is generally the case with pedestrians and vehicles), one can benefit from local aggregation as shown in the following subsection.

3.2 A Global Histogram Matching Algorithm

Imagine that the pairs of images are of known objects in *unknown* orientations. Further, imagine that the first images of the pair and the second images of the pair contain different numbers of pixels and are of different aspect ratios. In this case, a template comparison as described above could be arbitrarily poor. E.g., inverted or occluded pedestrians may not match any pixels with upright pedestrians in full view.

The most common approach in this difficult situation is to compare the global color distributions. This can be accomplished by taking the average likelihood of each pixel in the first image under a non-parametric density estimate of the entire second image. This is equivalent to having a TT which is uniform (and sums to one).

4 Results

In this section, we will discuss results from the previously mentioned example of pedestrian matching. We will also introduce and discuss a vehicle matching application. For each set of results we will show the performance of the algorithm as compared to the two models described in the previous section. The kernel function used for the other two approaches will be identical to the kernel used in the TT approach.

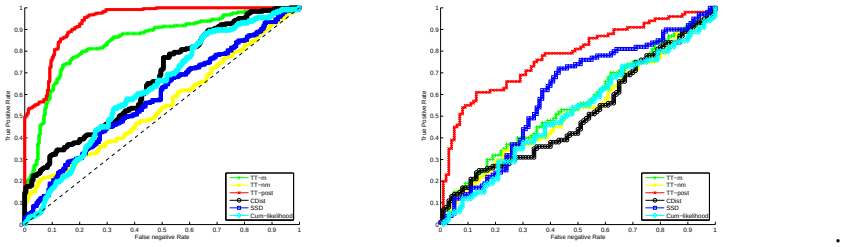

Figure 4: This figure contains a ROC-curve for an independent test set of 230 matching and 235 non-matching pedestrian images for our previous example and for a similar vehicle data set. The match likelihood (TT-m) tends to classify reasonably whereas the non-matching likelihood (TT-nm) is nearly a chance classifier over the range of classification thresholds. The resulting posterior probability (TT-post) shows very reasonable performance for this data set. Because of the difference in the views the global color histogram method (CDist) performs relatively poorly as does the cumulative likelihood (Cum-likelihood) and a weighted sum squared error measure (SSD), which was performed by scaling the images to the same bounding box.

Figure 5: This figure shows example matching pairs for our vehicle test bed. The images were taken along the same stretch of road by two different cameras less than a minute apart.

4.1 Pedestrian Results

Figure 4 shows the performance for the Translation Template score, the match likelihood, the non-match likelihood, and various other scoring methods. The template based approaches perform extremely poorly because of mis-registration, e.g., comparing the shirt color to the pants color. The color histogram-based methods are confounded by similar colors that are present in many of the objects, including khaki colors and various shaded regions. The color histogram-based methods are also confounded by pedestrians that have very non-descriptive color profiles and match other pedestrians better than the more unique pedestrians match themselves. Methods such as the Earth Movers Distance [9] may be more robust to this effect, but will still suffer due to severe occlusions and pose variation.

A partially occluded pedestrian is just one of many different situations where automatically aligning images from multiple cameras will be difficult. The proposed approach will also work for less complex cases, although the difference in performance relative to more naive approaches will not be as significant.

4.2 Vehicle Results

Figure 5 shows example image pairs from a controlled experiment. These images are from identical cameras from an adjacent stretch of road. Lighting is similar, but the view

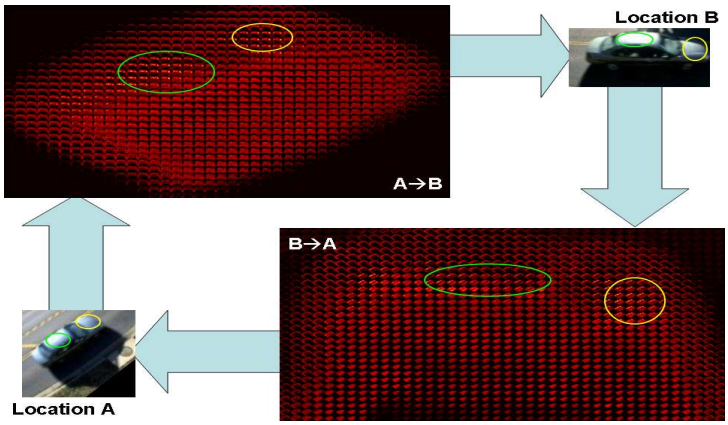

Figure 6: This figure shows a visualization of the posterior likelihood from location *a* to location *b* and vice versa for our vehicle test bed (see Figure 5). The front hood of the vehicles is highlighted in yellow and the roof is highlighted in green. In the corresponding region on the TTs, these regions show selectivity for both the hood and the roof of the other image, though at vastly different angles. The region between these two regions, which corresponds to the front windshield, shows almost no selectivity. The windshield doesn't generally match the windshield of the other car, because it generally reflects the sky or what it beyond the vehicle, which is generally different for the two locations.

angle is significantly different.

Figure 6 shows visualizations of the posterior defined in Equation 7. The similarity in the roof and hood area shows selectivity for the roof, hood, and tail area of the paired vehicle and vice versa. It is interesting to note that the hood and roof regions of vehicles from location *a* have higher selectivity towards the front hood than the roof of vehicles in location *b*. The results from the roof of vehicles in location *b* often reflecting direct sunlight at this time of day. The windshield is not selective towards any corresponding region as is expected from the unpredictible nature of reflections or sky and objects in the scene.

Figure 4 describes the performance for this method. It should be noted that for this data set, easily one in ten vehicles are nearly identical. Thus, though the performance shown in Figure 4 would not be useful to spot an particular vehicle out of hundreds of vehicles driving down a highway, it would be useful as a robust comparison to facilitate disambiguation for multi-camera tracking.

5 Future Work and Conclusions

The most obvious area for improvement is to investigate models that will be invariant to intra-camera color drift and lighting variations. First, the use of different color-spaces (hsv, yuv, color opponency) and different *k*-functions may increase the effectiveness of this approach. Later, simultaneously learning spatial relationships and modeling intra-camera variation, e.g., [3, 8], can be investigated.

Learning multiple models for particular sink and source location pairs may also be

useful to model different types of vehicles. For instance, the TT for pairs of sedan images may be significantly different from the TT for pairs of panel van images. By developing a mechanism to automatically detect the appropriate TT, estimating and using multiple TTs may be possible. We also plan to investigate using the latent region-pair decomposition to enable robust region aggregates and to enable very fast comparisons.

We have introduced Translation Templates, a matching model that represents regularities from one set of images to another set of images. TTs can be automatically estimated given pairs of images taken at two different locations. Using matching and non-matching pairs of images, it is possible to automatically learn to effectively compare images from two locations while exploiting the regularities that exist between these image pairs. It is also possible to visualize and decompose these images to enable better interpretation of TTs and to simplify computation. We show compelling preliminary results for pedestrians in indoor environments and vehicles in outdoor environments.

References

[1] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2005)*, San Diego, CA, 2005. IEEE Computer Society.

[2] Y. Boykov and D. Huttenlocher. A new bayesian framework for object recognition. In *Proc. of the Computer Vision and Pattern Recognition (CVPR1999)*, pages 2517–2523, Fort Collins, CO, 1999. IEEE Computer Society.

[3] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR2005)*, San Diego, CA, 2005.

[4] A. Mittal and L. Davis. M2tracker: A multi-view approach to segmentaing and tracking people in a cluttered scene. *IJCV*, 51:189–203, 2003.

[5] W. J. Rucklidge. Efficient visual recognition using the hausdorff distance. *Lecture Notes in Computer Science*, 1996.

[6] R. Schettini, G. Ciocca, and S. Zuffi. A survey on methods for colour image indexing and retrieval in image databases. In *Color Imaging Science: Exploiting Digital Media*. J. Wiley, 2001.

[7] Y. Shan, S. Sawhney, and R. Kumar. Unsupervised learning of discriminative edge measures for vehicle matching between non-overlapping cameras. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR2005)*, 2005.

[8] C. Stauffer. Learning a probabilistic similarity function for segmentation. In *IEEE Workshop on Perceptual Organization in Computer Vision (POCV2004)*, Washington, DC, 2001.

[9] Y. Rubner C. Tomasi and L. J. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision*, pages 59–66, 1998.

[10] A. Yuille, D. Cohen, and P. Hallinan. Feature extraction from faces using deformable templates. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR1989(*, pages 104–109, New York, 1989.