Similarity Measures for Occlusion, Clutter, and Illumination Invariant Object Recognition

Carsten Steger

MVTec Software GmbH
Neherstraße 1, 81675 München, Germany
Phone: +49 (89) 457695-0, Fax: +49 (89) 457695-55
steger@mvtec.com

Abstract Novel similarity measures for object recognition and image matching are proposed, which are inherently robust against occlusion, clutter, and nonlinear illumination changes. They can be extended to be robust to global as well as local contrast reversals. The similarity measures are based on representing the model of the object to be found and the image in which the model should be found as a set of points and associated direction vectors. They are used in an object recognition system for industrial inspection that recognizes objects under Euclidean transformations in real time.

1 Introduction

Object recognition is used in many computer vision applications. It is particularly useful for industrial inspection tasks, where often an image of an object must be aligned with a model of the object. The transformation (pose) obtained by the object recognition process can be used for various tasks, e.g., pick and place operations or quality control. In most cases, the model of the object is generated from an image of the object. This 2D approach is taken because it usually is too costly or time consuming to create a more complicated model, e.g., a 3D CAD model. Therefore, in industrial inspection tasks one is usually interested in matching a 2D model of an object to the image. The object may be transformed by a certain class of transformations, depending on the particular setup, e.g., translations, Euclidean transformations, similarity transformations, or general 2D affine transformations (which are usually taken as an approximation to the true perspective transformations an object may undergo).

Several methods have been proposed to recognize objects in images by matching 2D models to images. A survey of matching approaches is given in [3]. In most 2D matching approaches the model is systematically compared to the image using all allowable degrees of freedom of the chosen class of transformations. The comparison is based on a suitable similarity measure (also called match metric). The maxima or minima of the similarity measure are used to decide whether an object is present in the image and to determine its pose. To speed up the recognition process, the search is usually done in a coarse-to-fine manner, e.g., by using image pyramids [10].

The simplest class of object recognition methods is based on the gray values of the model and image itself and uses normalized cross correlation or the sum of squared or absolute differences as a similarity measure [3]. Normalized cross correlation is invariant to linear brightness changes but is very sensitive to clutter and occlusion as well as nonlinear contrast changes. The sum of gray value differences is not robust to any of these changes, but can be made robust to linear brightness changes by explicitly incorporating them into the similarity measure, and to a moderate amount of occlusion and clutter by computing the similarity measure in a statistically robust manner [6].

A more complex class of object recognition methods does not use the gray values of the model or object itself, but uses the object's edges for matching [2,8]. In all existing approaches, the edges are segmented, i.e., a binary image is computed for both the model and the search image. Usually, the edge pixels are defined as the pixels in the image where the magnitude of the gradient is maximum in the direction of the gradient. Various similarity measures can then be used to compare the model to the image. The similarity measure in [2] computes the average distance of the model edges and the image edges. The disadvantage of this similarity measure is that it is not robust to occlusions because the distance to the nearest edge increases significantly if some of the edges of the model are missing in the image.

The Hausdorff distance similarity measure used in [8] tries to remedy this shortcoming by calculating the maximum of the k-th largest distance of the model edges to the image edges and the l-th largest distance of the image edges and the model edges. If the model contains n points and the image contains m edge points, the similarity measure is robust to $100k/n$% occlusion and $100l/m$% clutter. Unfortunately, an estimate for m is needed to determine l, which is usually not available.

All of these similarity measures have the disadvantage that they do not take the direction of the edges into account. In [7] it is shown that disregarding the edge direction information leads to false positive instances of the model in the image. The similarity measure proposed in [7] tries to improve this by modifying the Hausdorff distance to also measure the angle difference between the model and image edges. Unfortunately, the implementation is based on multiple distance transformations, which makes the algorithm too computationally expensive for industrial inspection.

Finally, another class of edge based object recognition algorithms is based on the generalized Hough transform [1]. Approaches of this kind have the advantage that they are robust to occlusion as well as clutter. Unfortunately, the GHT requires extremely accurate estimates for the edge directions or a complex and expensive processing scheme, e.g., smoothing the accumulator space, to determine whether an object is present and to determine its pose. This problem is especially grave for large models. The required accuracy is usually not obtainable, even in low noise images, because the discretization of the image leads to edge direction errors that already are too large for the GHT.

In all of the above approaches, the edge image is binarized. This makes the object recognition algorithm invariant only against a narrow range of illumination changes. If the image contrast is lowered, progressively fewer edge points will be segmented, which has the same effects as progressively larger occlusion. The similarity measures proposed in this paper overcome all of the above problems and result in an object recognition strategy robust against occlusion, clutter, nonlinear illumination changes, and a relatively large amount of defocusing. They can be extended to be robust to global as well as local contrast reversals.

2 Similarity Measures

The model of an object consists of a set of points $p_i = (x_i, y_i)^T$ with a corresponding direction vector $d_i = (t_i, u_i)^T$, $i = 1, \ldots, n$. The direction vectors can be generated by a number of different image processing operations, e.g., edge, line, or corner extraction, as discussed in Section 3. Typically, the model is generated from an image of the object, where an arbitrary region of interest (ROI) specifies that part of the image in which the object is located. It is advantageous to specify the coordinates p_i relative to the center of gravity of the ROI of the model or to the center of gravity of the points of the model.

The image in which the model should be found can be transformed into a representation in which a direction vector $e_{x,y} = (v_{x,y}, w_{x,y})^T$ is obtained for each image point (x, y). In the matching process, a transformed model must be compared to the image at a particular location. In the most general case considered here, the transformation is an arbitrary affine transformation. It is useful to separate the translation part of the affine transformation from the linear part. Therefore, a linearly transformed model is given by the points $p'_i = Ap_i$ and the accordingly transformed direction vectors $d'_i = Ad_i$, where

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad .$$

As discussed above, the similarity measure by which the transformed model is compared to the image must be robust to occlusions, clutter, and illumination changes. One such measure is to sum the (unnormalized) dot product of the direction vectors of the transformed model and the image over all points of the model to compute a matching score at a particular point $q = (x, y)^T$ of the image, i.e., the similarity measure of the transformed model at the point q, which corresponds to the translation part of the affine transformation, is computed as follows:

$$s = \frac{1}{n} \sum_{i=1}^{n} \langle d'_i, e_{q+p'} \rangle = \frac{1}{n} \sum_{i=1}^{n} t'_i v_{x+x'_i, y+y'_i} + u'_i w_{x+x'_i, y+y'_i} \quad . \tag{1}$$

If the model is generated by edge or line filtering, and the image is preprocessed in the same manner, this similarity measure fulfills the requirements of robustness to occlusion and clutter. If parts of the object are missing in the image, there are no lines or edges at the corresponding positions of the model in the image, i.e., the direction vectors will have a small length and hence contribute little to the sum. Likewise, if there are clutter lines or edges in the image, there will either be no point in the model at the clutter position or it will have a small length, which means it will contribute little to the sum.

The similarity measure (1) is not truly invariant against illumination changes, however, since usually the length of the direction vectors depends on the brightness of the image, e.g., if edge detection is used to extract the direction vectors. However, if a user specifies a threshold on the similarity measure to determine whether the model is present in the image, a similarity measure with a well defined range of values is desirable. The following similarity measure achieves this goal:

$$s = \frac{1}{n} \sum_{i=1}^{n} \frac{\langle d'_i, e_{q+p'} \rangle}{\|d'_i\| \cdot \|e_{q+p'}\|} = \frac{1}{n} \sum_{i=1}^{n} \frac{t'_i v_{x+x'_i, y+y'_i} + u'_i w_{x+x'_i, y+y'_i}}{\sqrt{t'^2_i + u'^2_i} \cdot \sqrt{v^2_{x+x'_i, y+y'_i} + w^2_{x+x'_i, y+y'_i}}} \quad . \tag{2}$$

Because of the normalization of the direction vectors, this similarity measure is additionally invariant to arbitrary illumination changes since all vectors are scaled to a length of 1. What makes this measure robust against occlusion and clutter is the fact that if a feature is missing, either in the model or in the image, noise will lead to random direction vectors, which, on average, will contribute nothing to the sum.

The similarity measure (2) will return a high score if all the direction vectors of the model and the image align, i.e., point in the same direction. If edges are used to generate the model and image vectors, this means that the model and image must have the same contrast direction for each edge. Sometimes it is desirable to be able to detect the object even if its contrast is reversed. This is achieved by:

$$s = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\langle d'_i, e_{q+p'} \rangle}{\|d'_i\| \cdot \|e_{q+p'}\|} \right| \quad . \tag{3}$$

In rare circumstances, it might be necessary to ignore even local contrast changes. In this case, the similarity measure can be modified as follows:

$$s = \frac{1}{n} \sum_{i=1}^{n} \frac{|\langle d'_i, e_{q+p'} \rangle|}{\|d'_i\| \cdot \|e_{q+p'}\|} \quad . \tag{4}$$

The above three normalized similarity measures are robust to occlusion in the sense that the object will be found if it is occluded. As mentioned above, this results from the fact that the missing object points in the instance of the model in the image will on average contribute nothing to the sum. For any particular instance of the model in the image, this may not be true, e.g., because the noise in the image is not uncorrelated. This leads to the undesired fact that the instance of the model will be found in different poses in different images, even if the model does not move in the images, because in a particular image of the model the random direction vectors will contribute slightly different amounts to the sum, and hence the maximum of the similarity measure will change randomly. To make the localization of the model more precise, it is useful to set the contribution of direction vectors caused by noise in the image to zero. The easiest way to do this is to set all inverse lengths $1/\|e_{q+p'}\|$ of the direction vectors in the image to 0 if their length $\|e_{q+p'}\|$ is smaller than a threshold that depends on the noise level in the image and the preprocessing operation that is used to extract the direction vectors in the image. This threshold can be specified easily by the user. By this modification of the similarity measure, it can be ensured that an occluded instance of the model will always be found in the same pose if it does not move in the images.

The normalized similarity measures (2)–(4) have the property that they return a number smaller than 1 as the score of a potential match. In all cases, a score of 1 indicates a perfect match between the model and the image. Furthermore, the score roughly corresponds to the portion of the model that is visible in the image. For example, if the object is 50% occluded, the score (on average) cannot exceed 0.5. This is a highly desirable property because it gives the user the means to select an intuitive threshold for when an object should be considered as recognized.

A desirable feature of the above similarity measures (2)–(4) is that they do not need to be evaluated completely when object recognition is based on a threshold $s_{\min}$ for the similarity measure that a potential match must achieve. Let s_j denote the partial sum of the dot products up to the j-th element of the model. For the match metric that uses the sum of the normalized dot products, this is:

$$s_j = \frac{1}{n} \sum_{i=1}^{j} \frac{\langle d'_i, e_{q+p'} \rangle}{\|d'_i\| \cdot \|e_{q+p'}\|} \quad . \tag{5}$$

Obviously, all the remaining terms of the sum are all ≤ 1. Therefore, the partial score can never achieve the required score $s_{\min}$ if $s_j < s_{\min} - 1 + j/n$, and hence the evaluation of the sum can be discontinued after the j-th element whenever this condition is fulfilled. This criterion speeds up the recognition process considerably.

3 Object Recognition

The above similarity measures are applied in an object recognition system for industrial inspection that recognizes objects under Euclidean transformations, i.e., translation and rotation, in real time. Although only Euclidean transformations are implemented at the moment, extensions to similarity or general affine transformations are not difficult to implement. The system consists of two modules: an offline generation of the model and an online recognition.

The model is generated from an image of the object to be recognized. An arbitrary region of interest specifies the object's location in the image. Usually, the ROI is specified by the user. Alternatively, it can be generated by suitable segmentation techniques. To speed up the recognition process, the model is generated in multiple resolution levels, which are constructed by building an image pyramid from the original image. The number of pyramid levels $l_{\max}$ is chosen by the user.

Each resolution level consists of all possible rotations of the model, where thresholds $\phi_{\min}$ and $\phi_{\max}$ for the angle are selected by the user. The step length for the discretization of the possible angles can either be done automatically by a method similar to the one described in [2] or be set by the user. In higher pyramid levels, the step length for the angle is computed by doubling the step length of the next lower pyramid level.

The rotated models are generated by rotating the original image of the current pyramid level and performing the feature extraction in the rotated image. This is done because the feature extractors may be anisotropic, i.e., the extracted direction vectors may depend on the orientation of the feature in the image in a biased manner. If it is known that the feature extractor is isotropic, the rotated models may be generated by performing the feature extraction only once per pyramid level and transforming the resulting points and direction vectors.

The feature extraction can be done by a number of different image processing algorithms that return a direction vector for each image point. One such class of algorithms are edge detectors, e.g, the Sobel or Canny [4] operators. Another useful class of algorithms are line detectors [9]. Finally, corner detectors that return a direction vector, e.g., [5], could also be used. Because of runtime considerations the Sobel filter is used in the current implementation of the object recognition system. Since in industrial inspection the lighting can be controlled, noise does not pose a significant problem in these applications.

To recognize the model, an image pyramid is constructed for the image in which the model should be found. For each level of the pyramid, the same filtering operation that was used to generate the model, e.g., Sobel filtering, is applied to the image. This returns a direction vector for each image point. Note that the image is not segmented, i.e., thresholding or other operations are not performed. This results in true robustness to illumination changes.

To identify potential matches, an exhaustive search is performed for the top level of the pyramid, i.e., all precomputed models of the top level of the model resolution hierarchy are used to compute the similarity measure via (2), (3), or (4) for all possible poses of the model. A potential match must have a score larger than a user-specified threshold $s_{\min}$ and the corresponding score must be a local maximum with respect to neighboring scores. As described in Section 2, the threshold $s_{\min}$ is used to speed up the search by terminating the evaluation of the similarity measure as early as possible.

After the potential matches have been identified, they are tracked through the resolution hierarchy until they are found at the lowest level of the image pyramid. Various search strategies like depth-first, best-first, etc., have been examined. It turned out that a breadth-first strategy is preferable for various reasons, most notably because a heuristic for a best-first strategy is hard to define, and because depth-first search results in slower execution if all matches should be found.

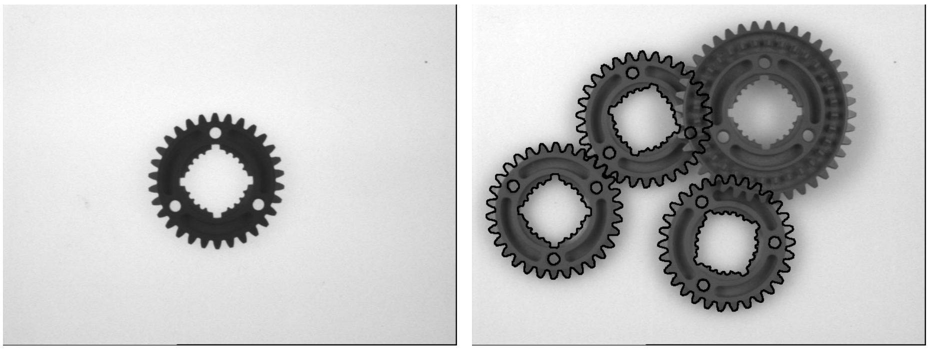
Once the object has been recognized on the lowest level of the image pyramid, its position and rotation are extracted to a resolution better than the discretization of the search space, i.e., the translation is extracted with subpixel precision and the angles with a resolution better than the angle step length. This is done by fitting a second order polynomial (in the three pose variables) to the similarity measure values in a $3 \times 3 \times 3$ neighborhood around the maximum score. The coefficients of the polynomial are obtained by convolution with 3D facet model masks. The corresponding 2D masks are given in [9]. They generalize to arbitrary dimensions in a straightforward manner.

4 Examples

Figure 1 displays an example of recognizing multiple objects. To illustrate the robustness against nonlinear illumination changes, the model image in Figure 1(a) was acquired using back lighting. Figure 1(b) shows that all three cog wheels have been recognized correctly despite the fact that front lighting is used and that a fourth cog wheel occludes two of the other cog wheels.

5 Conclusions

A new class of similarity measures for object recognition and image matching, which are inherently robust against occlusion, clutter, nonlinear illumination changes, and global as well as local contrast reversals, have been proposed. The similarity measures are used in an object recognition system for industrial inspection that is able to recognize objects under Euclidean transformations in video frame rate. The system is able to achieve an accuracy of 1/22 pixel and 1/12 degree on real images.



(a) Image of model object

(b) Found objects

Figure 1. Example of recognizing multiple objects. To illustrate the robustness against illumination changes, the model image uses back lighting while the search image uses front lighting.

Future work will focus on extending the object recognition system to handle at least similarity transformations and possibly general affine transformations.

References

1. D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.
2. Gunilla Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):849–865, November 1988.
3. Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, December 1992.
4. John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, June 1986.
5. Wolfgang Förstner. A framework for low level feature extraction. In Jan-Olof Eklundh, editor, *Third European Conference on Computer Vision*, volume 801 of *Lecture Notes in Computer Science*, pages 383–394, Berlin, 1994. Springer-Verlag.
6. Shang-Hong Lai and Ming Fang. Robust and efficient image alignment with spatially varying illumination models. In *Computer Vision and Pattern Recognition*, volume II, pages 167–172, 1999.
7. Clark F. Olson and Daniel P. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *IEEE Transactions on Image Processing*, 6(1):103–113, January 1997.
8. William J. Rucklidge. Efficiently locating objects using the Hausdorff distance. *International Journal of Computer Vision*, 24(3):251–270, 1997.
9. Carsten Steger. An unbiased detector of curvilinear structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(2):113–125, February 1998.
10. Steven L. Tanimoto. Template matching in pyramids. *Computer Graphics and Image Processing*, 16:356–369, 1981.