Object Recognition Using Flexible Groups of Local Features

Gustavo Carneiro and Allan D. Jepson

Technical Report CSRG-481
Department of Computer Science
University of Toronto ,Toronto, ON, Canada.
{carneiro,jepson}@cs.utoronto.ca
21 January 2004

Abstract. We propose a new object recognition system based on local features and their flexible spatial configuration within a probabilistic framework. Although the training set consists of only 1 image of the object, the system is able to recognize that same object under large rigid and/or non-rigid deformations. This is possible due to the robustness of the local features and of the pairwise geometric constraints between them that we introduce in this paper. These pairwise constraints are also used for grouping features during hypothesis generation. The new grouping method generally produces fewer groups, where each group has more inliers, than the commonly used alternative method. Furthermore, we also propose a novel filtering procedure to select the local features that have high distinctiveness, detectability and robustness to image deformations. As a result, we decrease the ambiguity of matching and, consequently, improve the scalability of the recognition process with respect to the size of the database. We show the viability of this system using a database of 15 objects, and several challenging test images containing rigid/non-rigid deformations, illumination changes, clutter, and partial occlusion. Moreover, we also show the system working in a long range motion problem using a challenging sequence of images.

1 Introduction

View-based object recognition using highly informative local image descriptors extracted from robustly detectable image locations has been intensively investigated lately (e.g., [9, 10, 17, 18]). The significance of this type of system lies on its ability to deal with clutter, occlusion, rigid and non rigid deformations, and multiple instances of the same object in an image. Furthermore, model acquisition is a simple task where the system is presented with one or several views of the same object, so there is no need of a geometric description of the model. However, a remaining problem that dampens the success of these systems is that the search for similar features in the database of models usually returns a relatively large set of correspondences where the number of inliers is very small. This happens due to a densely populated database of model features and background clutter.

Usually, systems of this type consist of a feature detection method, followed by a search procedure where the set of correspondences is built, and then, the verification checks if the hypotheses formed from the correspondences can be considered to be

matches. The critical point here is certainly the combinatorial explosion of the hypotheses generated due to the large size of the set of correspondences. Therefore, a grouping procedure to select suitable hypotheses is unavoidable. However, this issue is rarely addressed in the literature of local features similar to the one considered here, with the exception of [10], where the author uses the generalized Hough transform for the task. We noticed that this technique usually produces a large number of groups, where each group has few inliers in it, and that problem is aggravated in the presence of large non-rigid deformations. For efficient detection, we need a grouping approach that is able to collect as many inliers as possible within each group. Moreover, the number of groups returned should be kept to a minimum since each one must be verified whether it is a match or not. Therefore, we propose a grouping method based on flexible pairwise measurements that are robust to both rigid and non-rigid deformations that, in general, produces fewer groups and each group has more inliers than the Hough transform.

Another issue that influences the size of the set of correspondences is the size of the model database. The ambiguity of a database of model features grows with its size, which clearly affects the scalability of those systems. In order to alleviate this problem, we also present a method that reduces the database of model features based on the following 3 local feature properties: distinctiveness, detectability and robustness to image deformations. The system proposed here proves to be viable through experiments using a database of 15 models and a series of challenging images containing rigid/non-rigid deformations, clutter, partial occlusion and illumination changes. Also, we use the same system in a long range motion application where the system is faced with a challenging sequence of images containing all sorts of deformations.

We are particularly interested in object recognition systems based on local features and their spatial arrangements. Lades et al. [9] proposed a system based on elastic graph matching where local features represent nodes and links are distance between features. The use of graphs is also explored in [18], where objects are represented as a hierarchical graph. In [14], the authors select local features in the training phase based on their measure of detectability, reliability, and uniqueness, but no attempt is made to learn probability distributions of those features that can be used in the recognition step. Several works [1, 4, 5, 20]exploit local feature descriptors with deformable global geometry using learning techniques, but they are aimed at the problem of representing and recognizing categories of objects (e.g., faces) as opposed to the recognition of specific objects as we propose here. Systems that recognize specific objects based on local features and their spatial arrangements are proposed in [10, 15] where a critical feature added to the system [10] is a grouping stage before the matching so that the hypothesis space is drastically reduced. Schmid [17] proposed a system based on a probabilistic framework, but no grouping stage is described, possibly resulting in a large hypothesis space. Nevertheless, it is important to note that the latter system, similarly to [16], also uses semi-local constraints, where a *fixed* number of local features around a given feature is used to determine its semi-local structure. On the other hand, our approach considers all the features in a tunable neighborhood to group local features. Various other approaches using using local features and their semi-local constraints (e.g., [13, 22]) have been proposed, but the main difference with our system lies in the fact that these systems either add a spatial similarity term to the verification step or use semi-local constraints as an outlier rejection step. Therefore, no attempt is made to *group* features based on semi-local constraints.

2 Local Image Descriptor

Local image descriptors suitable for local image data representation must have 3 properties: a) distinctiveness, b) detectability, and c) robustness to image deformations. In [2, 3], it is empirically shown that the multi-scale phase-based features are suitable for this task since they generally have those properties. Other possible choices were later evaluated in [12]. However, each local individual point exhibits different probability distributions for each property aforementioned. Our goal in this section is to estimate the parameters of these probability distributions for each feature, and use that information in the verification stage of the recognition algorithm. Furthermore, in order to reduce the size of the database of model features, we also use these distributions to select only the best features.

The features proposed in [3] are extracted using the following 2 steps:

- 'where' step: selects interest points that are robustly localizable under common image deformations forming the set of locations $\mathcal{I}_i = \{\boldsymbol{x}_l\}$. Notice that the locations are detected at the following set of wavelengths (in pixels): $\{\Lambda_o = 4(2^{i/4}), i = 0, 1, .., 12\}$;
- 'what' step: extracts a feature vector describing the image structure in the neighborhood of an interest point, say $\boldsymbol{f}_l = \boldsymbol{f}(\boldsymbol{x}_l) = [m_l, \theta_l, \sigma_l, \boldsymbol{v}_l]$. Here m_l is the model identification, θ_l is the main orientation of the location $\boldsymbol{x}_l$ (see [7]), $\sigma_l = \frac{\lambda_l}{4,26}$ represents the feature scale (we use σ_l and λ_l interchangeably), and $\boldsymbol{v}_l = \boldsymbol{\rho}_l e^{i\boldsymbol{\phi}_l}$ is the vector of amplitudes ρ and phases ϕ of bandpass filter responses.

The features extracted from an image I_i is then represented by $\mathcal{O}_i = \{\boldsymbol{f}(\boldsymbol{x}_l) | \boldsymbol{x}_l \in \mathcal{I}_i\}$. The similarity between local features is computed using normalized phase correlation [6], as follows:

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

where $\cdot$ means dot product, and $\boldsymbol{v}_o^*$ is the complex conjugate of $\boldsymbol{v}_o$.

2.1 Feature Probability Distributions

First we describe the probability distribution for robustness $P_{\text{on}}(s(\boldsymbol{f}_l, \boldsymbol{f}_o); \boldsymbol{f}_l)$, i.e., the probability of observing phase correlation $s(\boldsymbol{f}_l, \boldsymbol{f}_o)$ given that the feature $\boldsymbol{f}_o$ is a true match for the feature $\boldsymbol{f}_l$, and distinctiveness $P_{\text{off}}(s(\boldsymbol{f}_l, \boldsymbol{f}_o); \boldsymbol{f}_l)$, i.e., the probability of observing phase correlation $s(\boldsymbol{f}_l, \boldsymbol{f}_o)$ given that the feature $\boldsymbol{f}_o$ is a false match for the feature $\boldsymbol{f}_l$. The detectability $P_{\text{det}}(\boldsymbol{x}_l)$ is the probability that an interest point is detected in the test image at the same object neighborhood location $\boldsymbol{x}_l$ of feature $\boldsymbol{f}_l$. In order to learn the distributions of each model feature, assume that we have a pool of training images $\{I_i\}_{i \in \{1,\ldots,n\}}$, and that we break that pool into 2 subsets, namely object and random images. The object images are image regions where an object can be found, and it is represented by images $\{I_i\}_{i=1}^{q}$, where $q < n$ is the number of objects, while random images are the remaining ones in the pool (i.e., $\{I_i\}_{i=q+1}^{n}$). The image deformations $\mathcal{DF}$ described in appendix A, when applied to an image I_i, produces $\tilde{I}_{i,d}$, where its interest points are transformed from $\mathcal{I}_i$ to $\tilde{\mathcal{I}}_{i,d}$, and its feature set from $\mathcal{O}_i$ to $\tilde{\mathcal{O}}_{i,d}$. The following databases and functions will be necessary hereafter:

– top k correspondences between a feature $\boldsymbol{f}_l \in \mathcal{O}_i$ and a database of features $\mathcal{O}_j$ in terms of phase correlation defined in (1): $\mathcal{K}(\boldsymbol{f}_l, \mathcal{O}_j, k) = \{\boldsymbol{f}_o \in \mathcal{O}_j\}_{o=1}^{k}$;
– correspondences between interest point maps given a known transformation d: $\mathcal{C}_{i,d} = \{(\boldsymbol{x}_l, \boldsymbol{x}_o) | \boldsymbol{x}_l \in \mathcal{I}_i, \boldsymbol{x}_o \in \tilde{\mathcal{I}}_{i,d}, \|\boldsymbol{x}_o - M(d)\boldsymbol{x}_l - \boldsymbol{b}(d)\| < \epsilon\}$, where $M(d)$ is the rotation/scale/shear, $b(d)$ is the translation suffered by I_i to produce $\tilde{I}_{i,d}$, and ϵ is an arbitrary constant (here, we consider $\epsilon = 2$);
– k nearest neighbors given a threshold τ_s for phase correlation: $\mathcal{N}_{ij} = \{(\boldsymbol{f}_l, \tilde{\boldsymbol{f}}_l) | \boldsymbol{f}_l \in \mathcal{O}_i, \tilde{\boldsymbol{f}}_l \in \mathcal{K}(\boldsymbol{f}_l, \mathcal{O}_j, k), s(\boldsymbol{f}_l, \tilde{\boldsymbol{f}}_l) > \tau_s\}$
– top feature correspondence of $\boldsymbol{f}_l \in \mathcal{O}_i$ given an interest map correspondence $(\boldsymbol{x}_l, \boldsymbol{x}_o) \in \mathcal{C}_{i,d}$: $c(\boldsymbol{f}_l, \tilde{\mathcal{O}}_{i,d}) = \{\boldsymbol{f}_o | \boldsymbol{f}_o \in \tilde{\mathcal{O}}_{i,d}, o = \arg\max_j \{s(\boldsymbol{f}_l, \boldsymbol{f}_j)\} | (\boldsymbol{x}_l, \boldsymbol{x}_j) \in \mathcal{C}_{i,d}$.

The $P_{\text{off}}(s, \boldsymbol{f}_l)$ of each feature $\boldsymbol{f}_l \in \mathcal{O}_i$ is computed from the histogram of phase correlations $\{s(\boldsymbol{f}_l, \boldsymbol{f}_o) | \boldsymbol{f}_o \in \mathcal{R}\}$, where $\mathcal{R} = \bigcup_{j=t+1}^{n} \mathcal{O}_j$ is the database of features from all random images. On the other hand, $P_{\text{on}}(s, \boldsymbol{f}_l)$ is computed from the histogram of phase correlations with respect to the set of image deformations $\mathcal{DF}$ (appendix A) between $\boldsymbol{f}_l \in \mathcal{O}_i$ and each $\tilde{\boldsymbol{f}}_{l,d} \in c(\boldsymbol{f}_l, \tilde{\mathcal{O}}_{i,d})$, forming $\{s(\boldsymbol{f}_l, \tilde{\boldsymbol{f}}_{l,d}) | d \in \mathcal{DF}\}$. These distributions can be adequately approximated by the *beta* parametric distribution,

$$P_\beta(x; a, b) = \begin{cases} \frac{1}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt} x^{a-1}(1-x)^{b-1} & \text{if } x \in (0,1) \text{ and } a, b > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

since this distribution is defined within the range $[0, 1]$ (i.e., the same range of phase correlation (1)), and, empirically, it represents a good fit to the robustness and distinctiveness distributions as shown in Fig. 1.

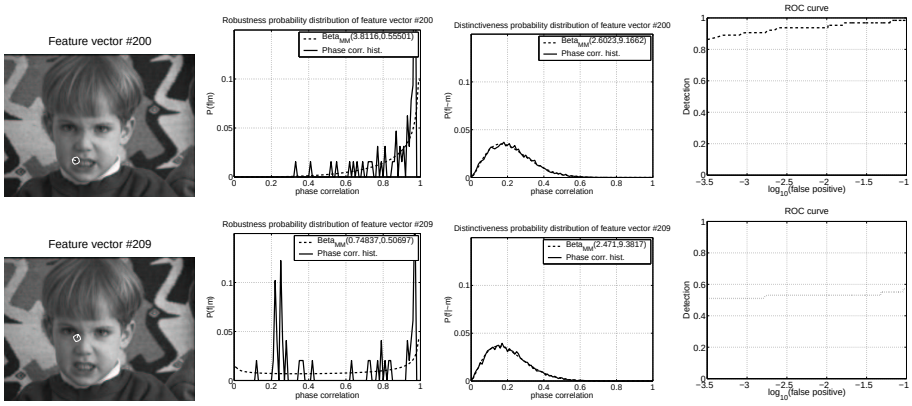

Fig. 1. Approximation of distinctiveness and robustness histograms using the beta function. On the rightmost column, we see the ROC curves of robustness vs. distinctiveness for the respective feature resulting from the histograms.

The method of moments (MM) provides a good estimate of the beta parameters a and b [21]. It is based on the first and second moments, namely μ_β and σ_β^2, of the

histograms for P_{off} and P_{on}. The parameters (a, b) of the fitted beta distribution are then

$$b = \frac{\mu_\beta(1-2\mu_\beta+\mu_\beta^2)}{\sigma_\beta^2} + \mu_\beta \quad \text{and } a = \frac{\mu_\beta b}{1-\mu_\beta}\,. \tag{3}$$

Fig. 1 shows two examples of the approximation of distinctiveness and robustness histograms using the beta distribution, and the last column displays the ROC curve resulting from the histograms. Fig. 2, 2^{nd} column, shows the mean and standard deviation of the ROC curves of all the features from the image of the object on the left. From this curve, we can determine the number of matches per image feature during a search procedure. For example, if our database has 10,000 features, we can expect to have the correct feature within the top 10 matches (0.1% of database size) with 75% to 95% of confidence, given that the feature is detected in both model and test images.

Finally, in order to determine P_{det} of a model feature position $\boldsymbol{x}_l$, where $\boldsymbol{f}_l \in \mathcal{O}_i$, we verify the number of times $c(\boldsymbol{f}_l, \tilde{\mathcal{O}}_{i,d}) \neq \emptyset$ with respect to the image deformation set $\mathcal{DF}$, producing the detectability probability:

$$P_{\text{det}}(\boldsymbol{x}_l) = \frac{\left|\bigcup_{d\in\mathcal{DF}} c(\boldsymbol{f}_l, \tilde{\mathcal{O}}_{i,d})\right|}{|\mathcal{DF}|} \tag{4}$$

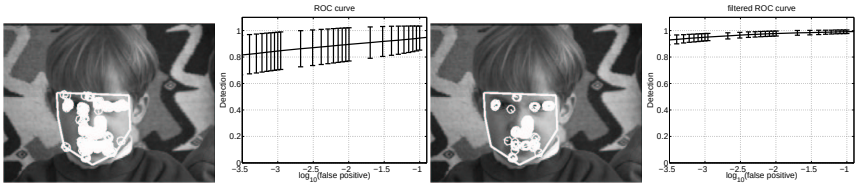

Fig. 2. ROC curve computed from all the features in the figure above. The 2^{nd} column shows the mean and standard deviation graph of the ROC curves computed from all the local descriptors at wavelength $\lambda_c = 8$ from the image shown in the 1^{st} column. The 4^{th} column shows the ROC curves with the points filtered by the procedure described in section 2.1. Notice the significant improvement in terms of robustness vs. distinctiveness, and also the reduction of the number of features detected.

2.2 Filtering Local Features

The number of features to be stored in the database of models can be reduced, diminishing the ambiguity in the database and improving the scalability of the system. This filtering consists of checking the following conditions: a) high robustness $a_{\text{on}}(\boldsymbol{f}) > \tau_{\text{on}} b_{\text{on}}(\boldsymbol{f})$ (distribution gets close to 1 with large τ_{on}); b) high distinctiveness $b_{\text{off}}(\boldsymbol{f}) > \tau_{\text{off}} a_{\text{off}}(\boldsymbol{f})$ (distribution gets close to 0 with large τ_{off}); and c) high detectability $P_{\text{det}}(\boldsymbol{x}) > p\%$. As a result, we obtain a subset of the interest points $\mathcal{I}_i^* \subseteq \mathcal{I}_i$. In Fig. 2, we see an example of the functionality of the filtering procedure, where $\tau_{\text{on}} = 7$, $\tau_{\text{off}} = 2$, and $p\% = 75\%$. Note the significant improvement of the ROC curve and the reduction of the number of features from 1.5% to 0.3% of total image size, which is roughly the same percentage produced in the approaches described in [10] and [11].

3 Semi-local Spatial Constraints

Spatial constraints of local image descriptors are important because they impose restrictions on the set of correspondences. These correspondences are built using only the feature similarities between the descriptors extracted from the test image and the ones from the model image. The use of spatial constraints allows for grouping descriptors that are likely to be part of the same model and for checking how well each of these groups matches the original model. There are 2 types of spatial constraints that can be envisaged for this task, namely global and semi-local. Here, we only explore the following semi-local constraints: pairwise relations and geometric predictions.

3.1 Pairwise Relations

The pairwise geometric relations 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 4):

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

(5)

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 $\mathbf{x}_o$.

Given the correspondences set $\mathcal{N}_{ij}$, 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

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

(6)

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:

$$\mathbf{A}(l, o) = \delta_{m_l m_o} \pi_{lo,g}\, g\left(\left[\Delta\mathcal{D}_{lo}(\mathcal{N}_{ij}), \Delta\mathcal{H}_{lo}(\mathcal{N}_{ij}), \Delta\mathcal{S}_{lo}(\mathcal{N}_{ij})\right]^T; \Sigma_\Delta\right), \qquad (7)$$

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, $\pi_{lo,g} = e^{-0.5\frac{\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o)}{\sigma_{\pi,g}^2}}$ 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 $\sigma_d^2 = \min(\kappa_{dist}, \max(p_{dist}\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o), 0.1))$ depends on the scaled original distance between model features $\mathbf{f}_l, \mathbf{f}_o \in \mathcal{O}_i$ (i.e., points that are far from each other in the model have a proportionally larger standard error for their relative distances).

3.2 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. Moreover, points that are close to the feature being predicted should have a higher influence than features far from it. In general, note that the following relations are true if the correspondence is correct: $\tilde{\mathbf{n}}_{lo}(\tilde{\mathbf{x}}_l - \tilde{\mathbf{x}}_o) \approx \|\mathbf{x}_l - \mathbf{x}_o\|$, 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 $\tilde{\sigma}_l - \tilde{\sigma}_o \approx \sigma_l - \sigma_o$. For position prediction, we build the linear system $(\mathbf{n}_{lo}\pi_{lo,p})\tilde{\mathbf{n}}_{lo} \cdot (\tilde{\mathbf{x}}_l^* - \tilde{\mathbf{x}}_o) = (\mathbf{n}_{lo}\pi_{lo,p})\|\mathbf{x}_l - \mathbf{x}_o\|$ 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 $\pi_{lo,p} = e^{-0.5\frac{\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o)}{\sigma_{\pi,p}^2}}$, 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 $\tilde{\theta}_l^* = \frac{1}{\sum_o \pi_{lo,p}} \sum_o \pi_{lo,p}(\theta_l - \vartheta_{lo} + \tilde{\vartheta}_{lo})$ and $\tilde{\sigma}_l^* = \frac{1}{\sum_o \pi_{lo,p}} \sum_o \pi_{lo,p}(\sigma_l - \sigma_o + \tilde{\sigma}_o)$, respectively.

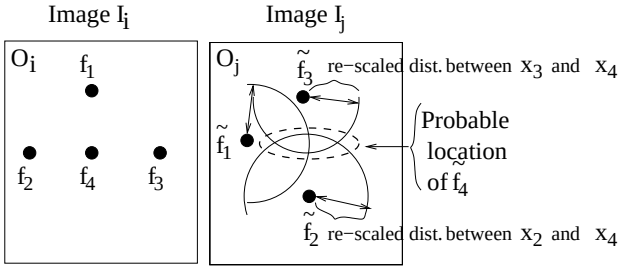

Fig. 3. Example of position prediction. Given the features of the model $\{\mathbf{F}_l\}$, including their positions, suppose we want to estimate the position of $\tilde{\mathbf{f}}_4$. The probable location of the feature (represented by a dotted ellipsoid in the Figure) is based on a Gaussian distribution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

The similarity between prediction and the observed position, main orientation and scale is computed as follows (see Fig. 3): $p(\mathbf{f}_l, \tilde{\mathbf{f}}_l) = g([\tilde{\mathbf{x}}_l, \tilde{\theta}_l, \tilde{\sigma}_l] - [\tilde{\mathbf{x}}_l^*, \tilde{\theta}_l^*, \tilde{\sigma}_l^*]; \Sigma_t)$, 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

$$\Sigma_D(\tilde{\mathbf{f}}_l) = \frac{1}{\sum_p \pi_{lo,p}} \left[B \left[\begin{array}{c} \vdots \\ \sigma_{\mathcal{D}}(\mathbf{f}_l, \mathbf{f}_o) \\ \vdots \end{array} \right] [\cdots \sigma_{\mathcal{D}}(\mathbf{f}_l, \mathbf{f}_o) \cdots] \, B^T \right],$$

where

$$\mathbf{B} = (\mathbf{A}^T \Pi \mathbf{A})^{-1} \mathbf{A}^T \Pi, \;\; \mathbf{A} = [\cdots \mathbf{n}_{lo} \cdots], \Pi = \left[\begin{array}{ccc} \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \pi_{lo,p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots \end{array} \right].$$

Also, the variances of the heading and scale estimates are
$\sigma^2_{\mathcal{H}}(\tilde{\mathbf{f}}_l) = \frac{1}{\sum_o \pi_{lo,p}} \sum_o \pi_{lo,p} \sigma^2_{\mathcal{H}}(\mathbf{f}_l, \mathbf{f}_o)$, and $\sigma^2_{\mathcal{S}}(\tilde{\mathbf{f}}_l) = \frac{1}{\sum_o \pi_{lo,p}} \sum_o \pi_{lo,p} \sigma^2_{\mathcal{S}}(\mathbf{f}_l, \mathbf{f}_o)$. The pairwise variances $\sigma_{\mathcal{D}}(\mathbf{f}_l, \mathbf{f}_o)$, $\sigma_{\mathcal{H}}(\mathbf{f}_l, \mathbf{f}_o)$, and $\sigma_{\mathcal{S}}(\mathbf{f}_l, \mathbf{f}_o)$ are estimated by the sample variances obtained by deforming I_i with the set of deformations $\mathcal{DF}$ defined in appendix A.

4 Grouping Based on Pairwise Relations

The set of correspondences formed from a typical search procedure (e.g., nearest neighbor) generates a large hypothesis space for the recognition system, where techniques like RANSAC [19] would be a poor choice (as noted in [10]) due to the extremely low ratio between inliers and outliers. This issue is rarely addressed in object recognition systems based on complex local features with the exception of [10], 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 (global spatial constraint) with large bin sizes selected to accommodate other spatial deformations. Because of these bin assignments, 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 non-rigid deformation, which aims at reducing the number of groups, where each group has a high number of inliers. This approach involves connected component analysis on an affinity matrix based on the pairwise relations described in (5). Given the correspondences $\mathcal{N}_{ij}$ between $\mathcal{O}_i$ and $\mathcal{O}_j$, we proceed as follows (see Fig. 4):

1. Build the affinity matrix based on the pairwise similarity measures $A(l, o)$ as defined in (7).
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. 4). 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 (7).

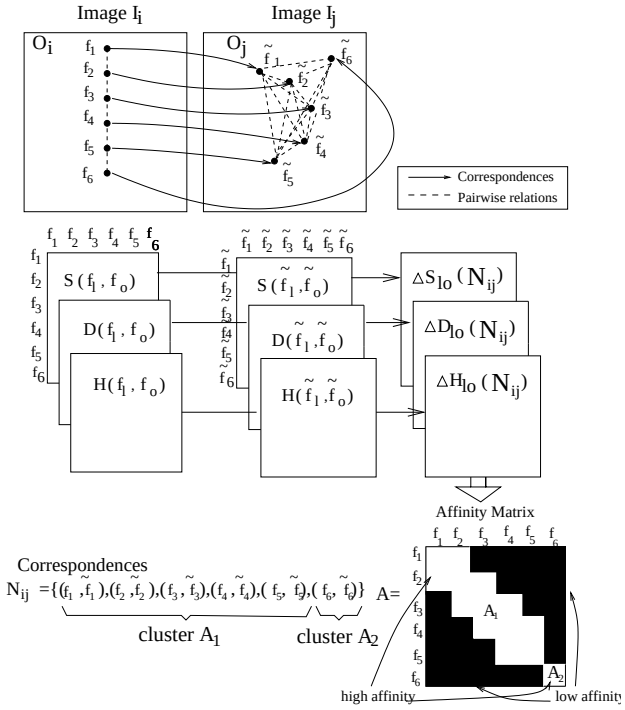

Fig. 4. Grouping based on pairwise relations. Notice in the figure that correspondences $1-5$ are semi-locally connected, while correspondence 6 is not. Therefore, we form 2 clusters.

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 3.2, 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 section 3.1). The parameters for our grouping method are $\sigma_h^2 = 0.1$, $\sigma_s^2 = 0.1$, $\kappa_{dist} = 1$, $p_{dist} = 0.1$, and $\sigma_{\pi,g} = \max(M/5, 10)$, where M is the maximum model diameter. The parameters for the geometric prediction are: $\tau_p = 10^{-32}$, and $\sigma_{\pi,p} = \max(M/50, 5)$.

For Hough clustering, we used the same parameters described in [10], 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 avoid 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 Figs. 5-8, where the model image is always presented on the right image, while the left image presents the test image. The table titled

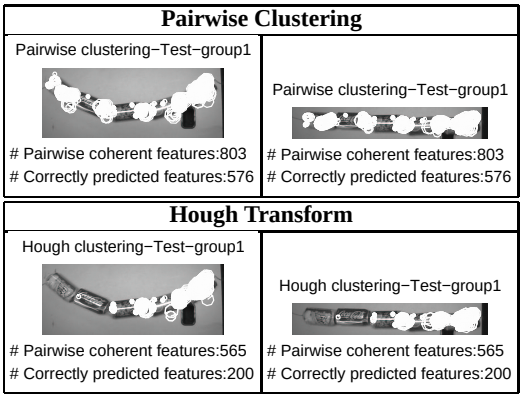

Fig. 5. Comparison between our grouping method and Hough clustering. The highlighted circles represent the feature correspondences that were grouped together by the respective method between the test image on the left, and the model image on the right. Note that while almost all features of the articulated object can be clustered in the same group using our method, Hough clustering can only group features that suffered a rough rigid deformation.

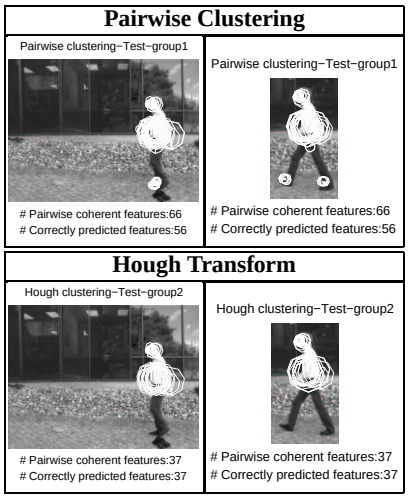

Fig. 6. Second example of our grouping method. Note that our method is able to cluster the feet features of the model in the same group as the upper body features. Since Hough transform assumes a rough rigid deformation, it again fails to place the feet features in the same group as the upper body features.

'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 also describe the total number of features inside the group, and the number of those features considered to be correctly predicted (i.e., features such that $p(\mathbf{f}_l, \tilde{\mathbf{f}}_l) > \tau_p$). For all the cases, we only show the group that has the highest number of features.

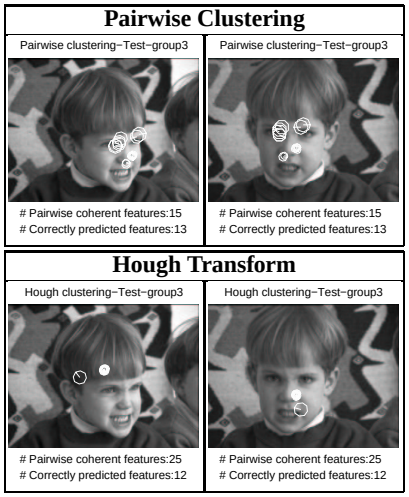

Fig. 7. Example of an extreme non-rigid deformation. Note that our method is capable of clustering the features in the same group while Hough clustering fails.

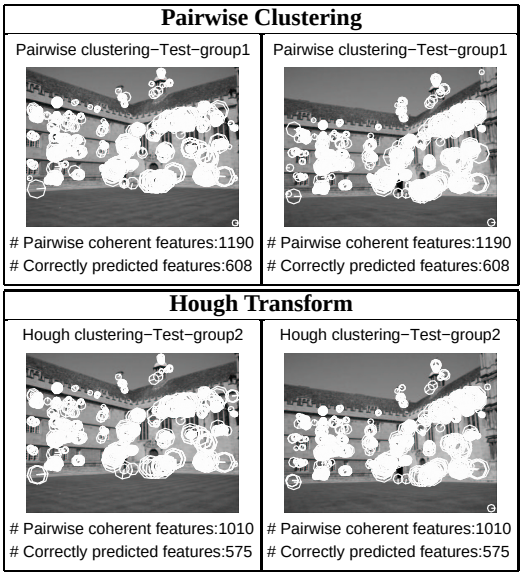

Fig. 8. Long range motion problem. Here, we compare our method with Hough clustering given a rigid deformation. Note that both methods have similar performances. However, according to our method of computing correctly predicted points, our method produces a higher percentage of inliers (see Table 1).

Fig. 5 shows the results for the grouping method proposed here where. The model is an object composed of a string built with soda cans (see the model 'snake of cans' in Fig. 9). This example shows the robustness of our method to deformations given by

articulated objects. Note that the Hough transform only matches a piece of the object that suffered a deformation that is close to a rigid transformation.

Fig. 6 shows another example with an articulated object. Specifically, given the model 'hedvig' in Fig. 9, we want to check if the semi-local spatial constraints are capable of dealing with the non-rigid deformations of a person walking. 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 feet in the same group as the upper part of her body.

We also show in Fig. 7 the robustness of our method non rigid deformation of a single body. The model is Kevin's face (see 'kevin' in Fig. 9), and the test image suffers a significant rotation in depth. Notice that Hough transform is unable to cluster the face's features in the same group as done by our method.

In order to show the effectiveness of our approach with respect to rigid deformation, we considered the long range motion problem. In this problem, we considered the groups formed by our approach and Hough transform to compute the **F** matrix [8]. We use RANSAC [19] 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. We also compute the number of trials necessary to make the probability $p < 0.05$ of choosing at least 1 outlier in every trial of the RANSAC algorithm for t trials. As a result, we want to have a small number of trials t so that **F** can be computed quickly. Therefore, $t = \log_2(0.05)/\log_2(1 - (\frac{in}{in+out})^8)$, where in is the number of inliers and out is the number of outliers. The correspondences used for this experiment are shown in Fig. 8. Table 1 shows the results of this experiment, where the first row 'Geometric prediction' shows the values for this experiment where the correspondences are formed by the the features correctly predicted as presented in section 3.2. The second row 'Pairwise relations' shows the results using the set of correspondences formed by the grouping method that uses the pairwise constraints described in section 3.1. Finally, the third row shows the results for the Hough clustering described in this section. In general, the percentage of inliers is higher (and consequently, t is smaller) for the correspondence set using the correctly predicted features than the correspondences formed by either the pairwise relations or the Hough clustering. Also, it is worthwhile to note that the same experiment was run with several other cases, and the results obtained were similar to the one described here.

Table 1. Comparison between our method and Hough clustering for computing the **F** matrix.

Grouping method	$\frac{in}{in+out}$	in	$in+out$	t
Geometric prediction	89%	543	608	5.77
Pairwise relations	70%	837	1190	48.49
Hough clustering	73%	740	1010	34.55

Finally, it is worth noting that the time complexity of our pairwise clustering algorithm is $O(n^2)$, where n is the maximum number of correspondences between features in the test image and features in a single model, and for Hough clustering, the complex-

ity is $O(\#bins)$. The running time of our algorithm is comparable to Hough transform when the relation $\#bins \approx n^2$ is true, and that happens to be the case with the configuration proposed in [10], so both grouping algorithms exhibited comparable running times.

5 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 it also checks for semi-local spatial configuration similarities. The object recognition method can be divided into the training and testing modes.

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 section 2.2), 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 object $\mathcal{O}_m$. From the computation of the affinity matrix (7), we know that all the features clustered in the same group belong to 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):

$$P(\mathcal{O}_m|\mathcal{E}_g, T) = \frac{P(\mathcal{E}_g|T, \mathcal{O}_m)P(T|\mathcal{O}_m)P(\mathcal{O}_m)}{P(\mathcal{E}_g|T, \mathcal{O}_m)P(T|\mathcal{O}_m)P(\mathcal{O}_m) + P(\mathcal{E}_g|T, \neg\mathcal{O}_m)P(T|\neg\mathcal{O}_m)P(\neg\mathcal{O}_m)}, \tag{8}$$

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 (8). The probabilistic formulation, based on [15], 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) $(\emptyset, \mathbf{f}_o) \in \mathcal{E}_g$:

$$P((\emptyset, \mathbf{f}_o) \in \mathcal{E}_g|T, \mathcal{O}_m) \approx (1 - P_{\text{det}}(\mathbf{x}_o)) + P_{\text{det}}(\mathbf{x}_o)P_{\text{on}}(s < \tau_s; \mathbf{f}_o), \tag{9}$$

   (b) $(\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g, [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]$:

$$\begin{aligned} &P((\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g|T, \mathcal{O}_m) = \\ &\quad P((\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g \text{ and} \\ &\qquad [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]|T, \mathcal{O}_m) \approx \\ &\quad P_{\text{det}}(\mathbf{x}_o)P_{\text{on}}(s(\mathbf{f}_k, \mathbf{f}_o); \mathbf{f}_o)p(\mathbf{f}_k, \mathbf{f}_o) \end{aligned} \tag{10}$$

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) $(\emptyset, \mathbf{f}_o) \in \mathcal{E}_g$:

$$P((\emptyset, \mathbf{f}_o) \in \mathcal{E}_g|T, \neg\mathcal{O}_m) \approx (1 - 0.015) + 0.015(1 - P_{\text{off}}(s(\mathbf{f}, \mathbf{f}_o) < \tau_s; \mathbf{f}_o)), \tag{11}$$

   where the number 0.015 represents the average number of interest points per test image divided by the size of the image (see [3]);
   (b) $(\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g, [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]$

$$\begin{aligned} &P((\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g|T, \neg\mathcal{O}_m) = \\ &\quad P((\mathbf{f}_k, \mathbf{f}_o) \in \mathcal{E}_g \text{ and} \\ &\qquad [\mathbf{x}_k^*, \theta_k^*, \sigma_k^*] = [\mathbf{x}_k, \theta_k, \sigma_k]|T, \mathcal{O}_m) \approx \\ &\quad (0.015) P_{\text{off}}(s(\mathbf{f}_k, \mathbf{f}_o); \mathbf{f}_o) \frac{1}{size(\mathcal{I})} \frac{1}{8} \frac{1}{2\pi}. \end{aligned} \tag{12}$$

   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).

6 Results

We considered a database of 15 objects shown in Fig. 9, and we use the same parameter values as described in section 4. 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 $\tau_{\mathcal{D}} = 20\%$ of the maximum model diameter. Our database has roughly 10,000 features, which were extracted from the objects in Fig. 9 during the learning stage. Our tests (see Figs. 10-13) 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. 14, 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 [3], but notice that these parameters are not used for verification.

7 Conclusions

We presented a new object recognition system based on higly distinctive image descriptors extracted from robustly detectable local features. A main issue in these type of systems is that the search for similar features in the database of models usually returns a relatively large set of correspondences where the number of inliers is very small.

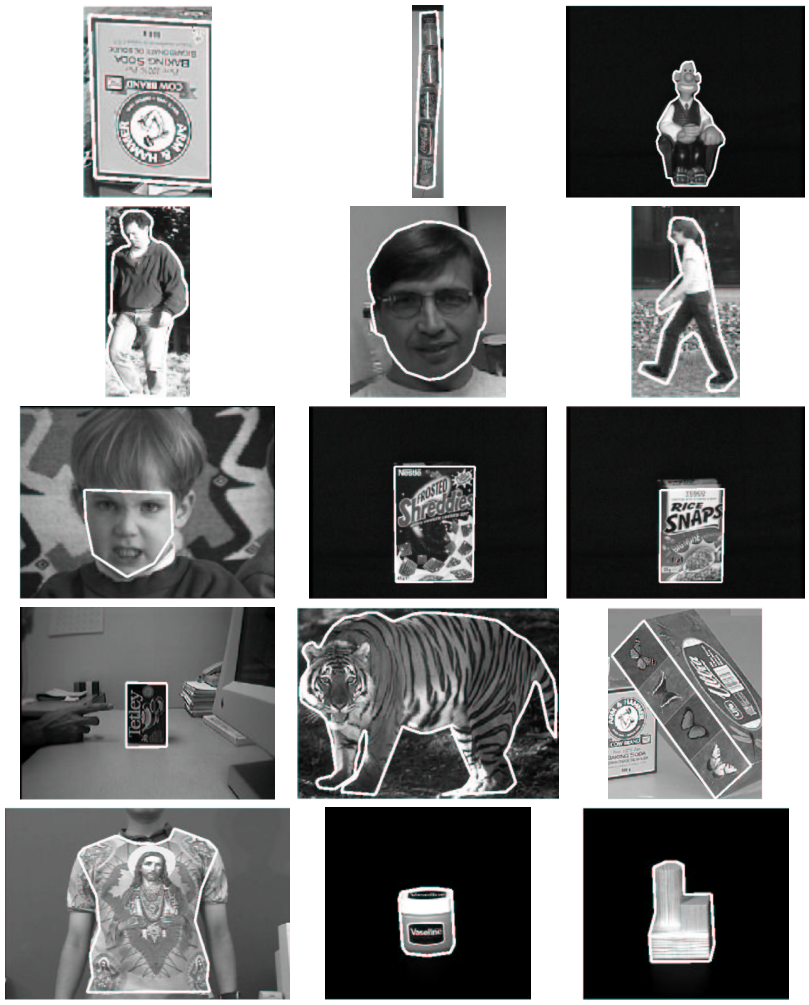

Fig. 9. Model database for object recognition. All the models are represented only by the features inside the white line around the object of interest.

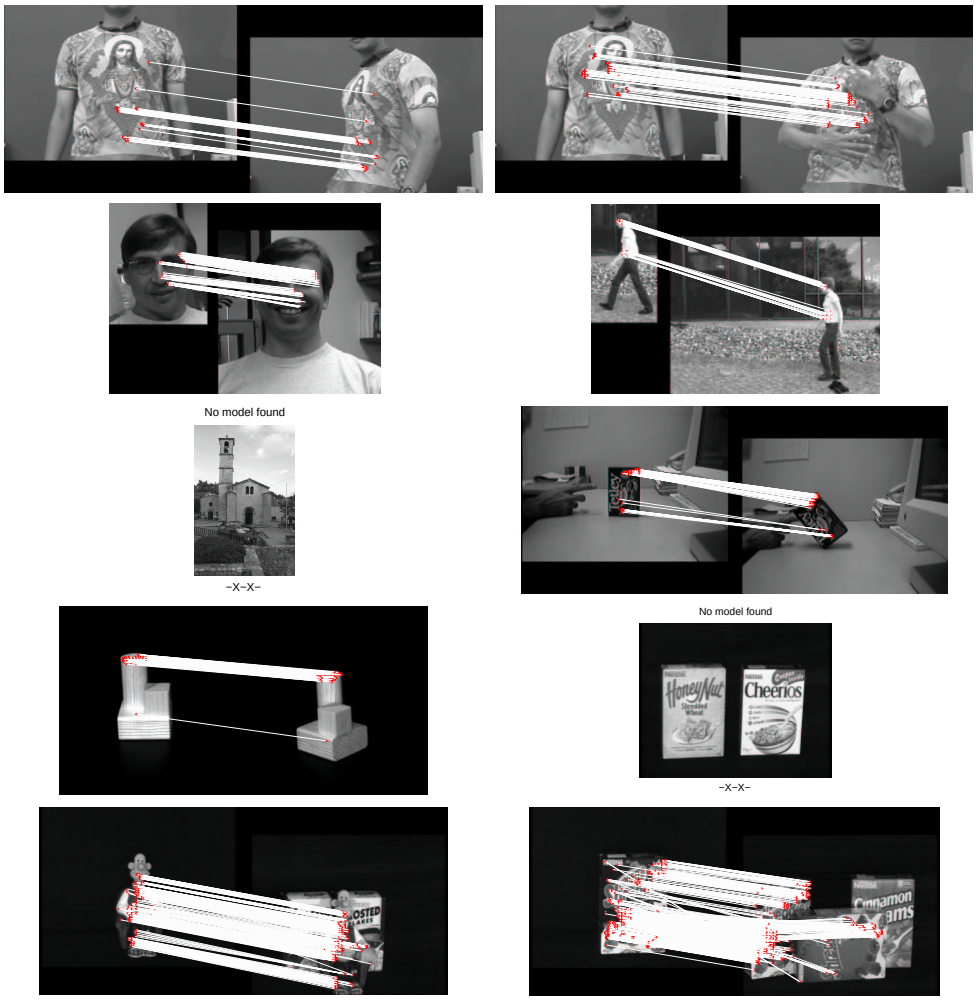

Fig. 10. Recognition results. The white lines between the model image on the left and the test image on the right show the correspondences used by the verification stage. Notice the ability of this system in dealing with clutter, rigid/non-rigid/illuminiation deformations, and partial occlusion. The system also demonstrates to be robust to false positives that are similar to the some of the models present in the database.

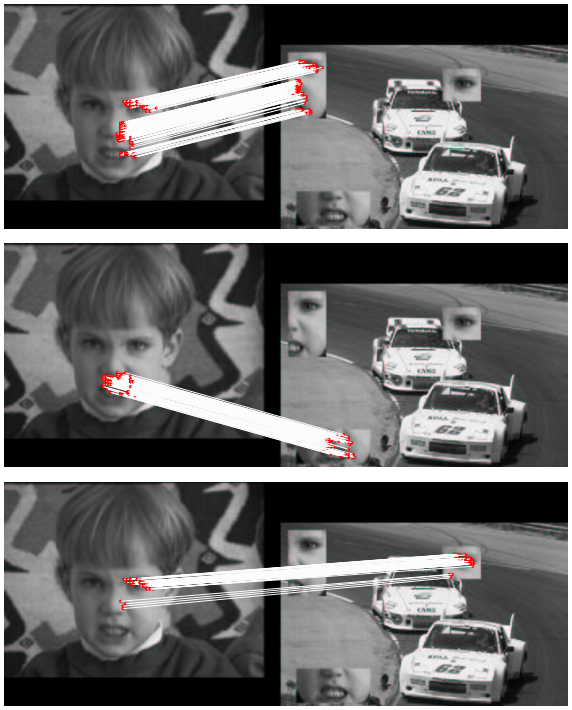

Fig. 11. Recognition results for a test image with multiple instances of the same object with severe partial occlusion.

Fig. 12. Recognition results for the model image suffering a large non-rigid deformation. Notice that the false positive detected seems to be coherent with the model.

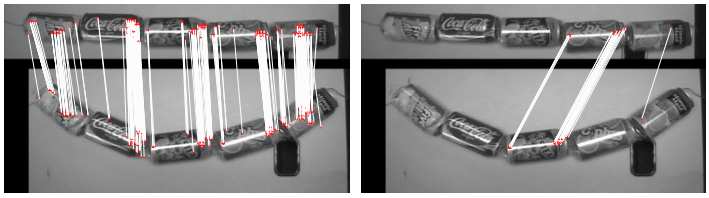

Fig. 13. Recognition results on an articulated model. Note that the false positive detected is reasonable with the model.

Fig. 14. 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 [3].

We address this problem by proposing two things: a) a new grouping procedure during hypothesis generation which reduces the number of hypotheses to investigate during the subsequent verification process, and also b) a novel filtering method that reduces the database of model features based on the empirical distinctiveness, detectability and robustness of local features. The new grouping method based on flexible pairwise measurements proves to be robust to both rigid and non-rigid deformations and, in general, produces fewer groups, where each group has more inliers than the Hough transform (the commonly used alternative method). The filtering procedure achieves a five fold reduction of the database of models by keeping only the most effective features. Finally the viability of this system is demonstrated in an object recognition system based on the local features described in [3] and their spatial configuration within a probabilistic framework. Even though only 1 image of the object is used during the training stage of the system, it is able to recognize that same object under significant image deformations (e.g., rigid, non-rigid, occlusion, clutter, and illumination changes).

References

1. Y. Amit and D. Geman. A computational model for visual selection. *Neural Computation*, 11:1691–1715, 1999.
2. G. Carneiro and A. Jepson. Phase-based local features. In *European Conference on Computer Vision*, pages 282–296, Copenhagen, Denmark, May 2002.
3. 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.
4. T.F. Cootes, G.F. Edwards, and C.J. Taylor. Active appearance models. In *European Conference on Computer Vision*, 2002.

5. 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.
6. D. Fleet. *Measurement of Image Velocity*. Kluwer Academic Publishers, 1992.
7. 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.
8. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
9. 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.
10. D. Lowe. Local feature view clustering for 3d object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.
11. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *IEEE International Conference on Computer Vision*, pages 525–531, Vancouver, Canada, July 2001.
12. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, June 2003.
13. K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. In *British Conference on Computer Vision*, 2003.
14. K. Ohba and K. Ikeuchi. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1043–1048, 1997.
15. A. Pope and D. Lowe. Probabilistic models of appearance for 3d object recognition. *International Journal of Computer Vision*, 40(2):149–167, 2000.
16. 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.
17. C. Schmid. A structured probabilistic model for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 485–490, 1999.
18. 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.
19. 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.
20. M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV (1)*, pages 18–32, 2000.
21. J. Wu. Bayesian estimation of stereo disparity from phase-based measurements. Master's thesis, Queen's Univeristy, Kingston, Ontario, Canada, 2000.
22. 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.

A Image Deformations Studied

The image deformations described in this section are used to compute probability distributions of local feature descriptors and of their geometric configurations. The set of image deformations $\mathcal{DF} = \{d\}$ considered here are (see [3]): a) two types of global brightness changes, b) non-uniform local brightness variations, c) additive noise, d) scale changes, e) 2D rotation, f) shear and g) sub-pixel translation. The non-uniform global

brightness changes are implemented by adding a constant to the brightness value, taking into account the gamma correction non-linearity: $\tilde{I}_d(\boldsymbol{x}) = 255 * \left[\max\left(0, \left(\frac{I(\boldsymbol{x})}{255}\right)^{\gamma} + k\right)\right]^{\frac{1}{\gamma}}$, where $\gamma = 2.2$, I is the original image, and $k \in [-.5, .5]$ controls the changes in brightness. The resulting image is linearly mapped to values between 0 and 255, and then quantized. The uniform brightness change is simply based on the division of gray values by a constant $c \in [1, 3]$.

For the non-uniform local brightness variations, a highlight at a specific location of the image is simulated by adding a Gaussian blob as follows: $\tilde{I}_d(\boldsymbol{x}) = I(\boldsymbol{x}) + 255 * g(\boldsymbol{x} - \boldsymbol{x}_0; \sigma)$, where $\sigma = 10$, $\boldsymbol{x}_0$ is a specific position in the image, and $g(\boldsymbol{x}; \sigma) = \exp\left(-x^2/(2\sigma^2)\right)$. Again, the resulting image is mapped to values between 0 and 255, and then quantized. For noise deformations, we simply add Gaussian noise with varying standard deviation ($\sigma = 255 * [10^{-3}, 10^{-1}]$), followed by normalization and quantization, as above. The geometric deformations are 2D rotations (from -90^o to $+90^o$ in intervals of 15^o), uniform scale changes (with expansion factors in the range $[0.25, 1]$), shear in the horizontal direction (so that a vertical line is perturbed by $\pm 26^o$), and sub-pixel translation (in the range [0,1]) pixel. The geometrically deformed images are quantized to $[0, 255]$ without normalization.