International Journal of Computer Vision 59(1), 61–85, 2004

Matching Widely Separated Views Based on Affine Invariant Regions

TINNE TUYTELAARS AND LUC VAN GOOL
University of Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
tinne.tuytelaars@esat.kuleuven.ac.be
luc.vangool@esat.kuleuven.ac.be

Received August 6, 2001; Revised August 6, 2001; Accepted August 14, 2003

Abstract. 'Invariant regions' are self-adaptive image patches that automatically deform with changing viewpoint as to keep on covering identical physical parts of a scene. Such regions can be extracted directly from a single image. They are then described by a set of invariant features, which makes it relatively easy to match them between views, even under wide baseline conditions. In this contribution, two methods to extract invariant regions are presented. The first one starts from corners and uses the nearby edges, while the second one is purely intensity-based. As a matter of fact, the goal is to build an opportunistic system that exploits several types of invariant regions as it sees fit. This yields more correspondences and a system that can deal with a wider range of images. To increase the robustness of the system, two semi-local constraints on combinations of region correspondences are derived (one geometric, the other photometric). They allow to test the consistency of correspondences and hence to reject falsely matched regions. Experiments on images of real-world scenes taken from substantially different viewpoints demonstrate the feasibility of the approach.

Keywords: wide baseline stereo, matching, invariance, local features, correspondence search, epipolar geometry, semi-local constraints

1. Introduction

Wide baseline stereo, i.e. stereo with the two cameras far apart or with a large vergence angle, has a number of important advantages: greater precision, wider applicability, and less effort by the user as fewer images can suffice. There also are important disadvantages though, like increased levels of occlusion and a correspondence search that is far more difficult. Especially the latter problem has hampered the use of wide baseline stereo setups.

In this paper, we propose a method to find a relatively sparse set of feature correspondences between wide baseline images. These initial matches yield the epipolar geometry and thus greatly facilitate the search for further correspondences. The initial features need not only be robust against the geometric distortions caused by a large change in viewpoint, but also against serious changes in color and intensity that may exist between views. Moreover, features should be quite local, as the risk of parts getting occluded in the other view increases with feature size.

As our goal is not dense correspondences but a set of seed matches, we can afford to restrict features to areas with characteristics that are benign to the task. One is that the local surface is almost planar. This simplifies the geometric distortions that are to be expected between the views. A second assumption is that these almost planar parts contain anchor points that remain stable under changing viewpoint. In particular, we will use corners as well as intensity extrema. It is not crucial that all such points can be retrieved robustly from different views—it suffices if this is the case for a sufficient number. To these anchor points, small patches will be attached as our features.

The principal contribution of this work is the construction of the patches as *invariant regions*: patches attached to the anchor points that have self-adaptive

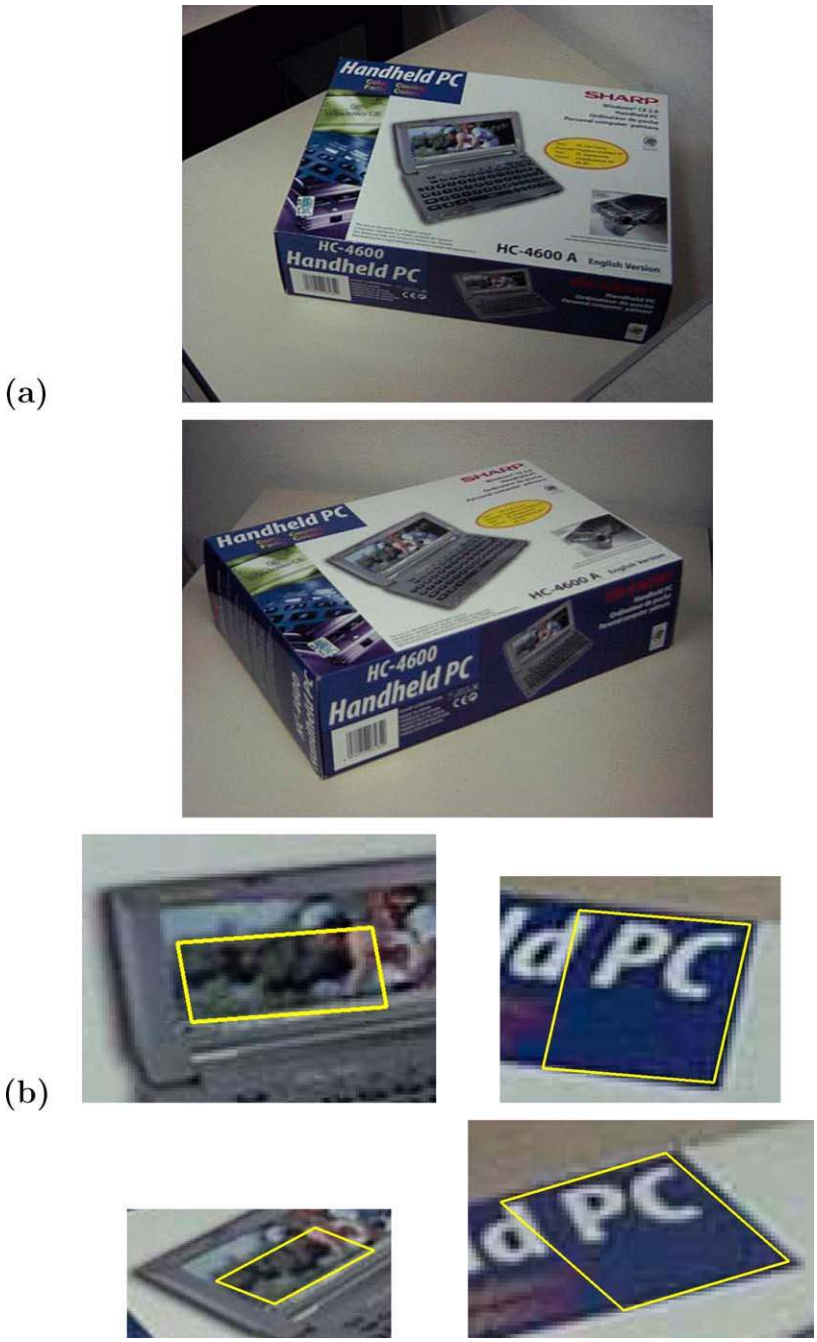

Figure 1. (a) Two images of the same object. (b) Two parallelogram-shaped patches as they are generated by the system: when the viewpoint changes the shapes of the patches are transformed automatically such that they cover the same physical part of the scene. Each of these local image patches has been extracted based on a single image.

shapes to cover the same, physical part of the scene independent of viewpoint (under the assumption of local planarity). With changing viewpoint, these invariant regions change their shape in the image. It is thanks to the viewpoint-dependency of their shape in the image that the regions' scene content can remain invariant. As an example, Fig. 1(b) shows two invariant regions for each of the two views shown in Fig. 1(a). The invariant

regions do indeed represent the same part of the box. The crux of the matter is that they were extracted from each of the views separately, i.e. without any information about the other view. This is important from both a computational and practical point of view, as no pairwise comparisons between regions are necessary for their extraction, and one is not limited to a predefined set of viewpoints.

Scenes can vary widely. In order to make sure that a sufficient number of invariant regions can be extracted, several types have been implemented. It is our intention to build an 'opportunistic' system that exploits several types of image structure, simply depending on what is on offer. This should maximize the applicability of the method and the number of invariant regions found. Here we propose a construction method based on corners and one based on intensity extrema. Others are currently being considered.

To achieve efficient matching of the invariant regions, their color pattern is characterized by a feature vector of moment invariants. They are invariant under both geometric and photometric changes. Finding corresponding invariant regions then boils down to the comparison of these vectors. Additional tests on the mutual consistency of matches are performed to increase robustness.

Both the regions and their feature vectors are invariant under geometric changes, which are modeled by affine transformations as the regions are small, i.e.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}$$

They are also both invariant under photometric changes, modeled by linear transformations with different scalings and offsets for each of the three color bands.

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} s_R & 0 & 0 \\ 0 & s_G & 0 \\ 0 & 0 & s_B \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} o_R \\ o_G \\ o_B \end{bmatrix}$$

Hence, correspondences can be found under a wide range of viewing conditions. Note that, contrary to the region description, the region extraction does not explicitly rely on color information: regions are extracted based on a single color band. Hence the same methods can equally well be applied to greyscale images.

The remainder of the paper is organized as follows. First, an overview of related work is given in Section 2. Section 3 describes the selection of anchor points. The next two sections discuss two different methods for extracting invariant regions: first a geometry-based method (Section 4) followed by an intensity-based method (Section 5). Section 6 describes how the actual correspondence search, based on affine moment invariants computed over these regions is carried out. Consistency checks that can be used to reject false matches are proposed in Section 7. Section 8 discusses some experimental results. Section 9 concludes the paper.

2. Related Work

An important source of inspiration for our approach has been the work of Schmid et al. (1997). They identify special 'points of interest' (in casu corners) and extract 2D translation and 2D rotation invariant features from the intensity pattern in fixed circular regions around these points (in casu the *local jet* as defined by Koenderink and Van Doorn (1987), based on Gaussian derivatives of image intensity). Invariance under scaling is handled by including circular regions of several sizes. Since the level of invariance in their method is limited, it is not really suited for wide baseline stereo applications. Nevertheless, they obtained remarkable results in the context of short baseline stereo, object recognition and database retrieval—for later versions of their system even in spite of very large scale changes (Dufournaud et al., 2000). Similar results have been reported for color images by Montesinos et al. (2000). Some extensions towards affine invariant regions have been reported as well. Lowe (1999) has extended these ideas to real scale-invariance, using circular regions that maximize the output of a difference of gaussian filters in scale space, while Hall et al. (1999) not only applied automatic scale selection (based on Lindeberg (1998)), but also retrieved the orientation of the circular region in an unambiguous way.

Wide Baseline Techniques

To cope with wider baselines, the affine geometric deformations in the image should fully be taken into account during the matching process. One approach is to deform a patch in the first image in an iterative way, until it more or less fits a patch in the

second image (Gruen, 1985; Super and Klarquist, 1997). However, the search that is involved reduces the practicality of this approach. In contrast, our method is based on the extraction and matching of invariant regions, and hence works on the two images separately, without searching over the entire image or applying combinatorics.

This is akin to the approach of Pritchett and Zisserman (1998) who start their wide baseline stereo algorithm by extracting quadrangles present in the image and match these based on normalized cross-correlation to find local homographies, which are then exploited in a search for additional correspondences. However, they use shapes that are explicitly present in the image, while ours are determined locally based on the color patterns around anchor points, so we are less dependent on the presence of specific structures in the scene. Hence, the applicability of our method is wider.

Tell and Carlsson (2000) also proposed a wide baseline correspondence method based on affine invariance. They extract an affine invariant Fourier description of the intensity profile along lines connecting two corner points. The non-local character of their method makes it more robust, but at the same time restricts its use to unoccluded planar objects, which limits the applicability of their method.

In summary, our system differs from other wide baseline stereo methods in that we do not apply a search between images but process each image and each local feature individually (Gruen, 1985; Super and Klarquist, 1997; Schaffalitzky and Zisserman, 2001), in that we fully take into account the affine deformations caused by the change in viewpoint (Lowe, 1999; Montesinos et al., 2000; Schmid and Mohr, 1997; Dufournaud et al., 2000) and in that we can deal with general 3D objects without assuming specific structures to be present in the image (Pritchett and Zisserman, 1998; Tell and Carlsson, 2000).

Affine Invariant Regions

Other approaches to extracting affine invariant regions described in literature are mainly situated in the context of texture analysis. Ballester and Gonzales (1998) have developed a method to find affine invariant regions in textured images. Implicitly, they use the fact that the second moment matrix remains more or less constant when varying the region parameters, which may be a reasonable assumption for textures but clearly does not hold for general image patches.

Lindeberg and Gårding (1997) on the other hand have developed a method to find blob-like regions using an iterative scheme, in the context of shape from texture. In the case of weak isotropy, the regions found by their algorithm correspond to rotationally symmetric smoothing and rotationally symmetric window functions in the tangent plane to the surface. However, in general, their method does not necessarily converge, as there are, in most cases, at least two additional attraction points.

Similar ideas have recently been used for wide baseline stereo by Schaffalitzky and Zisserman (2001). First, they roughly match textured regions in the image. Then, they use texture information (the second moment matrix) to lift some degrees of freedom, followed by an exhaustive search over all Harris corner points within that specific texture and over all possible 2D rotations to find point correspondences under wide baseline conditions. By exploiting texture information, they avoid having to delineate invariant regions, but at the same time this limits the applicability of their method to images containing stationary textures.

Baumberg (2000) proposed a wide baseline system that is based on a simplified version of the regions of Lindeberg and Gårding (1997). However, the regions Baumberg uses are only invariant under rotation, stretch and skew, while scale changes are dealt with by applying a scale space approach. The error on the scale also influences the other components of the transformation, such that the resulting invariant regions are probably not as accurate as ours.

Nevertheless, we believe that it could be beneficial to include the above region extraction methods into our system to further improve the performance of the system (i.e. more correspondences and a wider range of applicability).

3. Selection of Anchor Points

The first step in the extraction of affine invariant regions consists of selecting '*anchor points*', that serve as seeds for the subsequent region extraction. This allows to reduce the complexity of the problem and the needed computation time, since the attention can be focussed on regions around these points instead of examining every single pixel in the image. At the same time, extra assumptions can often be made concerning the regions based on the type of anchor point.

Good anchor points are points that result in stable invariant regions, are repeatable and easy-to-detect. With repeatability, we mean that there is a high probability that the same point will be found in another view as well—or at least, a point that would result in the same region.

Harris corner points (Harris and Stephens, 1983) are good candidates. Apart from the necessary properties of good anchor points mentioned above, they typically contain a large amount of information (Schmid and Mohr, 1998), resulting in a high distinctive power, and they are well localized, i.e. the position of the corner point is accurately defined (even up to sub-pixel accuracy) (Shi and Tomasi, 1994).

Instead of using corners, local extrema of image intensity can serve as anchor points as well. To this end, we first apply some smoothing to the image to reduce the effect of noise, causing too many unstable local extrema. Then, the local extrema are extracted with a non-maximum suppression algorithm. These points cannot be localized as accurately as corner points, since the local extrema in intensity are often rather smooth. However, they can withstand any monotonic intensity transformation and they are less likely to lie close to the border of an object resulting in a non-planar region. This last property is a major drawback when working with corner points.

Of course, which kind of anchor points perform best also depends on the method used for the region extraction, and how good this method deals with the shortcomings of the anchor points. For instance, for the corner points, the high chance of a non-planar region can be alleviated by constructing a region that is not centered around the corner point. Similarly, regions starting from local intensity extrema should not depend too much on the exact position of the extremum, to overcome the inaccurate localization of these points.

Other types of anchor points could be used as well. For instance, Lowe (1999) uses extrema of a difference of Gaussians filter.

4. Geometry-Based Method

The first method for affine invariant region extraction starts from Harris corner points (Harris and Stephens, 1983) and the edges that can often be found close to such a point (extracted using the Canny edge detector (Canny, 1986)). As this method so strongly relies on the presence and accurate detection of these geometric entities, we coined it the *geometry-based* method. Two different cases are considered: one method is developed for curved edges while a slightly different method is applied in case of straight edges.

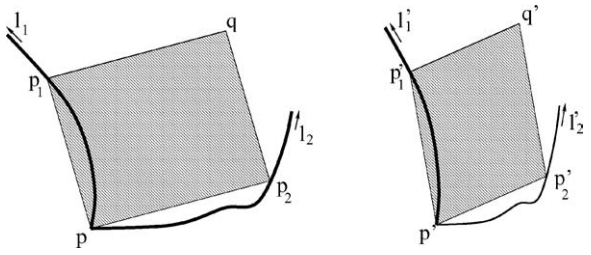

Figure 2. Based on the edges close to the corner point, an affine invariant region can be constructed.

4.1. Case 1: Curved Edges

Let $\mathbf{p} = (x_p, y_p)^T$ be a Harris corner point on an edge, as in Fig. 2. Two points $\mathbf{p_1}$ and $\mathbf{p_2}$ move away from the corner in both directions along the edge. Their relative speed is coupled through the equality of relative affine invariant parameters l_1 and l_2:

$$l_i = \int \text{abs}\big(|\mathbf{p_i}^{(1)}(s_i)\ \mathbf{p} - \mathbf{p_i}(s_i)|\big) ds_i \quad i = 1, 2$$

with s_i an arbitrary curve parameter, $\mathbf{p_i}^{(1)}(s_i)$ the first derivative of $\mathbf{p_i}(s_i)$ with respect to s_i, abs() the absolute value and $|..|$ the determinant. From now on, we simply use l when referring to $l_1 = l_2$. At each position, the two points $\mathbf{p_1}(l)$ and $\mathbf{p_2}(l)$ together with the corner $\mathbf{p}$ define a region Ω for the point $\mathbf{p}$ as a function of l: the parallelogram spanned by the vectors $\mathbf{p_1}(l) - \mathbf{p}$ and $\mathbf{p_2}(l) - \mathbf{p}$ (see Fig. 2). This gives us a one dimensional family of parallelogram-shaped regions. The points stop at positions where some photometric quantities of the texture covered by the parallelogram go through an extremum. We typically generate regions for a few extrema, which introduces a kind of scale concept as now regions of different sizes coexist for a single corner. Since it is not guaranteed that a single function will reach an extremum over the limited l-interval we are looking at, more than one function is tested. Taking extrema of several functions into account, we get a better guarantee that a high number of corners will indeed generate some regions.

Thanks to a good choice of the functions, the whole process can be made invariant under the aforementioned geometric and photometric changes. Examples

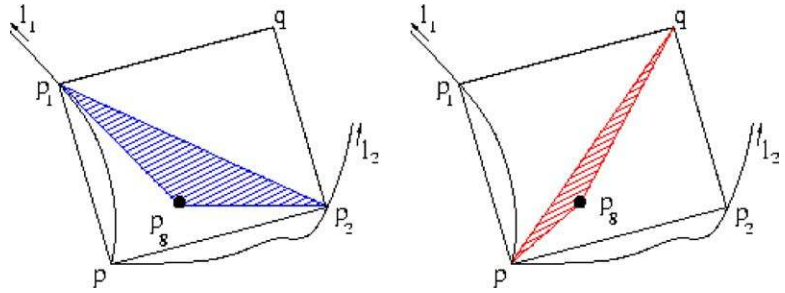

Figure 3. Physical interpretation of the functions $f_2(\Omega)$ (left) and $f_3(\Omega)$ (right).

of such functions are

$$f_1(\Omega) = \frac{M^1_{00}}{M^0_{00}}$$

$$f_2(\Omega) = \text{abs}\left(\frac{|\mathbf{p_1} - \mathbf{p_g} \quad \mathbf{p_2} - \mathbf{p_g}|}{|\mathbf{p} - \mathbf{p_1} \quad \mathbf{p} - \mathbf{p_2}|}\right) \times \frac{M^1_{00}}{\sqrt{M^2_{00}M^0_{00} - \left(M^1_{00}\right)^2}}$$

$$f_3(\Omega) = \text{abs}\left(\frac{|\mathbf{p} - \mathbf{p_g} \quad \mathbf{q} - \mathbf{p_g}|}{|\mathbf{p} - \mathbf{p_1} \quad \mathbf{p} - \mathbf{p_2}|}\right) \times \frac{M^1_{00}}{\sqrt{M^2_{00}M^0_{00} - \left(M^1_{00}\right)^2}}$$

$$\text{with } M^n_{pq} = \int_\Omega I^n(x, y)x^p y^q dxdy$$

$$\mathbf{p_g} = \left(\frac{M^1_{10}}{M^1_{00}}, \frac{M^1_{01}}{M^1_{00}}\right)$$

with M^n_{pq} the nth *order*, $(p+q)$th *degree* moment computed over the region $\Omega(l)$, $\mathbf{p_g}$ the center of gravity of the region, weighted with intensity $I(x, y)$ (one of the three color bands R, G or B), and $\mathbf{q}$ the corner of the parallelogram opposite to the corner point $\mathbf{p}$ (see Fig. 2).

The first function, $f_1(\Omega)$, represents the average intensity over the region $\Omega(l)$. It is not in itself invariant under the considered photometric transformations, but reaches its extrema in an invariant way. We do not use this function in our implementation though, since the minima of $f_2(\Omega)$ and $f_3(\Omega)$ tend to be better localized than the extrema of $f_1(\Omega)$, resulting in more stable regions. Nevertheless $f_1(\Omega)$ could be the better choice if the application needs high speed. $f_2(\Omega)$ and $f_3(\Omega)$ consist of two components each: first, a ratio of two areas, one of which depends on the center of gravity weighted with intensity and hence on the region pattern, and second, a factor that compensates for the dependence of the first component to offsets in the image intensity.[1] Figure 3 illustrates the geometrical interpretation of the first component for $f_2(\Omega)$ (left) and $f_3(\Omega)$ (right) respectively. It is twice the ratio of the marked area, divided by the total area of the region. By looking for local minima of these functions we favor more *balanced* regions, i.e. regions for which the center of gravity lies on or close to one of the diagonals of the parallelogram. In contrast to $f_1(\Omega)$, the functions $f_2(\Omega)$ and $f_3(\Omega)$ are invariant. Nevertheless, we still select the regions where the function reaches a minimum instead of selecting regions where the function reaches a specific value, hence avoiding the introduction of another (rather arbitrary) parameter. For a proof of the geometric and photometric invariance of the local minima of these functions, we refer to Appendix A.

Figure 4 shows two invariant parallelogram-shaped regions found for corresponding points in two widely separated views of the same object. Although there is a large image distortion between the two images (geometrically as well as photometrically), the affine invariant regions—which have been found for each image independently—cover similar physical parts of the scene. For clarity, the curved edges on which the extraction was based are added as well.

Note that the affine invariant regions found are not centered around the anchor point. A centered alternative is the parallelogram that has the non-centered parallelogram as one quadrant. Nevertheless, we prefer the non-centered regions, as—and experiments have borne that out—restricting the region to one quadrant (delineated by the edges) makes the assumption of planarity much more realistic, due to the fact that the anchor points we start from are corners, often lying close to a depth discontinuity (see Section 3).

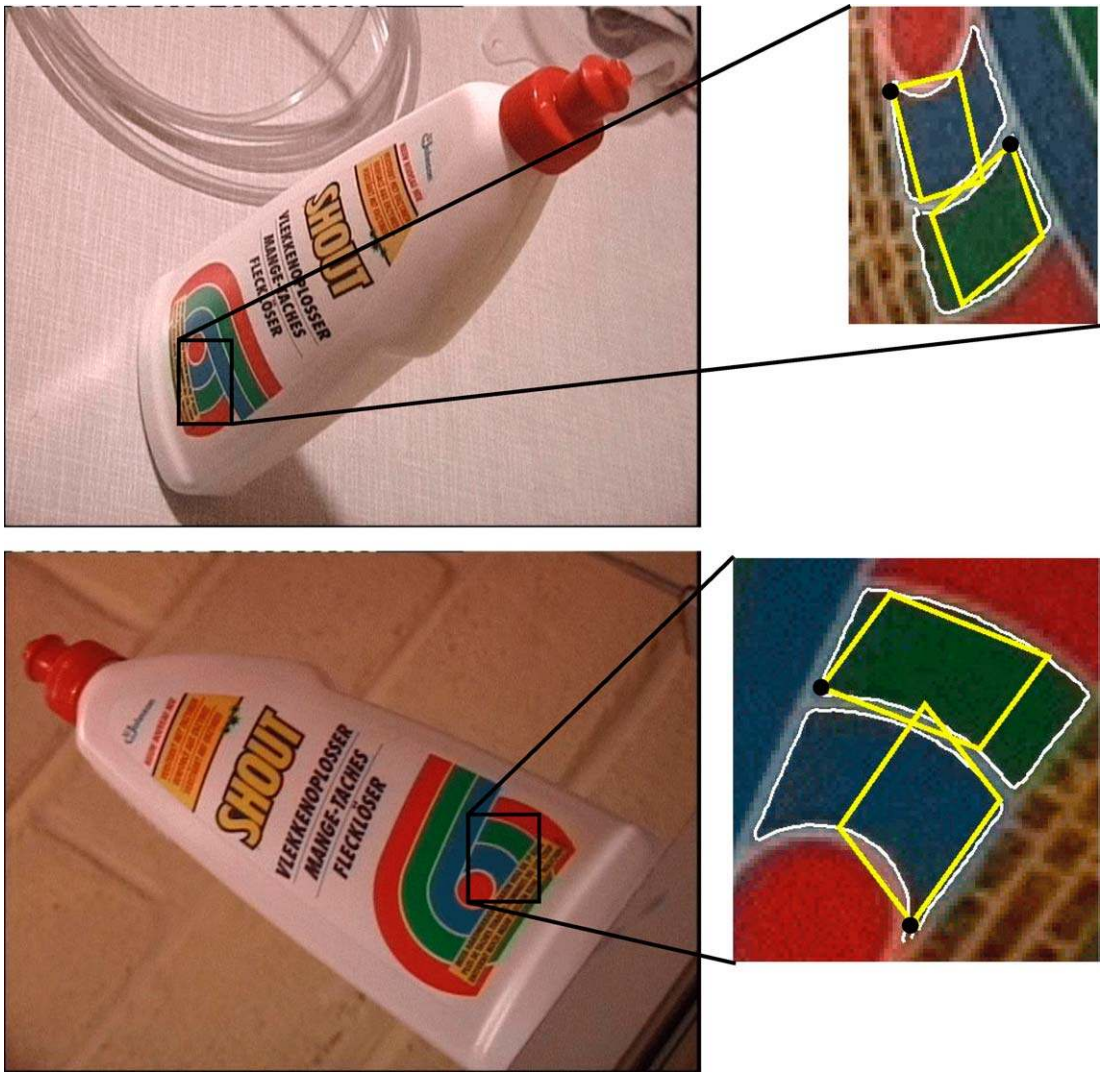

Figure 4. Affine invariant regions based on corners and curved edges.

4.2. *Case 2: Straight Edges*

In the case of straight edges, the method described above cannot be applied, since $l = 0$ along the entire edge. However, since straight edges occur quite often, we cannot simply neglect this case.

A straightforward extension of the previous technique would then be to search for local extrema in a 2D search-space spanned by two arbitrary parameters s_1 and s_2 for the two edges, instead of a 1D search-space over l. However, the functions $f_2(\Omega)$ and $f_3(\Omega)$ we used for the curved-edges case, do not show clear, well-defined extrema in the 2D case. Rather, we have some shallow *valleys* of low values (corresponding to cases where the center of gravity lies on or close to one of the diagonals). Instead of taking the inaccurate local extrema of one function, we combine the two functions and take the *intersections* of the two valleys, as shown in Fig. 5. The special case where the two valleys (almost) coincide must be detected and rejected, since the intersection is not accurate in that case. The regions so obtained proved to be much more stable than those based on a 2D local extremum.

Figure 6 shows some affine invariant regions extracted for the same images as in Fig. 4, but now using the method designed for straight edges.

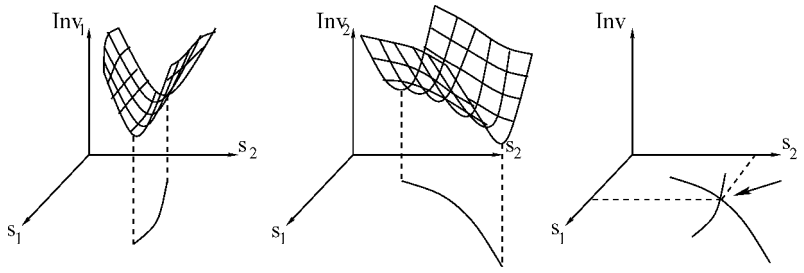

Figure 5. For the straight edges case, the intersection of the "valleys" of two different functions is used instead of a local extremum.

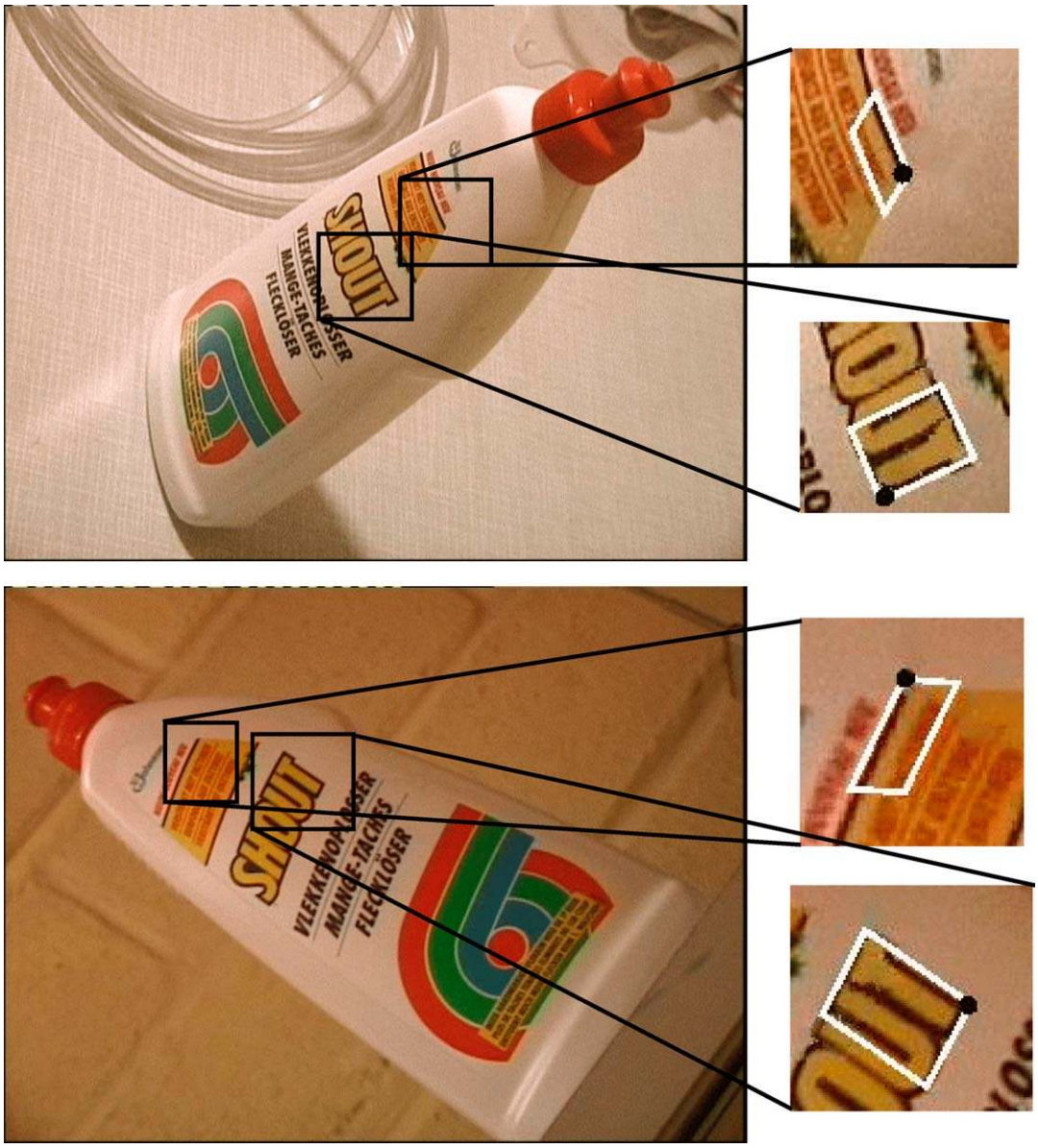

Figure 6. Affine invariant regions based on corners and straight edges.

Again, they clearly cover identical parts of the object.

5. Intensity-Based Method

A drawback of the method described in the previous section is that the edges it relies on are often a source of errors. Edges that were found in one image may be undetected, interrupted or connected in a different way in the second image. This section presents an alternative method for extracting invariant regions, that is directly based on the analysis of image intensity, without an intermediate step involving the extraction of features such as edges or corners. It turns out to complement the previous method very well, in that invariant regions are typically found at different locations in the image.

Instead of starting from corner points, this method uses local extrema in intensity as anchor points (cfr. Section 3). Given such a local extremum, the intensity function along rays emanating from the extremum is studied, as shown in Fig. 7. The following function is evaluated along each ray:

$$f_I(t) = \frac{\text{abs}(I(t) - I_0)}{\max\left(\frac{\int_0^t \text{abs}(I(t) - I_0)dt}{t}, d\right)}$$

with t the Euclidean arclength along the ray, $I(t)$ the intensity at position t, I_0 the intensity extremum and d a small number which has been added to prevent a division by zero. The point for which this function reaches an extremum is invariant under the aforementioned affine geometric and linear photometric transformations (given the ray). Typically, a maximum is reached at positions where the intensity suddenly increases or decreases dramatically compared to the intensity changes encountered on the line up to that point, for instance at the border of a more or less homogeneous area. $f_I(t)$ is in itself already invariant. Nevertheless, we again select the points where this function reaches an extremum for reasons of robustness.

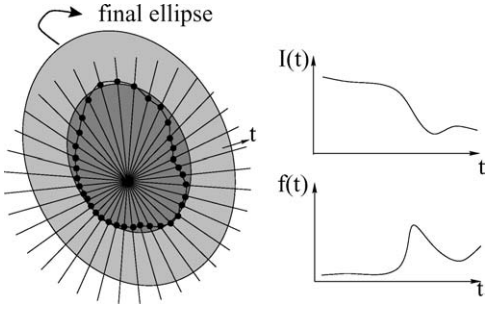

Figure 7. The intensity along 'rays' emanating from a local extremum are studied. The point on each ray for which a function $f_I(t)$ reaches a maximum is selected. Linking these points together yields an affine invariant region, to which an ellipse is fitted using moments.

Note that in theory, leaving out the denominator in the expression for $f_I(t)$ would yield a simpler function which still has invariant positions for its local extrema. In practice, however, this simpler function does not give as good results since its local extrema are more shallow, resulting in inaccurate positions along the rays and hence inaccurate regions. With the denominator added, on the other hand, the local extrema are in most cases more accurately localized.

Next, all points corresponding to maxima of $f_I(t)$ along rays originating from the same local extremum are linked to enclose an (affine invariant) region (see again Fig. 7). This often irregularly-shaped region is replaced by an ellipse having the same shape moments up to the second order. This ellipse-fitting is again affine invariant. Finally, we double the size of the ellipses found. This leads to more distinctive regions, due to a more diversified texture pattern within the region and hence facilitates the matching process, at the cost of a higher risk of non-planarity due to the less local character of the regions.

Problems may arise when more than one local extremum can be found along the ray. In such case, instead of choosing the global extremum, we select an extremum by imposing a *continuity constraint*: in case of multiple extrema, we select the extremum closest to the extrema found along the neighbouring rays.

Figure 8 shows some intensity-based regions (ellipses) and the linked points on which the region extraction is based.

Note that the resulting elliptical regions are not centered around the original anchor point (the intensity extremum). In fact, the whole procedure is pretty robust to the inaccurate localization of this point. In most cases (i.e. if the area enclosed by the linked points is more or less convex), small changes in its position have only a limited effect on the resulting region if the intensity profile is indeed showing a shallow extremum. This is illustrated in Fig. 9, where we repeated the region extraction starting from different anchor points lying close to the intensity extremum and having similar intensity values. Although the elliptical regions found are

Figure 8. Affine invariant regions found with the intensity-based region extraction method and the linked points used to extract them.

not identical, they are similar enough to be matched. To highlight the source of the deviations, we also added the linked points found along the rays, used in the region construction.

6. Finding Correspondences

Once local, invariant regions have been extracted, finding correspondences between two views becomes relatively easy. This is performed by means of a nearest neighbour classification scheme, based on feature vectors of invariants computed over the affine invariant regions. As in the region extraction step, we consider invariance both under affine geometric changes and linear photometric changes, with different offsets and different scale factors for each of the three color bands.

6.1. Normalization

Although it is very well possible to construct a feature vector that is in itself invariant to all the geometric and

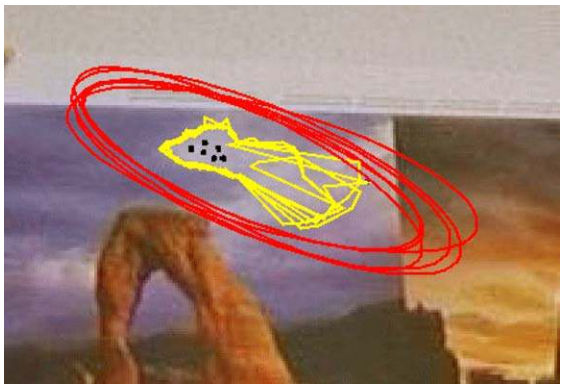

Figure 9. Robustness of the region extraction to the inaccurate localization of the intensity extremum.

photometric transformations we consider (e.g. Mindru et al., 1999), our experiments show that better results are obtained if one first compensates for (part of) the deformations through an extra normalization step, exploiting extra knowledge about the region.

For the geometry-based case, we first transform the parallelogram-shaped region to a square reference region of fixed size. Since we know a specific corner of the parallelogram (from the original anchor point) and since it is reasonable to assume that the clockwise order of the corners is preserved (i.e. the image is not being mirrored), the entire affine deformation can be compensated for in this way.

For the intensity-based case, the situation is slightly more complex. We can transform the elliptical region to a circular reference region of fixed size, but (again assuming the image is not being mirrored) this still leaves one degree of freedom to be determined (corresponding to a free rotation of the circle around its center). This last degree of freedom cannot be derived from purely geometric information gathered during the region extraction. Instead, we determine it based on a photometric invariant version of the axes of inertia. The major and minor axes of inertia are extracted as the lines passing through the center of the circular region with orientations $\theta_{\max}$, $\theta_{\min}$ defined by the solutions of:

$$\tan^2(\theta) + \frac{m_{20} - m_{02}}{m_{11}} \tan\theta - 1 = 0$$

with m_{pq} the $p+q$th order, first degree moment (see Section 4.1) *centered on the region's geometric center*. This equation differs from the usual definition of the axes of inertia by the use of these moments instead of moments centered on the center of gravity weighted with image intensity. This makes them invariant to linear intensity changes (including offsets). Based on these axes of inertia, one can apply an additional rotation, that brings the major axis of inertia into a horizontal position, hence fixing the last degree of freedom.

Instead of computing the axes of inertia to compensate for the last degree of freedom, one could also extract features that are invariant under rotation. This would probably give comparable results. However, retrieving the complete affine deformation not only allows to treat intensity-based and geometry-based regions in the same way but also allows to further compare the content of two matched regions in a pixelwise manner, based on normalized cross-correlation, independent of the geometric distortions (see Section 6.3).

Also the illumination variations can be compensated for in an extra normalization step. This is achieved by replacing each intensity value I (i.e. R, G or B) by $I' = aI + b$ with a and b such that the average intensity is 128 and with a spread on the intensities of 50.

6.2. *Region Description*

Each region is then characterized by a feature vector of moment invariants. The moments we use are *Generalized Color Moments*, which have been introduced in Mindru et al. (1999) to better exploit the multi-spectral nature of the data. They contain powers of the image coordinates and of the intensities of the different color channels.

$$M_{pq}^{abc} = \iint_{\Omega} x^p y^q [R(x,y)]^a [G(x,y)]^b [B(x,y)]^c \, dxdy$$

with *order* $p+q$ and *degree* $a+b+c$. In fact, they implicitly characterize the shape, the intensity and the color distribution of the region pattern in a uniform manner.

More precisely, we use 18 moment invariants, summarized in Table 1. These are invariant functions of moments up to second order and first degree (i.e. moments that use up to first order powers of intensities (R, G, B) and second order powers of (x, y) coordinates). Since we already normalized the regions with respect to viewpoint and illumination variations, *any* measurement can actually be used as an invariant measure, as all variations have been compensated for already. The reason why we still stick to moments is that these are more robust to noise. *inv*[1] to *inv*[3] are related to the

Table 1. Moment invariants used for comparing the patterns within regions after normalization against geometric and photometric deformations.

$inv[1] = M_{00}^{110}/M_{00}^{000}$	$inv[2] = M_{00}^{011}/M_{00}^{000}$	$inv[3] = M_{00}^{101}/M_{00}^{000}$
$inv[4] = M_{10}^{100}/M_{00}^{100}$	$inv[5] = M_{10}^{010}/M_{00}^{010}$	$inv[6] = M_{10}^{001}/M_{00}^{001}$
$inv[7] = M_{01}^{100}/M_{00}^{100}$	$inv[8] = M_{01}^{010}/M_{00}^{010}$	$inv[9] = M_{01}^{001}/M_{00}^{001}$
$inv[10] = M_{11}^{100}/M_{00}^{100}$	$inv[11] = M_{11}^{010}/M_{00}^{010}$	$inv[12] = M_{11}^{001}/M_{00}^{001}$
$inv[13] = M_{20}^{100}/M_{00}^{100}$	$inv[14] = M_{20}^{010}/M_{00}^{010}$	$inv[15] = M_{20}^{001}/M_{00}^{001}$
$inv[16] = M_{02}^{100}/M_{00}^{100}$	$inv[17] = M_{02}^{010}/M_{00}^{010}$	$inv[18] = M_{02}^{001}/M_{00}^{001}$

correlation between two color-bands. $inv[4]$ to $inv[6]$ and $inv[7]$ to $inv[9]$ are the x- and y-coordinates respectively of the centers of gravity weighted with one color-band, while $inv[10]$ to $inv[18]$ are combinations of higher order moments.

As an additional invariant, we use the region *type*. This value refers to the method that has been used for the region extraction. Only if the type of two regions corresponds, can they be matched.

6.3. Region Matching

Each region in the first image is then matched to the region in the second image for which the Mahalanobis-distance between the corresponding feature vectors is minimal and below a predefined threshold d. Then, all regions of the second image are matched in a similar way to the regions extracted from the first image. Only a mutual match is accepted as a real correspondence between the two views. The covariance matrix needed to compute the Mahalanobis-distance has been estimated by tracking representative regions over a set of images. Due to the different nature of the different region types, better results are obtained when different covariance matrices are computed for each region type separately. The comparison of feature vectors can be done in an efficient way using indexing-techniques. At this moment, only indexing based on the region type has been implemented.

Once corresponding regions have been found, the normalized cross-correlation between them is computed as a final check before accepting the region correspondence. This cross-correlation check is not performed on the raw image data, but after normalization of the two regions to a fixed-size square or circular reference region (depending on the region type), as described in Section 6.1. In this way, the effect of the geometric deformations on the normalized cross-correlation is annihilated.

7. Robustness—Rejecting Falsely Matched Regions

Due to the wide range of geometric and photometric transformations allowed and the local character of the regions, false correspondences are inevitable. These can be caused by symmetries in the image, or simply because the local region's distinctive power is insufficient. Semi-local or global constraints offer a way out: by checking the consistency between combinations of local correspondences (assuming a rigid motion), false correspondences can be identified and rejected. The best known constraint is checking for a consistent epipolar geometry in a robust way, e.g. using RANSAC (Fischler and Bolles, 1981), and rejecting all correspondences not conform with the epipolar geometry found. Although this method works fine in many applications, our experiments have shown that it may have difficulties in a typical wide baseline stereo setup, where false matches abound and may even outnumber the good ones while the total number of matches is rather low. In that case, many of the randomly selected seven-point samples contain outliers, resulting in large computation times (each time rejecting the sample and trying out a new combination), or even erroneous results (a sample containing outliers coincidentally yielding a reasonable amount of matches). The latter case happens more often than expected, since matches are in general *not* randomly spread over the image, but tend to clutter on linear or planar structures in the scene.

Here, two other semi-local constraints are proposed that may be used to reject outliers. Both work on a combination of *two* region correspondences only, hence the amount of combinatorics needed is limited. The first one tests the geometric consistency, while the second one is a photometric constraint. Checking these constraints first before testing the epipolar geometry with RANSAC can considerably improve the results under the hard conditions of wide baseline stereo. This is akin to the work of Carlsson (2000), who has recently proposed a view compatibility constraint for five points in two views based on a scaled orthographic camera model.

7.1. A Geometric Constraint

Each match between two affine invariant regions defines an affine transformation, matching the region in one image on the corresponding region in the second image. Such an affine transformation is in fact an approximation of the homography linking the projections of all points lying in the same plane.

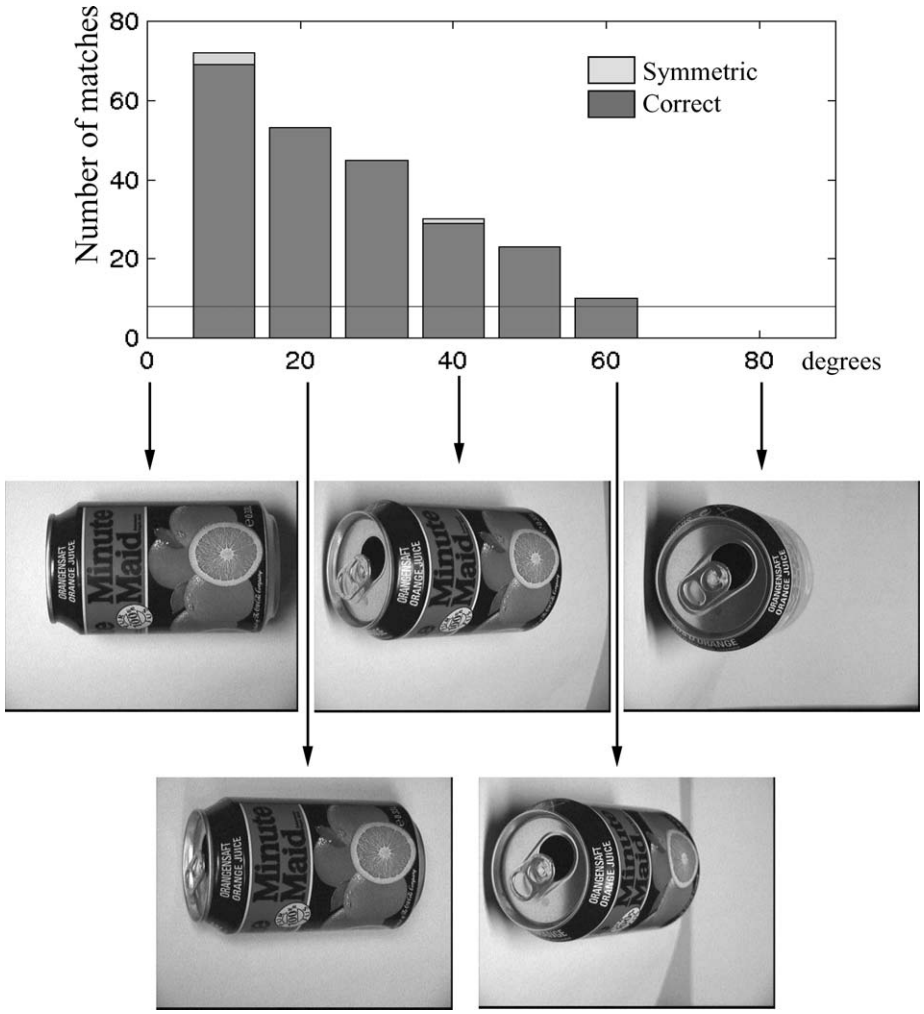

Figure 10. Viewpoint invariance of the region extraction and matching: number of correct, symmetric and false matches found as a function of the rotation angle with respect to the 0 degrees reference view.

Sinclair et al. (1995) proposed a method to test whether two rigid plane motions are compatible based on their homographies H_1 and H_2. Combining them as $H_1^{-1}H_2$ yields a planar homology, whose eigenanalysis reveals one fixed point (the epipole) and one line of fixed points (the common line of the two planes). They project this common line to the other image using H_1, and once again using H_2. If the two planes are indeed in rigid motion, the two resulting lines in the second image should coincide, which can easily be checked.

The geometric constraint we use here is a simple algebraic distance. As it only requires the evaluation of the determinant of a 3×3 matrix, it can be applied quite fast. This makes it well suited for applications like ours, where many consistency checks are performed on different combinations of planes (i.e. matches). To check whether two correspondences found are geometrically consistent with one another, it suffices to check whether

$$det \begin{pmatrix} a_{23} - b_{23} & b_{13} - a_{13} & a_{13}b_{23} - b_{13}a_{23} \\ a_{22} - b_{22} & b_{12} - a_{12} & a_{12}b_{23} - b_{13}a_{22} + a_{13}b_{22} - b_{12}a_{23} \\ a_{21} - b_{21} & b_{11} - a_{11} & a_{11}b_{23} - b_{13}a_{21} + a_{13}b_{21} - b_{11}a_{23} \end{pmatrix} \leq \delta_g$$

with δ_g a predefined threshold, $A = [a_{ij}]$ and $B = [b_{ij}]$ the affine transformations mapping the region in the first image to the region in the second image, for the first and second match respectively. For the derivation of this semi-local constraint, we refer to Appendix B.

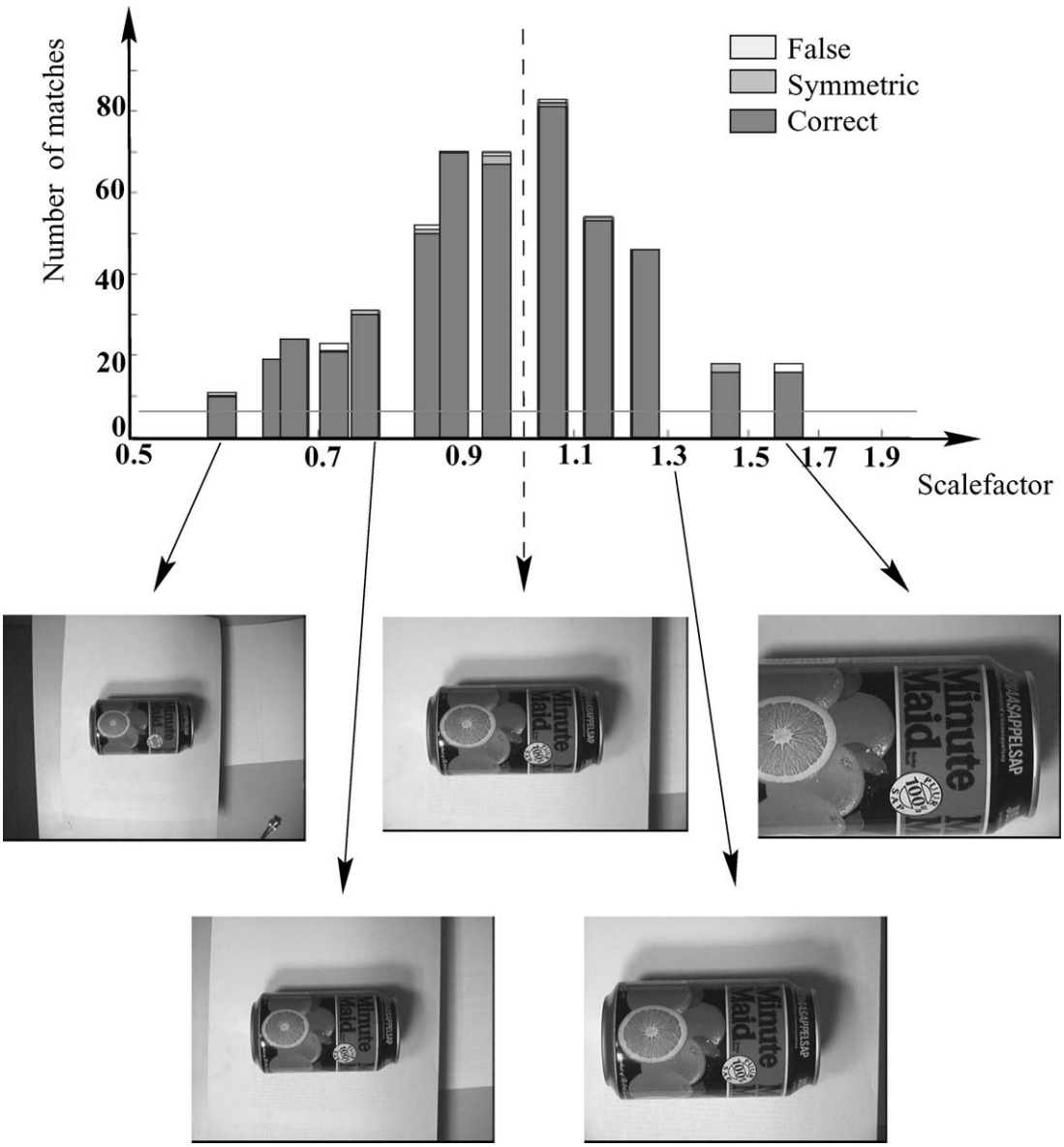

Figure 11. Scale invariance of the region extraction and matching: number of correct, symmetric and false matches found as a function of the scalefactor with respect to the reference image.

7.2. A Photometric Constraint

Apart from geometric constraints, photometric constraints can be derived as well. Although it is not necessarily true that the illumination conditions are constant over the entire image (due to shadows, multiple light sources, etc.), it is reasonable to assume that at least some parts of the images have similar illumination conditions.

First, we compute for each region correspondence the offsets and scalefactors of the photometric transformation using moments. Then, given a pair of region corespondences, we check for their photometric consistency by comparing their photometric transformations. For two region correspondences to be consistent, only an overall scale factor is allowed, to compensate for the different orientations of the regions.

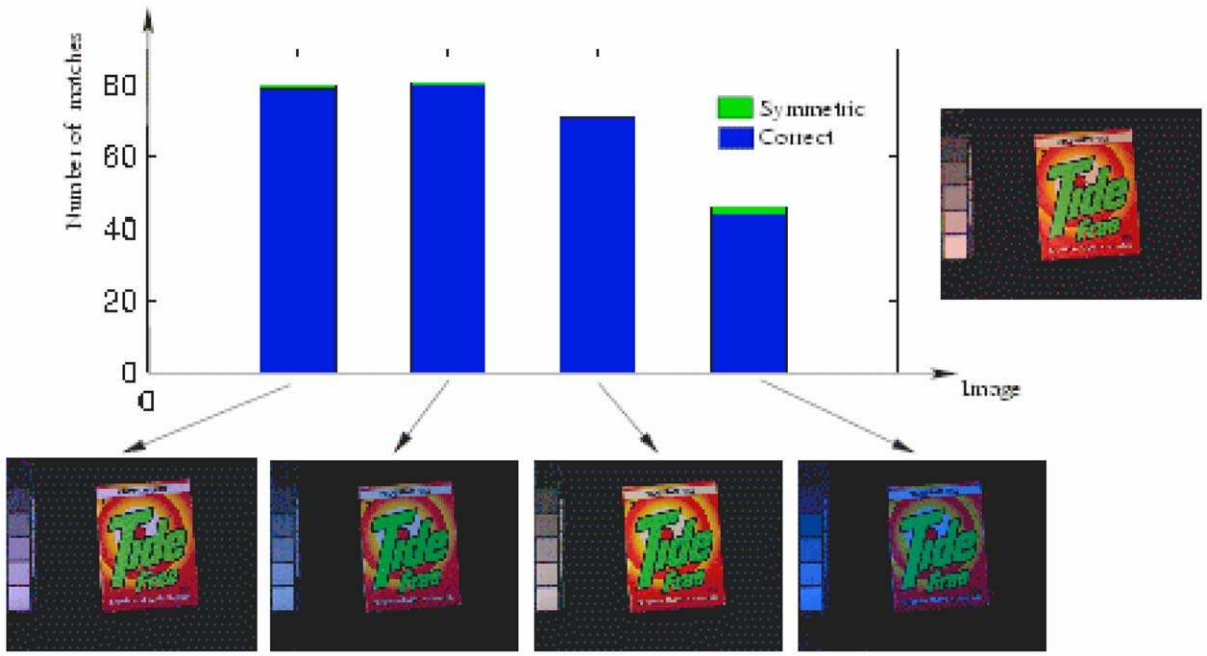

Figure 12. Illumination invariance of the region extraction and matching: number of correct and symmetric matches found between the images shown along the horizontal axis and the reference image shown on the right.

7.3. Rejecting False Matches

Suppose we have N correspondences, each linking a different local region in image I to a similar region in image I' by N different transformations. For each combination of two such correspondences, the above consistency constraints can be checked. A specific region correspondence is considered incorrect if it is consistent with less than n other correspondences (with n typically 8 for the geometric constraint and 4 for the photometric constraint). Hence each good correspondence should have at least n other consistent correspondences. This procedure may have to be repeated a number of times, since rejecting a correspondence may cause other correspondences to have their number of consistent correspondences decreased below the threshold as well.

After having rejected most false matches among the region correspondences using the geometric and photometric constraints described above, we apply RANSAC (Fischler and Bolles, 1981) (a robust method based on random sampling) to find a consistent epipolar geometry and to reject the remaining false correspondences. Since the number of false matches has already seriously been reduced, this process usually stops after a limited number of samples. One must note though that the computation of epipolar geometry is very sensitive to small misallignments in the data. The region matches we have found so far give in most cases only one stable point correspondence (e.g. the harris corner point in case of the geometry-based method). In theory, two more linearly independent point correspondences can be extracted from the invariant region. However, these additional point correspondences are insufficiently stable for the epipolar geometry computation, mainly due to deviations from our model, such as the object surface not being perfectly planar. This problem can be overcome by mapping one image onto the other using the affine transformation, and looking for more accurate point correspondences within the matched regions using small baseline matching techniques. RANSAC is then applied to the resulting set of point correspondences.

8. Experimental Results

8.1. Viewpoint Invariance

To quantitatively check the viewpoint invariance of our method, we took images of an object starting from head on and gradually increasing the viewing angle in steps of 10 degrees. All images were taken with our Sony

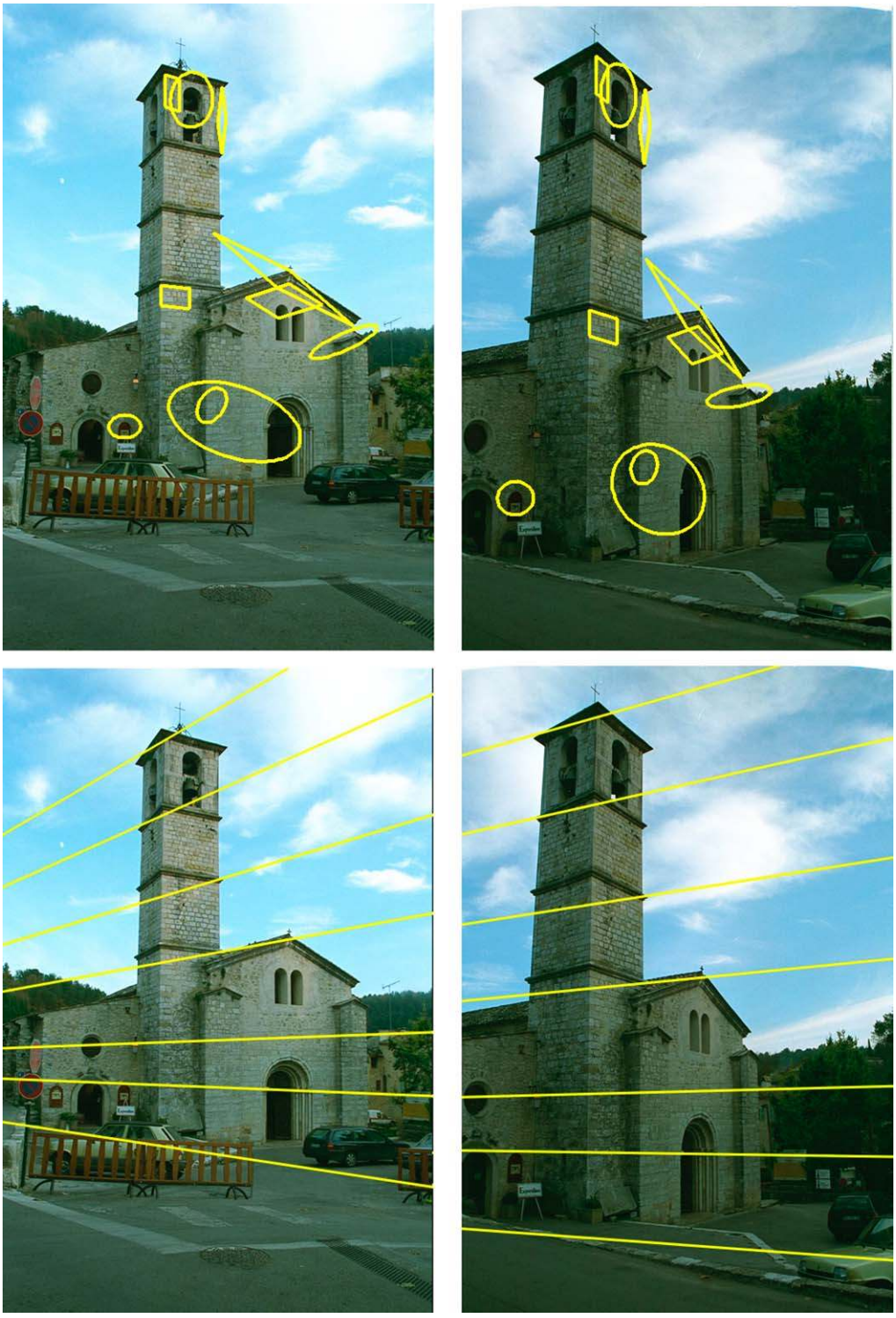

Figure 13. Example 1: Final region correspondences (top) and epipolar geometry (bottom).

digital camera, with a resolution of 768 × 576 pixels. The results of this experiment are shown in Fig. 10.

For each image, the affine invariant regions were extracted and matched to the regions found in the 0 degrees reference image. Next, the regions were filtered using the semi-local geometric and photometric constraints. Finally, we applied the epipolar test using RANSAC to automatically select the good matches, and verified these matches visually, subdividing them into three different categories: correct, symmetric and false.

Figure 14. Example 2: Final region correspondences (top) and epipolar geometry (bottom).

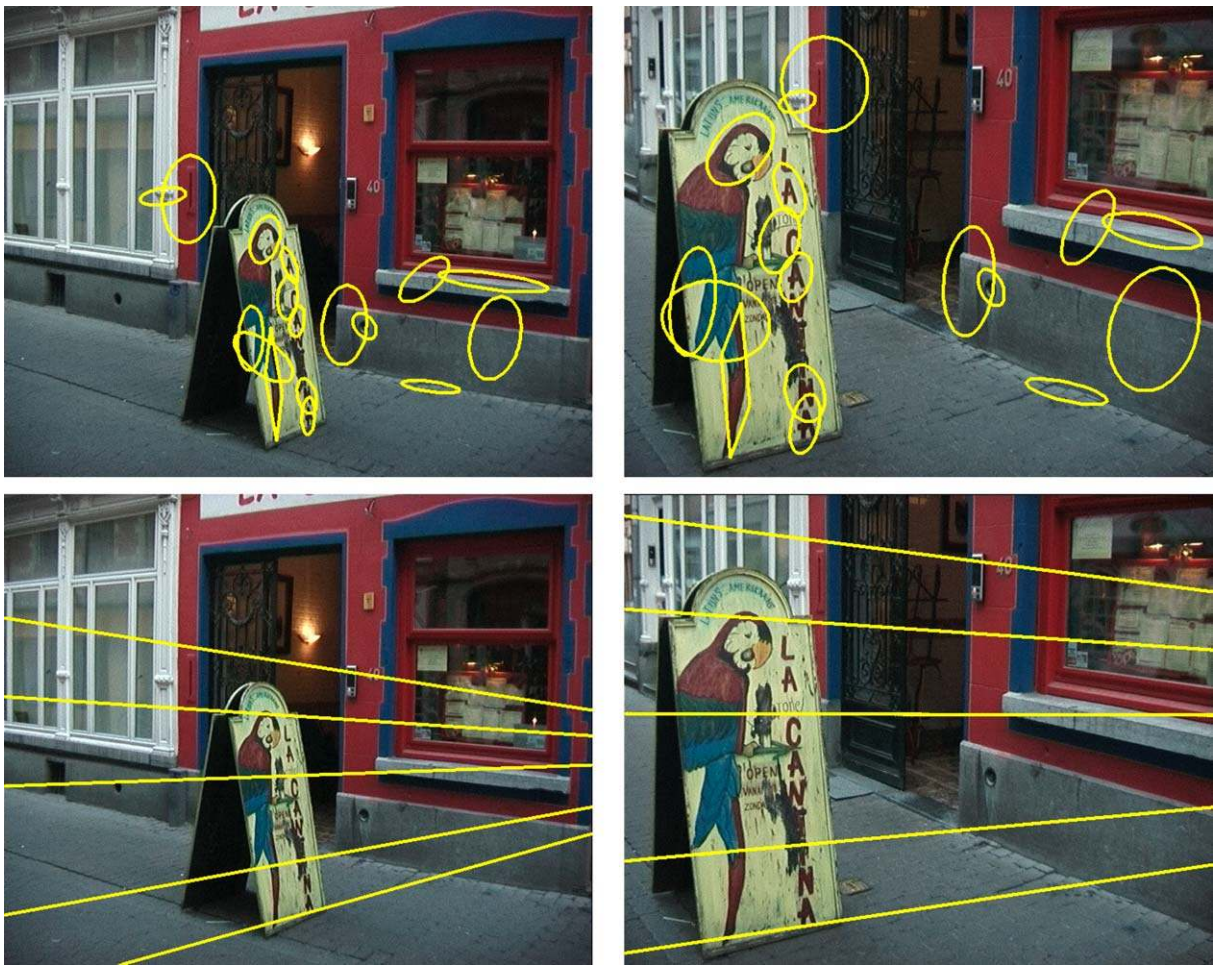

Figure 15. Example 3: Final region correspondences (top) and epipolar geometry (bottom).

With *symmetric* matches, we refer to those matches that do not link physically identical points, but points that can not be distinguished on a local scale due to a symmetry in the image. For instance, the text on the drink can used in this experiment contains twice the letter 'M'. Moreover, these letters are exactly below one another, so they lie more or less on the same epipolar line due to the chosen camera movement. As a result, there is no way for the system to distinguish between the regions found on these two letters.

From Fig. 10, one can see that the system can deal with changes in viewpoint up to 50 or 60 degrees. Only correct and symmetric matches were left. For larger angles, the geometric consistency test could no longer be applied, as the number of matches was too low (remember that we need at least $n = 8$ consistent matches to classify them as geometrically 'correct'). The horizontal line added to the figure indicates the minimum number of matches needed for this geometric filtering stage. It is mainly the change in scale due to the foreshortening of the object that causes problems, in combination with more and more specular reflection.

8.2. *Invariance to Scale Changes*

As scale changes seem to be the weakest point in the viewpoint invariance of the regions, we performed some extra experiments to specifically test for the invariance to scale changes. For the same test object, images with different scales were taken by zooming in and out with our digital camera. As can be seen from Fig. 11, the number of matches found decreases with increasing scale change. Nevertheless, one can conclude that the extraction and matching of affine

Figure 16. Example 4: Final region correspondences (top) and epipolar geometry (bottom).

invariant regions is able to withstand scalefactors ranging from 2/3 to 3/2. If larger scale changes are to be expected, a scale space approach should be adopted.

8.3. Illumination Invariance

Since changes in the illumination are harder to quantify than changes in scale or viewpoint, we decided to use the images provided by Funt et al. (1998) to test the illumination invariance of our system, as they provide very detailed information on the different illuminants used. Using these images, which are readily available through `ftp`,[2] allows for easy comparison of our results with other systems. Figure 12 shows the result. Each of the images shown below the horizontal axis was compared with the reference image taken under halogen illumination shown to the right. The left part of each image shows the white to black row of the Macbeth Color Checker, highlighting the large difference in illumination. Most of the 'symmetric' matches found were actually matches between these reference squares. For all images, plenty of correspondences were found, clearly showing the robustness of our region extraction and matching to changing illumination conditions.

8.4. Wide Baseline Stereo Examples

Figures 13–17 show some views of scenes taken from substantially different viewpoints. Note the large changes in scale in some parts of the images (e.g. Example 3), the serious occlusions (e.g. Example 4) and the extreme foreshortening (e.g. Example 5). Nevertheless, in all cases sufficient matches were found for

Figure 17. Example 5: Final region correspondences (top) and epipolar geometry (bottom).

Figure 18. Negative examples: Image pairs our system was not able to match.

an accurate determination of the epipolar geometry. Sometimes the number of matched regions is pretty low (e.g. Example 4). However, one must not forget that a single region correspondence yields three point correspondences. Each time, the upper part of the figure shows the regions that contributed to the epipolar geometry, i.e. those that were matched and survived both the geometric and photometric filtering as well as RANSAC. Some corresponding epipolar lines are shown in the lower half of the figures.

Finally, Fig. 18 shows some examples of scenes our system was *not* able to process. Although these scenes do not seem extraordinarily complex or difficult, the system failed, mainly due to the different backgrounds (car-example), the lack of texture on the objects (both examples), a large amount of specular reflection (car-example) and non-planarity (simpsons-example). These images clearly show some possible future research directions.

9. Conclusion

A new approach to the wide baseline stereo correspondence problem has been proposed, that extends the ideas of Schmid and Mohr on local invariant features towards more invariance and hence wider baselines. In each image, local image patches are extracted in an affine invariant way, such that they cover the same physical part of the scene (under the assumption of local planarity). These patches or 'invariant regions' are matched based on feature vectors of moment invariants that combine invariance under geometric and photometric changes. The consistency of the matches found is tested using semi-local constraints, followed by a test on the epipolar geometry using RANSAC. As shown in the experimental results, the feasibility of affine invariance even on a local scale has been demonstrated.

Robust matching is quite a generic problem in vision and several other applications can be considered. Object recognition is one, where images of an object can be matched against a small set of reference images of the same object. The sample set can be kept small because of the invariance. Moreover, as the features are local, recognition against variable backgrounds and under occlusion is supported by this method. Another application is grouping, where symmetries can be found as repeated structures. Image database retrieval can also benefit from these regions, where other pictures of the same scene or object can be found. Here, the viewpoint and illumination invariance gives the system the capacity to generalize to a great extent from a single query image. Finally, being able to match a current view against learned views can allow robots to roam extended spaces, without the need for a 3D model. Initial results for such applications can be found in Tuytelaars and Van Gool (1999), Tuytelaars et al. (1999) and Turina et al. (2001).

Appendix A: Affine Invariance of the Function Extrema

Suppose we have the following geometric and photometric deformations between two views:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}$$

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} s_R & 0 & 0 \\ 0 & s_G & 0 \\ 0 & 0 & s_B \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} t_R \\ t_G \\ t_B \end{bmatrix}$$

with (R, G, B) and $(R'G', B')$ the three different color-bands and (x', y') and (x, y) coordinates of corresponding points. In the sequel, we use I and I' to refer to either of the three color-bands R, G or B.

We now prove that the extrema of the functions given in Section 4.1 are invariant to the above deformations. In other words, for each region Ω in image I for which $f(\Omega)$ reaches an extremum, there must be a corresponding region Ω' in image I' for which $f'(\Omega')$ reaches an extremum as well, with $f(\Omega) = f_1(\Omega)$, $f_2(\Omega)$ or $f_3(\Omega)$.

Affine Invariant Extrema of $f_1(\Omega)$

As mentioned already earlier, the first function represents the average intensity over the region. The extrema of this function being invariant to the considered deformations, can easily be understood intuitively. Here, we give a more formal proof.

$$f_1(\Omega) = \frac{\int_\Omega I(x, y) dx dy}{\int_\Omega dx dy}$$

$$f_1'(\Omega') = \frac{\int_{\Omega'} I'(x', y') dx' dy'}{\int_{\Omega'} dx' dy'}$$

$$= \frac{\int_\Omega (sI(x,y)+t)(ad-bc)\,dxdy}{\int_\Omega (ad-bc)\,dxdy}$$
$$= s\frac{\int_\Omega I(x,y)dxdy}{\int_\Omega dxdy} + t = sf_1(\Omega) + t$$

In practice, s is always positive, such that

$$f_1(\Omega_1) > f_1(\Omega_2) \Leftrightarrow f_1'(\Omega_1') > f_1'(\Omega_2')$$

Hence, extrema of the function $f_1(\Omega)$ are preserved under the considered deformations. Even if s would have been negative, extrema would still be preserved, although maxima would be turned into minima and vice versa.

Effects of the Deformations on the Center of Gravity

For the other functions mentioned in Section 4.1, it is important to first fully understand the effect of the deformations on the center of gravity

$$\mathbf{p_g} = (x_g, y_g) = \left(\frac{\int_\Omega I(x,y)x dxdy}{\int_\Omega I(x,y)dxdy}, \frac{\int_\Omega I(x,y)y dxdy}{\int_\Omega I(x,y)dxdy}\right)$$

First, let us consider only geometric deformations. In that case, we get for $\mathbf{p_g'} = (x_g', y_g')$

$$\begin{aligned} x_g' &= \frac{\int_{\Omega'} I'(x',y')x'\,dx'dy'}{\int_{\Omega'} I'(x',y')\,dx'dy'} \\ &= \frac{\int_\Omega I(x,y)(ax+by+e)(ad-bc)\,dxdy}{\int_{\Omega'} I(x,y)(ad-bc)\,dxdy} \\ &= a\frac{\int_\Omega I(x,y)x\,dxdy}{\int_{\Omega'} I(x,y)\,dxdy} + b\frac{\int_\Omega I(x,y)y\,dxdy}{\int_{\Omega'} I(x,y)\,dxdy} + e \\ &= ax_g + by_g + e \\ y_g' &= \frac{\int_{\Omega'} I'(x',y')y'\,dx'dy'}{\int_{\Omega'} I'(x',y')\,dx'dy'} \\ &= \frac{\int_\Omega I(x,y)(cx+dy+f)(ad-bc)\,dxdy}{\int_{\Omega'} I(x,y)(ad-bc)\,dxdy} \\ &= c\frac{\int_\Omega I(x,y)x\,dxdy}{\int_{\Omega'} I(x,y)\,dxdy} + d\frac{\int_\Omega I(x,y)y\,dxdy}{\int_{\Omega'} I(x,y)\,dxdy} + f \\ &= cx_g + dy_g + f \end{aligned}$$

Hence, the center of gravity behaves as a normal point under the affine deformations.

Now, let us consider the effect of photometric deformations. Here, we investigate the coordinates of the center of gravity $\mathbf{p_g}$ relative to the coordinates of the region center $\mathbf{p_c}$.

$$\mathbf{p_c} = (x_c, y_c) = \left(\frac{M_{10}^0}{M_{00}^0}, \frac{M_{01}^0}{M_{00}^0}\right)$$

It can be shown that the effect of the photometric deformations on $\mathbf{p_g}$ is a shift towards $\mathbf{p_c}$:

$$\begin{aligned} x_g' - x_c' &= \frac{\int_{\Omega'} I'(x',y')x'dx'dy'}{\int_{\Omega'} I'(x',y')dx'dy'} - \frac{\int_\Omega x'dx'dy'}{\int_\Omega dx'dy'} \\ &= \dots \\ &= (x_g - x_c)\frac{\int_\Omega I(x,y)dxdy}{\int_\Omega (I(x,y)+\frac{t}{s})dxdy} \\ &= (x_g - x_c)\frac{M_{00}^1}{M_{00}^1 + \frac{t}{s}M_{00}^0} \\ y_g' - y_c' &= \frac{\int_{\Omega'} I'(x',y')y'dx'dy'}{\int_{\Omega'} I'(x',y')dx'dy'} - \frac{\int_\Omega y'dx'dy'}{\int_\Omega dx'dy'} \\ &= \dots \\ &= (y_g - y_c)\frac{\int_\Omega I(x,y)dxdy}{\int_\Omega (I(x,y)+\frac{t}{s})dxdy} \\ &= (y_g - y_c)\frac{M_{00}^1}{M_{00}^1 + \frac{t}{s}M_{00}^0} \end{aligned}$$

Affine Invariant Extrema of $f_2(\Omega)$ and $f_3(\Omega)$

$f_2(\Omega)$ and $f_3(\Omega)$ are both composed of two factors, a ratio of two areas, one of which depends on the center of gravity, and an expression of moments up to the second order.

$$\begin{aligned} f_2(\Omega) &= \text{abs}\left(\frac{|\mathbf{p_1}-\mathbf{p_g} \quad \mathbf{p_2}-\mathbf{p_g}|}{|\mathbf{p}-\mathbf{p_1} \quad \mathbf{p}-\mathbf{p_2}|}\right) \times \frac{M_{00}^1}{\sqrt{M_{00}^2 M_{00}^0 - \left(M_{00}^1\right)^2}} \\ f_3(\Omega) &= \text{abs}\left(\frac{|\mathbf{p}-\mathbf{p_g} \quad \mathbf{q}-\mathbf{p_g}|}{|\mathbf{p}-\mathbf{p_1} \quad \mathbf{p}-\mathbf{p_2}|}\right) \times \frac{M_{00}^1}{\sqrt{M_{00}^2 M_{00}^0 - \left(M_{00}^1\right)^2}} \\ &\text{with} \quad \mathbf{q} = \mathbf{p_1} + \mathbf{p_2} - \mathbf{p} \end{aligned}$$

The first factor is a ratio of two areas, defined by the points $\mathbf{p}$, $\mathbf{p_1}$ and $\mathbf{p_2}$ fixed to the region and the center of

gravity $\mathbf{p_g}$. As we have seen in the previous section, the center of gravity behaves as a normal, physical point under the affine geometric deformations, such that this first factor clearly is geometrically invariant. Also the second factor can easily be checked to be invariant to the geometrical deformations.

Next, we show that the effect of the photometric deformations on this first factor is similar to their effect on the coordinates of the center of gravity relative to the region center, namely a rescaling with the same scale-factor. This can be understood by the fact that the region center $\mathbf{p_c}$ lies on the diagonals of the parallelogram-shaped region, i.e. on the line connecting $\mathbf{p}$ and $\mathbf{q}$ on one hand and the line connecting $\mathbf{p_1}$ and $\mathbf{p_2}$ on the other hand, which also form one side of the areas in the numerator (see Fig. 3). Hence the shift in the position of the center of gravity causes a proportional rescaling of the area in the numerator:

$$\frac{|\mathbf{p}'_1 - \mathbf{p}'_g \;\; \mathbf{p}'_2 - \mathbf{p}'_g|}{|\mathbf{p}' - \mathbf{p}'_1 \;\; \mathbf{p}' - \mathbf{p}'_2|} = \frac{|\mathbf{p_1} - \mathbf{p_g} \;\; \mathbf{p_2} - \mathbf{p_g}|}{|\mathbf{p} - \mathbf{p_1} \;\; \mathbf{p} - \mathbf{p_2}|} \frac{M^1_{00}}{M^1_{00} + \frac{t}{s} M^0_{00}}$$

$$\frac{|\mathbf{p}' - \mathbf{p}'_g \;\; \mathbf{q}' - \mathbf{p}'_g|}{|\mathbf{p}' - \mathbf{p}'_1 \;\; \mathbf{p}' - \mathbf{p}'_2|} = \frac{|\mathbf{p} - \mathbf{p_g} \;\; \mathbf{q} - \mathbf{p_g}|}{|\mathbf{p} - \mathbf{p_1} \;\; \mathbf{p} - \mathbf{p_2}|} \frac{M^1_{00}}{M^1_{00} + \frac{t}{s} M^0_{00}}$$

This extra scale-factor must be compensated for by the second component in the expressions of $f_2(\Omega)$ and $f_3(\Omega)$. And indeed, the second component seems to have exactly the inverse scale-factor:

$$\begin{aligned}
&\frac{M'^1_{00}}{\sqrt{M'^2_{00} M'^0_{00} - \left(M'^1_{00}\right)^2}} \\
&= \frac{\int_{\Omega'} I'(x', y') dx' dy'}{\sqrt{\int_{\Omega'} I'^2(x', y') dx' dy' \int_{\Omega'} dx' dy' - \int_{\Omega'} I'(x', y') dx' dy' \int_{\Omega'} I'(x', y') dx' dy'}} \\
&= \cdots = \frac{M^1_{00}}{\sqrt{M^2_{00} M^0_{00} - M^1_{00} M^1_{00}}} \frac{M^1_{00} + \frac{t}{s} M^0_{00}}{M^1_{00}}
\end{aligned}$$

Appendix B: Derivation of a Geometric Semi-Local Constraint

Consider two images I and I'. Points in image I are denoted with homogeneous coordinates $\mathbf{p} = (x, y, z)^T$, while points in image I' are denoted with homogeneous coordinates $\mathbf{p}' = (x', y', z')^T$. For the coordinates of real world (3D) points, capital letters are used, such as $\mathbf{P} = (X, Y, Z)$. A homography H_i belonging to a plane Π_i defines the following relation between the projections in images I and I' of 3D points lying on the plane Π_i

$$\mathbf{p}' = H_i \mathbf{p}$$

with H_i a 3×3 matrix.

Take an arbitrary point $\mathbf{p} = (x, y, z)^T$ in image I, corresponding to the 3D point $\mathbf{P} = (X, Y, Z)^T$ and two homographies H_1 and H_2, corresponding to two different planes Π_1 and Π_2. Then, both $H_1\mathbf{p}$ and $H_2\mathbf{p}$ lie on the epipolar line corresponding to the point $\mathbf{p}$ in the second image. Hence, the following formula for the epipolar line corresponding to the point $\mathbf{p}$ can be derived

$$l = (H_1\mathbf{p}) \times (H_2\mathbf{p})$$

where $\times$ denotes the vector product.
All epipolar lines pass through the same point $\mathbf{e}$, the epipole.

$$\exists \mathbf{e} \forall \mathbf{p} : (H_1\mathbf{p} \times H_2\mathbf{p})^T \, \mathbf{e} = 0$$

From this property, we can derive a constraint on H_1 and H_2.
If $\mathbf{H_{ij}}$ denotes the j-th column of matrix H_i, this can be worked out as follows:

$$\begin{aligned}
\exists \mathbf{e} \forall (x, y, z) : [(x\mathbf{H_{11}} + y\mathbf{H_{12}} + z\mathbf{H_{13}}) \\
\times (x\mathbf{H_{21}} + y\mathbf{H_{22}} + z\mathbf{H_{23}})]^T \mathbf{e} = 0
\end{aligned}$$

This is a second-order equation in x, y and z with coefficients A, B, C, D, E and F functions of $\mathbf{e}$ and $\mathbf{H_{ij}}$.

$$\forall (x, y, z) : Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz = 0$$

Since this equation has to be fulfilled for all possible values x, y and z, all the coefficients in the equation have to be zero.

$$\begin{aligned}
A &= (\mathbf{H_{11}} \times \mathbf{H_{21}})^T \mathbf{e} = 0 \\
B &= (\mathbf{H_{12}} \times \mathbf{H_{22}})^T \mathbf{e} = 0 \\
C &= (\mathbf{H_{13}} \times \mathbf{H_{23}})^T \mathbf{e} = 0 \\
D &= (\mathbf{H_{11}} \times \mathbf{H_{22}} + \mathbf{H_{12}} \times \mathbf{H_{21}})^T \mathbf{e} = 0 \\
E &= (\mathbf{H_{11}} \times \mathbf{H_{23}} + \mathbf{H_{13}} \times \mathbf{H_{21}})^T \mathbf{e} = 0 \\
F &= (\mathbf{H_{12}} \times \mathbf{H_{23}} + \mathbf{H_{13}} \times \mathbf{H_{22}})^T \mathbf{e} = 0
\end{aligned}$$

In order for all the above equations to have a solution $\mathbf{e} \neq (0, 0, 0)^T$, the following matrix, which is a function of $\mathbf{H_{ij}}$, must be rank-deficient.

$$rank \begin{pmatrix} (\mathbf{H_{11}} \times \mathbf{H_{21}})^T \\ (\mathbf{H_{12}} \times \mathbf{H_{22}})^T \\ (\mathbf{H_{13}} \times \mathbf{H_{23}})^T \\ (\mathbf{H_{11}} \times \mathbf{H_{22}} + \mathbf{H_{12}} \times \mathbf{H_{21}})^T \\ (\mathbf{H_{11}} \times \mathbf{H_{23}} + \mathbf{H_{13}} \times \mathbf{H_{23}})^T \\ (\mathbf{H_{12}} \times \mathbf{H_{23}} + \mathbf{H_{13}} \times \mathbf{H_{22}})^T \end{pmatrix} \leq 2$$

Applied to Local Regions

For local regions, the perspective deformation is too small to be detected. As a result, only an affine transformation can be derived. In this case, the homographies can be approximated by affine transformations A and B of the following form:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

The rank-2 constraint derived in the previous section then becomes:

$$rank \begin{pmatrix} 0 & 0 & a_{11}b_{21} - b_{11}a_{21} \\ 0 & 0 & a_{12}b_{22} - b_{12}a_{22} \\ a_{23} - b_{23} & b_{13} - a_{13} & a_{13}b_{23} - b_{13}a_{23} \\ 0 & 0 & a_{11}b_{22} - b_{12}a_{21} + a_{12}b_{21} - b_{11}a_{22} \\ a_{22} - b_{22} & b_{12} - a_{12} & a_{12}b_{23} - b_{13}a_{22} + a_{13}b_{22} - b_{12}a_{23} \\ a_{21} - b_{21} & b_{11} - a_{11} & a_{11}b_{23} - b_{13}a_{21} + a_{13}b_{21} - b_{11}a_{23} \end{pmatrix} \leq 2$$

Rows (1), (2) and (4) force the epipole to lie at infinity. This corresponds to an orthographic projection model, which indeed leads to affine transformations between two views of a planar object. But also without forcing the epipole to infinity there is one constraint left:

$$rank \begin{pmatrix} a_{23} - b_{23} & b_{13} - a_{13} & a_{13}b_{23} - b_{13}a_{23} \\ a_{22} - b_{22} & b_{12} - a_{12} & a_{12}b_{23} - b_{13}a_{22} + a_{13}b_{22} - b_{12}a_{23} \\ a_{21} - b_{21} & b_{11} - a_{11} & a_{11}b_{23} - b_{13}a_{21} + a_{13}b_{21} - b_{11}a_{23} \end{pmatrix} \leq 2$$

The actual consistency constraint used in our experiments is then

$$det \begin{pmatrix} a_{23} - b_{23} & b_{13} - a_{13} & a_{13}b_{23} - b_{13}a_{23} \\ a_{22} - b_{22} & b_{12} - a_{12} & a_{12}b_{23} - b_{13}a_{22} + a_{13}b_{22} - b_{12}a_{23} \\ a_{21} - b_{21} & b_{11} - a_{11} & a_{11}b_{23} - b_{13}a_{21} + a_{13}b_{21} - b_{11}a_{23} \end{pmatrix} \leq \delta$$

with δ a predefined threshold.

Acknowledgments

We are grateful to RobotVis INRIA Sophia-Antipolis for providing the Valbonne images (Fig. 13) and for financial support from the EC project VIBES and the IUAP project 'Advanced Mechatronical Systems'. Tinne Tuytelaars is a postdoctoral researcher funded by the Fund for Scientific Research Flanders (Belgium).

Notes

1. Alternatively, one could leave out this second factor, and compensate for the offsets by an appropriate normalization of the intensities before computing the moments.
2. For more information about these images, see `http://www.cs.sfu.ca/~color/image_db/index.html`.

References

Ballester, C. and Gonzalez, M. 1998. Affine invariant texture segmentation and shape from texture by variational methods. *Journal of Mathematical Imaging and Vision*, 9:141–171.

Baumberg, A. 2000. Reliable feature matching across widely separated views. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 774–781.

Canny, J.F. 1986. A computational approach to edge detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

Carlsson, S. 2000. Recognizing walking people. In *Proc. European Conference on Computer Vision*, pp. 472–486.

Dufournaud, Y., Schmid C., and Horaud, R. 2000. Matching image with different resolutions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 612–618.

Fischler, M.A. and Bolles, R.C. 1981. Random sampling consensus—A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. Assoc. Comp. Mach*, 24(6):381–395.

Funt, B., Barnard, K., and Martin, L. 1998. Is colour constancy good enough? In *Proc. European Conference on Computer Vision*, pp. 445–459.

Gruen, A.W. 1985. Adaptive least squares correlation: A powerful image matching technique. *Journal of Photogrammetry, Remote Sensing and Cartography*, 14(3):175–187.

Hall, D., Colin de Verdière, V., and Crowley, L. Object recognition using coloured receptive fields. In *Proc. European Conference on Computer Vision*, pp.164–177.

Harris, C. J. and Stephens, M. 1983. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, pp. 147–151.

Koenderink, J.J. and Van Doorn, A. 1987. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375.

Lindeberg, T. 1998. Feature detection with automatic scale selection. *Int. Journal of Computer Vision*, 30(2):79–116.

Lindeberg, T. and Gårding, J. 1997. Shape-adapted smoothing in estimation of 3D depth cues from affine distortions of local, 2D brightness structure. *Image and Vision Computing*, 15:415–434.

Lowe, D. 1999. Object recognition from local scale-invariant features. In *Proc. Int. Conf. on Computer Vision*, pp. 1150–1157.

Mindru, F., Moons, T., and Van Gool, L. 1999. Recognizing color patterns irrespective of viewpoint and illumination. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 368–373.

Montesinos, P., Gouet, V., and Pele, D. 2000. Matching color uncalibrated images using differential invariants. *Image and Vision Computing*, Special Issue BMVC2000, Elsevier Science, 18(9):659–671.

Pritchett, P. and Zisserman, A. 1998. Wide baseline stereo. In *Proc. Int. Conf. on Computer Vision*, pp. 754–759.

Schaffalitzky, F. and Zisserman, A. 2001. Viewpoint invariant texture matching and wide baseline stereo. In *Proc. Int. Conf. on Computer Vision*, pp. 636–643.

Schmid, C., Mohr, R., and Bauckhage, C. 1997. Local grey-value invariants for image retrieval. *Int. Journal on Pattern Analysis and Machine Intelligence*, 19(5):872–877.

Schmid, C. and Mohr, R. 1998. Comparing and evaluating interests points. In *Proc. Int. Conf. on Computer Vision*, pp. 230–235.

Shi, J. and Tomasi, C. 1994. Good features to track. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pp. 593–600.

Sinclair, D., Christensen, H., and Rothwell, C. 1995. Using the relation between a plane projectivity and the fundamental matrix. In *Proc. Scandinavian Conf. on Image Analysis*, pp. 181–188.

Super, B.J. and Klarquist, W.N. 1997. Patch matching and stereopsis in a general stereo viewing geometry. *Int. Journal on Pattern Analysis and Machine Intelligence*, 19(3):247–253.

Tell, D. and Carlsson, S. 2000. Wide baseline point matching using affine invariants computed from intensity profiles. In *Proc. European Conf. on Computer Vision*, pp. 814–828.

Tuytelaars, T. and Van Gool, L. 1999. Content-based image retrieval based on local affinely invariant regions. In *Proc. Int. Conf. on Visual Information Systems*, pp. 493–500.

Tuytelaars, T., Van Gool, L., D'haene, L., and Koch, R. 1999. Matching affinely invariant regions for visual servoing. In *Proc. Int. Conf. on Robotics and Automation*, pp. 1601–1606.

Tuytelaars, T., Turina, A., and Van Gool, L. 2003. Non-combinatorial detection of regular repetitions under perspective skew. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(4):418–432.