Flexible Spatial Configuration of Local Image Features

Gustavo Carneiro, Allan D. Jepson

Abstract

Local image features have been designed to be informative and repeatable under rigid transformations and illumination deformations. Even though current state-of-the-art local image features present a high degree of repeatability, their local appearance alone usually does not bring enough discriminative power to support a reliable matching, resulting in a relatively high number of mismatches in the correspondence set formed during the data association procedure. As a result, geometric filters, commonly based on *global* spatial configuration, have been used to reduce this number of mismatches. However, this approach presents a trade off between the effectiveness to reject mismatches and the robustness to non-rigid deformations. In this paper, we propose two geometric filters, based on *semi-local* spatial configuration of local features, that are designed to be robust to non-rigid deformations and to rigid transformations, without compromising its efficacy to reject mismatches. We compare our methods to the Hough transform, which is an efficient and effective mismatch rejection step based on global spatial configuration of features. In these comparisons, our methods are shown to be more effective in the task of rejecting mismatches for rigid transformations and non-rigid deformations at comparable time complexity figures. Finally, we demonstrate how to integrate these methods in a probabilistic recognition system such that the final verification step uses not only the similarity between features, but also their semi-local configuration.

Index Terms

Local image feature, Feature clustering, Visual object recognition, Wide baseline matching, Long range matching.

I. Introduction

The field of computer vision has experienced an increasing interest in the use of local image features for the tasks of object recognition [25], image matching [33], object discovery and recognition [38], etc. When compared to image representations based on a large spatial support [29], local feature representations (based on a small spatial support) trade a poorer distinctiveness for a better robustness to brightness deformations and rigid transformations. Therefore, the search for similar features between the local features extracted from a test image and the features in the model database typically returns a correspondence set with high percentage of mismatches. The rejection of mismatches from this correspondence set is therefore one of the central issues in local feature based methods for recognition.

The rejection of mismatches is typically based on the spatial configuration of the model features. The global spatial configuration (e.g., [16], [23], [25], [32], [39], [43]) assumes that all model features suffered a rigid transformation. Usually, the more strict this assumption of global transform is, the more effective the method is to reject mismatches. As this assumption is relaxed, the method becomes more robust to non-rigid deformations, but allows more mismatches in the correspondence set. A more flexible scheme was introduced by Berg et al. [5], which alleviates this problem by allowing some flexibility to the initial rigid model through the use of thin plate splines, but the trade off mentioned above is still present. A method specifically designed to be robust to non-rigid deformations was presented by Ferrari et al. [17], where the authors propose an algorithm consisting of several steps of expansion and contraction of the correspondence set that slowly rejects mismatches and increases the number of correct correspondences. The main issue with the latter method is the high computational complexity of the whole algorithm. A method for real-time tracking of non-rigid surfaces is proposed by Pilet et al. [30], where the method is based on deformable 2-D meshes and the use of robust estimators. This systems produces impressive non-rigid matching results at relatively high frame rates (10 frames per second), but the main problem with the method is the difficulty in matching highly deformable objects because of issues involved in the minimization of the surface energy term. Here, we propose two efficient methods to reject mismatches that are designed to be robust to non-rigid deformations, but for which the rejection of mismatches from the correspondence set is less affected. Specifically, the following methods are considered: a) introduction of an intermediate grouping step using pairwise geometric relations [9], and b) improvement of the distinctiveness of the local feature using semi-local geometric information [10]. We also propose a novel probabilistic verification method based on feature similarity and semi-local geometric relations. This verification method can be combined with either mismatch rejection methods (a) or (b) above to increase the proportion of correct matches in the correspondence set and also to verify the correctness of the semi-local geometric configuration of the features.

We present a comparison between both mismatch rejection methods and Hough clustering, which is a common method to reject mismatches based on global spatial configuration. The results show that both methods lead to correspondence sets with a higher proportion of correct matches than Hough clustering for both rigid transformations and non-rigid deformations. We also show that our methods present a comparable time complexity when compared to Hough clustering for typical image matching tasks. The probabilistic verification method that uses semi-local geometric relations is shown to increase the ratio of correct matches in the final correspondence set without increasing the total time complexity for rejecting mismatches. Finally, we show how these methods can be combined in a recognition system, where we show results on wide baseline stereo and long range matching problems.

II. Literature Review

Systems that exploit pairwise relations to reject mismatches can be traced back to [3], [7], [24]. In [27], [45] pairwise relations were used to disambiguate matches, but both papers rely on a verification stage that is based on a global transformation, which is not suitable to handle non-rigid deformations. Yu et al. [44] exploit pairwise relations of parts, but compared to our approach, their method can work with five to ten parts only, while our method can handle hundreds of parts. The use of graphs is exploited by Dickinson et al. [14], [36], [37], where objects are represented as a hierarchical graph and the matching process takes into account the graph structure, the (semi-)local features, and their global spatial arrangements. Huet and Hancock [22] introduce an approach where the features are based only on pairwise geometric relations between lines in a structural representation of objects, and impressive recognition results are obtained, showing that pairwise relations alone can represent a powerful indexing feature. Even though the use of pairwise relations are generally associated with mismatch rejection methods, they can also be exploited in the verification stage, as implemented by Agarwal and Roth [1].

The use of semi-local information to enhance the discriminating power of local features has also been exploited in the literature. The most relevant work for our approach was presented by Schmid and Mohr [34], [35], where a fixed number of local features around a given feature is used to determine its semi-local structure. Also, a similar method to ours has been recently proposed by Mortensen et al. [28]. A slightly different approach to eliminate mismatches is proposed by Schaffalitzky and Zisserman [33], where a neighborhood consensus, formed by a fixed number of features, is imposed to reject mismatches. Semi-local constraints are also used by Tuytelaars and Van Gool [42] where an iterative method rejects mismatches based on homographies between matches of semi-local features. Tell and Carlsson [40] introduce a semi-local feature formed by a group of ordered local features which improves the discriminating power of the feature, but even though an optimal algorithm is used to avoid all possible combinations of neighboring local features, the method is still prohibitively complex. Parts and union of parts are exploited by Huang et al. [12], where parts are described as polynomial surfaces. This approach represents both semi-local and global features since the union of parts can represent the whole object, but articulated objects are not handled properly given that the relations are assumed to be fixed between parts. The use of pairwise relations to form a feature vector is also successfully used by Belongie et al. [4], where the authors propose the semi-local feature *shape context*. The method proposed by Amit and Geman [2] learns groups of a fixed number of local features (thus forming semi-local features) for recognition. Finally, Chum et al. [13] show that the use of three point correspondences (or regions) within a RANSAC loop to estimate the F matrix speeds up the estimation of the epipolar geometry and allows for a higher robustness to mismatches.

The novelty of our approaches lies in the use of semi-local configuration of features for rejecting mismatches *and* verifying hypotheses, which means that we *never* rely on the global configuration of local features. Both mismatch rejection methods proposed here build the semi-local configuration using *all* of the image features (as opposed to a fixed number of features) in a *tunable neighborhood* (the size of this neighborhood is a user-defined parameter). Moreover, the feature and semi-local similarity functions are combined in the verification step using probabilistic measures, thus avoiding the hard task of determining a reasonable similarity function involving these rather distinct similarity functions. Also, our methods are capable of handling correspondence sets containing thousands of pairings efficiently. Finally, similarly to [45], our approach weights the importance of a semi-local geometrical correspondence by its scale-invariant pairwise distance, meaning that nearby features are more likely to preserve such similarities than far away features.

III. Local Image Features

A local feature is represented by a geometric characterization of an image region plus a descriptor of the image function (photometry) of this region. More specifically, a local feature vector is described as $\mathbf{f}_l = [m_l, \mathbf{x}_l, \theta_l, \sigma_l, \mathbf{v}_l]$, where m_l is the model identification, $\mathbf{x}_l$ is the spatial position of the feature, θ_l represents the dominant orientation at position $\mathbf{x}_l$, σ_l denotes the feature scale, and $\mathbf{v}_l$ is the vector with the photometric values. The database of model features extracted from a model image I_m is then denoted as $\mathcal{O}_m = \mathcal{O}(I_m, \Lambda_o) = \{\mathbf{f}_l | \mathbf{x}_l \in \mathcal{I}(I_m, \Lambda_o)\}$, where $\Lambda_o = \left(2^{\frac{k}{4}}\right)$, with $k = 0, ..., 12$ representing the set of scales at which the image I_m was processed, and the set of interest points $\mathcal{I}(I_m, \Lambda_o)$ is defined as the set of positions in image I_m selected at each scale in Λ_o as interest points. Specifically, in this work we study the local phase feature [8] and the SIFT feature [25]. The local phase feature is computed from the responses of the second derivative of a Gaussian and its Hilbert transform [19], which form a local complex representation that can be denoted by amplitude and phase. The interest points for the local phase feature are based on the multi-scale Harris corner points [20], where the points presenting phase singularities [18] are filtered out [8]. The SIFT features [25] are computed using histograms of gradient values at several scales, and the interest points are locations at maxima and minima of a difference of Gaussian (DOG) function applied in scale space. Note that other types of local image features containing appearance and geometric information could also have been used in this work.

A. Correspondence Set

A correspondence set represents a data association between the set of model features $\mathcal{O}_m$ and a set of features $\mathcal{O}_t$ extracted from test image I_t. This set is denoted by

$$\mathcal{N}_{mt} = \{(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) | \tilde{\mathbf{f}}_\mathrm{l} \in \mathcal{O}_t, \mathbf{f}_l \in \mathcal{K}(\tilde{\mathbf{f}}_\mathrm{l}, \mathcal{O}_m, \kappa_{\mathcal{N}}), s_f(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) > \tau_s\}, \tag{1}$$

where the similarity function $s_f(.) \in [0, 1]$ represents the similarity between two features ($s_f(.) \approx 1$ means high similarity), and $\mathcal{K}(.)$ is the set of the top $\kappa_{\mathcal{N}}$ correspondences between test image feature $\tilde{\mathbf{f}}_\mathrm{l} \in \mathcal{O}_t$ and the database of model features $\mathcal{O}_m$ in terms of the similarity function.

IV. Methods to Reject Mismatches

In this section we present our methods to reject mismatches from a correspondence set, where the key idea exploited is the use of semi-local constraints. In Sec. IV-A we describe the grouping method based on pairwise relations between local image features and in Sec. IV-B we introduce our semi-local image feature.

A. Grouping Based on Pairwise Relations

One way of rejecting mismatches from the correspondence set is through a grouping stage. Typical grouping approaches for local features (e.g., Hough transform [25], or RANSAC [41]) rely on the global spatial configuration of features. Generally, these methods have become popular due to their efficiency and reasonably good performance for rejecting mismatches. However, a common property present in these approaches is the trade-off between the efficacy to reject mismatches and robustness to large deviations from the chosen class of transformations. Since the class of transformations is usually globally rigid (e.g., similarity, affine), any type of non-rigid deformation would cause these methods to reject correct matches, and to break large sets of appropriate matches up into several small-sized groups.

We propose a new grouping approach that aims at fixing these problems with a time complexity comparable to the methods based on global spatial configuration. Specifically, our grouping algorithm is designed to be robust to a broader class of deformations, which aims at reducing the number of groups, where each group has a higher percentage of correct matches and a higher number of correspondences. Our approach involves connected component analysis on an affinity matrix based on the pairwise relations.

1) Pairwise Relations: The pairwise geometric relations are composed of the following three measures between pairs of model features $\mathbf{f}_l, \mathbf{f}_o \in \mathcal{O}_m$:

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

(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 test image features $\tilde{\mathbf{f}}_{\mathrm{l}}, \tilde{\mathbf{f}}_{\mathrm{o}} \in \mathcal{O}_t$ such that $(\mathbf{f}_l, \tilde{\mathbf{f}}_{\mathrm{l}}), (\mathbf{f}_o, \tilde{\mathbf{f}}_{\mathrm{o}}) \in \mathcal{N}_{mt}(\ 1)$, thus forming $\mathcal{S}(\tilde{\mathbf{f}}_{\mathrm{l}}, \tilde{\mathbf{f}}_{\mathrm{o}})$, $\mathcal{D}(\tilde{\mathbf{f}}_{\mathrm{l}}, \tilde{\mathbf{f}}_{\mathrm{o}})$, and $\mathcal{H}(\tilde{\mathbf{f}}_{\mathrm{l}}, \tilde{\mathbf{f}}_{\mathrm{o}})$. The pairwise semi-local spatial similarity is then based on

scale	$\Delta\mathcal{S}_{lo}(\mathcal{N}_{mt}) = \mathcal{S}(\mathbf{f}_l, \mathbf{f}_o) - \mathcal{S}(\tilde{\mathbf{f}}_{\mathrm{l}}, \tilde{\mathbf{f}}_{\mathrm{o}})$
distance	$\Delta\mathcal{D}_{lo}(\mathcal{N}_{mt}) = \mathcal{D}(\mathbf{f}_l, \mathbf{f}_o) - \mathcal{D}(\tilde{\mathbf{f}}_{\mathrm{l}}, \tilde{\mathbf{f}}_{\mathrm{o}})$
heading	$\Delta\mathcal{H}_{lo}(\mathcal{N}_{mt}) = \mathcal{H}(\mathbf{f}_l, \mathbf{f}_o) - \mathcal{H}(\tilde{\mathbf{f}}_{\mathrm{l}}, \tilde{\mathbf{f}}_{\mathrm{o}})$

(3)

We define the similarity weight of the connection between $\tilde{\mathbf{f}}_{\mathrm{l}}$, $\tilde{\mathbf{f}}_{\mathrm{o}} \in \mathcal{O}_t$ in the test image based on the connection of their respective correspondences in the model $\mathbf{f}_l, \mathbf{f}_o \in \mathcal{O}_m$, as follows:

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

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, the pairwise weight $\pi_{lo,g}$ is defined as $\pi_{lo,g} = e^{-0.5\frac{\mathcal{D}^2(\mathbf{f}_l,\mathbf{f}_o)}{\sigma^2_{\pi,g}}}$ where $\sigma_{\pi,g} = \frac{D_M}{L_{\text{pair}}}$, with L_{pair} being a tuning variable, and D_M the maximum model diameter in pixels. Finally, $g(.)$ is the zero-mean 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 model features $\mathbf{f}_l, \mathbf{f}_o \in \mathcal{O}_m$ (i.e., points that are far from each other in the model have a proportionally larger standard error for their relative distances).

2) Grouping Algorithm: Given the correspondences $\mathcal{N}_{mt}$ (1) between the database of model features $\mathcal{O}_m$ and the set of test image features $\mathcal{O}_t$, we proceed as follows:

1) Build the affinity matrix based on the pairwise similarity measures $A(l,o)$ (see Eq. 4 and Step 1 in Fig. 1).
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 sub-matrix $\mathbf{A}_g$. We have then the sub-group of correspondences $\mathcal{L}_g(\mathcal{N}_{mt}) \subseteq \mathcal{N}_{mt}$ composed of the features grouped in $\mathbf{A}_g$. Note that a specific cluster of correspondences can only belong to a single model $\mathcal{O}_m$ due to the term δ_{m_l,m_o} in Eq. 4 (see Step 2 in Fig. 1).

The complexity of this grouping algorithm is $O(|\mathcal{N}_{mt}|^2)$, where $|\mathcal{N}_{mt}|$ denotes the size of the correspondence set. Thus a good strategy to keep the complexity of this algorithm manageable is to set τ_s at a high value and $\kappa_{\mathcal{N}}$ at a low value in (1), so that $|\mathcal{N}_{mt}|$ is reasonably small.

B. Semi-local Image Features

An intuitive method to improve the disambiguating power of local features is to group them in some pre-defined manner, and use these groups as indexes to the model database [14]. Although several cues for clustering visual features have been proposed in [6], [24], we only exploit local feature proximity in this work. More specifically, we propose a method to verify the correctness of a given correspondence using a variation of the shape context descriptor [4].

1) Variation of Shape Context: The shape context feature proposed in [4] is based on a log-polar space histogram as shown in Fig. 2. Although shown to be useful in some recognition

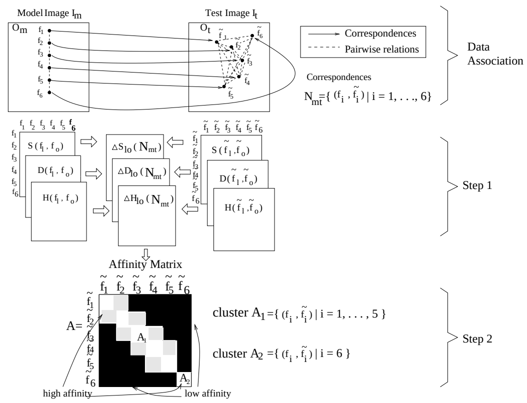


Fig. 1. Grouping based on pairwise relations. The data association consists of matching model and test features based solely on the similarity of their feature values (see Sec. III-A). Step 1 builds the pairwise similarity matrix as described in IV-A.1, and Step 2 comprises the clustering algorithm based on connected component analysis as defined in IV-A.2. Notice in the figure that correspondences $1-5$ are semi-locally connected, while correspondence 6 is not. Therefore, two clusters are formed.

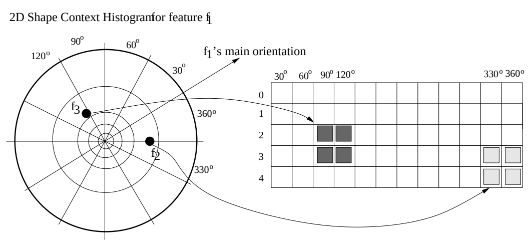


Fig. 2. Shape context of local feature $\mathbf{f}_1$. As in [4], we also use 5 bins for log(distance) and 12 bins for relative orientation. Note that we modify the original shape context method as explained in Sec. IV-B.1.

tasks, this image feature presents a few weaknesses in terms of robustness that needed to be addressed in order to improve the discriminating power of typical local features. Assuming that we are augmenting the feature $\mathbf{f}_l$, and that $\mathbf{f}_o$ is a neighboring feature, the modifications made to the original shape context are:

1) the robustness to non-rigid deformations is improved by weighting a vote in a specific histogram bin by

$$w(\mathbf{f}_l, \mathbf{f}_o) = e^{\frac{-0.5\mathcal{D}^2(\mathbf{f}_l,\mathbf{f}_o)}{L^2}}, \tag{5}$$

where $\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o)$ is defined in (2), and $L = \frac{D_M}{L_{\text{sc}}}$, with L_{sc} being a tuning variable, and D_M the maximum model diameter in pixels (in Fig. 2, darker cells in the histogram represent higher weight);

2) in order to reduce boundary effects in the histogram, each neighboring feature votes for the two closest bins in each dimension (see in Fig. 2 that each vote spans four bins);
3) we make the shape context robust to rotation changes by rotating the histogram axis according to the main orientation of the feature; and
4) the distance measures are scaled as in (2) in order to make them robust to scale changes.

The shape context similarity is computed using the $\mathcal{X}^2(h(\mathbf{f}_l), h(\tilde{\mathbf{f}}_\mathrm{l}))$ test statistic defined in [4], as follows

$$s_h(h(\mathbf{f}_l), h(\tilde{\mathbf{f}}_\mathrm{l})) = 1 - \mathcal{X}^2(h(\mathbf{f}_l), h(\tilde{\mathbf{f}}_\mathrm{l})) = 1 - \frac{1}{2}\sum_{k=1}^{K} \frac{[h_k(\mathbf{f}_l) - h_k(\tilde{\mathbf{f}}_\mathrm{l})]^2}{h_k(\mathbf{f}_l) + h_k(\tilde{\mathbf{f}}_\mathrm{l})} \in [0,1], \tag{6}$$

where $h(\mathbf{f}_l)$ and $h(\tilde{\mathbf{f}}_\mathrm{l})$ are the $K - bin$ normalized histograms of features $\mathbf{f}_l$ and $\tilde{\mathbf{f}}_\mathrm{l}$, respectively. Therefore, given an initial set of correspondences $\mathcal{N}_{mt}$ (1) built using a feature similarity function $s_f(.)$, we select the features belonging to a common model that also have shape context similarity above some value τ_c. This forms G groups $\mathcal{L}_g(\mathcal{N}_{mt}) = \{(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) | (\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) \in \mathcal{N}_{mt}, s_h(h(\mathbf{f}_l), h(\tilde{\mathbf{f}}_\mathrm{l})) > \tau_c\}$, where $\forall (\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}), (\mathbf{f}_o, \tilde{\mathbf{f}}_\mathrm{o}) \in \mathcal{L}_g(\mathcal{N}_{mt})$, $m_l = m_o$ (i.e., feature correspondences belonging to the same group $\mathcal{L}_g(\mathcal{N}_{mt})$ must belong to the same model). Hence, $\bigcup_{g=1}^{G} \mathcal{L}_g(\mathcal{N}_{mt}) \subseteq \mathcal{N}_{mt}$. Note that this system is able to detect only one instance per model stored in the database, so the maximum number of groups formed equals the number of models stored in the database.

The performance improvement of this new semi-local feature is assessed using the quantitative evaluation described in the Appendix I. For these comparisons, we use the local phase features where the similarity function is denoted by (see [8] for details):

$$s_f(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) = \frac{|\mathbf{v}_l \cdot \tilde{\mathbf{v}}_\mathrm{l}^*|}{1 + |\mathbf{v}_l||\tilde{\mathbf{v}}_\mathrm{l}|}, \tag{7}$$

where $\mathbf{v}_k$ is a complex-valued vector, $\mathbf{v}_k^*$ represents its complex conjugate for $k = l, \mathrm{l}$, and $\cdot$ denotes dot product. We also use SIFT [25], where the similarity function is $s_f(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) = \frac{1}{\|\mathbf{v}_l - \tilde{\mathbf{v}}_\mathrm{l}\|}$. Finally, $L_{\mathrm{sc}} = 100$ in (5).

We generate the ROC curves by varying the feature similarity threshold τ_s and then evaluating true positive (TP) and false positive (FP) using the threshold values $\tau_c \in \{0, .65, .75, .8, .9\}$ for the shape context similarity function such that $s_h(h(\mathbf{f}_l), h(\tilde{\mathbf{f}}_\mathrm{l})) > \tau_c$ (see Eq. 6). Notice that when $\tau_c = 0$, we are not using the shape context.

Fig. 3 shows the TP rates for a FP rate of 0.1% for the image deformations $d \in \mathcal{DF}$ described in the Appendix II. Note that the size of the error bars in the graphs is large due to a combination

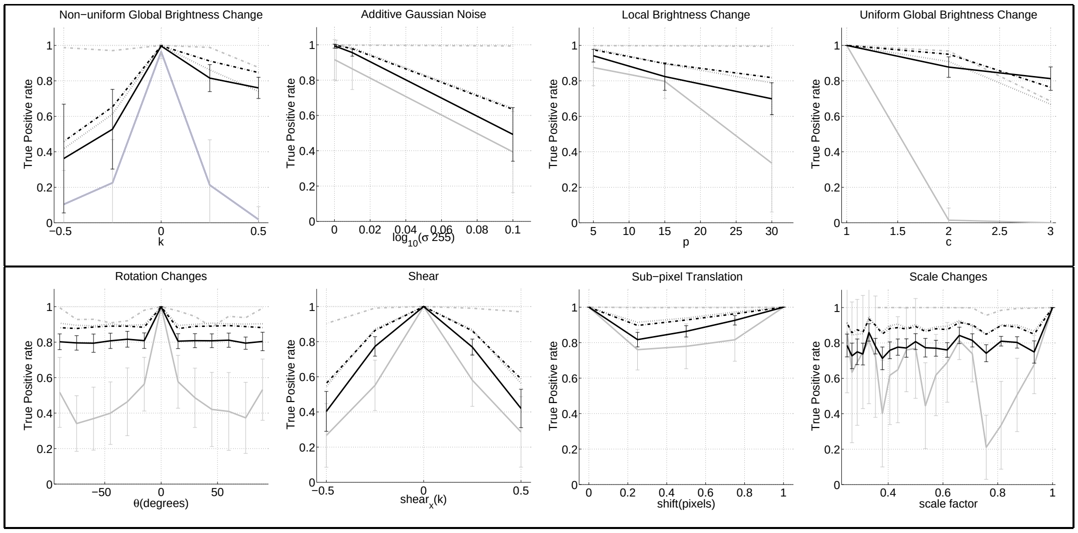


Fig. 3. The TP rate curves in terms of the image deformations $d \in \mathcal{DF}$ are obtained by holding the FP rate at 0.1% in the ROC curves generated by the evaluation experiment in Appendix I. Black curves are the phase local feature [8] without shape context (solid), with shape context such that $\tau_c = 0.65$ (dashed), and $\tau_c = 0.8$ (dotted). Gray curve shows the performance of SIFT [25] without shape context (solid), with shape context such that $\tau_c = 0.65$ (dashed), and $\tau_c = 0.8$ (dotted). Note that the error bars are omitted for the dashed and dotted curves for clarity, but are roughly the same size as the ones we show.

of two things: a) small number of descriptors present in some of the test images (especially for the SIFT descriptor), and b) large number of cases where the TP rate is zero for an FP rate = 0.1% . The correct matches and mismatches that are rejected from the correspondence set as τ_c increases (with FP=0.1%) are shown in Fig. 4. The correct match rejection is computed as $\frac{N_{in}(0)-N_{in}(\tau_c)}{N_{in}(0)}$, where $N_{in}(\tau_c)$ is the number of correct matches (see computation of TP rate above) for a given τ_c, while the mismatches rejection is calculated as $\frac{(N_{tot}(0)-N_{in}(0))-(N_{tot}(\tau_c)-N_{in}(\tau_c))}{(N_{tot}(0)-N_{in}(0))}$, where $N_{tot}(\tau_c)$ is the total number of features in the correspondence set for a given τ_c. From these curves it is clear that the use of shape context rejects many mismatches while keeping most of the correct matches in the correspondence set. It is interesting to notice in Figures 3 and 4 that the use of shape context is more effective to remove mismatches in correspondences set of SIFT features than in sets of local phase features. A possible reason for this is the combination of a relatively smaller number of SIFT features detected in an image and the robustness of the interest point detector DOG to the image deformations studied. It can also be seen in Fig. 3 that the local phase feature alone performs better than SIFT. This happens not only because the local phase information is robust to geometric transformations and brightness variations [18],

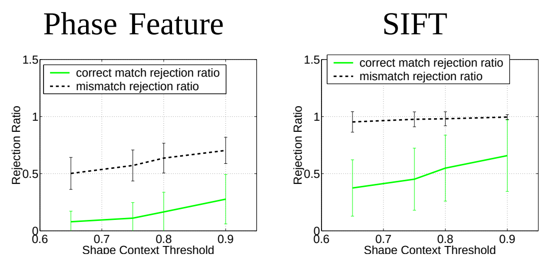


Fig. 4. Correct match and mismatch rejection ratios for our local phase feature and SIFT [25] using shape context to reject mismatches.

but also because the relatively higher number of local phase descriptors per image (compared to the number of SIFT descriptors) increases the chances of a successful match in the deformed test image.

Similarly to the pairwise clustering, the time complexity to build the semi-local feature is $O(|\mathcal{N}_{mt}|^2)$, where $|\mathcal{N}_{mt}|$ denotes the size of the correspondence set. Again, a good strategy to keep the complexity of this algorithm manageable is to set τ_s at a relatively high value and $\kappa_{\mathcal{N}}$ at a low value in (1), so that $|\mathcal{N}_{mt}|$ is reasonably low.

C. *Performance Evaluation*

A comparison between our mismatch rejection methods described above and the generalized Hough transform is provided next. The reason for comparing our methods against the Hough transform resides in its attractive properties, which include: a) low time complexity, b) reasonably high accuracy, and c) wide availability and acceptance. We intend to show that our methods prune the initial correspondence set more accurately than the Hough transform, generating groups with a higher rate of correct matches not only in terms of non-rigid, but also rigid transformations. We also illustrate that the efficiency of our method is comparable to the one presented by the Hough transform for typical matching problems.

In the experiments below, we used the phase-based local feature for the model representation with the feature similarity defined by (7). For the pairwise clustering scheme, we assumed the following values for the constants in (4): the standard deviation of heading, scale, and distance are respectively $\sigma_h^2 = 0.2$, $\sigma_s^2 = 0.2$, $\sigma_d^2 = \min(\kappa_{\text{dist}}, \max(p_{\text{dist}}\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o), 0.1))$ with $\kappa_{\text{dist}} = 2$ and $p_{\text{dist}} = 0.2$, and $L_{\text{pair}} = 5$ for the computation of pairwise weight (4). In order to generate the graphs below, we vary the parameter $\tau_{\text{CCA}} = k/10$ for $k = \{1, 2, \cdots, 9\}$, which is the threshold for the connected component analysis algorithm described in Sec. IV-A. For the semi-local feature, the parameter $L_{\text{sc}} = 100$ is used in the computation of (5). We vary the threshold for shape context similarity between corresponding features in order to generate the graphs using $\tau_c = k/10$ for $k = \{4, 4.5, 5, ..., 9\}$. Therefore, this mismatch rejection method discards any correspondence $(\mathbf{f}_l, \tilde{\mathbf{f}}_{\text{l}}) \in \mathcal{N}_t$ that $s_h(h(\mathbf{f}_l), h(\tilde{\mathbf{f}}_{\text{l}})) < \tau_c$ (see Eq. 6).

The Hough clustering algorithm builds a transform space (e.g., similarity, affine) and using each element of $\mathcal{N}_{mt}$ in (1) as a point in this space, it finds groups of points that move coherently according to the transformation being modeled. For the experiments in this section, we use a space of similarity transform in the Hough clustering algorithm with the following bin sizes for translation: $\{0.3, 0.15, 0.05\}$ times D_M (i.e., the maximum model diameter). For rotation, the bin sizes studied are $\{30^o, 15^o, 5^o\}$. The bin sizes above are varied in order to produce the results for the experiments in the next section. We did not vary the scale bin sizes since the examples considered do not present much variability in terms of scale. Instead, the histogram for scale changes has the following fixed bin values: $[0.125, 0.25, 0.5, 1, 2, 4, 8, 16]$. Finally, each hypothesis is hashed into the two closest bins in each dimension in order to reduce boundary effects. Also, in order to avoid a high number of groups we run a non-maximum suppression when searching for local maxima in this space. Note that the complexity of Hough transform is simply the number of bins in this transformation space.

1) Rigid Transformation: In order to show the effectiveness of our approaches with respect to rigid transformation, we consider the wide baseline matching problem. Using the set provided by each mismatch rejection method, we compute the $\mathbf{F}$ matrix as presented in [21] using RANSAC [41]. We are interested in computing the proportion of inliers given the size of this set. An inlier is considered to be a feature that lies within four pixels (approximately the spatial resolution of the local features used) of the epipolar lines computed from the $\mathbf{F}$ matrix. For this experiment, we used two sequences available from Oxford's Visual Geometry Group's webpage, namely Wadham and Merton College sequences (see Figures 19 and 20). In Fig. 5, we present the graphs of each matching. Note that the proportion of inliers for correspondences set of the same size is, for the cases studied, always higher for our methods than for Hough. These results show that for correspondence sets containing on the order of $1,000$ matches there are around 90% to 95% of inliers. This means that point prediction estimates might be affected by the remaining 5% to 10%

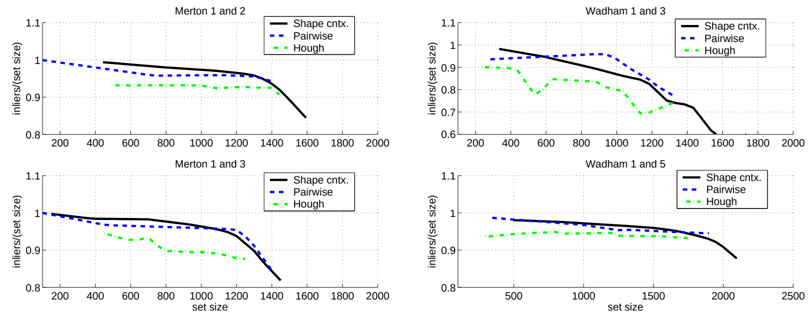


Fig. 5. Quantitative comparisons between our mismatch rejection methods and Hough transform for rigid transformations. The comparisons show the proportion of correct matches from the correspondence sets of varying size provided by each of the mismatch rejection methods.

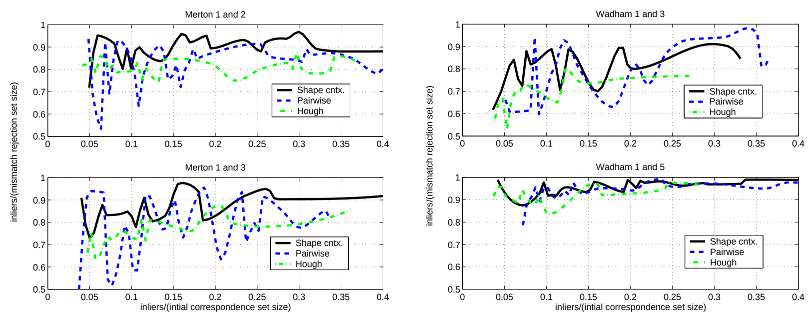


Fig. 6. Quantitative comparisons that show the proportion of correct matches as a function of the percentage of inliers present in the initial correspondence set built from feature similarity search.

mismatches. In Section V we propose a method to eliminate the remaining mismatches. Fig. 6 shows the robustness of each mismatch rejection method to high percentages of mismatches present in the initial correspondence set (the variation of the correspondence set size is obtained by varying the threshold in Eq. 1). Notice that the semi-local feature presents the best robustness since its performance is relatively stable even with the presence of high percentage of mismatches, while both the pairwise clustering and Hough start to present an unstable behavior when the initial proportion of correct matches falls below 15%.

For the experiments in this section, the number of operations carried out by the pairwise grouping and the semi-local feature algorithms is about 10^6, which is proportional to $|\mathcal{N}_{mt}|^2$. Moreover, the number of operations of the Hough transform varies between 10^5 and 10^7, depending on the number of bins used in the transformation space.

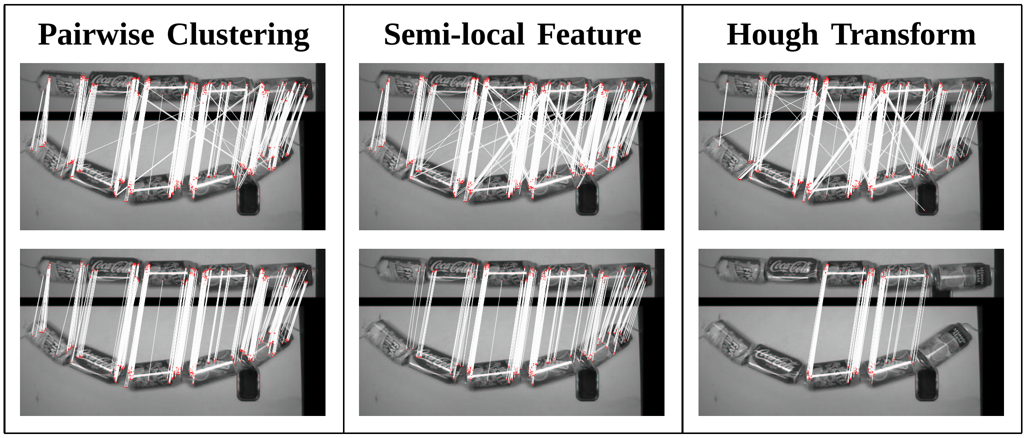


Fig. 7. Comparison between our mismatch rejection methods and Hough clustering for non-rigid deformation. The lines represent the feature correspondences that were grouped together by the respective method between the test image on the bottom, and the model image on the top. The first row shows the results where the parameters of each method were set to be extremely tolerant to mismatches, while the second row shows the results where the parameters were set such that the group formed had the highest number of correspondences without any visible mismatch.

2) Non-rigid Deformation: Two comparisons are presented in Figures 7-8, where, for the pairwise clustering and Hough transform, only the group that clustered the highest number of features is shown in each case. Note that for the case of the semi-local feature only one group per correspondence set can be formed, and this is the group shown in the experiments. Fig. 7 shows the results of our mismatch rejection methods proposed here and of Hough transform where the model is an object composed of a string built with soda cans. For each method, two results are shown. In the first row, the parameters of each method are set to be extremely tolerant to mismatches, while the second row depicts the case where each method produces the largest correspondence set without any visually detectable mismatch. Notice that the Hough transform only matches a piece of the object that suffered a deformation close to a rigid transformation when its parameters are set to be robust to mismatches, while our methods tend to be more robust to non-rigid deformations even when they are very resistant to mismatches.

Finally, in the experiments above the number of operations carried out by the pairwise grouping and the semi-local feature algorithms is around 10^6, while that of the Hough transform varies between 10^5 and 10^7.

3) Discussion: Although both methods are shown to be effective at reducing the mismatches in correspondence sets, each one has advantages and disadvantages. The grouping based on

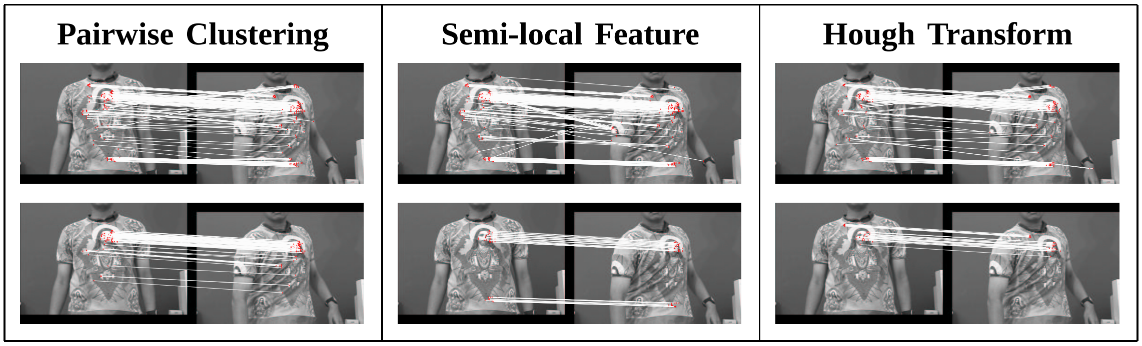


Fig. 8. Second comparison between our mismatch rejection algorithms and Hough transform. Please refer to the caption of Fig. 7 for details.

pairwise relations shows a slightly higher robustness to non-rigid deformations, but it needs neighboring model points to be neighbors in the test image, which means that a large gap of neighboring matches in the correspondence set can potentially break the initial group into sub-groups. One advantage of the semi-local method is its high robustness to mismatches in the correspondence sets as depicted in Fig. 6 Another advantage of the semi-local feature is in terms of efficiency, where the computation of the shape feature can be performed in parallel to that of the local feature after the location and orientation of the interest points are determined, but the fact that it can form only one group per model may represent a problem in recognition tasks involving the detection of several instances of a model in a test image.

V. Geometric Predictions

The mismatch rejection methods presented in Sections IV-A and IV-B can be made arbitrarily robust to mismatches by varying the threshold τ_{CCA} in (4) and τ_c in (6). Generally, it is desirable to be tolerant at this stage, and let the next stages in the system do the fine tuning by rejecting mismatches that remained in the correspondence set. The main reason for letting the system accept a few mismatches at this first stage is to make it less prone to false negatives. Moreover, once we a have a correspondence set relatively free of mismatches, the system has to determine whether this set represents an instance of a model. Therefore, the geometric predictions that we present now have two objectives: a) further reject mismatches from the correspondences set, and b) provide a measure of likelihood of model presence in the correspondence set.

Consider again the set of correspondences $\mathcal{N}_{mt}$ defined in (1) between model features $\mathcal{O}_m$

and test image features $\mathcal{O}_t$. The idea is to predict $\tilde{\mathbf{x}}_\mathrm{k}$, $\tilde{\theta}_\mathrm{k}$, and $\tilde{\sigma}_\mathrm{k}$ for each test image feature $\tilde{\mathbf{f}}_\mathrm{k} \in \mathcal{O}_t$ that has a correspondence in $\mathcal{N}_{mt}$, and compare those predicted values with the actual values of the feature $\tilde{\mathbf{f}}_\mathrm{k}$. This comparison is then used to measure the likelihood of the presence of $\tilde{\mathbf{f}}_\mathrm{k}$ assuming the model presence. In general, note that the following relations are true if the correspondence is correct:

$$\begin{aligned}
&\tilde{\mathbf{n}}_\mathrm{lo}^T(\tilde{\mathbf{x}}_\mathrm{l} - \tilde{\mathbf{x}}_\mathrm{o}) \approx \|\mathbf{x}_l - \mathbf{x}_o\|, \text{ where } \tilde{\mathbf{n}}_\mathrm{lo} = \tfrac{\tilde{\mathbf{x}}_\mathrm{l} - \tilde{\mathbf{x}}_\mathrm{o}}{\|\tilde{\mathbf{x}}_\mathrm{l} - \tilde{\mathbf{x}}_\mathrm{o}\|}, \\
&\tilde{\theta}_\mathrm{l} - \tilde{\vartheta}_\mathrm{lo} \approx \theta_l - \vartheta_{lo}, \\
&\tfrac{\tilde{\sigma}_\mathrm{l} - \tilde{\sigma}_\mathrm{o}}{\tilde{\sigma}_\mathrm{o}} \approx \tfrac{\sigma_l - \sigma_o}{\sigma_o}.
\end{aligned} \tag{8}$$

Let us consider the position prediction first. Assuming that the observed position $\tilde{\mathbf{x}}_\mathrm{l}$ is affected by additive Gaussian noise, we have

$$\pi_{lo,p}\tilde{\mathbf{n}}_\mathrm{lo}^T\tilde{\mathbf{x}}_\mathrm{l} = \pi_{lo,p}\|\mathbf{x}_l - \mathbf{x}_o\| + \pi_{lo,p}\tilde{\mathbf{n}}_\mathrm{lo}^T\tilde{\mathbf{x}}_\mathrm{o} + \pi_{lo,p} r_\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o) \tag{9}$$

for all $(\mathbf{f}_o, \tilde{\mathbf{f}}_\mathrm{o}) \in \mathcal{N}_{mt} - (\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l})$, where $r_\mathcal{D}(.)$ is a Gaussian noise with zero mean and variance $\sigma_\mathcal{D}^2(\mathbf{f}_l, \mathbf{f}_o)$, which is defined later in Section VI-B. Here, $\pi_{lo,p} = e^{-0.5\frac{\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o)}{\sigma_{\pi,p}^2}}$ is the pairwise weight, meaning that neighboring points to $\mathbf{f}_l$ within a range of roughly $\sigma_{\pi,p}$ pixels have a higher weight in predicting the position of the test feature $\tilde{\mathbf{f}}_\mathrm{l}$ than neighboring points that are farther away. We set the value of $\sigma_{\pi,p}$ as a fraction of the model diameter in pixels. Equation 9 can be re-written as

$$\Pi\mathbf{K}^T\tilde{\mathbf{x}}_\mathrm{l} = \Pi\mathbf{b} + \Pi\mathbf{r}_\mathcal{D}, \tag{10}$$

where $\mathbf{K} \in \Re^{2\times N-1}$ is a matrix with the vectors $\tilde{\mathbf{n}}_\mathrm{lo} \in \Re^{2\times 1}$ in its columns with N being the number of correspondences in $\mathcal{N}_{mt}$, $\Pi \in \Re^{N-1\times N-1}$ is a diagonal matrix with the values $\pi_{lo,p}$ for all $o \neq l$, $\mathbf{b} \in \Re^{N-1\times 1}$ with $\mathbf{b} = \|\mathbf{x}_l - \mathbf{x}_o\| + \mathbf{n}_{lo}^T\tilde{\mathbf{x}}_\mathrm{o}$ for all $o \neq l$, and $\mathbf{r}_\mathcal{D} \in \Re^{N-1\times 1}$ is the vector with the Gaussian noise mentioned above. From (10), we have

$$\tilde{\mathbf{x}}_\mathrm{l} = \mathbf{B}\mathbf{b} + \mathbf{B}\mathbf{r}_\mathcal{D}, \tag{11}$$

where $\mathbf{B} = (\mathbf{K}\Pi\mathbf{K}^T)^{-1}\mathbf{K}\Pi$. Note that we do not know the specific values of $r_\mathcal{D}(.)$, but only their distribution, so we approximate the position $\tilde{\mathbf{x}}_\mathrm{l}$ by the following prediction (see Fig. 9):

$$\tilde{\mathbf{x}}_\mathrm{l}^* = E[\tilde{x}_\mathrm{l}] = \mathbf{B}\mathbf{b}. \tag{12}$$

In order to compute the similarity between the observed position $\tilde{\mathbf{x}}_\mathrm{l}$ and its prediction $\tilde{\mathbf{x}}_\mathrm{l}^*$, we have to compute the position covariance, as follows:

$$\Sigma_\mathcal{D}(\tilde{\mathbf{f}}_\mathrm{l}) = E[(\tilde{\mathbf{x}}_\mathrm{l} - E[\tilde{\mathbf{x}}_\mathrm{l}])(\tilde{\mathbf{x}}_\mathrm{l} - E[\tilde{\mathbf{x}}_\mathrm{l}])^T] = E[\mathbf{B}\mathbf{r}_\mathcal{D}\mathbf{r}_\mathcal{D}^T\mathbf{B}^T] = \mathbf{B}\mathrm{diag}(\sigma_\mathcal{D}^2(\mathbf{f}_l, \mathbf{f}_o))\mathbf{B}^T, \tag{13}$$

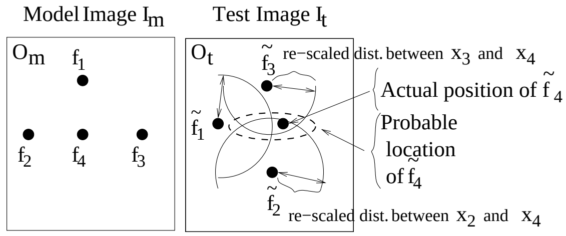


Fig. 9. Example of position prediction. Given the set of model features $\{\mathbf{f}_l\}_{l \in \{1,2,3,4\}}$, suppose we want to estimate the position of test image feature $\tilde{\mathbf{f}}_4$. The probable location of the feature (represented by a dotted ellipsoid in the Figure) is based on a Gaussian distribution computed using the position of the correspondences in the test and model images and the pairwise variances $\sigma^2_{\mathcal{D}}(\mathbf{f}_l, \mathbf{f}_o)$ estimated in the learning stage.

where $\sigma^2_{\mathcal{D}}(\mathbf{f}_l, \mathbf{f}_o)$ is assumed to be independent for all $o \neq l$. Finally, the similarity between $\tilde{\mathbf{x}}_{\mathrm{l}}$ and $\tilde{\mathbf{x}}^*_{\mathrm{l}}$ is computed as $\mathcal{G}(\tilde{\mathbf{x}}_{\mathrm{l}} - \tilde{\mathbf{x}}^*_{\mathrm{l}}; \Sigma_{\mathcal{D}}(\tilde{\mathbf{f}}_{\mathrm{l}}))$, where $\mathcal{G}(.)$ is the normalized zero mean Gaussian function.

Following the same reasoning, the The similarity between $\tilde{\theta}_{\mathrm{l}}$ and $\tilde{\theta}^*_{\mathrm{l}}$ is defined as $\mathcal{G}(\tilde{\theta}_{\mathrm{l}} - \tilde{\theta}^*_{\mathrm{l}}; \sigma^2_{\mathcal{H}}(\tilde{\mathbf{f}}_{\mathrm{l}}))$, with $\mathcal{G}(.)$ is, again, the normalized zero mean Gaussian function, and $\sigma^2_{\mathcal{H}}(\tilde{\mathbf{f}}_{\mathrm{l}}) = \left(\frac{1}{\sum_{o \neq l} \pi_{lo,p}}\right)^2 \left(\sum_{o \neq l} \pi^2_{lo,p} \sigma^2_{\mathcal{H}}(\mathbf{f}_l, \mathbf{f}_o)\right)$, where $\sigma^2_{\mathcal{H}}(\mathbf{f}_l, \mathbf{f}_o)$ is defined in Sec. VI-B. Finally, the similarity between $\tilde{\sigma}_{\mathrm{l}}$ and $\tilde{\sigma}^*_{\mathrm{l}}$ is computed as $\mathcal{G}(\tilde{\sigma}_{\mathrm{l}} - \tilde{\sigma}^*_{\mathrm{l}}; \sigma^2_{\mathcal{S}}(\tilde{\mathbf{f}}_{\mathrm{l}}))$, with $\mathcal{G}(.)$ is the normalized zero mean Gaussian function, and $\sigma^2_{\mathcal{S}}(\tilde{\mathbf{f}}_{\mathrm{l}}) = \left(\frac{1}{\sum_{o \neq l} \pi_{lo,p}}\right)^2 \left(\sum_{o \neq l} \pi^2_{lo,p} \sigma^2_{\mathcal{S}}(\mathbf{f}_l, \mathbf{f}_o)\right)$, where $\sigma^2_{\mathcal{S}}(\mathbf{f}_l, \mathbf{f}_o)$ is also defined in Sec. VI-B.

Therefore, the similarity between the predicted and observed position, main orientation, and scale is computed in just one step as follows:

$$p(\mathbf{f}_l, \tilde{\mathbf{f}}_{\mathrm{l}}) = \mathcal{G}([\tilde{\mathbf{x}}_{\mathrm{l}}, \tilde{\theta}_{\mathrm{l}}, \tilde{\sigma}_{\mathrm{l}}] - [\tilde{\mathbf{x}}^*_{\mathrm{l}}, \tilde{\theta}^*_{\mathrm{l}}, \tilde{\sigma}^*_{\mathrm{l}}]; \Sigma_t), \tag{14}$$

where $\mathcal{G}(.)$ is the normalized Gaussian function with zero mean, and $\Sigma_t = \mathrm{diag}(\Sigma_{\mathcal{D}}(\tilde{\mathbf{f}}_{\mathrm{l}}), \sigma^2_{\mathcal{H}}(\tilde{\mathbf{f}}_{\mathrm{l}}), \sigma^2_{\mathcal{S}}(\tilde{\mathbf{f}}_{\mathrm{l}}))$.

The likelihood of the correspondence between $\mathbf{f}_l$ and $\tilde{\mathbf{f}}_{\mathrm{l}}$ represented by $p(.)$ in (14) is used for two goals. The first is to form the final set of correspondences by thresholding $p(.)$ and forming the following set: $\tilde{\mathcal{L}}_g(\mathcal{N}_{mt}) = \{(\mathbf{f}_l, \tilde{\mathbf{f}}_{\mathrm{l}}) | (\mathbf{f}_l, \tilde{\mathbf{f}}_{\mathrm{l}}) \in \mathcal{L}_g(\mathcal{N}_{mt}), p(\mathbf{f}_l, \tilde{\mathbf{f}}_{\mathrm{l}}) > \tau_p\}$ [1]. The second goal is to use the value provided by $p(.)$ to determine the likelihood of the correspondence between $\mathbf{f}_l$

[1]Notice that we intentionally gave the same name for the sets of hypotheses to be verified $\tilde{\mathcal{L}}_g(\mathcal{N}_{mt})$ built from both mismatch rejection methods (i.e., semi-local features and grouping based on pairwise relations).

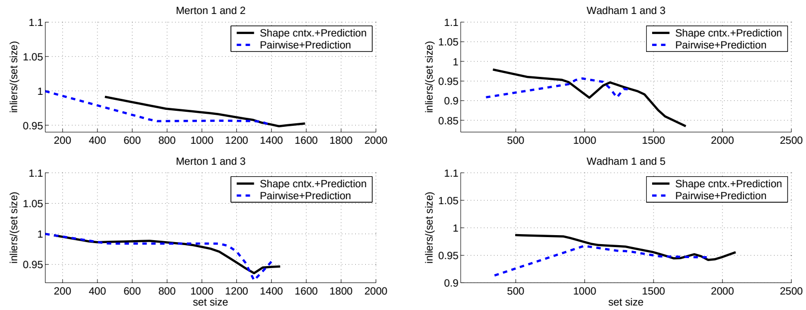


Fig. 10. Percentage of correct matches versus correspondence set after geometric prediction. This graph represents an extension of the graphs in Fig. 5 where, here, the geometric prediction filters out mismatches from the group formed by the respective mismatch rejection method. Note that the vertical scale is slightly different from that of Fig. 5.

and $\tilde{\mathbf{f}}_l$. The time complexity of this algorithm is, like the mismatch rejection methods above, $O(|\mathcal{N}_{mt}|^2)$, so it does not deteriorate the complexity of the system.

A. Performance Evaluation

In this section we demonstrate the efficacy of the geometric prediction algorithm for the task of rejecting remaining mismatches left by the mismatch rejection methods presented in Sec. IV.

The geometric prediction has two parameters to set. The first is the weight that a feature $\mathbf{f}_o$ has in predicting the position, orientation, and scale of a feature $\mathbf{f}_l$. We use $\pi_{lo,p} = e^{\frac{-0.5\mathcal{D}^2(\mathbf{f}_l,\mathbf{f}_o)}{\sigma^2_{\pi,p}}}$, where we set $\sigma_{\pi,p} = \frac{D_M}{10}$. The other parameter is the correct match threshold τ_p, which is set at 10^{-16}.

1) Rigid Deformation: The experimental setup introduced in Sec. IV-C.1 is used here and we show the results in the final correspondence set after the geometric prediction rejected remaining outliers from the groups formed by both mismatch rejection methods. Fig. 10 shows the inlier percentage versus correspondence set size for the respective graphs of Fig. 5. Note that the main difference is that the inlier percentage rarely falls below 90% to 95% even for large correspondence sets. Also, Fig. 11 illustrates the consistent robustness of geometric prediction combined with the mismatch rejection methods to extremely high percentage of mismatches in the initial correspondence set. Even in cases with less than 10% of initial correct matches, both methods return a final correspondence set with generally more than 90% of inliers.

2) Non-rigid Deformation: We extend the experiment presented in Section IV-C.2, where the geometric prediction is used to reject the mismatches from the groups built by both mismatch

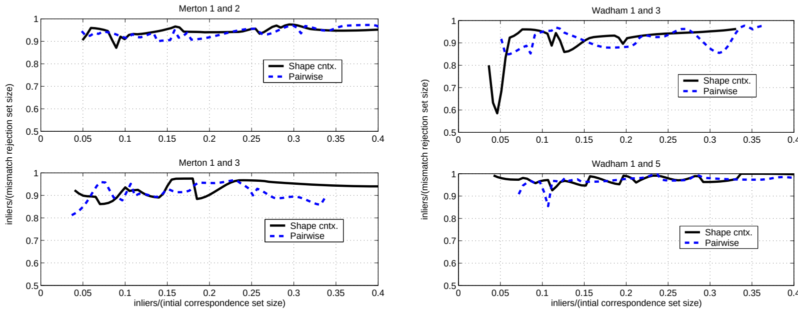


Fig. 11. Quantitative comparisons that shows the proportion of correct matches as a function of the percentage of inliers present in the initial correspondence set. This graph represents an extension of the graphs in Fig. 6 where, here, the geometric prediction filters out mismatches from the group formed by the respective mismatch rejection method.

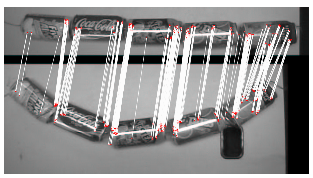
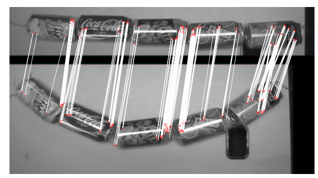
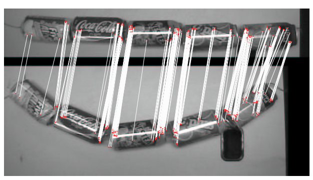
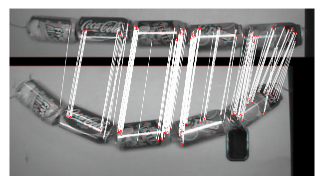
Pairwise Clustering + Geometric Prediction	**Semi-local Feature + Geometric Prediction**
	
	

Fig. 12. Correspondence set after the geometric prediction method filtered out the mismatches present in the groups of Fig. 7.

rejection methods. Fig. 12 shows the results from the geometric prediction on the groups of Fig. 7, while Fig. 13 presents the final correspondence sets from Fig. 8.

VI. Probabilistic Formulation for Verification

In this section, we introduce the probabilistic formulation for the hypothesis verification stage, which is based on [31], but we make somewhat less restrictive assumptions that may improve the verification performance.

The problem of constructing a probabilistic method for the verification of hypotheses has been intensively studied lately. Similar probabilistic verification methods to recognize limited categories of objects are presented in [1], [2], [15], [16], [43], where the systems generally work with a small set of parts (substantially fewer than 100 parts). It is worth noting that among the

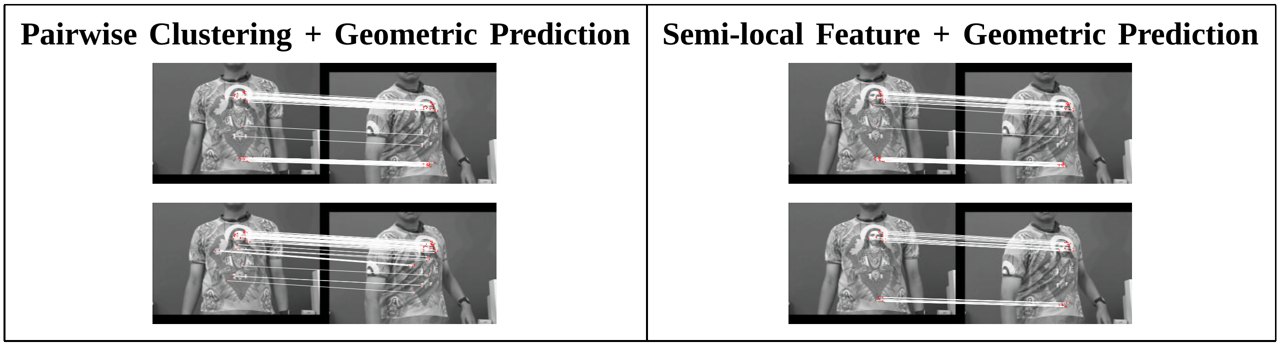


Fig. 13. Correspondence set after the geometric prediction algorithm filtered out the remaining mismatches present in the groups of Fig. 8.

papers cited before, only the work described in [1] uses a flexible spatial coherence based on pairwise relations for the verification. Systems more closely related to ours are described in [26], [31], [34]. Lowe [26] relies on a probabilistic verification that takes into account the global shape of the model and information about the distinctiveness of the model as a whole. Schmid [34] describes a probabilistic verification that uses semi-local coherence, where a learning approach to estimate the feature appearance variation is described. However, it is likely that this system suffers from the presence of mismatches in large hypothesis sets.

In order to assess the hypothesis that a particular object is present in a test image, we propose a probabilistic formulation framework that involves the feature correspondences and the semi-local spatial configuration similarities. Assuming that $\mathcal{O}_m$ represents the hypothesis that an instance of the model m is present in the test image, $\mathcal{E}$ is a set of correspondences, and T represents the global geometric configuration of features (i.e., their position $\mathbf{x}$, scale σ, and main orientation θ). We define the posterior $P(\mathcal{O}_m|\mathcal{E},T)$ as (using the Bayes rule):

$$P(\mathcal{O}_m|\mathcal{E},T) = \frac{P(\mathcal{E}|T,\mathcal{O}_m)P(T|\mathcal{O}_m)P(\mathcal{O}_m)}{\sum_{\mathcal{O}=\mathcal{O}_m,\neg\mathcal{O}_m} P(\mathcal{E}|T,\mathcal{O})P(T|\mathcal{O})P(\mathcal{O})}. \tag{15}$$

In [31], three assumptions are made:

1) $P(\mathcal{E},T) = P(\mathcal{E})P(T)$, i.e., the correspondences are independent of their global geometrical configuration;
2) $P(T|\mathcal{O}_m) = P(T)$, which means that the global configuration is conditionally independent of the hypothesized model; and
3) $\frac{P(\mathcal{E}|T,\mathcal{O}_m)}{P(\mathcal{E})} = \prod_i \frac{P(e_i|T,\mathcal{O}_m)}{P(e_i)}$, where e_i's are the individual elements of set $\mathcal{E}$.

On the other hand, we have the following two assumptions:

1) $P(T|\mathcal{O}_m) = P(T|\neg\mathcal{O}_m) = P(T)$, or the global geometrical configuration can be assumed to be conditionally independent of the hypothesized model; and
2) $P(\mathcal{E}|T, \mathcal{O}_m) = \prod_i P(e_i|T, \mathcal{O}_m)$.

Our first assumption above is necessary to remove the global spatial configuration of features from the posterior calculation, which is straightforward from the mismatch rejection methods proposed. Even though we know that our second assumption is unrealistic, it is necessary since the estimation of the joint probability $P(\mathcal{E}|T, \mathcal{O}_m)$ would require an extremely large number of training cases.

A. Probabilistic Correspondences Based on Feature Similarity

Using the image deformations of Appendix II and the database of random features of Appendix I, it is possible to determine three properties of each model feature $\mathbf{f}_l \in \mathcal{O}_m$ (please refer to [11] for more details): a) the probability distribution of feature similarities given a correct correspondence $P_{\text{on}}(s_f(.); \mathbf{f}_l)$, b) the probability distribution of feature similarities given a false correspondence $P_{\text{off}}(s_f(.); \mathbf{f}_l)$, and c) the probability of feature detection $P_{\text{det}}(\mathbf{f}_l)$. Using these properties we compute the probabilistic correspondence, as explained later in Sec. VI-C.

B. Probabilistic Correspondences based on Semi-local Geometry

The likelihood terms $P(e_i|T, \mathcal{O}_m)$ and $P(e_i|T, \neg\mathcal{O}_m)$ of each correspondence e_i in $\mathcal{E}$ also involves feature value and semi-local geometric similarity. Since we assume that the pairwise relations are affected by a zero-mean Gaussian noise (see Sec. V), only the variance of each pairwise relation in the model needs to be learned. Ideally, these variances should be estimated from real images of the same object, but that would require strong supervision in order to determine the locations of each model feature in each training image. Instead of that, we resorted to a simpler training procedure, where we use a single training image and artificially deform it (see deformations in Appendix II), so that the exact position of each model feature can be computed precisely. Let $\mathcal{O}_m$ represent the model features from model image I_m, and $\tilde{\mathcal{O}}_{m,d}$ be the features detected from the deformed version of image I_m, namely $\tilde{I}_{m,d}$, using a deformation $d \in \mathcal{DF}$. The correspondence set between these two sets is given by

$$\mathcal{N}_{m,d} = \{(\mathbf{f}_l, \tilde{\mathbf{f}}_{\text{l}}) | \tilde{\mathbf{f}}_{\text{l}} \in \mathcal{O}_{m,d}, \|\tilde{\mathbf{x}}_{\text{l}} - M(d)\mathbf{x}_l - \mathbf{b}(d)\| < \epsilon, \mathbf{f}_l \in \mathcal{K}(\tilde{\mathbf{f}}_{\text{l}}, \mathcal{O}_m, \kappa_{\mathcal{N}})\},$$

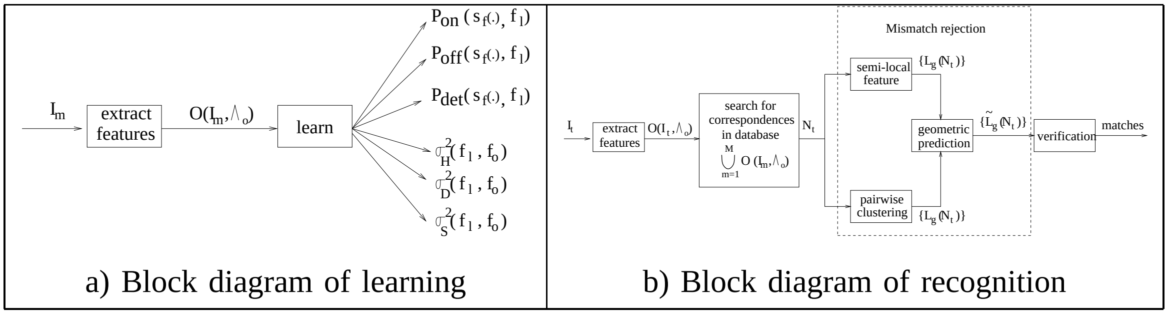


Fig. 14. Block diagrams of the learning and recognition procedures.

where $\mathcal{K}$, defined in (1), is the top $\kappa_{\mathcal{N}}$ correspondences (here, $\kappa_{\mathcal{N}} = 1$), ϵ was fixed at $\frac{\lambda_d}{4} = 2.0$ pixels (as measured in the image $\tilde{I}_{m,d}$), $\mathbf{x}_l$ is the position of feature $\mathbf{f}_l$, $\tilde{\mathbf{x}}_\mathrm{l}$ is the position of feature $\tilde{\mathbf{f}}_\mathrm{l}$, and the transformation parameters $M(d)$ and $\mathbf{b}(d)$ are obtained from the deformation $d \in \mathcal{DF}$. Assuming that the uncertainties of the pairwise relations are normally distributed, we have

$$\begin{aligned} \sigma^2_{\mathcal{S}}(\mathbf{f}_l, \mathbf{f}_o) &= var(\{\Delta\mathcal{S}_{lo}(\mathcal{N}_{m,d})\}_{d\in\mathcal{DF}}), \\ \sigma^2_{\mathcal{D}}(\mathbf{f}_l, \mathbf{f}_o) &= var(\{\Delta\mathcal{D}_{lo}(\mathcal{N}_{m,d})\}_{d\in\mathcal{DF}}), \text{ and} \\ \sigma^2_{\mathcal{H}}(\mathbf{f}_l, \mathbf{f}_o) &= var(\{\Delta\mathcal{H}_{lo}(\mathcal{N}_{m,d})\}_{d\in\mathcal{DF}}), \end{aligned} \tag{16}$$

for all deformations $d \in \mathcal{DF}$, where var is the sample variance of the values in the set, and the pairwise relations between an object and its deformed version are provided by $\Delta\mathcal{S}_{lo}(\mathcal{N}_{m,d})$, $\Delta\mathcal{D}_{lo}(\mathcal{N}_{m,d})$, and $\Delta\mathcal{H}_{lo}(\mathcal{N}_{m,d})$ (see Eq. 3). Therefore, for the term $P(e_i|T, \mathcal{O}_m)$, the idea is to use the geometric predictions defined in Sec. V to determine the likelihood of the correspondence e_i, and for $P(e_i|T, \neg\mathcal{O}_m)$, we simply assume a uniform distribution of the geometric configuration error.

C. Final Verification

Given a model $\mathcal{O}_m$, learned using the algorithm described above (see block diagram (a) in Fig. 14), the model presence in a test image I_t is determined as follows (see block diagram (b) in Fig. 14). First, build the set of local features $\mathcal{O}_t$ from I_t, then search for similar local features in the database of models, thus forming the $\mathcal{N}_{mt}$ (1). Note that each test image feature is matched to $\kappa_{\mathcal{N}}$ model features and that it is possible that a model feature is matched to more than one test feature. We handle this kind of multiple correspondences originating from one feature in the model image by representing them as separate entities in the correspondence set. Given these correspondences, the mismatch rejection step forms a set of G clusters $\{\tilde{\mathcal{L}}_g(\mathcal{N}_t)\}_{g=1}^{G}$ (see Sections IV-A and IV-B.1). 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 $\tilde{\mathcal{L}}_g(\mathcal{N}_t)$ actually represents an instance of the object $\mathcal{O}_m$. Let us first define the set of pairings for all model features $\mathbf{f}_l \in \mathcal{O}_m$ from group $\tilde{\mathcal{L}}_g(\mathcal{N}_t)$, as $\mathcal{E}_g = \tilde{\mathcal{L}}_g(\mathcal{N}_t) \bigcup \{(\mathbf{f}_l, \emptyset) | \mathbf{f}_l \in \mathcal{O}_m, \neg\exists \tilde{\mathbf{f}}_\mathrm{l} \in \mathcal{O}_t \text{ s.t. } (\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) \in \tilde{\mathcal{L}}_g(\mathcal{N}_t)\}$. Therefore, we compute the posterior (15) as follows:

1) $P(\mathcal{O}_m)$ is the prior expectation of model presence, and $P(\neg\mathcal{O}_m) = 1 - P(\mathcal{O}_m)$ (here, we assume that $P(\mathcal{O}_m) = 0.001$).
2) $P(\mathcal{E}_g | T, \mathcal{O}_m) \approx \prod_{(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) \in \mathcal{E}_g} P((\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) | T, \mathcal{O}_m)$, where we have the following two cases:
   a) $P((\mathbf{f}_l, \emptyset) \in \mathcal{E}_g | T, \mathcal{O}_m) \approx (1 - P_{\mathrm{det}}(\mathbf{x}_l)) + P_{\mathrm{det}}(\mathbf{x}_l) P_{\mathrm{on}}(s_f < \tau_s; \mathbf{f}_l)$, where τ_s is the threshold in (1), and $P_{\mathrm{det}}(\mathbf{f}_l)$ and $P_{\mathrm{on}}(.)$ are defined in Sec. VI-A. The intuition is that if the model feature is not matched to a test feature, either it was not detected (first term of the sum), or it was detected, but not included in the correspondence set (second term);
   b) $P((\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) \in \mathcal{E}_g | T, \mathcal{O}_m) \approx P_{\mathrm{det}}(\mathbf{f}_l) P_{\mathrm{on}}(s_f(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}); \mathbf{f}_l) p(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l})$, where $p(.)$ is defined in (14). Here, we consider that for $(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) \in \mathcal{E}_g$, the feature has to be detected in the test image (first term of the multiplication), with a certain similarity value (second term), and geometric configuration (third term).
3) $P(\mathcal{E}_g | T, \neg\mathcal{O}_m) = \prod_{(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) \in \mathcal{E}_g} P((\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) | T, \neg\mathcal{O}_m)$, where we have the following two cases:
   a) $P((\mathbf{f}_l, \emptyset) \in \mathcal{E}_g | T, \neg\mathcal{O}_m) \approx (1 - 0.032) + 0.032(P_{\mathrm{off}}(s_f < \tau_s; \mathbf{f}_l))$, where the number 0.032 represents the average number of interest points per test image divided by the size of the image (see [8]), and $P_{\mathrm{off}}(.)$ is defined in Sec. VI-A. Similarly to the case above, the likelihood of having an unmatched model feature assuming that the model is not present is approximated by the probability of general detection failure (first term) plus the likelihood of detection times the likelihood of not including the match in $\mathcal{E}_g$;
   b) $P((\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}) \in \mathcal{E}_g | T, \neg\mathcal{O}_m) \approx (0.032) P_{\mathrm{off}}(s_f(\tilde{\mathbf{f}}_\mathrm{l}, \mathbf{f}_l); \mathbf{f}_l) \frac{1}{size(\mathcal{I})} \frac{1}{8} \frac{1}{2\pi}$. In the last term, we assume uniform distribution of position, main orientation, and scale given a background feature. The intuition is that the likelihood of matching a model to a test feature in

this case is related to the general feature detection, high similarity to false positive matches, and arbitrary geometric configuration.

Finally, we accept a hypothesis if $P(\mathcal{O}_m|\mathcal{E}_g, T)$ is above a probability value, the number of correctly predicted matches (using Eq. 14) is above a threshold, and the maximum distance between test image features is bigger than a threshold , i.e., assuming $\tilde{\mathbf{x}}_\mathrm{l}$ and $\tilde{\mathbf{x}}_\mathrm{o}$ are the positions of test image features $\tilde{\mathbf{f}}_\mathrm{l}$ and $\tilde{\mathbf{f}}_\mathrm{o}$, respectively, with $(\mathbf{f}_l, \tilde{\mathbf{f}}_\mathrm{l}), (\mathbf{f}_o, \tilde{\mathbf{f}}_\mathrm{o}) \in \mathcal{E}_g$, we require $\max_{\forall l,o}\left(\frac{\|\mathbf{x}_l - \mathbf{x}_o\|}{\sqrt{\sigma_l^2+\sigma_o^2}}\right) > \tau_{\mathcal{D}}$ (this is done to avoid a large number of features all in a small area of the image).

VII. Experiments

In this section we show the qualitative and quantitative performance of our recognition algorithm using the phase-based feature [8], both mismatch rejection methods, and the probabilistic verification. The following tasks are considered: a) wide baseline stereo matching, and b) long range motion matching. The main difference between the wide baseline stereo and long range motion experiments is that the former always involves the computation of the epipolar geometry given a pair of images presenting a significant 3-D rigid transformation, while the latter concerns matching pairs of images that might have suffered not only 3-D rigid but also non-rigid deformations.

A. Recognition Parameters

Referring to the block diagram (b) of Fig. 14, the search for similar features (see Eq. 1) in the model database involves the following two parameters: a) phase correlation threshold (here, $\tau_s = 0.75$), and the maximum number of nearest neighbors ($\kappa_{\mathcal{N}} = 1$). The following step is the mismatch rejection based either on the pairwise clustering or the semi-local feature. The parameters used for the mismatch rejection methods are the same as described in Section IV-C, where $\tau_{\text{CCA}} = 0.2$ for the pairwise grouping method (see Eq. 4), and $\tau_c = 0.5$ for the semi-local feature (see Eq. 6). The acceptance of a hypothesis is evaluated in the verification step, which depends upon: the posterior $P(\mathcal{O}_m|\mathcal{E}_g, T) > 0.5$, the maximum distance between test image features being at least 20% of the maximum model diameter in pixels, and the number of correctly predicted matches being at least 3% of the total number of features of the model. The

Fig. 15. Models used for long range motion. All the models are represented only by the features inside the contour around the object of interest.

parameters above are found to provide a good balance between robustness to image deformations and to false positives, and they are kept fixed throughout the experiments below.

B. Long Range Motion Results

The long range motion application is likely to be the most appropriate application for the system presented in this work. In fact, any task that involves the recognition and rough localization of textured objects that suffered severe 3-D rigid and non-rigid deformations (including articulation) is well suited for this system. In this section the model (see Fig. 15) is always represented by only one view of the object, and the system tries to find it throughout the sequence. We also provide a comparison using Hough transform as a baseline method for eliminating mismatches in combination with the verification stage based on geometric prediction.

The sequences of the Torso, Hedvig, Kevin, and Dudek models (see samples in Fig. 16) are quite challenging due to the presence of non-rigid, brightness, 3-D rigid transformations, and partial occlusion. Fig. 16 shows the verification results using either the pairwise grouping or the semi-local feature methods to reject mismatches. Although we only show the most severely deformed samples in each sequence, it is interesting to see the quantitative performance of this system in each sequence shown in Table I. We do not show the number of true negatives since that number would be related to all possible data associations between the set of model and test features, which is equal to $|\mathcal{O}_m|^{|\mathcal{O}_t|}$, where both $|\mathcal{O}_m|$ and $|\mathcal{O}_t|$ are in the order of 10^3. Also in this table we show the performance of the system using Hough transform to reject mismatches followed by the geometric prediction in the verification stage.

The snake of cans in Fig. 17 represents another challenging set of images that shows the articulated object in several different poses. Illumination changes are also present due to the

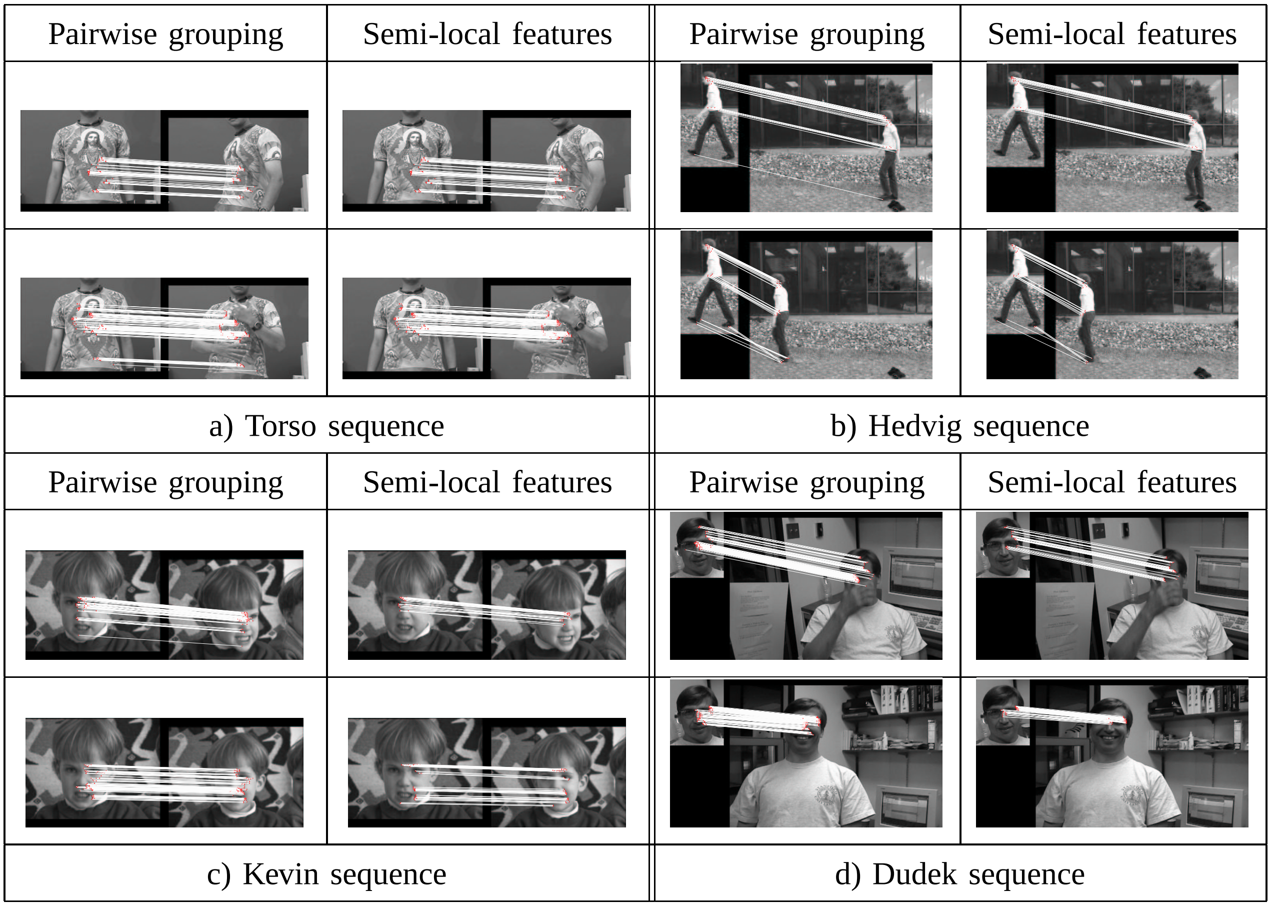


Fig. 16. Matchings for the Torso, Hedvig, Kevin, and Dudek models. First and third columns show the verification results using pairwise grouping for rejecting mismatches, and second and fourth columns use the phase-based semi-local features. White lines are the correspondences between model and test images after verification.

TABLE I

PERFORMANCE OF THE RECOGNITION ALGORITHM IN EACH SEQUENCE.

Sequence	Length	True positives			False positives			False negatives		
		Pairwise	Semi-local	Hough	Pairwise	Semi-local	Hough	Pairwise	Semi-local	Hough
Dudek	140	138	130	105	0	0	9	2	10	35
Kevin	120	120	109	111	0	0	12	0	11	9
Hedvig	33	30	31	28	0	0	1	3	2	5
Torso	148	148	147	148	0	0	8	0	1	0

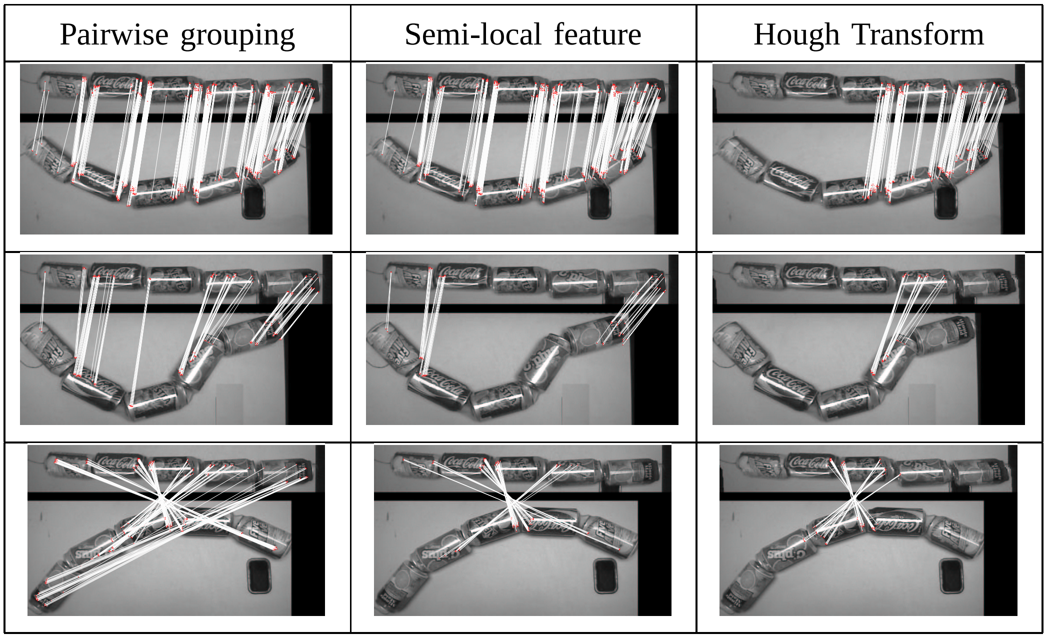


Fig. 17. Matchings for snake of cans model.

highlights in the metal cans. Notice that both methods are quite robust in terms of articulate deformations. In contrast, Hough transform provides a poorer performance in these cases.

Finally, Fig. 18 shows the most challenging cases (in terms of non-rigid deformation) from the database of images designed by Ferrari et al. [17]. In general the pairwise grouping and local features are more robust to non-rigid deformations than the Hough transform, and consequently tend to include more correct matches in the final correspondence set. The running time for the tasks of searching for correspondences, mismatch rejection, and verification varies between five to ten seconds in non-optimized Matlab code for all the cases presented in this section.

C. Wide Baseline Stereo Results

A wide baseline stereo problem involves two images where a significant 3-D rigid transformation took place between them, and the goal is to reliably compute their epipolar geometry. In order to robustly compute this epipolar geometry, we need a reasonably large number of matches situated on different planes of the scene. Using the same experimental setup introduced in Sec. IV-C.1, we focus on the computation of the $\mathbf{F}$ matrix and also on the number of trials t necessary to make the probability of choosing at least one outlier in every trial of the RANSAC

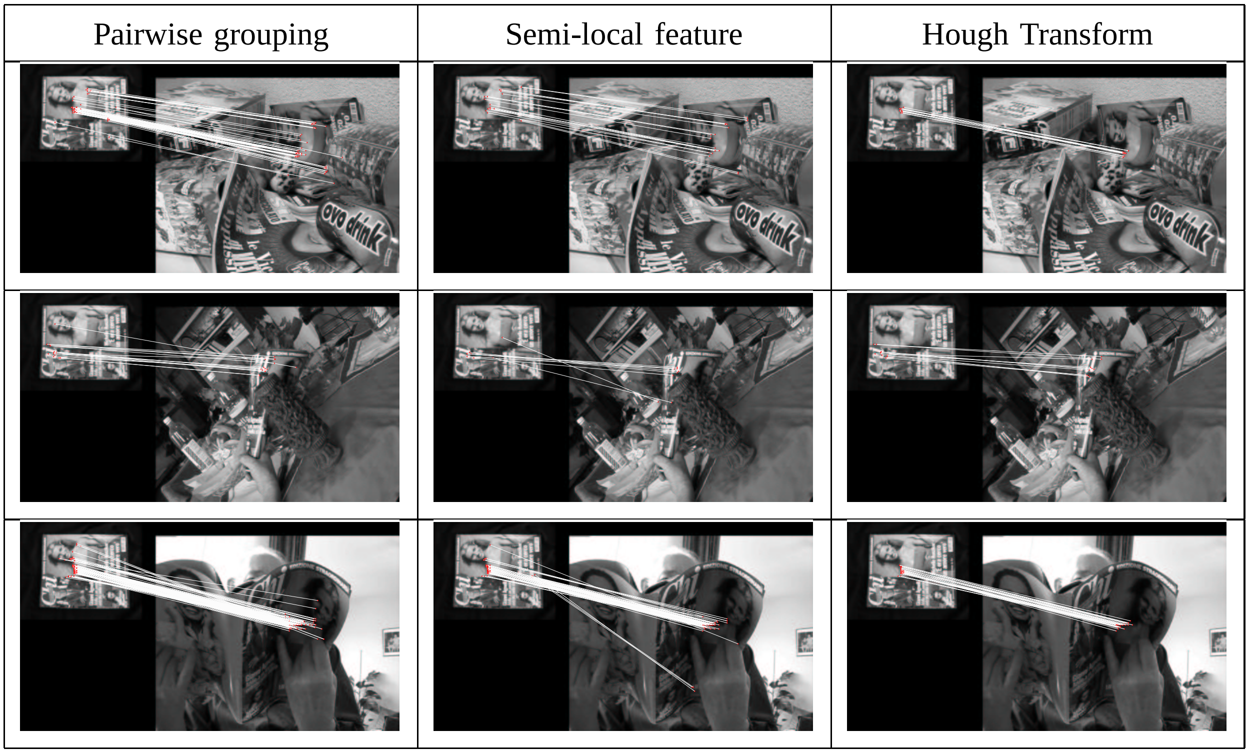


Fig. 18. Matchings for the michelle model [17].

algorithm smaller than 5%. We assume that the percentage of inliers is $p_{crtmtc} = \frac{in}{in+out}$, where in is the number of inliers in the set and out the number of outliers, and that the matrix $\mathbf{F}$ has seven degrees of freedom. Using eight point correspondences to estimate F, the probability of finding at least one mismatch in a randomly selected subset of eight correspondences from the initial set is $p_{error} = 1 - p_{crtmtc}^{8}$. As a result, the number of trials t to make the probability of choosing at least one outlier in every trial of the RANSAC algorithm smaller than 5% is defined as $p_{error}^{t} \leq 0.05$, so t can be determined by $t \leq \lceil \frac{\log_2(0.05)}{\log_2(p_{error})} \rceil$.

Figures 19 and 20 show the wide baseline stereo pairs for the Merton and Wadham sequences. Notice that both outlier rejection methods return a correspondence set with a high percentage of inliers, which is between 93% and 99%. This large proportion of true correspondences is likely to reduce the complexity of the algorithm to compute the $\mathbf{F}$ matrix. The average running time for the tasks of searching for correspondences, mismatch rejection, and verification is around ten seconds in non-optimized Matlab code.

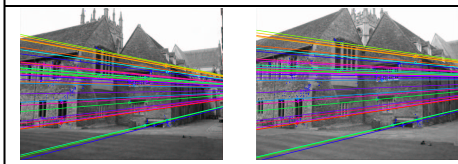
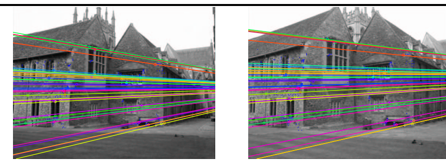
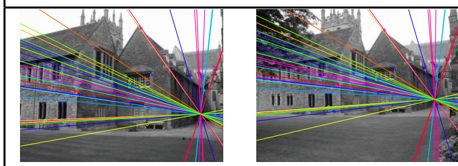
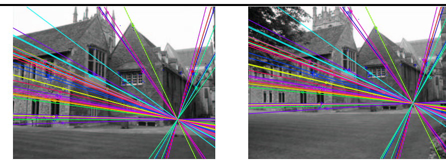
Pairwise grouping	Semi-local feature
	
% inliers= 96%, # inliers=628, $t \leq 3$	% inliers= 95%, # inliers=629, $t \leq 3$
	
% inliers= 96%, # inliers=336, $t \leq 3$	% inliers= 96%, # inliers=328, $t \leq 3$

Fig. 19. Epipolar geometry for Merton sequence. In the caption we show the proportion of correct matches given the **F** matrix computed ('% correct matches'). Also, '# correct matches' shows the total number of correct matches used, and 't' is the number of trials necessary to make the probability $p < 0.05$ of choosing at least 1 mismatch in every trial of the RANSAC algorithm.

VIII. Conclusions

The use of spatial configuration of local features aims at reducing the number of mismatches in the correspondence set. This is desirable in order to decrease the complexity of the verification stage and to reduce the likelihood of false positives and false negatives. We proposed two methods to reject mismatches based on semi-local spatial information and another method to reject mismatches and to verify hypotheses based on the prediction of the geometric information of local features. We presented comparisons between our methods and Hough clustering, which is a common mismatch rejection method based on global spatial configuration of features, and the results show that our approaches are more robust to rigid transformations and non-rigid deformations. Also, our mismatch rejection methods are shown to have a time complexity roughly similar to that of Hough transform. We also propose a new probabilistic verification that takes into account the semi-local spatial configuration of each feature and the feature similarity. Results on long range matching and wide baseline stereo matching show the efficacy of the proposed method.

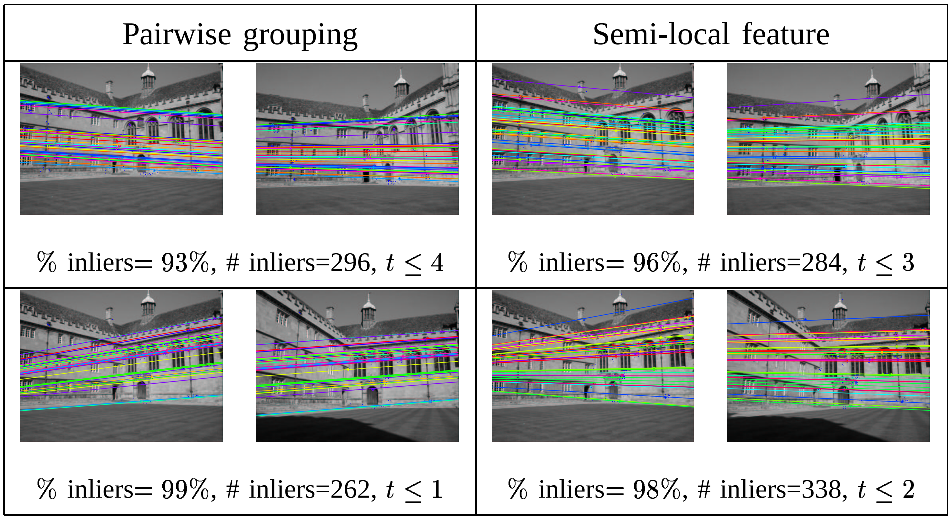


Fig. 20. Epipolar geometry for Wadham sequence. See Fig. 19 for details on the captions.

REFERENCES

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

[2] Y. Amit and D. Geman. A computational model for visual selection. *Neural Computation*, 11:1691–1715, 1999.

[3] H. Barrow and R. Popplestone. Relational descriptions in picture processing. *Machine Intelligence*, 6, 1971.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(24):509–522, 2002.

[5] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005.

[6] K. L. Boyer and S. Sarkar. *Perceptual Organization for Artificial Vision Systems*. Boston: Kluwer Academic Publishers, 2000.

[7] R. Brooks. Model-based 3-d interpretations of 2-d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):140–150, 1983.

[8] G. Carneiro and A. Jepson. Multi-scale phase-based local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, June 2003.

[9] G. Carneiro and A. Jepson. Flexible spatial models for grouping local image features. In *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, USA, June 2004.

[10] G. Carneiro and A. Jepson. Pruning local feature correspondences using shape context. In *IEEE International Conference on Pattern Recognition*, Cambridge, UK, August 2004.

[11] G. Carneiro and A. Jepson. The distinctiveness, detectability, and robustness of local image features. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 2005.

[12] C.Huang, O.Camps, and T.Kanungo. Object recognition using appearance-based parts and relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 877–883, Puerto Rico, 1997.

[13] O. Chum, J. Matas, and S. Obdrzalek. Epipolar geometry from three correspondences. In *Computer Vision - CVWW*, 2003.

[14] S. Dickinson, A. Pentland, and A. Rosenfeld. 3-d shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992.

[15] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. pages 66–75, 2002.

[16] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[17] V. Ferrari, T. Tuytelaars, and Luc Van Gool. Simultaneous object recognition and segmentation by image exploration. In *European Conference on Computer Vision*, Prague, Czech Republic, 2004.

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

[19] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[20] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.

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

[22] B. Huet and E. Hancock. Relational object recognition from large structural libraries. *Pattern Recognition*, 35:1895–1915, 2002.

[23] 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.

[24] D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.

[25] D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, September 1999.

[26] D. Lowe. Local feature view clustering for 3d object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[27] K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. In *British Machine Vision Conference*, 2003.

[28] E. Mortensen, H. Deng, and L. Shapiro. A sift descriptor with global context. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005.

[29] H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[30] J. Pilet, V. Lepetit, and P. Fua. Real-time non-rigid surface detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

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

[32] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *International Conference on Computer Vision*, pages 754–760, 1998.

[33] F. Schaffalitzky and A. Zisserman. Automated scene matching in movies. In *Proceedings of the Challenge of Image and Video Retrieval, London*, LNCS 2383, pages 186–197. Springer-Verlag, 2002.

[34] C. Schmid. A structured probabilistic model for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 485–490, 1999.

[35] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[36] A. Shokoufandeh, S. Dickinson, C. Jonsson, L. Bretzner, and T. Lindeberg. On the representation and matching of qualitative shape at multiple scales. In *European Conference on Computer Vision*, 2002.

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

[38] J. Sivic, F. Schaffalitzky, and A. Zisserman. Object level grouping for video shots. In *European Conference on Computer Vision*, Prague, Czech Republic, 2004.

[39] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *European Conference on Computer Vision*, pages 814–828, 2000.

[40] D. Tell and S. Carlsson. Combining appearance and topology for wide baseline matching. In *European Conference on Computer Vision*, pages 68–81, Copenhagen, Denmark, 2002.

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

[42] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *British Machine Vision Conference*, 2000.

[43] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *European Conference on Computer Vision*, pages 18–32, 2000.

[44] S. Yu, R. Gross, and J. Shi. Concurrent object recognition and segmentation by graph partitioning. In *Neural Information Processing Systems*, Vancouver,B.C., Canada, December 2002.

[45] Z. Zhang, R. Deriche, O. D. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.