Flexible Spatial Models for Grouping Local Image Features

Gustavo Carneiro and Allan D. Jepson
Department of Computer Science
University of Toronto ,Toronto, ON, Canada.
{carneiro,jepson}@cs.utoronto.ca

Abstract

A key step for the effective use of local image features (i.e., highly distinctive and robust features) for recognition or image matching is the appropriate grouping of feature matches. Spatial constraints are important in this grouping because, during a recognition process, they allow for the reduction of the number of hypotheses that must be verified and also reduce the number of false positives present in each of these hypotheses. A common choice for this grouping task is to use the Hough transform on the global spatial transformation parameters of the hypothesized matches. Here, instead, we use semi-local spatial constraints which allow for a greater range of shape deformations. A comparison with Hough transform shows that our method is more robust to both rigid and non-rigid deformations. Its functionality is demonstrated in an exemplar-based object recognition system that deals well with severe non-rigid deformations. We also show the efficacy of our flexible spatial grouping for long range motion problems.

1. Introduction

The complexity of the image descriptor (also called indexing primitive) used for image representation in an object recognition system has a great impact on the design of a recognition system (for a thorough discussion, see [7]). Complex global/semi-local image descriptors (e.g., generalized cylinders [3], geons [2], superquadrics [17], among others) reduce the complexity of the model by decreasing the number of descriptors necessary for the representation. This allows for a sparsely populated database of model features, which causes a reduction in the complexity of the search and verification steps. However, these image descriptors are difficult to extract and sensitive to partial occlusion. Alternatively, simple local image descriptors (e.g., 2D points [13]) are easy to extract, robust to rigid deformations and partial occlusion, but sensitive to background clutter and non-rigid deformation. Unfortunately, their low distinctiveness typically results in an overpopulated database of model descriptors due to the large number of descriptors necessary to form a model. Therefore, systems based on simple local descriptors have complex search and verification steps, where the latter step depends strongly on global pose determination.

In this context, there is a recent surge of interest in more complex local descriptors that aim at finding a good balance between detectability, robustness to image deformation, and distinctiveness. The goal is to increase the robustness to background clutter and to reduce the complexity of the search and verification steps. For example, in the literature we find descriptors based on: principal components analysis of image patches [8, 16], Gabor filter responses [12], wavelet coefficients [23], differential invariants [22], local phase features [4], and histograms of local filter responses [14, 20].

Nevertheless, as the size of the database of object models grows, the false detection rates for correspondences between test image features and database features also increases. As a result, pose determination is still a necessary step for the grouping and verification stages in systems based on complex local image descriptors. The use of pose in the grouping stage stems from the fact that the search for similar descriptors in the database of models usually returns a relatively large set of correspondences where the number of inliers tends to be small. The critical point here is certainly the explosion of the number of hypotheses generated due to the large size of the set of possible correspondences. Furthermore, the detection of multiple instances of an object depends on the pose determination (i.e., each different instance will be grouped separately based on its pose). Finally, the verification step also uses pose in order to reduce the number of false positive detections. The overall system therefore relies on both correspondences and spatial structure to accept a hypothesis.

Pose can be represented using global and semi-local models. Global pose determination is based on some underlying transform (e.g., rigid, affine, etc.), where, usually, the positions estimated for the correspondences are relaxed a bit so that the system can accept small deformations from the chosen class of transforms (see [1, 8, 12, 14, 26]). These methods impose a limitation on the type of objects suitable for recognition. Specifically, objects that can suffer a greater range of deformations are not suitable.

An alternative approach to global pose determination is based on semi-local pose determination, which is capable of dealing with a larger range of deformations. Thus, it provides an appropriate framework for both rigid and flexible objects. In [21, 22], the authors use semi-local geometric constraints, but its use is limited to the verification stage. Semi-local constraints are explored in an iterative grouping stage in [24], but the system relies upon global constraints for the final verification.

In this paper, we present new methods for feature grouping and verification based on semi-local spatial constraints. Hence, we *do not* use global constraints in any step of our recognition system. The method involves two components, namely pairwise constraints and geometric predictions. The first component represents pairwise geometric constraints amongst neighboring features. The second component generates predictions of the location, scale, and orientation of each feature, based on these pairwise constraints. This method not only enables the grouping of image descriptors that underwent severe non-rigid deformation, but it also allows for the verification of multiple instances of the same object in an image. A comparison with the Hough transform, which is a classical grouping method based on global spatial coherence, shows that our method provides groups that are considerably more robust to rigid and non-rigid deformation, and typically returns groups with a greater percentage of inliers. An exemplar-based recognition system was developed to demonstrate the efficacy of the semi-local spatial constraints proposed here, and the results show impressive results with respect to extreme non-rigid deformations, in addition to robustness to illumination changes, partial occlusion, and rigid deformation. This approach has applications in other areas, such as long range motion problems, which is also demonstrated below.

2. Semi-local Spatial Constraints

Here we introduce the specific semi-local constraints we use and then, in section 2.2, show how these constraints can be used to make geometric predictions. The pairwise relations are used to form groups of features from the correspondences set, and geometric predictions are used to eliminate remaining outliers from those groups (see section 3), and also to verify the correctness of the hypothesis provided by each group (see section 4).

2.1. Pairwise Relations

Suppose that the local image descriptors are extracted from interest points $\mathcal{I}_i = \{\mathbf{x}_l\}$ detected in an image I_i according to a local image feature method. In particular, each local image descriptor forms a feature vector $\mathbf{f}_l = \mathbf{f}(\mathbf{x}_l) = [m_l, \theta_l, \sigma_l, \mathbf{v}_l]$, where $\mathbf{x}_l$ is the interest point location, m_l is the model identification from which this feature was extracted, θ_l is the main orientation, σ_l is the scale, and the vector $\mathbf{v}_l$ contains the feature values. Here, we use the local image descriptor proposed in [5], where $\mathbf{v}_l = \rho_l e^{i\phi_l}$ is the vector of amplitudes ρ and phases ϕ of bandpass filter responses. The features extracted from a model image I_i are then stored in the database of model features $\mathcal{O}_i = \{\mathbf{f}(\mathbf{x}_l) | \mathbf{x}_l \in \mathcal{I}_i\}$. The similarity between local features is computed using normalized phase correlation [9], as follows:

$$s(\mathbf{f}_l, \mathbf{f}_o) = \frac{|\mathbf{v}_l \cdot \mathbf{v}_o^*|}{1 + \rho_l \cdot \rho_o} \in [0, 1), \tag{1}$$

where $\cdot$ means dot product, and $\mathbf{v}_o^*$ is the complex conjugate of $\mathbf{v}_o$. We wish to know if a subset of $\mathcal{O}_i$ is present in the set of test image features $\mathcal{O}_j = \{\mathbf{f}(\mathbf{x}_l) | \mathbf{x}_l \in \mathcal{I}_j\}$ extracted from image I_j. The set of correspondences is represented by $\mathcal{N}_{ij} = \{(\mathbf{f}_l, \tilde{\mathbf{f}}_l) | \tilde{\mathbf{f}}_l \in \mathcal{O}_j, \mathbf{f}_l \in \mathcal{K}(\tilde{\mathbf{f}}_l, \mathcal{O}_i, k), s(\mathbf{f}_l, \tilde{\mathbf{f}}_l) > \tau_s\}$, where $\mathcal{K}(.)$ is the top k correspondences between a feature $\tilde{\mathbf{f}}_l \in \mathcal{O}_j$ and the database of model features $\mathcal{O}_i$ in terms of phase correlation.

The pairwise geometric relations are computed the same way for both the test image and the model image. They are composed of the following 3 measures between pairs of features from the same image $\mathbf{f}_l, \mathbf{f}_o \in \mathcal{O}_i$ (see Figure 2):

$$\begin{array}{|c|c|}
\hline
\textbf{scale} & \mathcal{S}(\mathbf{f}_l, \mathbf{f}_o) = \frac{(\sigma_l - \sigma_o)}{\sqrt{\sigma_l^2 + \sigma_o^2}} \\
\hline
\textbf{distance} & \mathcal{D}(\mathbf{f}_l, \mathbf{f}_o) = \frac{\|\mathbf{x}_l - \mathbf{x}_o\|}{\sqrt{\sigma_l^2 + \sigma_o^2}} \\
\hline
\textbf{heading} & \mathcal{H}(\mathbf{f}_l, \mathbf{f}_o) = \Delta_\theta \left(\theta_l - \vartheta_{lo}\right) \\
\hline
\end{array} \tag{2}$$

where σ_k is the scale of image feature $\mathbf{f}_k$, $\mathbf{x}_k$ is the image position of $\mathbf{f}_k$, $\Delta_\theta(.) \in [-\pi, +\pi]$ denotes the principal angle, θ_k is the main orientation of feature $\mathbf{f}_k$ for $k = l, o$, and $\vartheta_{lo} = \tan^{-1}(\mathbf{x}_l - \mathbf{x}_o)$. The heading measurement considers the main orientation θ_l of feature vector $\mathbf{f}_l$ relative to the displacement between $\mathbf{x}_l$ and $\mathbf{x}_o$.

We can build the same pairwise relations between $\tilde{\mathbf{f}}_l$ and $\tilde{\mathbf{f}}_o$ such that $(\mathbf{f}_l, \tilde{\mathbf{f}}_l), (\mathbf{f}_o, \tilde{\mathbf{f}}_o) \in \mathcal{N}_{ij}$, thus forming $\mathcal{S}(\tilde{\mathbf{f}}_l, \tilde{\mathbf{f}}_o)$, $\mathcal{D}(\tilde{\mathbf{f}}_l, \tilde{\mathbf{f}}_o)$, and $\mathcal{H}(\tilde{\mathbf{f}}_l, \tilde{\mathbf{f}}_o)$. The pairwise semi-local spatial similarity is then based on

$$\begin{array}{|c|c|}
\hline
\textbf{scale} & \Delta\mathcal{S}_{lo}(\mathcal{N}_{ij}) = \mathcal{S}(\mathbf{f}_l, \mathbf{f}_o) - \mathcal{S}(\tilde{\mathbf{f}}_l, \tilde{\mathbf{f}}_o) \\
\hline
\textbf{distance} & \Delta\mathcal{D}_{lo}(\mathcal{N}_{ij}) = \mathcal{D}(\mathbf{f}_l, \mathbf{f}_o) - \mathcal{D}(\tilde{\mathbf{f}}_l, \tilde{\mathbf{f}}_o) \\
\hline
\textbf{heading} & \Delta\mathcal{H}_{lo}(\mathcal{N}_{ij}) = \mathcal{H}(\mathbf{f}_l, \mathbf{f}_o) - \mathcal{H}(\tilde{\mathbf{f}}_l, \tilde{\mathbf{f}}_o) \\
\hline
\end{array} \tag{3}$$

Given that small values denote high similarities, we can define the weight of the connection between $\tilde{\mathbf{f}}_l$, $\tilde{\mathbf{f}}_o \in \mathcal{O}_j$ in the test image based on the connection of their respective correspondences $\mathbf{f}_l, \mathbf{f}_o \in \mathcal{O}_i$, as follows:

$$\begin{aligned} \mathbf{A}(l, o) = {} & \delta_{m_l m_o} \pi_{lo,g} \\ & g\left([\Delta\mathcal{D}_{lo}(\mathcal{N}_{ij}), \Delta\mathcal{H}_{lo}(\mathcal{N}_{ij}), \Delta\mathcal{S}_{lo}(\mathcal{N}_{ij})]^T ; \Sigma_\Delta \right), \end{aligned} \tag{4}$$

where m_l is the model index of feature $\mathbf{f}_l$ matched to deformed feature $\tilde{\mathbf{f}}_l$ and similarly for m_o, and $\delta_{m_l, m_o} = 1$ if $m_l = m_o$ and 0 otherwise. Also, $\pi_{lo,g} = e^{-0.5 \frac{\mathcal{D}^2(\mathbf{f}_l, \mathbf{f}_o)}{\sigma_{\pi,g}^2}}$ is the pairwise weight, which means that neighboring points to $\mathbf{f}_l$ within a range of roughly $\sigma_{\pi,g}$ pixels in the model have higher weight in the geometric pairwise similarity, where $\sigma_{\pi,g}$ is determined based on the maximum model diameter (in pixels). Finally, $g(.)$ is the unnormalized Gaussian function defined as $g(\mathbf{v};\Sigma) = e^{-\mathbf{v}^T\Sigma^{-1}\mathbf{v}/2}$, where the covariance matrix Σ_Δ is a 3×3 diagonal matrix with distance, scale, and heading variances, namely σ_d^2, σ_h^2, and σ_s^2, respectively, such that σ_h^2, σ_s^2 are pre-defined constants, and $\sigma_d^2 = \min\left(\kappa_{dist}, \max(p_{dist}\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o), 0.1)\right)$ depends on the scaled original distance between features in the model database (i.e., $\mathbf{f}_l, \mathbf{f}_o \in \mathcal{O}_i$), which means that points that are far from each other in the model have a proportionally larger standard error for their relative distances.

2.2. Geometric Predictions

Consider again the set of correspondences $\mathcal{N}_{ij}$ between $\mathcal{O}_i$, and $\mathcal{O}_j$, and that $(\mathbf{f}_l, \tilde{\mathbf{f}}_l), (\mathbf{f}_o, \tilde{\mathbf{f}}_o) \in \mathcal{N}_{ij}$ where $\mathbf{f}_k = \mathbf{f}(\mathbf{x}_k) = [m_k, \theta_k, \sigma_k, \mathbf{v}_k]$, and $\tilde{\mathbf{f}}_k = \tilde{\mathbf{f}}(\tilde{\mathbf{x}}_k) = [\tilde{m}_k, \tilde{\theta}_k, \tilde{\sigma}_k, \tilde{\mathbf{v}}_k]$ with $k = l, o$. The idea is to predict $\tilde{\mathbf{x}}_k$, $\tilde{\theta}_k$, and $\tilde{\sigma}_k$ for each feature $\tilde{\mathbf{f}}_k \in \mathcal{O}_j$ using the information available in the correspondences set and the semi-local spatial constraints for the database of model features. Moreover, points that are close to the feature being predicted should have a higher influence on this prediction than features far from it. In general, note that the following relations are true if the correspondence is correct: $\tilde{\mathbf{n}}_{lo}^T(\tilde{\mathbf{x}}_l - \tilde{\mathbf{x}}_o) \approx \|\mathbf{x}_l - \mathbf{x}_o\|$, where $\tilde{\mathbf{n}}_{lo} = \frac{\tilde{\mathbf{x}}_l - \tilde{\mathbf{x}}_o}{\|\tilde{\mathbf{x}}_l - \tilde{\mathbf{x}}_o\|}$, $\tilde{\theta}_l - \tilde{\vartheta}_{lo} \approx \theta_l - \vartheta_{lo}$, and $\frac{\tilde{\sigma}_l - \tilde{\sigma}_o}{\tilde{\sigma}_o} \approx \frac{\sigma_l - \sigma_o}{\sigma_o}$. For position prediction, we therefore build the linear system $\pi_{lo,p}\tilde{\mathbf{n}}_{lo}^T(\tilde{\mathbf{x}}_l^* - \tilde{\mathbf{x}}_o) = \pi_{lo,p}\|\mathbf{x}_l - \mathbf{x}_o\|$ for all $(\mathbf{f}_o, \tilde{\mathbf{f}}_o) \in \mathcal{N}_{ij} - (\mathbf{f}_l, \tilde{\mathbf{f}}_l)$ and solve it for $\tilde{\mathbf{x}}_l^*$, which is the prediction of feature position $\tilde{\mathbf{x}}_l$. Here, $\mathbf{n}_{lo} = \frac{\mathbf{x}_l - \mathbf{x}_o}{\|\mathbf{x}_l - \mathbf{x}_o\|}$, and $\pi_{lo,p} = e^{-0.5\frac{\mathcal{D}^2(\mathbf{f}_l, \mathbf{f}_o)}{\sigma_{\pi,p}^2}}$, is the pairwise weight, meaning that neighboring points to $\mathbf{f}_l$ within a range of roughly $\sigma_{\pi,p}$ pixels have higher weight in predicting the position of the test feature. We set the value of $\sigma_{\pi,p}$ as a small fraction of the model diameter in pixels. Similarly, the main orientation and scale predictions are defined as $\tilde{\theta}_l^* = \frac{1}{\sum_{o\neq l}\pi_{lo,p}}\sum_{o\neq l}\pi_{lo,p}(\theta_l - \vartheta_{lo} + \tilde{\vartheta}_{lo})$ and $\tilde{\sigma}_l^* = \frac{1}{\sum_{o\neq l}\pi_{lo,p}}\sum_{o\neq l}\pi_{lo,p}\left(\frac{\sigma_l - \sigma_o}{\sigma_o} + 1\right)\tilde{\sigma_o}$, respectively.

The similarity between the predicted and observed position, main orientation and scale is computed as follows (see Fig. 1): $p(\mathbf{f}_l, \tilde{\mathbf{f}}_l) = g([\tilde{\mathbf{x}}_l, \tilde{\theta}_l, \tilde{\sigma}_l] - [\tilde{\mathbf{x}}_l^*, \tilde{\theta}_l^*, \tilde{\sigma}_l^*]; \Sigma_t)$, where $g(.)$ is the Gaussian function, and $\Sigma_t = \text{diag}(\Sigma_{\mathcal{D}}(\tilde{\mathbf{f}}_l), \sigma_{\mathcal{H}}^2(\tilde{\mathbf{f}}_l), \sigma_{\mathcal{S}}^2(\tilde{\mathbf{f}}_l))$. Here, $\Sigma_D(\tilde{\mathbf{f}}_l)$ is an estimate for the spatial variance of the predicted location $\tilde{\mathbf{x}}_l^*$, namely

$$\Sigma_D(\tilde{\mathbf{f}}_l) = B diag(\sigma_{\mathcal{D}}^2(\mathbf{f}_l, \mathbf{f}_o))B^T,$$

where

$$\mathbf{B} = (\mathbf{K}\Pi\mathbf{K}^T)^{-1}\mathbf{K}\Pi, \qquad \mathbf{K} = [\cdots \mathbf{n}_{lo} \cdots],$$

$$\Pi = \begin{bmatrix} \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \pi_{lo,p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots \end{bmatrix}.$$

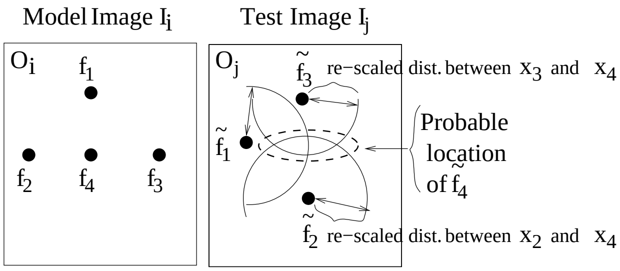
Model Image I_i
Test Image I_j
O_i
f_1
f_2
f_4
f_3
O_j
f̃_3
re−scaled dist. between x_3 and x_4
f̃_1
Probable location of f̃_4
f̃_2
re−scaled dist. between x_2 and x_4

Figure 1. Example of position prediction. Given the features from the model $\{\mathbf{f}_l\}$ and their correspondences $\{\tilde{\mathbf{f}}_l\}$, for $l = \{1, 2, 3\}$, we want to estimate the position of $\tilde{\mathbf{f}}_4$. Its probable location (represented by a dotted ellipsoid) is based on a Gaussian distribution computed using the position of the correspondences in the test and model images and the pairwise variances $\sigma_{\mathcal{D}}^2(\mathbf{f}_l, \mathbf{f}_o)$ estimated in the learning stage.

Also, the variances of the heading and scale estimates are $\sigma_{\mathcal{H}}^2(\tilde{\mathbf{f}}_l) = \left(\frac{1}{\sum_{o\neq l}\pi_{lo,p}}\right)^2\sum_{o\neq l}\pi_{lo,p}^2\sigma_{\mathcal{H}}^2(\mathbf{f}_l, \mathbf{f}_o)$, and $\sigma_{\mathcal{S}}^2(\tilde{\mathbf{f}}_l) = \left(\frac{1}{\sum_{o\neq l}\pi_{lo,p}}\right)^2\sum_{o\neq l}\pi_{lo,p}^2\sigma_{\mathcal{S}}^2(\mathbf{f}_l, \mathbf{f}_o)$. The pairwise variances $\sigma_{\mathcal{D}}^2(\mathbf{f}_l, \mathbf{f}_o)$, $\sigma_{\mathcal{H}}^2(\mathbf{f}_l, \mathbf{f}_o)$, and $\sigma_{\mathcal{S}}^2(\mathbf{f}_l, \mathbf{f}_o)$ are estimated by the sample variances obtained by deforming the model image I_i with the set of deformations $\mathcal{DF}$ defined in [6].

3. Grouping Based on Pairwise Relations

Given a set of test image features, the set of correspondences formed from the search for matching features in the database (e.g., using nearest neighbor) usually generates a large hypothesis space for the recognition system. Typical grouping and verification stages rely on the global spatial configuration of features to constrain this hypothesis space. An example of such a grouping method is RANSAC [25], which estimates the global spatial deformation of features. This is a poor choice for our purposes here due to the extremely low ratio between inliers and outliers in the correspondences set, as also noted in [15]. This issue is rarely addressed in object recognition systems which use complex local features, with the exception of [15], where Lowe selects the generalized Hough transform for the task. The key problem is that the Hough space which is used is a similarity transform space (i.e., a global spatial constraint) with large bin sizes selected to accommodate other spatial deformations. Due to the large bin sizes, Hough clustering for local features usually produces a large number of groups, where each group has a low number of true inliers (especially given a non-rigid deformation). Here, we propose a new grouping approach that is more robust to a broader class of deformations, which aims at reducing the number of groups, where each group has a higher number of in-

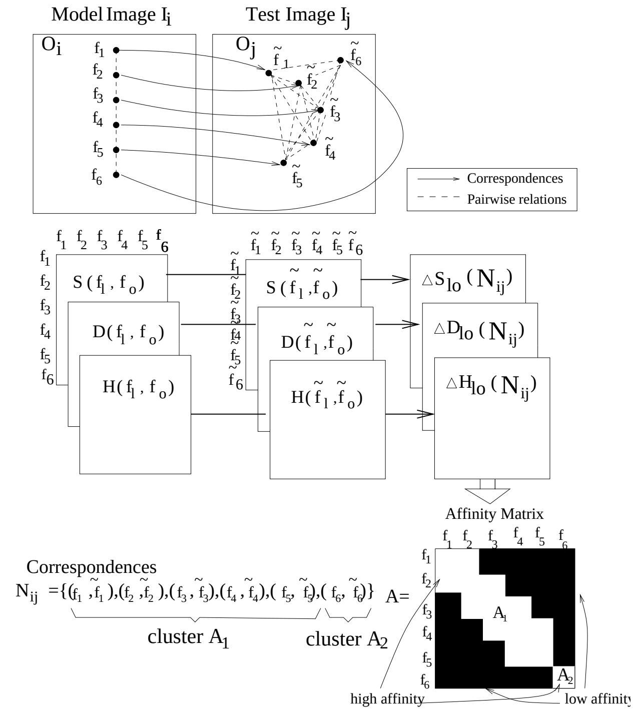

Figure 2. Grouping based on pairwise relations. Notice in the figure that correspondences $1-5$ are semi-locally connected, while correspondence 6 is not. Therefore, we form 2 clusters, A_1 and A_2.

liers. This approach involves connected component analysis on an affinity matrix based on the pairwise relations described in (2). Given the correspondences $\mathcal{N}_{ij}$ between the database of model features $\mathcal{O}_i$ and the set of test image features $\mathcal{O}_j$, we proceed as follows (see Fig. 2):

1. Build the affinity matrix based on the pairwise similarity measures $A(l,o)$ as defined in (4).

2. Perform a Connected Component Analysis (CCA). The strategy here is to select a weak threshold τ_{CCA} and connect every pair of points l and o for which $\mathbf{A}(l,o) \geq \tau_{\text{CCA}}$, thus forming $|G|$ connected clusters represented by the submatrix $\mathbf{A}_g$ (see Fig. 2). We have then the sub-group of correspondences $\mathcal{L}_g(\mathcal{N}_{ij}) \in \mathcal{N}_{ij}$ composed of the features grouped in $\mathbf{A}_g$. Note that a specific cluster of correspondences can only belong to a single model $\mathcal{O}_i$ due the term δ_{m_l,m_o} in (4).

Finally, an intermediate step between the grouping and verification procedures is a deletion of features that are loosely clustered to a group $\mathbf{A}_g$. This is done by checking the geometric predictions computed in section 2.2, and thresholding $p(\mathbf{f}_l, \tilde{\mathbf{f}}_l)$, thus forming the final sets of feature correspondences: $\tilde{\mathcal{L}}_g(\mathcal{N}_{ij}) = \{(\mathbf{f}_l, \tilde{\mathbf{f}}_l) | (\mathbf{f}_l, \tilde{\mathbf{f}}_l) \in \mathcal{N}_{ij}, p(\mathbf{f}_l, \tilde{\mathbf{f}}_l) > \tau_p\}$.

A comparison between our approach and the generalized Hough transform is provided next. Here the feature correspondences between the features of 2 images I_i and I_j are given by the set $\mathcal{N}_{ij}$, where $k = 2$, and $\tau_s = 0.75$ (see first paragraph, sec. 2.1). The parameters for our grouping method are $\sigma_h^2 = 0.2$, $\sigma_s^2 = 0.2$, $\kappa_{dist} = 2$, $p_{dist} = 0.2$, $\tau_{\text{CCA}} = 0.2$, and $\sigma_{\pi,g} = \max(M/5, 10)$, where M is the maximum model diameter. The parameters for the geometric prediction are: $\tau_p = 10^{-16}$, and $\sigma_{\pi,p} = \max(M/50, 5)$.

For Hough clustering, we used the same parameters described in [15], where bin sizes are set as follows: 30^o for rotation, factor of 2 for scale, and 0.25 times the maximum model diameter for translation, and each hypothesis is hashed into the 2 closest bins in each dimension in order to reduce bin boundary effects. For both cases, the minimum number of correspondences to form a group is set at 2% of the total number of features extracted from the model.

The comparisons are presented in Fig. 3, where the model image is presented either on top or left of the image, while the bottom/right image shows the test image. The table titled ‘Pairwise Clustering’ shows the results for our method, and the ‘Hough Transform’ table presents the result for the same image pair using the Hough clustering method. We show the correspondences formed by each grouping method as lines between the model and test images. For all the cases, we only show the group that clustered the highest number of features.

Fig. (3-a) shows the robustness of our method to deformations produced by articulated objects. Note that the Hough transform only matches a piece of the object whose deformation is close to a similarity transformation. Fig. (3-b) shows an example with the articulated model ‘hedvig’ (see Fig. 6). Notice that while the Hough transform can only deal with roughly rigid transform (upper part of the Hedvig’s body), our method is capable of clustering Hedvig’s foot in the same group as the upper part of her body. We also show in Fig. (3-c) the robustness of our method to non-rigid deformation with the model ‘kevin’ (Fig. 6). Here, the Hough transform is unable to correctly cluster the face’s features in the group with the highest number of features.

In order to show the efficacy of our approach with respect to rigid deformation, we considered the long range motion problem using the Wadham and Merton college sequences downloaded from the U. Oxford’s Visual Geometry Group’s web page. In this problem, we considered the groups formed by our approach and Hough transform to compute the **F** matrix [10]. We use RANSAC [25] in order to estimate **F**, and apply the following error measure to calculate the number of inliers: a feature is considered an inlier if its location is within 4 pixels of the epipolar line computed with the **F** matrix.

Fig. 4 illustrates an example of the epipolar lines computed from the image pair Wadham 1 and 5 using both clustering methods. In Fig. 5, we present the proportion of inliers in terms of the set size produced by each grouping

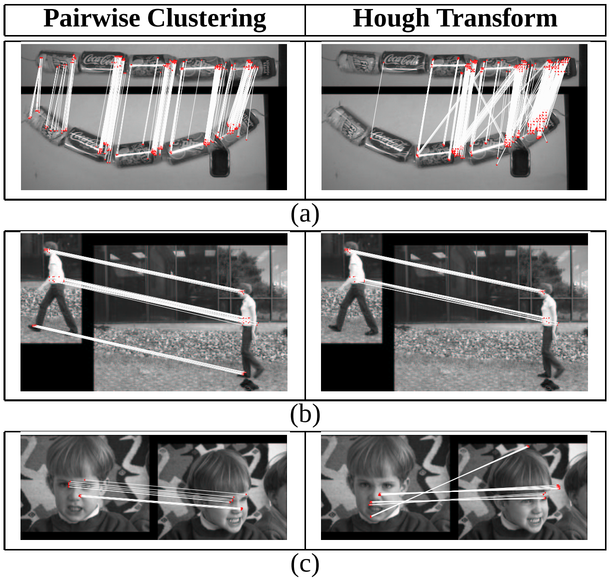

Figure 3. Comparison between our grouping method (left column) and Hough clustering (right column). The lines represent the feature correspondences that were grouped together by each method. (a) Note that while almost all features between the model (top) and the deformed model (bottom) can be clustered in the same group using our method, Hough clustering can only group features that suffered a roughly rigid deformation. (b) Our method is able to cluster the foot features of the model (left) in the same group as the upper body features. Since Hough transform assumes a roughly rigid deformation, it fails to place the foot features in the same group as the upper body features. (c) While our method is capable of clustering the features in the same group, Hough clustering fails.

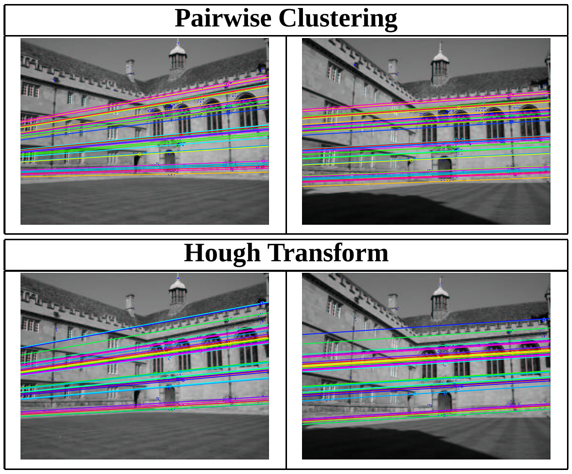

Figure 4. Epipolar lines computed from the algorithm described in [10] using the initial set of correspondences given by each clustering method (i.e., pairwise clustering and Hough).

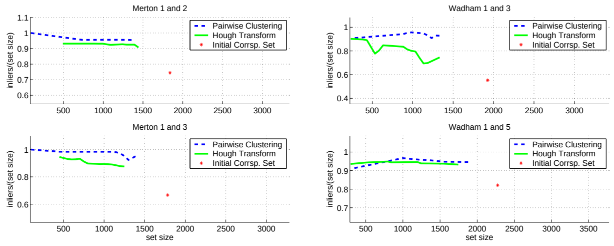

Figure 5. Proportion of inliers from the sets (of varying size) provided by each of the clustering methods.

method. The curves were obtained by varying all the parameters of our grouping method and varying the bin sizes of the Hough transform. Notice that for sets of equal size, the use of pairwise clustering for rejecting outliers generally provides a higher inlier ratio than Hough transform, which indicates a better robustness to rigid deformations.

Finally, it is worth noting that the time complexity of our clustering algorithm is $O(n^2)$, where n is the maximum number of correspondences between features in the test image with features in a single model, and for Hough clustering, the complexity is $O(\#bins)$. For the examples shown above, we had $\#bins \approx n^2$, and both grouping algorithms exhibited comparable running times.

4. Verification

In order to assess the hypothesis that a particular object is present in an image, we propose a verification stage based on a probabilistic framework that uses not only the correspondences in terms of phase correlation, but also the semi-local spatial constraints. The object recognition method can be divided into the training and testing modes. Assume that there is a pool of images $\{I\}_{i\in\{1,\ldots,n\}}$ that is divided into 2 sets, namely the model and random image sets. The model images set is $\{I_i\}_{i\in\{1,\ldots,q\}}$, while the random images set is $\{I_i\}_{i\in\{q+1,\ldots,n\}}$. During the training mode, we take each model image and learn the following feature distributions: a) $P_{\text{on}}(s(\mathbf{f}_l,\mathbf{f}_o);\mathbf{f}_l)$, i.e., the probability of observing phase correlation $s(\mathbf{f}_l,\mathbf{f}_o)$ given that the feature $\mathbf{f}_o$ is a true match for the feature $\mathbf{f}_l$; b) $P_{\text{off}}(s(\mathbf{f}_l,\mathbf{f}_o);\mathbf{f}_l)$, i.e., the probability of observing phase correlation $s(\mathbf{f}_l,\mathbf{f}_o)$ given that the feature $\mathbf{f}_o$ is a false match for the feature $\mathbf{f}_l$; and c) feature position, main orientation, and scale uncertainties. We also learn the feature detectability $P_{\text{det}}(\mathbf{x}_l)$, which is the probability that an interest point is detected in the test image at the same object neighborhood location $\mathbf{x}_l$ of feature $\mathbf{f}_l$. P_{on}, P_{det}, and the uncertainties are learned using a set of image deformations as described in [6], while P_{off} is learned using the random images set.

From the training mode, we build the database of models, namely $\bigcup_{i=1}^{q} \mathcal{O}_i$, where the model features are formed by the filtered set of features $\mathcal{I}_i^*$ (see [6]), for example $\mathcal{O}_i = \{\mathbf{f}_l(\mathbf{x}_l) | \mathbf{x}_l \in \mathcal{I}_i^*\}$. In the testing mode, we take a test image I_j, where $j \notin \{1, 2, ..., n\}$ (i.e., I_j is not in the pool of images used in the learning stage), extract its local features $\mathcal{O}_j = \{\mathbf{f}_l(\mathbf{x}_l) | \mathbf{x}_l \in \mathcal{I}_j\}$, search for similar local features in the database of features, thus forming the set of correspondences $\bigcup_{i=1}^{q} \mathcal{N}_{ji}$. Given the correspondences, we perform the grouping procedure forming the set of clusters $\{\tilde{\mathcal{L}}_g(\mathcal{N}_{jm})\}_{g=1}^{|G|}$. Each cluster is a hypothesis that a particular object is present in the image, so our goal is to determine if any of the clusters $\tilde{\mathcal{L}}_g$ represents an instance of the object $\mathcal{O}_m$. From the computation of the affinity matrix (4), we know that all the features clustered in the same group match features from the same object $\mathcal{O}_m$. We only process groups $\tilde{\mathcal{L}}_g(\mathcal{N}_{jm})$ with a minimum number of correspondences. Let us first define the set of pairings for all model features $\mathbf{f}_o \in \mathcal{O}_m$ from group $\tilde{\mathcal{L}}_g(\mathcal{N}_{jm})$, as $\mathcal{E}_g = \tilde{\mathcal{L}}_g \bigcup \{(\emptyset, \mathbf{f}_o) | \mathbf{f}_o \in \mathcal{O}_m, \neg\exists \mathbf{f}_k \in \mathcal{O}_j \text{ s.t. } (\mathbf{f}_k, \mathbf{f}_o) \in \tilde{\mathcal{L}}_g\}$. Therefore, we want to define the posterior $P(\mathcal{O}_m | \mathcal{E}_g, T)$, where T represents the geometric configuration of features (i.e., their position $\mathbf{x}$, scale σ, and main orientation θ), which can be defined as (using Bayes rule):

$$P(\mathcal{O}_m|\mathcal{E}_g,T) = \frac{P(\mathcal{E}_g|T,\mathcal{O}_m)P(T|\mathcal{O}_m)P(\mathcal{O}_m)}{P(\mathcal{E}_g|T,\mathcal{O}_m)P(T|\mathcal{O}_m)P(\mathcal{O}_m)+P(\mathcal{E}_g|T,\neg\mathcal{O}_m)P(T|\neg\mathcal{O}_m)P(\neg\mathcal{O}_m)}, \tag{5}$$

where $P(\mathcal{O}_m)$ means our prior expectation that a specific model is present, and $P(\neg\mathcal{O}_m) = 1 - P(\mathcal{O}_m)$. Notice that $P(T|\mathcal{O}_m)$ represents the global spatial configuration given $\mathcal{O}_m$, which we treat to be similar to $P(T|\neg\mathcal{O}_m)$ and cancel these terms from (5). The probabilistic formulation, based on [18], is as follows:

1. $P(\mathcal{E}_g|T,\mathcal{O}_m) \approx \prod_{(\mathbf{f},\mathbf{f}_o)\in\mathcal{E}_g} P((\mathbf{f},\mathbf{f}_o)|T,\mathcal{O}_m)$, where we have the following 2 cases:
   (a) $(\emptyset, \mathbf{f}_o) \in \mathcal{E}_g$:
   $$P((\emptyset,\mathbf{f}_o) \in \mathcal{E}_g|T,\mathcal{O}_m) \approx (1 - P_{\text{det}}(\mathbf{x}_o)) + P_{\text{det}}(\mathbf{x}_o)P_{\text{on}}(s < \tau_s; \mathbf{f}_o), \tag{6}$$
   (b) $(\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g, [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]$:
   $$P((\mathbf{f}_k,\mathbf{f}_o) \in \mathcal{E}_g|T,\mathcal{O}_m) = P((\mathbf{f}_k,\mathbf{f}_o) \in \mathcal{E}_g \text{ and } [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]|T,\mathcal{O}_m) \approx P_{\text{det}}(\mathbf{x}_o)P_{\text{on}}(s(\mathbf{f}_k,\mathbf{f}_o);\mathbf{f}_o)p(\mathbf{f}_k,\mathbf{f}_o) \tag{7}$$
   where $[\mathbf{x}_k^*, \theta_k^*, \sigma_k^*]$ is the vector of position, main orientation, and scale predicted for test image feature $\mathbf{f}_k \in \mathcal{O}_j$ given its correspondence $\mathbf{f}_o \in \mathcal{O}_m$ such that $(\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g$.
2. $P(\mathcal{E}_g|T,\neg\mathcal{O}_m) = \prod_o P((\mathbf{f},\mathbf{f}_o)|T,\neg\mathcal{O}_m)$, where we have the following 2 cases:
   (a) $(\emptyset, \mathbf{f}_o) \in \mathcal{E}_g$:
   $$P((\emptyset,\mathbf{f}_o) \in \mathcal{E}_g|T,\neg\mathcal{O}_m) \approx (1 - 0.015) + 0.015(1 - P_{\text{off}}(s(\mathbf{f},\mathbf{f}_o) < \tau_s;\mathbf{f}_o)), \tag{8}$$
   where the number 0.015 represents the average number of interest points per test image divided by the size of the image (see [5]);
   (b) $(\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g, [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]$
   $$P((\mathbf{f}_k,\mathbf{f}_o) \in \mathcal{E}_g|T,\neg\mathcal{O}_m) = P((\mathbf{f}_k,\mathbf{f}_o) \in \mathcal{E}_g \text{ and } [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]|T,\mathcal{O}_m) \approx (0.015)P_{\text{off}}(s(\mathbf{f}_k,\mathbf{f}_o);\mathbf{f}_o)\frac{1}{size(\mathcal{I})}\frac{1}{8}\frac{1}{2\pi}. \tag{9}$$
   In the last term, we assume uniform distribution of position, main orientation, and scale given a background feature.

Finally, we accept a hypothesis if $P(\mathcal{O}_m|\mathcal{E}_g,T) > 0.9$, and the maximum distance between test image features is bigger than a threshold , i.e., assuming $\mathbf{x}_l$ is the position of test image feature $\mathbf{f}_l$ with $(\mathbf{f}_l, \mathbf{f}_p) \in \tilde{\mathcal{L}}_g$, we require $\max_{\forall l,k}\left(\frac{\|\mathbf{x}_l - \mathbf{x}_k\|}{\sqrt{\sigma_l^2+\sigma_k^2}}\right) > \tau_{\mathcal{D}}$ (this is done to avoid a large number of features all in a small area of the image).

5. Results

We considered the problem of exemplar-based recognition using a database of 15 objects shown in Fig. 6, and we use the same parameter values as described in section 3. Also, the prior expectation that a specific model is present $P(\mathcal{O}_m) = 0.0001$, and the maximum distance between test image features must be at least $\tau_{\mathcal{D}} = 0.2M$, where M is the maximum model diameter. Our database has roughly 10,000 features, which were extracted from the objects in Fig. 6 during the learning stage. Our tests (Fig. 7) were conceived to demonstrate the ability of our system to deal with non-rigid/rigid deformations, partial occlusion, and brightness changes. Finally, we also show an experiment on the long range motion problem, where the model ‘fleet’ is being filmed by a hand held camera. Given the image on the top-left corner of Fig. 8, we try to find the model throughout the sequence. In this case, we used the match correspondences to estimate the parameters of the affine transform of the model silhouette [5], but note that these parameters are used only for display and not for verification.

6. Conclusions

We presented novel methods for feature clustering and verification based on semi-local spatial constraints. The use of spatial constraints is necessary to reduce the number of object matching hypotheses to investigate and also to increase the number of inliers in each hypothesis. Although less restrictive than global spatial constraints, semi-local spatial constraints are shown to be adequate for systems

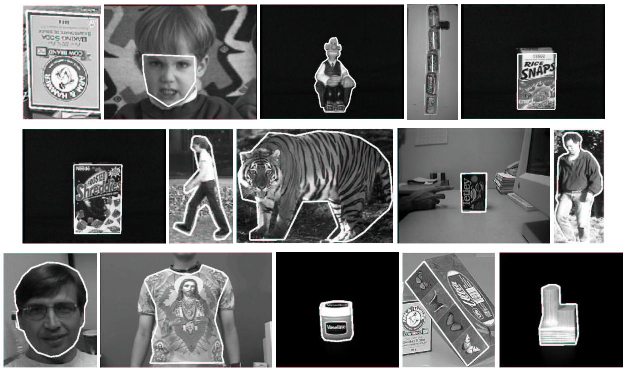

Figure 6. Model database. From top to bottom, left to right: baking soda, kevin, plastic toy [11], snake of cans, rice snaps [11], nestle shreddies [11], hedvig, tiger, tetley tea box, fleet, dudek, torso, vaseline [19], tissue box, and wooden toy [19].

based on complex local features. Moreover, semi-local constraints are able to cope with a broader range of deformations.

The feature clustering method proposed here is shown to be consistently better than the Hough transform when dealing with rigid and non-rigid deformations. The functionality of this method is shown with an exemplar-based recognition and long range motion tasks, which illustrate its robustness in terms of a wide range of image deformations. It is interesting to note that this system might also be adapted to categorization problems given the false positive detected in Fig. (7-e), which will be considered for future research.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, pages 113–130, 2002.

[2] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics Image Processing*, 32:29–73, 1985.

[3] R. Brooks. Model-based 3-d interpretations of 2-d images. *IEEE PAMI*, 5(2):140–150, 1983.

[4] G. Carneiro and A. Jepson. Phase-based local features. In *ECCV*, pages 282–296, Copenhagen, Denmark, May 2002.

[5] G. Carneiro and A. Jepson. Multi-scale phase-based local features. In *IEEE CVPR*, Madison, Wisconsin, USA, June 2003.

[6] G. Carneiro and A. Jepson. Object recognition using flexible groups of local features. Technical report, Department of Computer Science, University of Toronto, January 2004.

[7] S. Dickinson, A. Pentland, and A. Rozenfeld. From volumes to views: An approach to 3-d object recognition. *VGIP: Image Understanding*, 55(2):130–154, 1992.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE CVPR*, 2003.

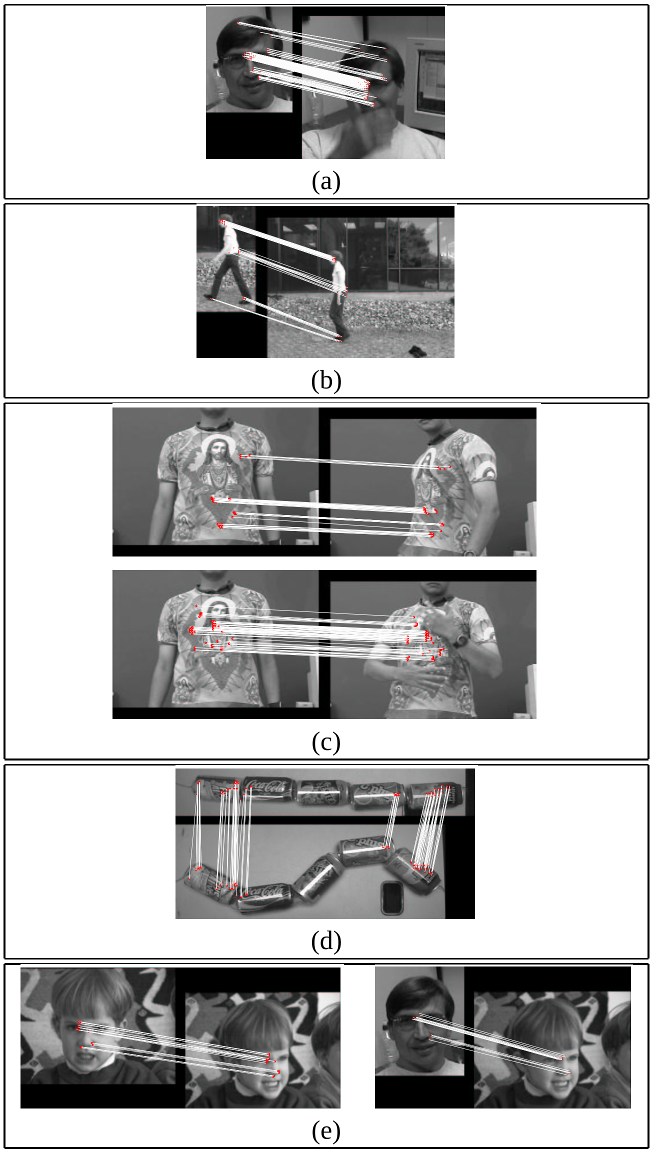

Figure 7. Recognition experiments using database of models in Fig. 6. The white lines represent the correspondences between the model found (left or top) and the test image (right or bottom) used by the verification stage. (a) Model ‘dudek’ matched despite partial occlusion, motion blur, rotation in depth, etc. (b) Articulated object ‘hedvig’ is recognized. (c) Significant non-rigid deformation and partial occlusion (right) of model ’torso’ on the left. (d) Recognition of articulated object ’snake of cans’. (e) Model ‘kevin’ under significant deformation. Note that we also detected a false positive (dudek) that might suggest that this method can be useful for the categorization problem.

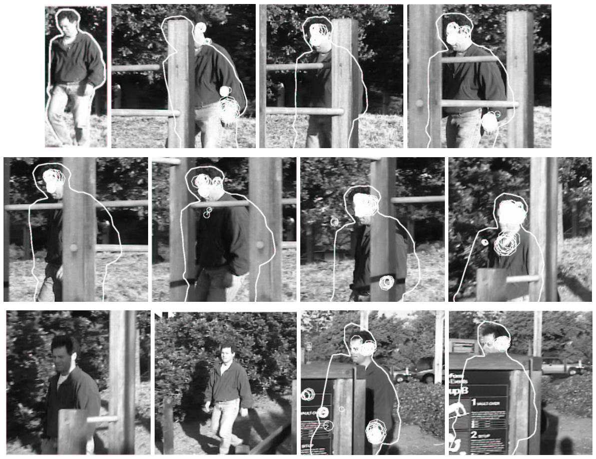

Figure 8. Long range motion problem. The model in the top left figure is searched throughout the sequence using the grouping and verification methods described in this paper. Note that the system shows a good robustness in terms of non-rigid deformations, brightness changes, and partial occlusion. The silhouette shown is computed using the robustly estimated affine parameters of the affine transformation from the model to the test image [5], and the circles in each test image represent the final set of correspondences accepted by our verification step as a match.

[9] D. Fleet. *Measurement of Image Velocity*. Kluwer Academic Publishers, 1992.

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[11] D. Koubaroulis, J. Matas, and J. Kittler. Evaluating colour object recognition algorithms using the soil-47 database. In *ACCV*, 2002.

[12] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C. v.d.Malsburg, R.P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.

[13] Y. Landam and H. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *ICCV*, pages 238–249, 1988.

[14] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, Corfu, Greece, September 1999.

[15] D. Lowe. Local feature view clustering for 3d object recognition. In *IEEE CVPR*, 2001.

[16] K. Ohba and K. Ikeuchi. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *IEEE PAMI*, 19(9):1043–1048, 1997.

[17] A. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331, 1986.

[18] A. Pope and D. Lowe. Probabilistic models of appearance for 3d object recognition. *IJCV*, 40(2):149–167, 2000.

[19] S. K. Nayar S. A. Nene and H. Murase. Columbia object image library (coil-20). Technical report, Department of Computer Science, Columbia University, February 1996.

[20] B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31–50, 2000.

[21] C. Schmid. A structured probabilistic model for recognition. In *IEEE CVPR*, pages 485–490, 1999.

[22] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE PAMI*, 19(5):530–535, 1997.

[23] A. Shokoufandeh, I. Marsic, and S. Dickinson. View-based object recognition using saliency maps. *Image and Vision Computing*, 17:445–460, 1999.

[24] D. Tell and S. Carlsson. Combining appearance and topology for wide baseline matching. In *ECCV*, pages 68–81, Copenhagen, Denmark, 2002.

[25] P. Torr and D. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24(3):271–300, 1997.

[26] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV (1)*, pages 18–32, 2000.