Simultaneous Object Recognition and Segmentation by Image Exploration*

Vittorio Ferrari[1], Tinne Tuytelaars[2], and Luc Van Gool[1,2]

[1] Computer Vision Group (BIWI), ETH Zuerich, Switzerland
{ferrari,vangool}@vision.ee.ethz.ch
[2] ESAT-PSI, University of Leuven, Belgium
Tinne.Tuytelaars@esat.kuleuven.ac.be

Abstract. Methods based on local, viewpoint invariant features have proven capable of recognizing objects in spite of viewpoint changes, occlusion and clutter. However, these approaches fail when these factors are too strong, due to the limited repeatability and discriminative power of the features. As additional shortcomings, the objects need to be rigid and only their approximate location is found. We present a novel Object Recognition approach which overcomes these limitations. An initial set of feature correspondences is first generated. The method anchors on it and then gradually explores the surrounding area, trying to construct more and more matching features, increasingly farther from the initial ones. The resulting process covers the object with matches, and simultaneously separates the correct matches from the wrong ones. Hence, recognition and segmentation are achieved at the same time. Only very few correct initial matches suffice for reliable recognition. The experimental results demonstrate the stronger power of the presented method in dealing with extensive clutter, dominant occlusion, large scale and viewpoint changes. Moreover non-rigid deformations are explicitly taken into account, and the approximative contours of the object are produced. The approach can extend any viewpoint invariant feature extractor.

1 Introduction

Recently, object recognition (OR) approaches based on local invariant features have become increasingly popular [8,5,2,4,7]. Typically, local features are extracted independently from both a model and a test image, then characterized by invariant descriptors and finally matched. The success of these approaches is twofold. First, the feature extraction process and description are viewpoint invariant. Secondly, local features bring tolerance to clutter and occlusion, de facto removing the need for prior segmentation. In this respect, global methods, both contour-based [9] and appearance-based [10], are a step behind.

In spite of their success, the robustness and generality of these approaches are limited by the repeatability of the feature extraction, and the difficulty of

* This research was supported by EC project VIBES and the Fund for Scientific Research Flanders.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 40–54, 2004.

matching correctly, in the presence of large amounts of clutter and challenging viewing conditions. Large scale or viewpoint changes considerably lower the probability that any given model feature is re-extracted in the test image (e.g.: figure 2, left). Simultaneously, occlusion reduces the number of visible model features. The combined effect is that only a small fraction of model features has a correspondence in the test image. This fraction represents the maximal number of features that can be correctly matched. Unfortunately, at the same time extensive clutter gives rise to a large number of non-object features, which disturb the matching process. As a final outcome of these combined difficulties, only a few, if any, correct matches are produced. Because these often come together with many mismatches, recognition tends to fail.

Even in easier cases, to suit the needs for repeatability in spite of viewpoint changes, only a sparse set of *distinguished* features [7] are extracted. As a result, only a small portion of the object is typically covered with matches. Densely covering the visible part of the object is desirable, as it increases the *evidence* for its presence, which results in higher discriminative power.

In this paper, we face these problems by no longer relying solely on matching viewpoint invariant features. Instead, we propose to anchor on an initial set thereof, and then *look around* them trying to construct more matching features. As new matches arise, they are exploited to construct even more, in a process which gradually *explores* the test image, recursively constructing more and more matches, increasingly farther from the initial ones. As the number and extent of matched features increases, so does the information available to judge their individual correctness. Gradually the system's confidence in the presence of the object grows.

We build upon a multi-scale extension of the affine invariant region extractor of [2]. An initial large set of unreliable region correspondences is generated through a process tuned to maximize the amount of correct matches, at the cost of producing many mismatches (section 2). Additionally, we generate a grid of circular regions homogeneously covering the model image. The core of the method iteratively alternates between *expansion* phases, where correspondences for these coverage regions are constructed, and *contraction* phases, which attempt to remove mismatches. In the first expansion phase (section 3), we try to propagate the coverage regions based on the geometric transformation of nearby initial matches. By *propagating* a region, we mean constructing the corresponding one in the test image. The propagated matches and the initial ones are then passed through a novel local filter, during the first contraction phase (section 4). The processing continues by alternating faster expansion phases (section 5), where coverage regions are propagated over a larger area, with contraction phases based on a global filter (section 6). The filter exploits both topological arrangements and appearance information, and tolerates *non-rigid deformations*. During the expansion phases, the shape of each new region is adapted to the local surface orientation, thus allowing the exploration process to follow curved surfaces and deformations (e.g. a folded magazine). At each iteration, the presence of the newly propagated matches helps the filter to take better removal decisions. In turn, the cleaner set of supports makes the next expansion more effective. As a

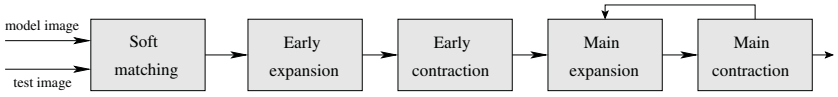


Fig. 1. Scheme of the system

result, the amount, and the percentage, of correct matches grows every iteration. The algorithm is getting a clearer idea about the object's presence and location. The two closely cooperating processes of expansion and contraction gather more evidence about the presence of the object and separate correct matches from wrong ones *at the same time*. This results in the simultaneous recognition and segmentation of the object. By constructing matches for the coverage regions, the system succeeds in covering also image areas which are not interesting for the feature extractor or not discriminative enough to be correctly matched by traditional techniques.

The basic advantage of the approach is that each single correct initial match can expand to cover a contiguous surface with *many* correct matches, even when starting from a large amount of mismatches. This leads to filling the visible portion of the object with matches. Some interesting *direct* advantages derive from it. First, robustness to scale, viewpoint, occlusion and clutter are greatly enhanced, because most cases where the traditional approach generated only a few correct matches are now solvable. Second, discriminative power is increased, because decisions about the object's identity are based on information densely distributed over the entire portion of the object visible in the test image. Third, the approximate boundary of the object in the test image is directly suggested by the final set of matched regions (section 8). Fourth, non-rigid deformations are explicitly taken into account.

2 Soft Matches

The feature extraction algorithm [2] is applied to both a *model image* I_m and a *test image* I_t independently, producing two sets of regions Φ_m, Φ_t.

Tentative Matches

For each test region $T \in \Phi_t$ we compute the Mahalanobis distance of the invariant descriptors [2] to all model regions $M \in \Phi_m$. An appearance similarity measure $\text{sim}(T, M)$ is computed between T and each of the 10 closest regions. The measure is a linear combination of grey-level normalized cross-correlation (NCC) and the average Euclidean distance in RGB space, after geometric and photometric normalization. This mixture is more discriminant than NCC alone, while keeping invariance to brightness changes. We consider each of the 3 most similar regions above a low threshold t_1. Repeating this operation for all regions $T \in \Phi_t$, yields a first set of *tentative matches*. At this point, every test region could be matched to either none,1,2 or 3 model regions.

Refinement and Re-thresholding

Since all regions are independently extracted from the two images, the geometric registration of a correct match might not be optimal, which lowers its similarity. The registration of the tentative matches is *refined* using our recently proposed algorithm [1], that efficiently looks for the affine transformation that maximizes the similarity. After refinement, the similarity is re-evaluated and only matches scoring above a second, higher threshold t_2 are kept. Refinement tends to raise the similarity of correct matches much more than that of mismatches. The increased *separation* between the similarity distributions makes the second thresholding more effective.

The obtained set of matches usually still contains *soft-matches*, i.e. more than one region in Φ_m corresponding to the same region in Φ_t, or vice-versa. This contrasts with classic matching methods [7,2,5,11,8], but there are two good reasons for it. First, the scene might contain repeated, or visually similar elements. Secondly, large viewpoint and scale changes cause loss of resolution which results in a less accurate correspondence and a lower similarity. When there is also extensive clutter, it might be impossible, based *purely* on local appearance [14], to decide which of the top-3-matches is correct, as several competing regions might appear very similar, and score higher than the correct match.

The proposed process outputs a large set of plausible matches, all with a reasonably high similarity. The goal is to maximize the amount of correct matches, even at the cost of accepting a substantial fraction of mismatches. In difficult

Fig. 2. Left: case-study (top: model image, bottom: test image). Middle: a closer view with 3 initial matches. The two model regions on the left are both matched to the same region in the test image. Note the small occluding rubber on the spoon. Right-top: the homogeneous coverage Ω. Right-bottom: a support region (dark), associated sectors (lines) and candidates (bright)

cases this is important, as each correct match can start an expansion which will cover significant parts of the object.

Figure 2 shows a case-study, for which 3 correct matches out of 217 are found (a *correct-ratio* of 3/217). The large scale change (factor 3.3), combined with the modest resolution (720x576), causes heavy image degradation which corrupts edges and texture. In such conditions only a few model regions are re-extracted and many mismatches are inevitable. In the remainder of the paper, we refer to the current set of matches as the *configuration* Γ.

How to proceed ? Global, robust geometry filtering methods, like detecting outliers to the epipolar geometry through RANSAC [3] fail, as they need a minimal amount of inliers of about 30% [8]. Initially, this may very well not be the case. Even if we could separate out the few correct matches, they would not be sufficient to draw reliable conclusions about the presence of the object. In the following we explain how to gradually increment the number of correct matches and simultaneously decrease the number of mismatches.

3 Early Expansion

Coverage of the Model Image

We generate a grid Ω of overlapping circular regions densely covering the model image I_m (figure 2, top-right). The expansion phases will try to construct in I_t as many regions corresponding to them as possible.

Propagation Attempt

We now define the concept of *propagation attempt* which is the basic building-block of the expansion phases and will be used later. Consider a region C_m in model image I_m without match in the test image I_t and a nearby region S_m, matched to S_t. If C_m and S_m lie on the same physical facet of the object, they will be mapped to I_t by similar affine transformations. The *support* match (S_m, S_t) *attempts to propagate* the *candidate* region C_m to I_t as follows:

1. Compute the affine transformation A mapping S_m to S_t.
2. Project C_m to I_t via $A : C_t = AC_m$.

The benefits of exploiting previously established geometric transformations was also noted by [13].

Early Expansion

Propagation attempts are used as follows. Consider as supports $\{S^i = (S^i_m, S^i_t)\}$ the soft-matches configuration Γ, and as candidates Λ the coverage regions Ω. For each support region S^i_m we partition I_m into 6 circular sectors centered on the center of S^i_m (figure 2, bottom-right). Each S^i_m attempts to propagate the closest candidate region in each sector. As a consequence, each candidate C_m has an associated subset $\Gamma_{C_m} \subset \Gamma$ of supports that will *compete* to propagate it. For a candidate C_m and each support S^i in Γ_{C_m} do:

1. Generate C_t^i by attempting to propagate C_m via S^i.
2. Refine C_t^i. If C_t^i correctly matches C_m, this adapts it to the local surface orientation (handles curved and deformable objects) and perspective effects (the affine approximation is only valid on a local scale).
3. Evaluate the quality of the refined propagation attempt: $sim_i = \mathrm{sim}(C_m, C_t^i)$

We retain C_t^{best}, with $best = \arg\max_i sim_i$, the best refined propagation attempt. C_m is considered successfully propagated to C_t^{best} if $sim_{best} > t_2$ (the matching threshold). This procedure is applied for all candidates $C_m \in \Lambda$.

Most support matches may actually be mismatches, and many of them typically lie around each of the few correct ones (e.g.: several matches in a single soft-match, figure 2, middle). In order to cope with this situation, each support concentrates its efforts on the nearest candidate in each direction, as it has the highest chance to undergo a similar geometric transformation. Additionally, every propagation attempt is refined before evaluation. Refinement raises the similarity of correctly propagated matches much more than the similarity of mispropagated ones, thereby helping correct supports to win. This results in a limited, but controlled growth, maximizing the chance that each correct match propagates, and limiting the proliferation of mispropagations. The process also restricts the number of refinements to at most 6 per support (contains computational cost).

For the case-study, 113 new matches are generated and added to the configuration Γ. 17 of them are correct and located around the initial 3. The correct-ratio of Γ improves to 20/330 (figure 4, left), but it is still very low.

4 Early Contraction

The early expansion guarantees high chances that each initial correct match propagates. As initial filter, we discard all matches that did not succeed in propagating any region. The correct-ratio improves to 20/175 (no correct match is lost), but it is still too low for applying a global filter. Hence, we have developed the following local filter.

A local group of regions in the model image have uniform shape, are arranged on a grid and intersect each other with a specific pattern. If all these regions are correctly matched, the same regularities also appear in the test image, because the surface is contiguous and smooth (regions at depth discontinuities can't be matched correctly anyway). This holds for curved or deformed objects as well, because the affine transformation varies slowly and smoothly across neighboring regions (figure 3, left). On the other hand, mismatches tend to be located elsewhere in the image and to have different shapes. We propose a novel, local filter based on this observation. Let $\{N_m^i\}$ be the neighbors of a region R_m in the model image. Two regions A, B are considered neighbors if they intersect, i.e.: if $\mathrm{Area}(A \bigcap B) > 0$. Only neighbors which are actually matched to the test image are considered. Any match (R_m, R_t) is removed from Γ if

$$\sum_{\{N_m^i\}} \left| \frac{\mathrm{Area}(R_m \bigcap N_m^i)}{\mathrm{Area}(R_m)} - \frac{\mathrm{Area}(R_t \bigcap N_t^i)}{\mathrm{Area}(R_t)} \right| > t_s$$

Fig. 3. Left: the regular arrangement of the regions is preserved. Middle: top: a candidate (thin) and 2 of 20 supports (thick) within the large circular area. bottom: the candidate is propagated to the test image using the affine transformation of the support on the right. Refinement adapts the shape to the perspective (brighter). Right: sidedness constraint. R^1 is on the same side of the line in both images

with t_s some threshold. The filter tests the preservation of the pattern of intersections between R and its neighbors (the ratio of areas is affine invariant). Hence, a removal decision is based solely on *local* information. As a consequence, this filter is unaffected by the current, low overall ratio of correct matches. Shape information is integrated in the filter, making it capable of spotting insidious mismatches which are roughly correctly located, yet have a wrong shape. This is an advantage over the (semi-)local filter proposed by [6], and later also used by others [14], which verifies if a minimal amount of regions in an area around R_m in the model image also match near R_t in the test image.

The input regions need not be arranged in a regular grid, the filter applies to a general set of (intersecting) regions. Note that incorrectly matched regions with no neighbors will not be detected. The algorithm can be implemented to run in $O(|\Gamma| + x)$, with $x \ll |\Gamma|^2$ the number of region intersections.

Applying this filter to the case-study brings the correct-ratio of Γ to 13/58, thereby greatly reducing the number of mismatches.

5 Main Expansion

The first 'early' expansion and contraction phases brought several additional correct matches and removed many mismatches, especially those that concentrated around the correct ones. Since Γ is cleaner, we can now try a faster expansion.

All matches in the current configuration Γ are removed from the candidate set $\Lambda \leftarrow \Lambda \backslash \Gamma$, and are used as supports. All support regions S_m^i in a circular area [1] around a candidate C_m compete to propagate it:

1. Generate C_t^i by attempting to propagate C_m via S^i
2. Evaluate $sim_i = \mathrm{sim}(C_m, C_t^i)$

We retain C_t^{best}, with $best = \arg\max_i sim_i$ and refine it, yielding C_t^{ref}. C_m is considered successfully propagated to C_t^{ref} if $\mathrm{sim}(C_m, C_t^{ref}) > t_2$ (figure 3, middle). This scheme is applied for each candidate.

In contrast to the early expansion, many more supports compete for the same candidate, and no refinement is applied *before* choosing the winner. However, the presence of more correct supports, now tending to be grouped, and fewer mismatches, typically spread out, provides good chances that *a* correct support will win a competition. In this process each support has the chance to propagate many more candidates, spread over a larger area, because it offers help to all candidates within a wide circular radius. This allows the system to grow a *mass* of correct matches. Moreover, the process can jump over small occlusions or degraded areas, and costs only one refinement per candidate. 185 new matches, 61 correct, are produced for the case-study, thus lifting the correct-ratio of Γ to 74/243 (30.5%, figure 4, middle).

6 Main Contraction

At this point the chances of having a sufficient number of correct matches to try a global filter are much better. In contrast to the local filter of section 4, the following global filter is capable of finding also isolated mismatches. The algorithm extends our topological filter in [1] to include also appearance similarity.

Figure 3 (right) illustrates the property on which the filter is based. The center of a region R^1 should be on the same side of the directed line going from the center of a second region R^2 to the center of a third region R^3 in both the model and test images (noted $\mathrm{side}(R^1, R^2, R^3)$). This *sidedness constraint* holds for all correctly matched triples of coplanar regions and also for most non-coplanar ones [1]. It does not hold for non-coplanar triples in presence of strong parallax in a few cases, coined *parallax-violations* [1].

A triple including any mismatched region has higher chances to violate the constraint. When this happens, we can only conclude that probably at least one of the matches is incorrect, but we do not yet know which. However, by integrating the weak information each triple provides, it is possible to robustly discover mismatches. Hence, we check the constraint for all unordered triples and we expect wrong matches to be involved in a higher share of violations:

$$\mathrm{err}_{\mathrm{topo}}(R^i) = \frac{1}{v} \sum_{R^j, R^k \in \Gamma \backslash R^i, j>k} |\mathrm{side}(R_m^i, R_m^j, R_m^k) - \mathrm{side}(R_t^i, R_t^j, R_t^k)| \qquad (1)$$

[1] In all experiments the radius is set to 1/6 of the image size.

with $v = (n-1)(n-2)/2, n = |\Gamma|$. $\mathrm{err}_{\mathrm{topo}}(R^i) \in [0,1]$ because it is normalized w.r.t. the maximum number of violations v any region can be involved in. As a novel extension to [1], the topological error share (1) is combined with an appearance term, giving the total error

$$\mathrm{err}_{\mathrm{tot}}(R^i) = \mathrm{err}_{\mathrm{topo}}(R^i) + (t_2 - \mathrm{sim}(R^i_m, R^i_t))$$

The filtering algorithm goes as follows:

1. (Re-)compute $\mathrm{err}_{\mathrm{tot}}(R^i)$ for all $R^i \in \Gamma$.
2. Find the worst match R^w, with $w = \arg\max_i \mathrm{err}_{\mathrm{tot}}(R^i)$
3. If $\mathrm{err}_{\mathrm{tot}}(R^w) > 0$, remove R^w: $\Gamma \leftarrow (\Gamma \backslash R^w)$, and iterate to 1, else stop.

The idea of the algorithm is that at each iteration the most probable mismatch R^w is removed and the error of correct matches decreases, because they are involved in less triples containing any mismatch. After several iterations, ideally only correct matches are left and the algorithm stops. The second term of $\mathrm{err}_{\mathrm{tot}}$ decreases with increasing appearance similarity, and it vanishes when $\mathrm{sim}(R^i_m, R^i_t) = t_2$, the matches acceptance threshold. The removal criteria $\mathrm{err}_{\mathrm{tot}} > 0$ expresses the idea that topological violations are accepted up to the degree to which they are compensated by high similarity. This helps finding mismatches which can hardly be judged by only one cue. A typical mismatch with similarity just above t_2, will be removed unless it is perfectly topologically located. Conversely, correct matches with $\mathrm{err}_{\mathrm{topo}} > 0$ due to parallax-violations are in little danger, because they typically have good similarity. Including appearance makes the filter more robust to low correct-ratios, and remedies the drawback (parallax-violations) of the purely topological filter [1].

The proposed method offers two main advantages over rigid-motion filters, traditionally used in the matching literature [2,5,4,13,7,14], e.g.: detecting outliers to the epipolar geometry through RANSAC [3]. First, it allows for non-rigid deformations, like the bending of paper or cloth, because the structure of the spatial arrangements, captured by the sidedness constraints, is stable under these transformations. Second, it is much less sensitive to inaccurate localizations, because $\mathrm{err}_{\mathrm{topo}}$ varies slowly and smoothly for a region departing from its ideal location.

Topological configurations of points and lines are also used in [15], which enforces the cyclic ordering of line segments connecting corners as a mean for steering the matching process.

In the case-study, the filter starts from 74/243 and returns 54/74, which is a major improvement. 20 correct matches are lost, but many more mismatches (149) are removed. The further processing will recover the lost correct matches and generate even more.

7 Exploring the Test Image

The processing continues by iteratively alternating main expansion and main contraction phases:

1. Do a main expansion phase. This produces a set of propagated region matches Υ, which are added to the configuration: $\Gamma \leftarrow (\Gamma \bigcup \Upsilon)$.
2. Do a main contraction phase on Γ.
3. If at least one newly propagated region survives the contraction, i.e. $|\Upsilon \bigcap \Gamma| > 0$, then iterate to 1, after updating the candidate set to contain $\Lambda \leftarrow (\Omega \backslash \Gamma)$, all original candidate regions Ω which are not yet in the configuration.

In the first iteration, the expansion phase generates some correct matches, along with some mismatches, thereby increasing the correct-ratio. The first main contraction phase removes mostly mismatches, but might also lose several correct matches: the amount of noise could still be high and limit the filter's performance. In the next iteration, this cleaner configuration is fed into the expansion phase again which, less distracted, generates more correct matches and fewer mismatches. The new correct matches in turn help the next contraction stage in taking better removal decisions, and so on. As a result, the amount, percentage and spatial extent of correct matches increase at every iteration, reinforcing the confidence about the object's presence and location. The two goals of separating correct matches and gathering more information about the object are achieved *at the same time*.

Correct matches erroneously killed by the contraction step in an iteration get another chance during the next expansion phase. With even fewer mismatches present, they are probably regenerated, and this time have higher chances to survive the contraction (higher correct-ratio, more positive evidence present).

Thanks to the refinement, each expansion phase adapts the shape of the newly created regions to the local surface orientation. Thus the whole exploration process follows curved surfaces and deformations.

The exploration procedure tends to 'implode' when the object is not in the test image, typically returning 0, or at most a few matches. Conversely, when the object is present, the approach fills the visible portion with many high confidence matches. This yields high discriminative power and the qualitative shift from only *detecting* the object to knowing its extent in the image and which parts are occluded. Recognition and segmentation are intensely intertwined.

Fig. 4. Case-study. Left: 20 correct matches (dark) out of 330 after early expansion. Middle: 74/243 after the first main expansion. Right: contour of the final set of matches. Note the segmentation quality, in particular the detection of the occluding rubber

In the case-study, the second main expansion propagates 141 matches, 117 correct, which is better than the previous 61/185. The second main contraction starts from 171/215 and returns 150/174, killing a lower percentage of correct matches than the first main contraction. After the 11th iteration 220 matches cover the whole visible part of the object (figure 4, right).

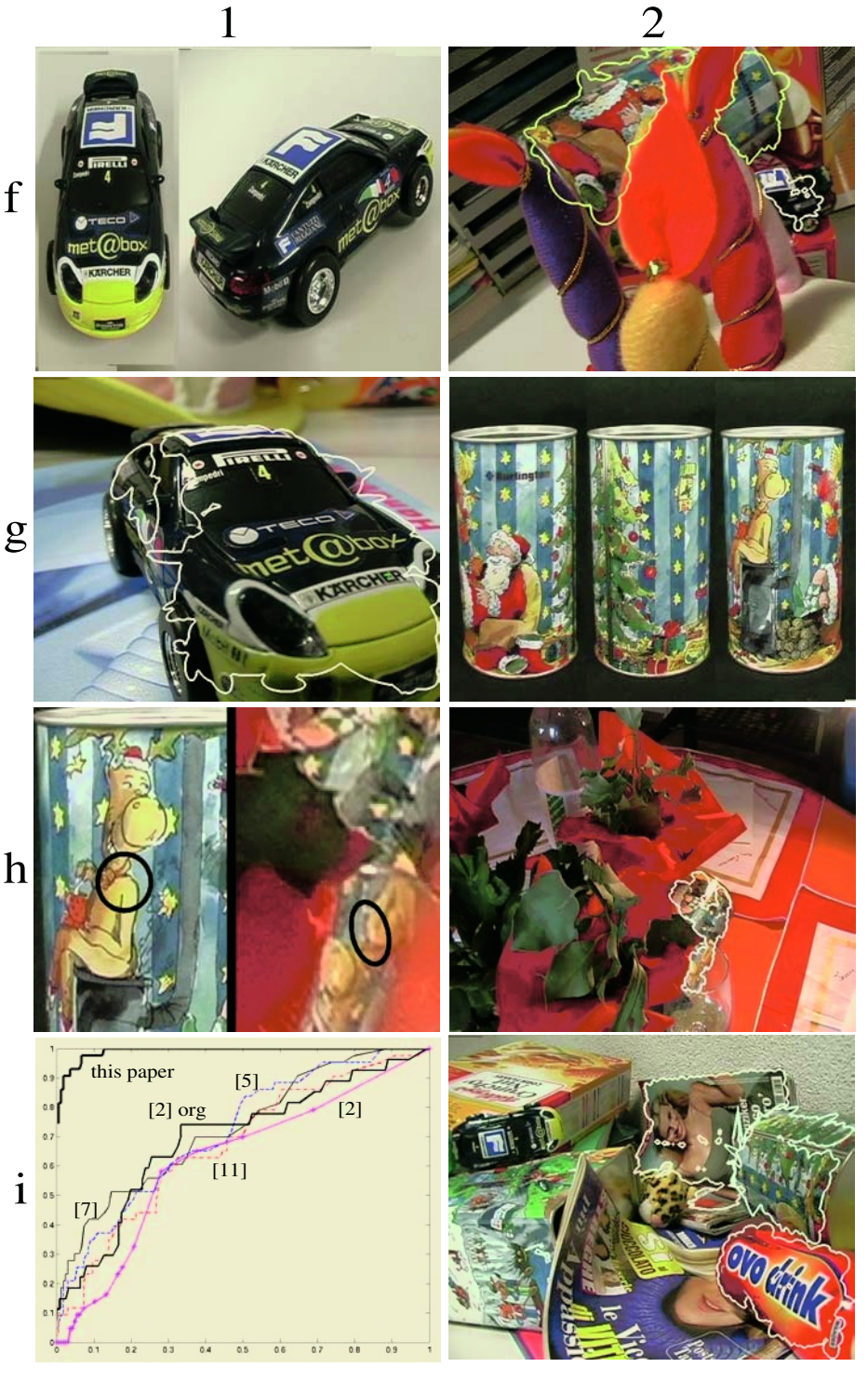
8 Results and Conclusion

We report results for a set of 9 model objects and 23 test images. In total, the objects appear 43 times, as some test images contain several objects. There are 3 planar objects, each modeled by a single view, including a *Kellogs* box (figure 2), and two magazines *Michelle* (figure b2) and *Blonde* (analog model view). Two objects with curved shapes, *Xmas* (g2) and *Ovo* (e1), have 6 model views. *Leo* (c1), *Car* (f1), *Suchard* (d2) feature more complex 3D shape and have 8 model views. Finally, one frontal view models the last 3D object, *Guard* (a1). Multiple model views are taken equally spaced around the object. The contributions from all model views are integrated by superimposing the area covered by the final set of matched regions (to find the contour), and by summing their number (recognition criteria). All images are shot at a modest resolution (720x576) and all experiments are conducted with the same set of parameters. In general, in the test cases there is considerable clutter and the objects appear smaller than in the models (all models are shown at the same scale as the test images).

Tolerance to deformations is shown in a2, where *Michelle* is simultaneously strongly folded and occluded. The contours are found with a good accuracy, extending to the left until the edge of the object. Note the extensive clutter. High robustness to viewpoint changes is demonstrated in b1, where *Leo* is only half visible and captured in a considerably different pose than any of the model views, while *Michelle* undergoes a very large out-of-plane rotation of about 80 degrees. *Guard*, occluding *Michelle*, is also detected in the image, despite a scale change of factor 3. In d1, *Leo* and *Ovo* exhibit significant viewpoint change, while *Suchard* is simultaneously scaled factor 2.2 and 89% occluded. This very high occlusion level makes this case challenging even for a human observer. A scale change of factor 4 affecting *Suchard* is illustrated in e2. In figure f2, *Xmas* is divided in two by a large occludor. Both visible parts are correctly detected by the presented method. On the right size of the image, *Car* is found even if half occluded and very small. *Car* is also detected in spite of considerable viewpoint change in g1. The combined effects of strong occlusion, scale change and clutter make h2 an interesting case. Note how the boundaries of *Xmas* are accurately found, and in particular the detection of the part behind the glass. As a final example, 8 objects are detected at the same time in i2 (for clarity, only 3 contours are shown). Note the correct segmentation of the two deformed magazines and the simultaneous presence of all the aforementioned difficulty factors.

Figure h1 presents a close-up on one of 93 matches produced between a model view of *Xmas* (left) and test case h2 (right). This exemplifies the great appearance variation resulting from combined viewpoint, scale and illumination


1
2
a
b
c
d
e
Michelle Hunziker
E ora cerco
un amore vero
Chi
Gianni Agnelli
Michelle Hunziker
E ora cerco
un amore vero
EDIZIONE STRAORDINARIA
ovo drink


1
2
f
g
h
i
this paper
[5]
[2] org
[2]
[11]
[7]
met@box
KARCHER
TECO
PIRELLI
ovo drink

changes, and other sources of image degradation (here a glass). In these cases, it is very unlikely for the region to be detected by the initial region extractor, and hence traditional methods fail. This figure also illustrates the accuracy of the correspondences generated by the expansion phases.

As a proof of the method's capability of following deformations, we tried to process the case in c2 starting with only one match (dark). 356 regions, covering the whole object, were produced. Each region's shape fits the local surface orientation (for clarity, only 3 regions are shown).

The discriminative power of the system was assessed by processing all pairs of model-object and test images, and counting the resulting amount of region matches. The highest ROC curve in figure i1 depicts the detection rate versus false-positive rate, while varying the detection threshold from 0 to 200 matches. The method performs very well, and can achieve 98% detection with 6% false-positives. For comparison, we processed the dataset also with 4 state-of-the-art affine region extractors [7,5,11,2], and described the regions with the SIFT [8] descriptor [2], which has recently been demonstrated to perform best [12]. The matching is carried out by the 'unambiguous nearest-neighbor' approach [3] advocated in [11,8]: a model region is matched to the region of the test image with the closest descriptor if it is closer than 0.7 times the distance to the second-closest descriptor (the threshold 0.7 has been empirically determined to optimize results). Each of the central curves in i1 illustrates the behavior of a different extractor. As can be seen, none is satisfactory, which demonstrates the higher level of challenge offered by the dataset and therefore suggests that our approach can broaden the range of solvable OR cases. Closer inspection reveals the source of failure: typically only very few, if any, correct matches are produced when the object is present, which in turn is due to the lack of repeatability and the inadequacy of a simple matcher under such difficult conditions. The important improvement brought by the proposed method is best quantified by the difference between the highest curve and the central thick curve, representing the system we started from [2] (labeled '[2] org' in the plot).

The experiments confirm the power of the presented approach in solving very challenging cases. Moreover, non-rigid deformations are explicitly taken into account, and the approximate boundaries of the object is found, two features lacking in competing approaches [4,8,2,7,11,5,14]. The method is of general applicability, as it works with any affine invariant feature extractor. Future work aims at better exploiting the relationships between multiple model-views, at extending the scope to less richly textured objects, and at improving computational efficiency (currently, a 1.4 Ghz computer takes some minutes to process a pair of model and test images).

[2] All region extractors and the SIFT descriptor are implementations of the respective authors. We are grateful to Jiri Matas, Krystian Mikolajczyk, Andrew Zisserman, Cordelia Schmid and David Lowe for providing the programs.

[3] We have also tried the standard approach, used in [7,5,2,12], which simply matches two nearest-neighbors if their distance is below a threshold, but it produced slightly worse results.

References

1. V. Ferrari, T. Tuytelaars and L. Van Gool, Wide-baseline Multiple-view Correspondences *IEEE Comp. Vis. and Patt. Rec.*, vol I, pp. 718–725, 2003.
2. T. Tuytelaars and L. Van Gool Wide Baseline Stereo based on Local, Affinely invariant Regions *Brit. Mach. Vis. Conf.*, pp. 412–422, 2000.
3. P.H.S. Torr and D. W. Murray The development and comparison of robust methods for estimating the fundamental matrix *IJCV*, 24(3), pp. 271–300, 1997.
4. F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, 3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints, *IEEE Comp. Vis. and Patt. Rec.*, vol II, pp. 272–277, 2003.
5. K.Mikolajczyk and C.Schmid, An affine invariant interest point detector *European Conf. on Comp. Vis.*, vol. 1, pp. 128–142, 2002.
6. C.Schmid, Combining greyvalue invariants with local constraints for object recognition *IEEE Comp. Vis. and Patt. Rec.*, pp. 872–877, 1996.
7. S. Obdzalek and J. Matas, Object Recognition using Local Affine Frames on Distinguished Regions *Brit. Mach. Vis. Conf.*, pp. 414–431, 2002.
8. D. Lowe, Distinctive Image Features from Scale-Invariant Keypoints *submitted to Intl. Journ. of Comp. Vis.*, 2004
9. C. Cyr, B. Kimia, 3D Object Recognition Using Similarity-Based Aspect Graph *Intl. Conf. on Comp. Vis.*, 2001
10. H. Murase, S. Nayar, Visual Learning and Recognition of 3D Objects from Appearence *Intl. Journ. of Comp. Vis.*, 14(1), 1995
11. A. Baumberg, Reliable feature matching across widely separated views *IEEE Comp. Vis. and Patt. Rec.*, pp. 774–781, 2000
12. K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors *IEEE Comp. Vis. and Patt. Rec.*, vol II, pp. 257–263, 2003
13. F. Schaffalitzky, A. Zisserman Multi-view matching for unordered image sets *European Conf. on Comp. Vis.*, pp. 414–431, 2002.
14. F. Schaffalitzky, A. Zisserman Automated Scene Matching in Movies *CIVR*, 2002.
15. D. Tell, S. Carlsson Combining Appearance and Topology for Wide Baseline Matching *European Conf. on Comp. Vis.*, pp. 68–81, 2002.