Wide Baseline Stereo Matching based on Local, Affinely Invariant Regions

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

[1] University of Leuven, ESAT-PSI,
Kard. Mercierlaan 94,
B-3001 Leuven, Belgium

[2] ETH, D-ELEK/IKT
Gloriastrasse 35
CH-8092 Zürich, Switzerland

[tuytelaa, vangool]@esat.kuleuven.ac.be

Abstract

'Invariant regions' are image patches that automatically deform with changing viewpoint as to keep on covering identical physical parts of a scene. Such regions are then described by a set of invariant features, which makes it relatively easy to match them between views and under changing illumination. In previous work, we have presented invariant regions that are based on a combination of corners and edges. The application discussed then was image database retrieval.

Here, an alternative method for extracting (affinely) invariant regions is given, that does not depend on the presence of edges or corners in the image but is purely intensity-based. Also, we demonstrate the use of such regions for another application, which is wide baseline stereo matching. 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 even further, 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.

1 Introduction

Local, invariant features are powerful tools for finding correspondences between different views of an object or scene. The local character yields robustness against occlusions and changing backgrounds. The invariance makes them immune against changes in viewpoint or illumination. An excellent example is the work by Schmid and Mohr [10]. They use relatively small, circular patches around corners. The surface textures they cover are characterized with invariant combinations of Gaussian derivatives. Invariance is under rotations, while invariance under scaling is handled by using circular neighborhoods of several sizes. Lowe *et al.* extended these ideas to real scale-invariance [5], using circular regions that maximize the output of a difference of Gaussians (DOG) filter in scale space. Special attention has been given to the efficiency of their implementation.

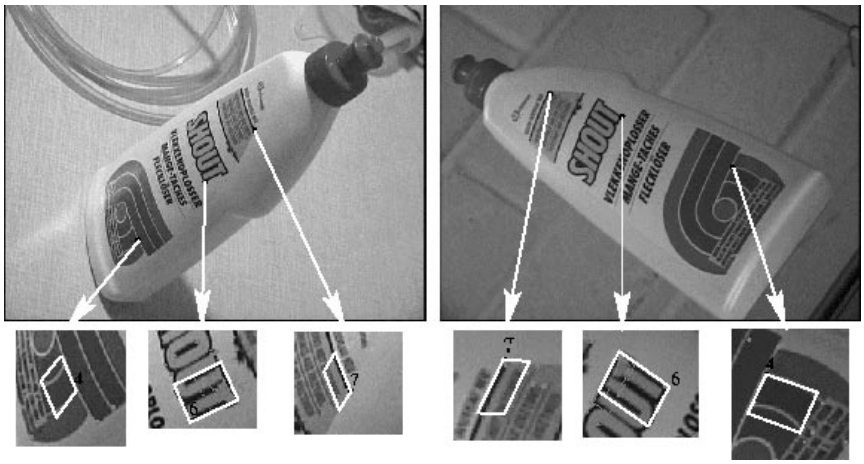

Figure 1: *Affinely invariant regions based on corners and edges.*

In previous work [14], we have extended this approach to affine invariance, by introducing a method to delineate parallelogram shaped regions that automatically adopt different shapes for different viewpoints, such that they systematically cover the same, physical part of a surface. The crux of the matter is that these corresponding shapes are determined solely on the basis of a single image, i.e. no other views are necessary. Corresponding regions are formed automatically in different images separately, without any knowledge about the other images. Once such invariant regions have been extracted, the texture they enclose is characterized with generalized color moment invariants. Both the region extraction and their invariant description are invariant under affine geometric changes and different scalings and offsets in the three color bands. The affine geometric invariance subsumes that the rather small regions correspond to almost planar surface patches. Such invariance implies that correspondences can be found under wider changes in viewpoint.

A disadvantage of the method is that it heavily relies on the accurate detection of geometric features such as corners and edges. It starts from a corner and its nearby edges. Two points move away from the corner in both directions along the edges. Their relative speed is coupled through the equality of relative affinely invariant parameters. At each position, the two points together with the corner define a parallelogram. The points stop at positions where simple photometric quantities of the texture covered by the parallelogram go through an extremum. Such quantity can be for instance the average value in one of the color bands. The whole procedure is invariant under the aforementioned geometric and photometric changes.

Figure 1 shows the invariant parallelograms for three pairs of corresponding points. Although there is a large image distortion between the two images, the affinely invariant parallelograms – which have been found for these images independently – cover similar, physical regions.

A first contribution in this paper is that we propose an alternative way of constructing affinely invariant regions, which does not rely on the presence of corners or edges. They are derived from image intensities directly. In contrast to a number of existing tex-

ture oriented approaches [1, 4] the process is non-iterative. The idea is not to replace the above parallelogram type regions, but rather to complement them with other types. This should result in an opportunistic system, that exploits a wide diversity of invariant regions depending on what is on offer. This should increase robustness and the number of correspondences found.

The focus in the work of Schmid and Mohr and our previous work was on object recognition and image database retrieval. In this paper we use invariant regions for extracting the epipolar geometry of wide baseline stereo setups. Examples are shown where the cameras have very different orientations. Our approach is akin to that of Pritchett and Zisserman [8] who start their wide baseline stereo algorithm by extracting quadrangles present in the image and match these based on a normalized cross-correlation to find local homographies, which are then extended to larger parts of the image. The difference is that here no special patterns, like quadrangles, are assumed to be directly visible in the image. Hence, the applicability is much wider. Recently, Tell and Carlsson [13] also proposed a wide baseline correspondence method based on affine invariance. They extract an affinely invariant Fourier description of the intensity profile along a line connecting two corner points. The non-local character of their method makes it more robust, but only suited for planar objects, which is a serious limitation on the applicability of their method.

A third contribution is the introduction of geometric and photometric constraints to check the consistency of potential correspondences. As will be shown, these constraints filter out false matches. We have found that the application of these constraints before using RANSAC [3] to extract epipolar geometry yields important improvements, certainly in the case where false matches strongly outnumber the good ones.

The remainder of the paper is organized as follows. First, the new, intensity-based method for extracting affinely invariant regions is discussed in section 2. Then, section 3 explains in more detail how the actual correspondence search, based on affine moment invariants computed over these regions, is carried out. Several consistency checks that can be used to reject false matches and hence to increase the overall robustness of the system are proposed in section 4. Finally, section 5 discusses some experimental results obtained with our system. Section 6 concludes the paper.

2 Intensity-based method

A major difficulty when extracting local, affine invariants is that they have to be computed over corresponding image regions. When the camera rotates about other axes than the optical axis, the shape of the region in the image should necessarily change with the viewpoint. This section presents a way of extracting such self-adaptive 'invariant regions'. The method is directly based on the analysis of intensity, without extraction of features such as edges or corners. It turns out to complement our previous method based on corners and edges well, in that invariant regions are typically found at other locations in the image.

Instead of taking corners as anchor points, the method starts from local extrema in the image intensity, extracted with a non-maximum suppression algorithm. Such points cannot be localized as accurately as corners, since the local extrema in intensity are often rather smooth, but they can withstand quite some changes in illumination and they are less likely to lie on the edge of an object resulting in a non-planar neighborhood. Besides, slight changes in their position do not affect the construction of the regions too badly. Of

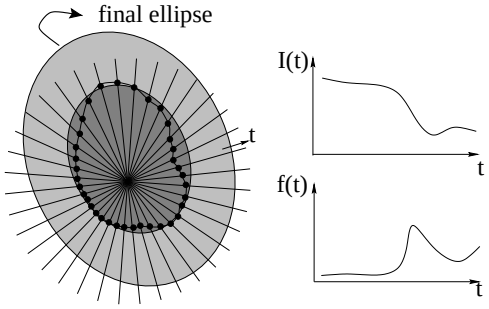

Figure 2: *The intensity along "rays" emanating from a local extremum are examined. The point on each ray for which a function f(t) reaches an extremum is selected. Linking these points together yields an affinely invariant region, to which an ellipse is fitted using moments.*

course, some illumination effects will defy our construction, e.g. specular highlights.

Given such a local extremum, the intensity function along rays emanating from the extremum is studied, as shown in figure 2. The following function is evaluated along each ray:

$$f(t) = \frac{|I(t) - I_0|}{max(\frac{\int_0^t |I(t) - I_0| dt}{t}, d)}$$

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 photometric transformations (given the ray). Typically, such extrema occur at positions where the intensity suddenly increases or decreases dramatically compared to the intensity changes we encountered on the line up to that point. Although $f(t)$ as such is not invariant to the geometric and photometric transformations we consider, the positions of its extrema are invariant. Note that in theory, leaving out the denominator in the expression for $f(t)$ would yield a simpler function which still has invariant positions for its local extrema. In practice, however, this simpler function does not give good results since its local extrema are too shallow, resulting in inaccurate positions along the rays and hence inaccurate regions. With the denominator added, on the other hand, the local extrema are localized quite accurately.

Next, all points corresponding to extrema of $f(t)$ along rays originating from the same local extremum are linked to enclose an (affinely invariant) region (see figure 2). This often irregularly-shaped region is then replaced by an ellipse having the same shape moments up to the second order. This ellipse-fitting is affinely invariant as well.

Note that the resulting region is not centered around the original anchor point (the intensity extremum). In fact, the whole procedure is quite robust to the inaccurate localization of this point. In most cases, small changes in its position will have no effect on the resulting region if the intensity profile is indeed showing a shallow extremum.

Finally, we double the area of the ellipses found. This leads to a higher distinctive power of the 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

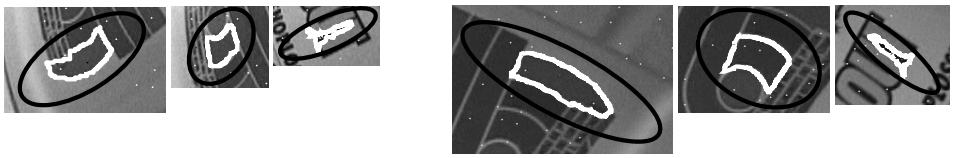

Figure 3: *Affinely invariant regions based on intensities only (black) and the linked points used to extract them (white).*

less local character of the regions.

Figure 3 again shows some details of the two very different views shown in fig. 1, with some corresponding invariant regions that were extracted using the intensity-based method (black) and the linked points on which the region extraction is based (white).

3 Finding Correspondences

3.1 Region Description

Once local, invariant regions have been extracted, finding correspondences between two views becomes much simpler. This is achieved by means of a nearest neighbor classification scheme, based on feature vectors containing moment invariants computed over the affinely invariant image regions. As in the region finding 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.

Each region is characterized by a feature vector of moment invariants. The moments we use are "Generalized Color Moments', which have been introduced in [7] 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* $a+b+c$. They yield a broader set of features to build the moment invariants from and, as a result, moment invariants that are simpler and more robust than the classical moment invariants. 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. These are invariant functions of moments up to the first order and second degree (i.e. moments that use up to second order powers of intensities (R,G,B) and first order powers of (x,y) coordinates). In [7] it has been proven that these 18 invariants form a basis for all geometric/photometric invariants involving this kind of moments. For an overview of the invariants used, see table 1. As an additional invariant – and, as shown by our experiments, quite a distinctive one – we use the region "type". This value refers to the method that has been used for the region extraction, i.e. is it a parallelogram-shaped region found on the basis of edges and corners, or an elliptic region found on the basis a an intensity extremum. We plan to extend the number of types in the future. Only if the type of two regions corresponds, can they be matched.

Table 1: Moment invariants used for comparing the patterns within an invariant region.

$inv[1] = S_{12}^{R} = \dfrac{\left\{ \begin{array}{l} M_{10}^{200} M_{01}^{100} M_{00}^{000} - M_{10}^{200} M_{00}^{100} M_{01}^{000} - M_{01}^{200} M_{10}^{100} M_{00}^{000} \\ +M_{01}^{200} M_{00}^{100} M_{10}^{000} + M_{00}^{200} M_{10}^{100} M_{01}^{000} - M_{00}^{200} M_{01}^{100} M_{10}^{000} \end{array} \right\}^2}{(M_{00}^{000})^2 \left[M_{00}^{200} M_{00}^{000} - (M_{00}^{100})^2 \right]^3}$
$inv[2] = S_{12}^{G} \qquad (similar)$
$inv[3] = S_{12}^{B} \qquad (similar)$
$inv[4] = D_{02}^{RG} = \dfrac{\left[M_{00}^{110} M_{00}^{000} - M_{00}^{100} M_{00}^{010} \right]^2}{\left[M_{00}^{200} M_{00}^{000} - (M_{00}^{100})^2 \right] \left[M_{00}^{020} M_{00}^{000} - (M_{00}^{010})^2 \right]}$
$inv[5] = D_{02}^{GB} \qquad (similar)$
$inv[6] = D_{02}^{BR} \qquad (similar)$
$inv[7] = D_{12}^{1RG} = \dfrac{\left\{ \begin{array}{l} M_{10}^{100} M_{01}^{010} M_{00}^{000} - M_{10}^{100} M_{00}^{010} M_{01}^{000} - M_{01}^{100} M_{10}^{010} M_{00}^{000} \\ +M_{01}^{100} M_{00}^{010} M_{10}^{000} + M_{00}^{100} M_{10}^{010} M_{01}^{000} - M_{00}^{100} M_{01}^{010} M_{10}^{000} \end{array} \right\}^2}{(M_{00}^{000})^4 \left[M_{00}^{200} M_{00}^{000} - (M_{00}^{100})^2 \right] \left[M_{00}^{020} M_{00}^{000} - (M_{00}^{010})^2 \right]}$
$inv[8] = D_{12}^{1GB} \qquad (similar)$
$inv[9] = D_{12}^{1BR} \qquad (similar)$
$inv[10] = D_{12}^{2RG} = \dfrac{\left\{ \begin{array}{l} (M_{00}^{000})^2 M_{10}^{100} M_{01}^{020} - M_{00}^{000} M_{10}^{100} M_{01}^{000} M_{00}^{020} \\ -2M_{00}^{000} M_{01}^{010} M_{00}^{010} M_{10}^{100} + 2M_{01}^{000} (M_{00}^{010})^2 M_{10}^{100} \\ -M_{00}^{000} M_{10}^{000} M_{00}^{100} M_{01}^{020} + 2M_{10}^{000} M_{00}^{010} M_{00}^{100} M_{01}^{010} \\ -(M_{00}^{000})^2 M_{01}^{100} M_{10}^{020} + M_{00}^{000} M_{01}^{100} M_{10}^{000} M_{00}^{020} \\ +2M_{00}^{000} M_{10}^{010} M_{00}^{010} M_{01}^{100} - 2M_{10}^{000} (M_{00}^{010})^2 M_{01}^{100} \\ +M_{00}^{000} M_{01}^{000} M_{00}^{100} M_{10}^{020} - 2M_{10}^{010} M_{00}^{100} M_{01}^{000} M_{00}^{010} \end{array} \right\}^2}{(M_{00}^{000})^4 \left[M_{00}^{200} M_{00}^{000} - (M_{00}^{100})^2 \right] \left[M_{00}^{020} M_{00}^{000} - (M_{00}^{010})^2 \right]^2}$
$inv[11] = D_{12}^{2GB} \qquad (similar)$
$inv[12] = D_{12}^{2BR} \qquad (similar)$
$inv[13] = D_{12}^{3RG} = \dfrac{\left\{ \begin{array}{l} (M_{00}^{000})^2 M_{10}^{010} M_{01}^{200} - M_{00}^{000} M_{10}^{010} M_{01}^{000} M_{00}^{200} \\ -2M_{00}^{000} M_{01}^{100} M_{00}^{100} M_{10}^{010} + 2M_{01}^{000} (M_{00}^{100})^2 M_{10}^{010} \\ -M_{00}^{000} M_{10}^{000} M_{00}^{010} M_{01}^{200} + 2M_{10}^{000} M_{00}^{100} M_{00}^{010} M_{01}^{100} \\ -(M_{00}^{000})^2 M_{01}^{010} M_{10}^{200} + M_{00}^{000} M_{01}^{010} M_{10}^{000} M_{00}^{200} \\ +2M_{00}^{000} M_{10}^{100} M_{00}^{100} M_{01}^{010} - 2M_{10}^{000} (M_{00}^{100})^2 M_{01}^{010} \\ +M_{00}^{000} M_{01}^{000} M_{00}^{010} M_{10}^{200} - 2M_{10}^{100} M_{00}^{100} M_{01}^{000} M_{00}^{010} \end{array} \right\}^2}{(M_{00}^{000})^4 \left[M_{00}^{200} M_{00}^{000} - (M_{00}^{100})^2 \right]^2 \left[M_{00}^{020} M_{00}^{000} - (M_{00}^{010})^2 \right]}$
$inv[14] = D_{12}^{3GB} \qquad (similar)$
$inv[15] = D_{12}^{3BR} \qquad (similar)$
$inv[16] = D_{12}^{4RG} = \dfrac{\left\{ \begin{array}{l} (M_{00}^{000})^2 M_{10}^{100} M_{01}^{110} - M_{00}^{000} M_{10}^{100} M_{01}^{000} M_{00}^{110} \\ -M_{00}^{000} M_{10}^{100} M_{00}^{100} M_{01}^{010} + M_{10}^{100} M_{00}^{100} M_{01}^{000} M_{00}^{010} \\ -M_{00}^{000} M_{10}^{000} M_{00}^{100} M_{01}^{110} + M_{10}^{000} (M_{00}^{100})^2 M_{01}^{010} \\ -M_{10}^{000} M_{00}^{100} M_{00}^{010} M_{01}^{100} - (M_{00}^{000})^2 M_{01}^{100} M_{10}^{110} \\ +M_{00}^{000} M_{01}^{100} M_{10}^{000} M_{00}^{110} + M_{00}^{000} M_{01}^{100} M_{00}^{100} M_{10}^{010} \\ +M_{00}^{000} M_{01}^{000} M_{00}^{100} M_{10}^{110} - M_{01}^{000} (M_{00}^{100})^2 M_{10}^{010} \end{array} \right\}^2}{(M_{00}^{000})^4 \left[M_{00}^{200} M_{00}^{000} - (M_{00}^{100})^2 \right]^2 \left[M_{00}^{020} M_{00}^{000} - (M_{00}^{010})^2 \right]}$
$inv[17] = D_{12}^{4GB} \qquad (similar)$
$inv[18] = D_{12}^{4BR} \qquad (similar)$

3.2 Matching Regions

Each region in the first image is then matched to the region in the other image for which the Mahalanobis-distance is minimal and below a predefined threshold d. Then, all regions of the second image are matched in a similar way to the regions of the first image. Only a mutual match is accepted as a real correspondence between the two views. The covariances needed to compute the Mahalanobis-distance are estimated based on all the regions found. Due to the different nature of the different region types, better results are obtained when different covariances are computed for each region type separately (based on all the regions of that type). The comparison of feature vectors can be done in an efficient way using hashing-techniques. At this moment, only hashing based on the region type has been implemented.

Once corresponding regions have been found, the 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 unit square or circle (depending on the region type) [1]. In this way, the effect of the geometric deformations on the cross-correlation is annihilated.

4 Robustness - Rejecting false matches

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, e.g. based on RANSAC [3], a robust method based on random sampling and rejecting all correspondences not conform with the found epipolar geometry. Although this method works fine in many applications, our experiments have shown that this approach may have difficulties in a typical wide baseline stereo setup, where false matches may outnumber the good matches. In that case, many of the randomly selected seven-point samples will 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 an outlier coincidentally yielding a reasonable amount of matches).

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 [2], who has recently proposed a view compatibility constraint for five points in two views based on a scaled orthographic camera model.

[1] For the circular regions, the correct relative 'orientation' is found by maximizing the cross-correlation.

4.1 A Geometric Constraint

Each match between two image regions defines an affine transformation, which is, in turn, an approximation of the homography linking the projections of all points lying in the same plane. All possible plane-related homographies between two images span a four-dimensional subspace of the nine-dimensional space of 3×3 matrices [6, 11]. However, this does not mean that a combination of more than four different homographies (i.e. four region correspondences) is needed before one is able to derive a constraint. Due to the special structure of this four-dimensional subspace, a constraint can already be derived given two different homographies. It is even possible to derive the fundamental matrix starting from two homographies [9].

Suppose we have two homographies H_1 and H_2, belonging to planes Π_1 and Π_2 respectively. 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 planes Π_1 and Π_2). This line of fixed points is used by Sinclair *et al* [12] to test whether two rigid planar motions are compatible. 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 derive here is an 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.

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 the appendix.

Suppose we have N correspondences, each linking a different local region in image I to a similar region in image I' by N different affine transformations. For each combination of two such correspondences, the above consistency constraint can be checked. A specific region correspondence is considered incorrect if it is consistent with less than n_g other correspondences (with n_g typically 8). Each good correspondence should have at least n_g 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.

4.2 A Photometric Constraint

Apart from geometric constraints, photometric constraints may be derived as well. Although it is not necessarily true that the illumination conditions are constant over whole the 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 as the region correspondence under consideration. So for each region correspondence, one should be able to

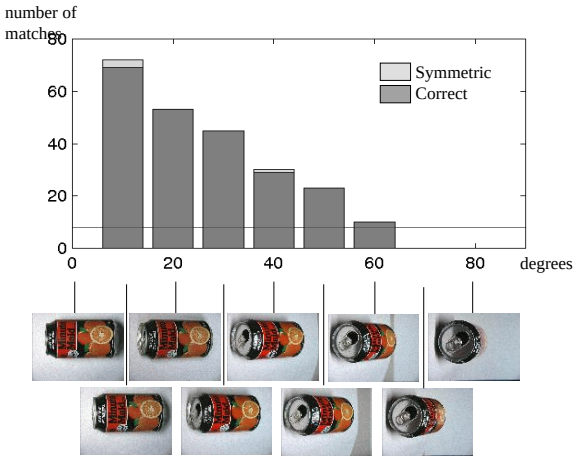

Figure 4: *Viewpoint invariance of the region extraction and matching: number of correct and symmetric matches found as a function of the rotation angle with respect to the reference view (0 degrees).*

find at least n_{ph} other region correspondences with a similar transform in the intensities (with n_{ph} typically 4).

First, the linear transformations linking the intensities in both images are computed for the region correspondences using moments. Then, these transformations are compared. To be consistent, only an overall scale factor is allowed, to compensate for different orientations. If not enough consistent region correspondences can be found, the region correspondence is rejected. Again, a few iterations may be needed.

5 Experimental Results

5.1 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. The results of this experiment are shown in figure 4.

For each image, the affinely invariant regions were extracted, and matched to the regions found in the 0 degrees reference image. Next, the regions were fine-tuned to optimize the cross-correlation and 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 classes: correct, symmetric or false. With 'symmetric' matches, we refer to those matches that do not link physically identical points between the two images, 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 two times the letter 'M'. Moreover, these letters are exactly below one another, such that they lie more or less on the same epipolar line as well due to the chosen camera movement. So there is no way for the system to distinguish between the regions on these two letters. The horizontal line added

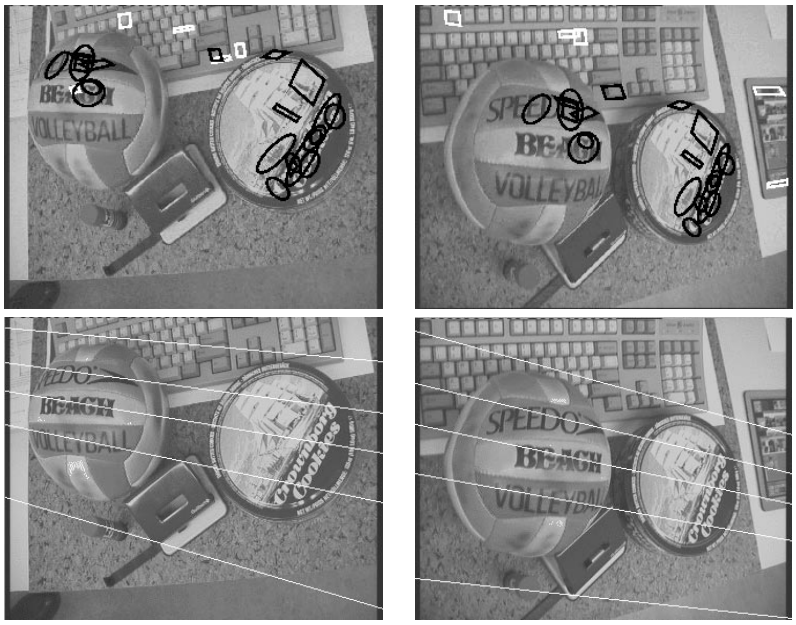

Figure 5: *Example 1: final region correspondences (top) (black for correct matches, white for false matches) and epipolar geometry (bottom).*

to the figure indicates the lower threshold of 8 correct matches needed for computing the epipolar geometry in a simple (linear) way.

Clearly, the system can deal very well with the changes in viewpoint up to 60 degrees. Only correct and symmetric matches were left. For larger angles, the epipolar test could no longer be applied, as the number of matches was too low. 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.

5.2 Wide Baseline Stereo Examples

As a first wide baseline example, consider the images shown in figure 5. The repetition on the keyboard caused many wrong matches at first, since more or less the same regions were found for most of the keys. Based on the semi-local constraints though, many of these wrong matches (17) were rejected. From the 26 matches left, 21 were correct, 3 were still caused by symmetries on the keyboard and 2 were completely false matches. All of these are shown superimposed on the images of figure 5. Applying the epipolar test using RANSAC to those 26 matches, allowed to reject the five outliers, and gave a quite accurate epipolar geometry, shown also in figure 5.

Figure 6 shows a second example. Although the distance between the two cameras is about 4 or 5 meter, and the change in orientation is over 50 degrees, several correct region correspondences could be found (after the filtering, using the geometric and photometric constraints and testing the epipolar geometry using RANSAC). These are shown superimposed on the upper images in figure 6. All these regions are matched correctly to

Figure 6: *Example 2: final region correspondences (all matches correct) and epipolar geometry.*

the corresponding region in the other image. As can be seen in the figures, not all region correspondences match perfectly. This is due to the lack of texture within these regions (e.g. the quadrangle on the wall), or due to the non-planarity of the neighborhood of the interest point. Nevertheless, the similarity between the regions was good enough for these match to be found and to be maintained throughout the filtering process.

Again, from those region correspondences, the epipolar geometry was derived using RANSAC. Some epipolar lines are shown at the bottom of figure 6. These clearly correspond very well over whole the image.

As a third and last wide baseline stereo example, look at the two images shown in figure 7. Note the large change in viewpoint, resulting in large changes in scale in some parts of the image, and extreme foreshortening in other parts. Nevertheless, sufficient matches were found for an accurate determination of the epipolar geometry.

6 Conclusion

In this contribution, some refinements to the system described in [14] are proposed. First, the robustness is substantially increased by introducing a new method for extracting local affinely invariant regions. This method is not meant to replace the one proposed in [14], but rather to be used as a complementary method. The final goal is to obtain an opportunistic system that exploits several types of invariant regions simultaneously. Other kind of regions might be developed as well, both general ones as more specific ones tuned towards specific features or applications. The more methods included in the system, the higher the number of correspondences found, the more accurate the resulting epipolar

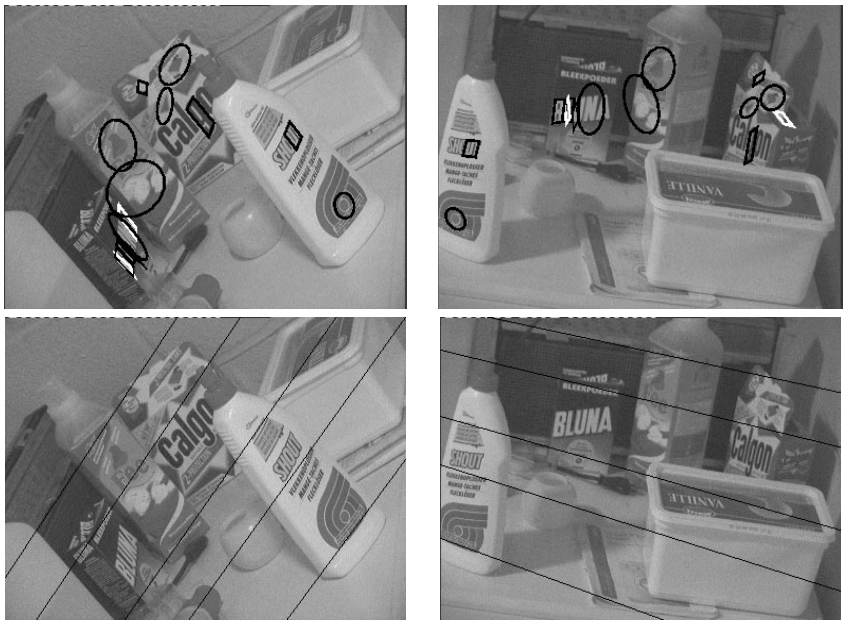

Figure 7: *Example 3: final region correspondences (top) (black for correct matches, white for false matches) and epipolar geometry (bottom).*

geometry and the wider the range of images to which the method will be applicable.

Second, two semi-local constraints have been proposed, that allow to test the geometric and photometric consistency of combinations of correspondences. This allows to reject falsely matched correspondences at an early stage. This is vital in such wide baseline stereo setups, where the number of false matches may be quite large.

Appendix: Derivation of the 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 $\mathbf{P} = (X, Y, Z)^T$. Then, both $H_1\mathbf{p}$ and $H_2\mathbf{p}$ lie on the epipolar line corresponding to the point $\mathbf{p}$. 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:

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

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 (from now on referred to as A and B) are of the following form:

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

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

The authors gratefully acknowledge support from the Flemish Fund for Scientific Research FWO and the Belgian IUAP project 'Intelligent Mechatronic Systems'.

References

[1] C. Ballester, M. Gonzalez, *Affine Invariant Texture Segmentation and Shape from Texture by Variational Methods*, Journal of Mathematical Imaging and Vision 9, pp. 141-171, 1998.

[2] S. Carlsson, *Recognizing walking people*, to appear in European Conference on Computer Vision, 2000.

[3] M.A.Fischler, R.C.Bolles *Random Sampling Consensus - a paradigm for model fitting with applications to image analysis and automated cartography*, Commun. Assoc. Comp. Mach, vol. 24, nr.6 :381-395, 1981.

[4] J. Gårding and T. Lindeberg *Direct computation of shape cues using scale-adapted spatial derivative operators*, Int'l Journal of Computer Vision, Vol. 17, no. 2, pp. 163-191, 1996.

[5] D. Lowe, *Object Recognition from Local Scale-Invariant Features*, Int'l Conference on Computer Vision, pp. 1150-1157, 1999.

[6] Q. T. Luong, O. D. Faugeras, *The Fundamental Matrix: theory, algorithms, and stability analysis*, Int'l. Journal on Computer Vision, vol. 17, nr. 1, pp. 43-75, 1996.

[7] F. Mindru, T. Moons and L. Van Gool *Recognizing color patterns irrespective of viewpoint and illumination*, IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 368-373, 1999.

[8] P. Pritchett and A. Zisserman, *Wide Baseline Stereo*, Int'l Conference on Computer Vision, pp. 754-759, 1998.

[9] P. Pritchett and A. Zisserman, *Matching and Reconstruction from Widely Separated Views*, Proc. SMILE Workshop, LNCS 1506, pp.138-153, Springer-Verlag, 1998.

[10] C. Schmid, R. Mohr, C. Bauckhage *Local Grey-value Invariants for Image Retrieval*, Int'l Journal on Pattern Analysis and Machine Intelligence Vol. 19, no. 5, pp. 872-877, 1997.

[11] Shashua, Avidan *The rank 4 constraint in multiple view geometry*, European Conference on Computer Vision, Vol. 2, pp. 196-206, 1996.

[12] D. Sinclair, H. Christensen, C. Rothwell. *Using the Relation between a Plane Projectivity and the Fundamental Matrix*, Proc. SCIA, pp. 181-188, 1995.

[13] D. Tell, S. Carlsson, *Wide baseline point matching using affine invariants computed from intensity profiles*, to appear in European Conference on Computer Vision, 2000.

[14] T. Tuytelaars and L. Van Gool, *Content-based Image Retrieval based on Local Affinely Invariant Regions*, Int'l Conference on Visual Information Systems, pp. 493-500, 1999.