Spiders as Robust Point Descriptors

Adam Stanski and Olaf Hellwich

Computer Vision & Remote Sensing Group, Technical University of Berlin,
FR 3-1, Franklinstr. 28/29, 10587 Berlin, Germany
Email: {astanski, hellwich}@cs.tu-berlin.de

Abstract. This paper introduces a new operator to characterize a point in an image in a distinctive and invariant way. The robust recognition of points is a key technique in computer vision: algorithms for stereo correspondence, motion tracking and object recognition rely heavily on this type of operator. The goal in this paper is to describe the salient point to be characterized by a constellation of surrounding anchor points. Salient points are the most reliably localized points extracted by an interest point operator. The anchor points are multiple interest points in a visually homogenous segment surrounding the salient point. Because of its appearance, this constellation is called a spider. With a prototype of the spider operator, results in this paper demonstrate how a point can be recognized in spite of significant image noise, inhomogeneous change in illumination and altered perspective. For an example that requires a high performance close to object / background boundaries, the prototype yields better results than David Lowe's SIFT operator.

1 Introduction

Numerous algorithms in computer vision require the recognition of the same point in images that differ in perspective, illumination, position of objects, image noise, etc. Homologous points are needed for depth reconstruction of two stereo images, for motion tracking and in most object recognition methods (e.g. [1]). Regardless of the application, the problem always remains the same: How can the position of a point be described in a distinctive and invariant way? Distinctive in the sense that the description is strongly differing from that of other points. Invariant in a way that a change in the surrounding area of a point caused by rotation, scaling, variation of illumination, etc. does not alter its description beyond recognition.

In this paper the problem is solved with spiders. A spider characterizes a salient point with a constellation of anchor points, whose positions are stored relative to the salient point. The following three steps summarize the approach:

1. *Interest Point Localization:* An interest point operator is utilized to localize points and associates each one with a reliability to be repeatedly found in an altered image. Some points with high reliability are chosen as the salient point of spiders.
2. *Extraction of Image Segment*: For each salient point, a segment of the image is determined that is visually homogenous. All points extracted by the interest point operator, which are also part of this segment, are the anchor points be-

longing to this salient point. A salient point with multiple surrounding anchor points forms a spider.

3. *Comparison of Spiders*: To determine the similarity of two salient points, the constellation of their anchor points has to be compared. However, when observed from a different perspective, the geometric constellation of the anchor points is distorted. Therefore, an affine transformation to compensate for this distortion is estimated before comparing two spiders.

Figure 1 gives an example of a typical spider in an image, determined by the prototype of the spider operator:

Fig. 1. An example of the determination of a spider. From left to right: 1) The original image (a part of the Lena image). 2) Interest points extracted with Lowe's scale space extrema operator are highlighted. A reliable point is marked, which is used as a salient point for the spider in the following. 3) A segment of pixels with a locally averaged brightness similar to the salient point is highlighted. Interest points in this area are used as the anchor points of the spider. 4) The final spider with 14 anchor points

2 Related Research

The concept that is most often used to describe a point is the simplest one, too: the area surrounding a point is compared pixel by pixel. Usually this area is square or circular in shape, with the salient point at its center. Two points are compared by aggregating the differences of every corresponding pair of pixels. There are several ways to do this. For example, the difference between two pixels can be described by the difference of their intensities (absolute intensity differences) or their squared intensity (squared intensity differences). The aggregation may be based on simple addition (e.g. sum-of-squared-differences) or on additional normalization (e.g. normalized-grayscale-correlation). This class of methods is successfully used in, for example small baseline stereo matching (e.g. Scharstein [6]) and object recognition (e.g. Leibe [2]). The algorithms are fast, simple, robust with respect to image noise and partly invariant in relation to homogenous changes of illumination or contrast. However, other changes of the image, such as changes in perspective or rotation, change the description of a point so much that it can no longer be identified by these methods.

Another concept for the description of a point is based on calculations on a group of pixels. The calculation aims at generating features that identify a point more distinctively and with increased invariance than possible by direct comparison of pixels. Two examples of current methods that calculate features from scaled round or affine transformed squared regions are proposed by Mikolaiczyk and Schmid [4] and Lowe [3]. The method in the first paper begins with determining interest points with an af-

fine-invariant Harris operator. In a second step their scale is calculated by searching for local scale space extremas. Finally, the affine shape of the surrounding area of a point is estimated by a second moment matrix.

Lowe's method can also be divided into multiple steps. First, the scale of a point is determined by detecting local scale space extremas. In a second step, a quadratic function is fitted to the results, so that the position and scale of the points can be calculated more precisely. In a third step the orientation of points is determined by local image gradients. Finally, the surrounding area is normalized relative to scale and orientation in order to describe the point with the SIFT descriptor – a feature based on several local gradients in the surrounding area.

Lowe's method provides a higher density of distinctive points than the approach of Mikolaiczyk and Schmid. However, the interest points calculated by Lowe in the first step are only scale-invariant, which diminishes the tolerance of his method compared to a full affine transformation. Both approaches cannot maintain their generally good performance when it comes to object / background boundaries. This is impossible for a point on the silhouette of an object as there are too many pixels in the surrounding area that belong to the background, which falsifies the calculation. For example, points in images containing trees, animals or people cannot be detected in a reliable way, if the parts of the objects (here branches and limbs) are only a few pixels wide.

Two approaches that solve this problem are proposed by Mikolajczyk, Zisserman und Schmid [5] and by Tuytelaars und Van Gool [8]. The first paper uses a description related to Lowe's SIFT Descriptor. However, the initial points are determined on edges and their scale is calculated relative to the opposite edge. Every point is now described by two independent descriptors: each describing the pixel on one side of the edge on which the point is situated. In case a point is indeed close to the silhouette of an object, one descriptor will refer only to the object and the other one only to the background. In the second paper from Tuytelaars and Van Gool, edges are also utilized to determine the area on which the description of a point is based. A parallelogram is fit to every two edges that form a corner. The description of a point is now calculated based on the pixels inside this parallelogram. As the edges the parallelogram was fit to usually do not extend beyond object boundaries, it rarely crosses an object border. Both approaches work with Canny edge detection, which utilizes two thresholds. Slight changes in an image can therefore cause an interruption of edges; they may not be detected at all or connected differently. This lack of stability has a negative effect on the performance of both methods.

3 Interest Point Localization

The first step of the proposed spider operator is to detect stable points. The most important property of these points is that they can be localized in spite of changed light conditions, viewpoint on a scene, etc. Multiple candidates with this characteristic were examined, such as points detected with the Harris operator or the ends of edges. The prototype examined in this paper utilizes scale space extremas, as described by Lowe in [3]. The interest points detected by Lowe's operator are the centers of round areas of different size that are lighter or darker than their environment. Basically, they

are light or dark blobs in the image. Utilizing multiple interest point operators is another possibility, as long as all of them yield stable points.

The implemented version of the scale space extrema operator yields more stable points than the Harris operator and a newly developed approach looking for the ends of edges in an image. However, it does not achieve quite the same good results as given by Lowe for unknown reasons. While trying to obtain a higher performance, a new way to determine the reliability of the points was identified that enhances their stability. Instead of using the absolute value of a scale space extrema as a criterion for its stability, its relative value is used in the prototype: The reliability of a maximum depends on its value minus the highest value of the 98 pixels, which have a chessboard distance of 2 to this maximum of the 3d scale space. The reliability of a minimum found in scale space depends on the lowest of these values minus the value of this minimum.

4 Extraction of Image Segment

After a salient point is located, the segment which belongs to this point must be determined. A segment is a connected area of the image, in which each pixel has a high reliability of being on the same plain as the salient point. A segment is approximated by extracting an area, which is visually similar to the salient point. Multiple techniques known from image segmentation based on color gradients, textures or boundary smoothness could be applied here.

The prototype of the spider operator simply aggregates adjacent pixels starting from the salient point, as long as the difference of the brightness between the pixel and the salient point is below a threshold. This aggregation is performed on a Gaussian blurred image to reduce the influence of single pixels.

All points that are extracted by the interest point operator and are also part of the segment of a salient point form its anchor points. Usually, a salient point can be recognized even if its segment is extracted differently in an altered image, as long as the majority of its anchor points is detected in both images.

5 Comparison of Spiders

If two images show the same point from an altered perspective, the spiders calculated on the two views of the point will be different also. An example of this can be seen in the two leftmost images of figure 2, of which one is a rotated and scaled version of the other. To identify the two points as the same, the perspective distortion of one spider has to be eliminated by a transformation, thereby changing the shape of the spider, so that its appearance becomes similar to the other. The method for matching two spiders is illustrated in figure 2 and described in the remainder of this section.

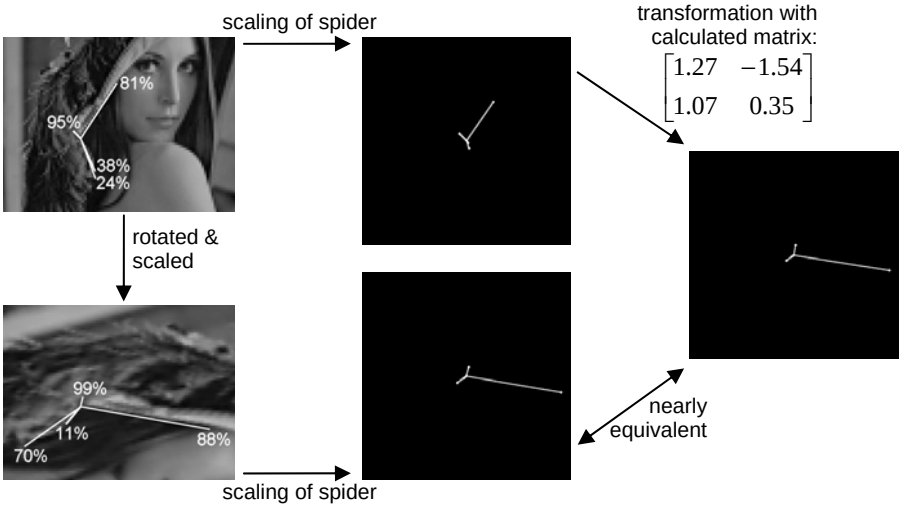

Fig. 2. Simplified example for the comparison of two spiders. Upper left: the original image with a spider and the reliabilities of all four anchor points. Lower left: a version of the original image scaled by factor 1.3 horizontally and 2.0 vertically, than rotated clockwise by 68.4°. The spider in this image is distorted, because of the transformation of the image. Middle column: the spiders from the two left images, whose anchor points are scaled with equation 2 according to their reliability. The two most unreliable anchor points (with 38% and 24% in the upper and 11% and 70% in the lower image) are shortened strongly, so that one of them is barely visible. Right image: By transformation with the given matrix, which was determined with equation 3, the upper spider becomes nearly equivalent with the lower spider. This makes it possible to determine reliably that both salient points are actually the same

To calculate this transformation in practice the following three assumptions are made: 1) The salient point and all anchor points are assumed to be coplanar. This should be the case in the segment of a salient point and reduces the necessary calculations as well as the number of anchor points required. 2) The perspective distortion is estimated with an affine transformation, which also reduces the computational effort. 3) As it is assumed that the salient point of both spiders is identical and that the coordinates of the anchor points are given relative to the salient point, translations need not to be considered. This leaves us with a matrix containing only four unknowns, which transforms an anchor point (x_1, y_1) of a spider S_1 into a point (x_2, y_2) of a spider S_2:

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \tag{1}$$

The determination of anchor points is error-prone. Thus, to match two spiders with this transformation, the error of the position of its anchor points has to be estimated. As they should be well localized, the probability of detecting an anchor point decreases with an increased distance to its correct position. Therefore, a least-squares optimal solution is chosen. However, this is a very poor estimate for some cases, e.g. if the search for an anchor point reveals that multiple positions in a large area are

likely to be its correct position. Hence, an additional reliability is determined for each anchor point, which can be interpreted as the probability that its detected position is close to its correct location. The lower the reliability of a point, the weaker the influence of its detected location on the matching of spiders should be. As an anchor point is matched with a corresponding one, both of their reliabilities determine their influence on the whole matching. This is achieved by scaling a pair of corresponding anchor points P_1 and P_2 with their combined reliabilities r_1 and r_2:

$$\begin{bmatrix} x_1' \\ y_1' \end{bmatrix} = r_1 \quad r_2 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_2' \\ y_2' \end{bmatrix} = r_1 \quad r_2 \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \tag{2}$$

It is possible to determine an optimal solution for the stated problem (see e.g. Strang [7]). First, equations 1 and 2 are converted into a linear equation. Each anchor point contributes to two more rows in the matrix and the last vector (at least two points are required):

$$\begin{bmatrix} r_1 r_2 x_1 & r_1 r_2 y_1 & 0 & 0 \\ 0 & 0 & r_1 r_2 x_1 & r_1 r_2 y_1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{bmatrix} = \begin{bmatrix} r_1 r_2 x_2 \\ r_1 r_2 y_2 \\ \vdots \end{bmatrix} \tag{3}$$

This inhomogeneous linear equation can be written as:

$$Ax = b \tag{4}$$

By rewriting this formula and calculating the pseudo inverse with a singular value decomposition the parameters x of an affine transformation that converts spider S_1 into spider S_2 are determined (Moore-Penrose inverse):

$$x = \left[A^T A\right]^+ A^T b \tag{5}$$

6 Comparison of Results

In the following, the spider prototype is compared with the SIFT operator using an example where object / background segmentation is of the essence. Figure 3 shows two images with a fork moved in front of a fixed background. The first column displays lines connecting each pair of points matched by the SIFT operator. The original version of the SIFT operator as provided by David Lowe on his website is used (Version 3, August 2004). Matching points on the background were discarded for simplicity. 3 of the 6 point pairs are correct. The false correspondences are due to the low information content in the area between the spikes of the fork.

The second column displays the results of the spider operator. 7 of the 8 pairs are correct matches. As the spider operator uses information from a visually similar seg-

ment to describe a point, it can utilize the structure of the fork instead of being limited to local information. This is exemplified in the last column, which displays one of the correctly matched spiders.

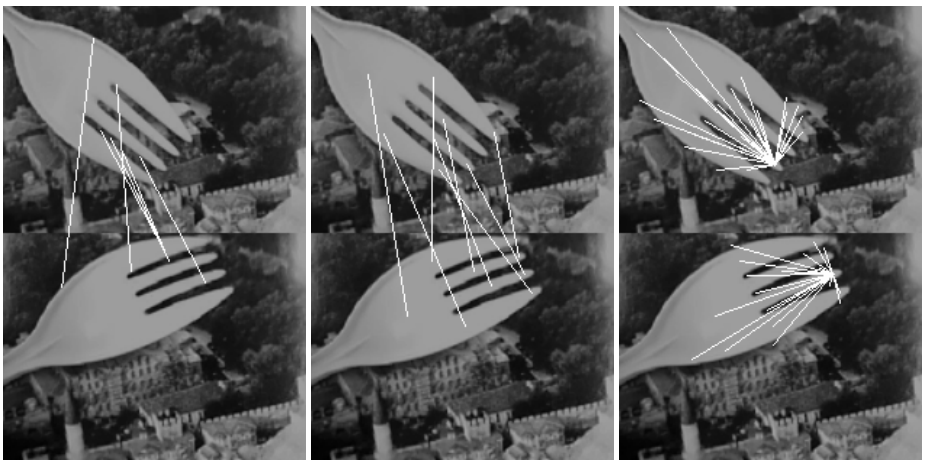

Fig. 3. Comparison of the SIFT and the spider operator. From left to right: 1) The SIFT operator matches 6 points on the object, of which 3 are correct. 2) The spider operator matches 8 points, of which 7 are correct. 3) An example of two of the spiders that are matched correctly. The upper has 28 anchor points; the lower is formed by 22 points

7 Conclusions

In this paper, a new operator for recognizing a point under varying perspective and lighting conditions was introduced. It provides a basis for various applications, like wide-baseline stereo, motion tracking or object recognition. The operator characterizes a salient point with a constellation of surrounding anchor points, which are interest points detected in a visually homogenous neighborhood of the salient point. An algorithm for matching spiders determined from altered perspectives on a scene was developed. In an example, the prototype of the spider operator yields a better result than the SIFT operator when applied to a jagged object. As this is only a single example, additional experiments have to be performed to make a statement about the general quality of the spider operator.

References

1. Forsyth, D. A. and Ponce, J.: Computer Vision – A Modern Approach (2003)
2. Leibe, B. and Schiele, B.: Scale Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search. DAGM'04 Annual Pattern Recognition Symposium. Springer LNCS, Volume 3175 (2004)

3. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, Volume 60 (2004)
4. Mikolajczyk, K. and Schmid, C.: Scale & Affine Invariant Interest Point Detectors. International Journal of Computer Vision, Volume 60 (2004)
5. Mikolajczyk, K., Zisserman, A., and Schmid, C.: Shape recognition with edge-based features. Proceedings of the British Machine Vision Conference, Norwich, U.K. (2003)
6. Scharstein, D. and Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision, Volume 47 (2002)
7. Strang, G.: Introduction to Linear Algebra, Third Edition (2003)
8. Tuytelaars, T. and Gool, L.V.: Matching Widely Separated Views Based on Affine Invariant Regions. International Journal of Computer Vision, Volume 59 (2004)