Visual detection of business cards: study of interest key-point detectors

Domen Tabernik
Faculty of Computer and Information Science, Ljubljana
November 2015

***Abstract*—This study examines the use of interest key-point detector for the application of detecting and recognizing business cards that are slowly waved in front of the camera. We focus on two interest point detectors Scale Invariant Feature Transform and Maximally Stable Extremal Regions. Both are presented in detail with main emphasis on SIFT detector. Characteristics of both are examined with respect to the detection of business cards and SIFT is selected as suitable detector for this problem. A stability of SIFT key-points is also experimentally evaluated on a newly created database of business cards for this purpose.**

I. Introduction

The main goal of our application is localization and recognition of business card which user shows and waves in front of the camera. As a first step towards building this application a detection of position of business card in each frame of the image sequence needs to be performed. For this task we focus on using interest points detectors which detect a set of key-points that are stable across different scales, rotations and small 3D geometric transformations. This characteristics allows interest points to be detected and matched across multiple frames in which business card will change its position and viewing orientation. Matching across multiple frames requires certain level of stability of key-points. This study examines two interest point detectors, namely SIFT [6] and MSER [8], and its stability properties on a dataset of business cards.

This report is organized as follows: in Section II two intereset key-point detectors are presented and their comparision w.r.t. our domain is performed in Section III. In Section IV our dataset of buissnies cards is presented and experimental evaluation of stability of SIFT key-point detector is carried out. Concluding remarks are given in Section V.

II. Interest point detectors

This section describes in detail two types of interest points detectors:

a) edge based key-point detector (SIFT) and
b) region based key-point detector (MSER).

Both types of detectors find key-points that are stable under scale, rotation and small 3D geometric transformation, but utilize different properties of local images patches in the process. Edge based detector utilizes approximation of Laplacian of Gaussian to find points that are invariant to scale while Hessian matrix is used to find corner points with rotational invariance. On the other hand region based detector finds extremal regions that are stable under multiple thresholding levels of watershed algorithm. In the following subsections we further describe each detector in detail.

A. Scale Invariant Feature Transform

Scale Invariant Feature Transform (SIFT) was proposed by Lowe [7] as a process to transform image into a set of distinctive features suitable for matching of object or scenes seen from different viewpoints. This approach comprises of two stages. Firstly, properties of scale and orientation invariance around local image patches are utilized to find key-points that are repeatable in images when changes are made in scale and rotation. Features are designed to be invariant to scale and rotations, however they also exhibit invariance to small degree of changes in illumination and 3D geometric transformation. We describe interest key-point detection in detail in Subsection II-A1.

Secondly, each local patch around invariant key-point is encoded into SIFT descriptor. Descriptor utilizes direction of key-point to extract gradient information in rotation invariant way. This allows object or scene to be encoded with hundreds of small local descriptors which as a group are distinctive enough to enable good object or scene matching. At the same time extraction of descriptors is efficient and has low computational and storage requirements. We describe descriptor in detail in Subsection II-A2.

1) Scale-invariant and rotation-invariant key-points: We describe three criteria that are considered for detection of scale and rotation invariant key-points. Lowe [6] proposes that if key-point is to be repeatable when viewing under different scale or rotation changes then such key-point must meet following criteria:

a) key-point needs to be detected at a location and scale that is extrema in scale-space representation,
b) key-point needs to describe an edge where eigenvalues ratio of second-moment matrix are low i.e. key-point needs to be a corner, and
c) key-point needs to be oriented along the principal axis of second-moment matrix.

Satisfying the first criteria can be performed as shown by Lindeberg [5]. He showed that using Gaussian kernel $G(x, y, \sigma)$ as scale-space representation and searching for 3D maxima of the scale normalized Laplacian-of-Gaussian $LoG(x, y, \sigma)$ function returns scale position that can be repeatably detectable under different scale changes. Lowe [6]

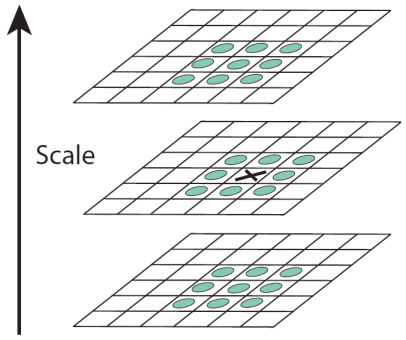

Figure 1: In SIFT scale is selected as extrema point between neigbooring pixels and neigboring scales of DoG. Image source: [6].

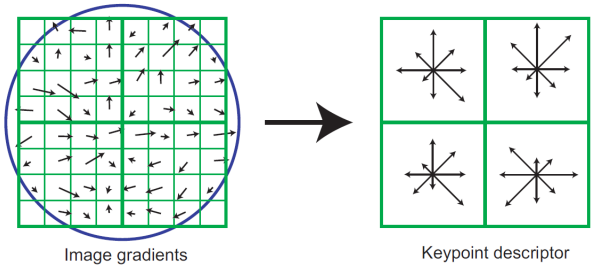

Figure 2: Example of SIFT descriptor being constrcuted by histograming orientations into 2x2 grid where each histogram covers 4x4 pixel area. Circle on the left represent gaussian window used to weight gradient magnitudes. Note, in the actual implementation 16 histograms are created instead of 4 in a grid of 4x4. Image source: [6].

further improved this process by approximating Laplacian-of-Gaussian with difference-of-Guassian (DoG)$D(x, y, \sigma)$:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \\ &\approx LoG(x, y, \sigma) \end{aligned}$$

where $I(x, y)$ is image and $L(x, y, \sigma)$ is convolution of image with Gaussian kernel $G(x, y, \sigma)$. He showed that using DoG approximation does not reduce accuracy and is also computationally efficient as it can be computed by subtracting two subsequent convolution of Gaussian which are already needed for scale-space representation. The process of finding extrema in scale-space is then performed by finding local maxima and local minima over 8 neighboring location in the same scale and 9 neighboring locations at two neighbor scales of DoG as shown in Figure 1.

Lowe [6] further refined interest point position by Brown and Lowe [1] method using a quadratic Taylor expansion of scale-space function $D(x, y, \sigma)$ around the sample point. Correction offset x around sample point $\hat{x}$ can be estimated using Hessian where derivative of D are approximated based on neighboring sample points. Additionally, using $D(\hat{x})$ at the extremum is used to eliminate points with low contrast. To remove non-corner points and satisfy the second criteria the second-moment matrix H can be used which estimates principal curvatures of D.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix}$$

Any key-point where ratio between eigenvalues of H is high is eliminated as non-corner point. As shown by Harris and Stephens [2] and Lowe [6] this can be accomplished as:

$$\frac{\text{Tr}(H)^2}{\text{Det}(H)} = \frac{(r+1)^2}{r},$$

where $\text{Tr}(H)$ is trace, $\text{Det}(H)$ is determinant and $r = 10$ is threshold value.

The third criteria enables rotational invariance of points and allows key-points descriptor to represents interest points based on neighboring gradients in rotationally invariant manner. Direction of key-point is determined based on gradient magnitude $M(x, y, \sigma)$ and orientation $\Theta(x, y, \sigma)$ of image $I(x, y)$ at scale-space σ, which are precomputed for all image locations for fast computation. Direction is calculated by collecting a histogram of gradient orientations $\Theta(x, y, \sigma)$ weighted by gradient magnitude $M(x, y, \sigma)$ and by 1.5 times σ of the location scale. Histogram is discretized to 36 bins with global peeks representing a dominant direction of a key-point. Additionally, any neighboring values with at least 80% of peek value are used to fit parabola and improve accuracy of the peek.

2) Key-point descriptor: Around each key-point a descriptor is extracted to transform each image into a representation as a set of features. Each descriptor needs to consistently describe a local patch around the key-points regardless of viewpoint and illumination changes. This makes it possible to describe an object or scene with the same representation even under viewpoint and illumination changes. Rotational invariance is already achieved by key-point detector and can be used to consistently align local patch around any key-point into the same direction

Descriptor is constructed from gradient magnitudes $M(x, y, \sigma)$ and orientation $\Theta(x, y, \sigma)$ of image $I(x, y)$ at scale-space σ, where magnitudes are additionally weighted by Gaussian function $G(x - \bar{x}, y - \bar{y}, \sigma/2)$ centered at key-point location $(\bar{x}, \bar{y})$ with half of location scale σ. Gaussian weighting has an effect of preferring sample points closer to key-point location and ignoring samples further away. Orientations around key-point location are sampled into 16 non-overlapping histograms in 4x4 gird, where each histogram covers 4x4 pixels. Example in Figure 2 shows the same principle but with using 4 hisograms in a grid of 2x2. Orientations are discretized into 8 bins and each sample contributes with Gaussian weighted magnitudes $G(x-\bar{x}, y-\bar{y}, \sigma/2) \cdot M(x, y, \sigma)$ to discretized bin as well as to adjacent bins using trilinear interpolation. SIFT descriptor is composed by concatenating histograms into $4 \times 4 \times 8 = 128$ dimensional feature.

Invariance to illumination is additionally achieved by normalizing descriptor to unit length, then clipping values to max value of 0.2 and finally re-normalizing to unit length. Clipping to 0.2 achieves invariance to non-linear illumination changes

from camera saturation or illumination changes due to 3D viewpoint change.

B. Maximally Stable Extremal Regions

Next, we explain second type of key-point detector which is based on regions. Maximally Stable Extremal Regions algorithm has been introduced by Matas et al. [8] as a method for detecting regions, deemed distinguished regions, that have high repeatability and posses some distinguishing, invariant and stable properties. Method introduces new type of distinguished regions, named extremal regions, whose desirable properties include closed boundary under perspective geometric transformation and under monotonic image intensity transformation. They present an efficient algorithm for finding extremal regions that are maximally stable, which they apply for matching correspondences in wide-baseline stereo matching.

Regions that have stable local binarization over a wide range of thresholds have properties that are useful for extraction of interest key-points. Such regions are invariant to affine transformation, they are stable as their supporting points are unchanged over a range of threshold, they can be detected over multiple scales since no smoothing is involved and finding them is computationally efficient.

The algorithm for finding maximally stable extremal regions proceeds as follows. Image intensities are thresholded from smaller intensities to higher. At each level of threshold a list of connected components is established. Area of connected components are stored as a function of image intensities and maximally stable regions are defined at local minima of the rate of change of the area function. This algorithm is identical to watershed thresholding [11], with the exception that connected components whose area is maximally stable, i.e. does not change over range of thresholds, is the output of the algorithm. Extremal regions are found by thresholding image first from smaller to higher intensities and then from higher to smaller intensities. This is accomplished by running the algorithm on the original and on the inverted image.

Utilizing MSER for the matching correspondences requires a selection of measurement region that is large enough to be distinctive and small enough to exclude the background pixels. The method proposes to use four measurement regions, one at the size of the region, one at 1.5, 2 and 3 times the size of convex hull around the region. Additionally, invariance to rotation to make description independent of the viewing position is normally accomplished with complex moments.

III. Comparison

Out of both presented detectors, we selected SIFT detector as our main subject of evaluation due to its advantages over MSER detector. Principal advantage of SIFT is selection of scale for each key-point through the search of local extrema in Difference-of-Gaussian. This ensures that only points which can be found and are stable across multiple scales are selected. MSER on the other hand does not perform any scale selection, but searches for both smaller and bigger regions and relies on difference between inside and outside region intensity to be stable across scale. Mikolajczyk et al. [9] showed that MSER regions are not as stable as SIFT key-points and perform poorly in scale change and image blur due to lack of scale selection. Robustness to scale and blur is important criteria for our problem as motion of the business card in front of a camera will induce certain amount of blur.

Comparing orientational invariance, SIFT achieves it with local gradient information in a robust way. Finding peeks in a histogram of orientations around a key-point determine principal direction, while eliminating edges and focusing only on corner points ensures that key-point will be detectable even under viewpoint changes. This is ensured by the main characterizing of corners that their position remains unchanged under affine and perspective transformations. In MSER, orientational invariance is not explicitly searched for but is achieved due to pixel intensity preservation in region borders under geometric transformation. Invariance to rotation is achieved by finding second-moment matrix of region borders and aligning them to the principal curvature. Compared to SIFT the lack of using gradient orientation in MSER make it slightly less robust when image intensities will change due to illumination changes. However, advantage of using regions border is expressed in invariance to higher degree of affine and perspective deformations. In SIFT, invariance to the same degree of affine geometric changes can only be achieved with additional modification as in A-SIFT [10]. In our problem, the movement of business card will induce changes in translation, rotation and perspective transformation. However, changes in perspective transformation are not expected to be as prominent as are changes in rotation and translation. This makes SIFT still suitable for our application.

Main advantage of MSER is computational efficiency as thresholding image intensities and finding connected components can be faster then performing Difference-of-Gaussian over many scales, calculating Hessian matrix and histograming orientations. However, computational efficiency is traded for lower robustness to scale changes.

IV. Evaluation

In this section we describe a database used in evaluation and the experimental evaluation of the detectors stability. Due to advantage of using scale selection in SIFT this section focuses only on SIFT detector being applied to the domain of our problem.

A. Database

The evaluation of SIFT descriptor was performed on our own database, which consists of ten samples of business cards, five non-textured cards and five cards with more texture. We captured samples under realistic conditions with background clutter being visible. Each sample is captured as video recording and as still images. In both cases, we captured samples with multiple changes. In particular, we varied viewpoint change by tilting the card in horizontal and vertical direction maximally to around 60 - 80 degree. Samples captured with video also contain around 15 degree of rotation over the third axis (depth axis) as well as higher changes in translation as the card was being waved in front of the camera.

Figure 3: Three examples of buissnis cards each different simulated under of condition changes. First column is unchanged image and remaing columns are modified images. From top to bottom the changes applied are: Gaussian blur (at $\sigma = \sqrt{2}$, $\sigma = \left(\sqrt{2}\right)^3$ and $\sigma = \left(\sqrt{2}\right)^5$), perspective viewpoint change (20, 40, 60 and 70 degrees) and zero-mean Gaussian noise (4 dB, 16 dB and 128 dB). Note, not all changes evaluated are show here.

Camera calibration: Database was captured with Logitech c920 webcam in resolution of 1920x1080 pixels, with samples being positioned around 15 - 20 cm away from the camera. By specification camera is reported to have a focal length of 3.67 mm and diagonal field of view 78 degrees. We performed camera calibration using based on [12], [3] using Camera Calibration Toolbox for Matlab [4]. The focal length measured by calibration process was reported as $(3.71, 3.71) \pm (0.005)$ mm with pixel error being 0.51 in horizontal direction and 0.52 in vertical direction. The effect of radial distortion was insignificant with coefficient of estimating distortion measuring only to:

$$\begin{bmatrix} 0.09745 & -0.16889 & 0.00092 & -0.00058 & 0.00000 \end{bmatrix}.$$

B. Stability of SIFT key-points

Stability of detector was evaluated only on still images from our database. From each business card only first image sample was used. In total 10 samples of business card were used. The selected image contained close to zero viewpoint deformation with as little as possible rotation around depth axis, i.e. business card was in parallel and aligned with the camera image plane. Samples were additionally cropped to eliminate any background clutter. By cropping we eliminated source of any error that would arise from poor detection on background clutter as we prefer to focus on stability of detector only with respect to the business cards. We evaluated descriptor by artificially adding several types of changes to the image: (a) geometric transformation, (b) Gaussian bluing and (c) Gaussian white noise. Some examples of evaluated images are shown in Figure 3.

We measure stability of detector using the repeatability measure [7], [6], [9]. The measure is defined as a percent of all detected points in image which are also detected at the same location in the second image being matched. This can be computed by knowing the homography between two images. We use the implementation from [9], which recognizes positive detections only if location of detected key-point is less then 1.5 pixels away from the projected key-point location of the second image and by having the ellipse describing the interest point region overlap at least 60%. Distance and overlap are computed as an average of projecting key-points in both direction, i.e. from first image to second and from second image to first. We additionally evaluated detector with an overlap of ellipses lowered to at least 40% to asses the accuracy of position and ignore error from improper orientation changes.

In all experiments we report results averaged into two repeatability values, one for textured and one for non-textured samples. Additionally, we report standard deviation over each group as error bars in plots.

Geometric transformations: Geometric transformations were simulated using projective transformation of viewing position around the vertical axis and repeatability measure was computed under each viewpoint. Seven different viewpoints were evaluated ranging from 10 degree up to 70 degree by an increment of 10 degree change. We report performance of detector in Figure 4a. Results indicate that up to 10 degrees of viewpoint change repeatability can be maintained above 80%. At viewing angle of 20 degree repeatability drops only

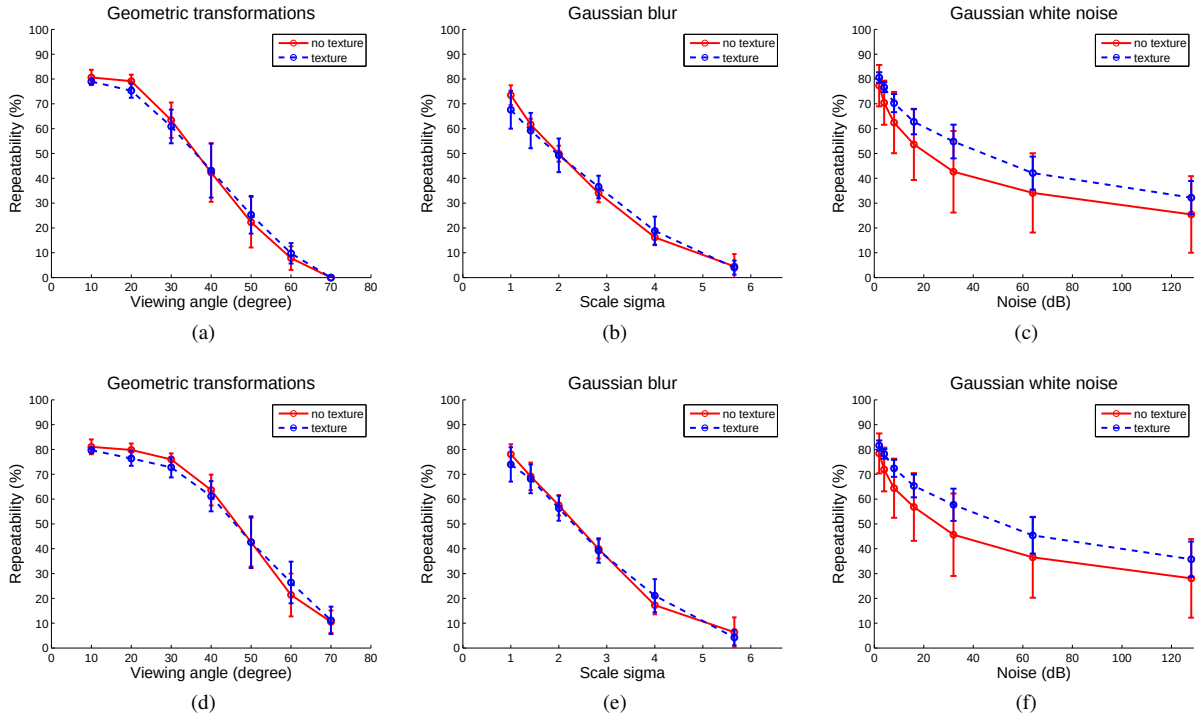

Figure 4: Repeatability results for three different types of changes being evaluated. Top row is repeatability at minimally 60% overlap allowed and bottom row is at 40% overlap allowed. Left column captures stability of key-points under changing viewpoint perspective, middle column under Gaussian blur and the third column under Gaussian white noise. Repeatability is calculated as in [9].

slightly below 80%, however, after viewing angle increased to 30 or higher repeatability drops quicker. At viewing angle of 40 degrees repeatability drops to slightly above 40% and at 60 degree or more it drops below 10%. The same level of performance is shown to be consistent with both textured and non-textured samples.

When overlap between regions of key-points is allowed to be at least 40% then repeatability at up to 30 - 40 degree change can be maintained at 70 - 80% and the trend of dropping with higher degrees of viewing angles is slightly slowed down as can be seen from Figure 4d. This is an indication of source of error at higher viewing angle changes being the ellipse matching and an indicator that locations of key-points are quite accurate.

Gaussian blur: Next, we evaluate stability of the detector under scale changes and blurring. We simulate this effect by adding Gaussian blur to the image. Six levels of blur were evaluated with $\sigma = \left(\sqrt{2}\right)^k$, where $k = [0,..,5]$. We report results in Figure 4b and 4e. The results for both textured and non-textured images are similar. With $\sigma \approx 1$ repeatability is between 70% - 80%, however it quickly drops to 50% at $\sigma \approx 2$ and to 20% at $\sigma = 4$. This drop is understandable as $\sigma = 4$ is already significant change of scale. Using overlap error of eclipses of at least 40% increases performance by around 5 percent points.

Gaussian white noise: Lastly, we evaluated stability of the detector under varying Gaussian white noise. We added zero-mean noise to image from a range of 2 dB to 120 dB in an increment of 2 dB based on 2^k, where $k = [1,..7]$, thus evaluating at seven different levels of white noise. The results are reported in Figure 4c and 4f, and can be seen to quickly drop to around 80% repeatability with low noise below 10 dB, while higher noise slowly reduced repeatability to 40% at 120 dB level. With overlap error of ellipses of at least 40% repeatability dropped only by 5 percent point lower.

V. Conclusion

In this study we examined two interest point detectors for the application of detecting and recognizing business cards that are waved in front of the camera. We presented SIFT [6] and MSER [8] key-point detector. In particular, we presented SIFT in more detail, as it was selected as appropriate detector. We arrived at this conclusion through theoretical examination of each detector and related work, and found that SIFT has higher robustness to scale changes and blur due to automatic selection of scale, while MSER lacks this mechanism. We have pointed out that each detector integrates invariance to rotation, and that due to stability of regions that are found by MSER they are invariant to slightly higher degree of affine and perspective orientation changes then SIFT features.

We experimentally evaluated stability of SIFT detector on proposed database of business cards and have show that under low geometric transformations SIFT features have 80% repeatability rate. We also evaluated and reported susceptibility of key-points to different levels of Gaussian blur and Gaussian white noise. Under extreme levels of all three types of changes

the stability of key-points severely drops, however, at moderate and low level of changes SIFT key-points have repeatability of around 80-50%. We estimated that this should be good enough for our problem where only low level of changes in affine transformation and blur are expected. Detector also performed similar for both textured and non-textured samples, except in the case of white noise, where textured samples performed by almost 10 percent points better. This can be contributed to lower robustest of point around non-textured images where gradient values are so small that minor level of noise changes them.

REFERENCES

[1] Matthew Brown and David G Lowe. Invariant Features from Interest Point Groups. In *In British Machine Vision Conference*, pages 656–665, 2002.

[2] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Procedings of the Alvey Vision Conference 1988*, pages 147–151, 1988.

[3] Olli Heikkilä, Janne ; Silvén. A Four-step Camera Calibration Procedure with Implicit Image Correction. In *Computer Vision and Pattern Recognition*, pages 1106 – 1112, 1997.

[4] Jean-Yves Bouguet. Camera Calibration Toolbox for Matlab.

[5] Tony Lindeberg. Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 30:79 – 116, 1998.

[6] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.

[8] J Matas, O Chum, M Urban, and T Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. In *Image and Vision Computing*, pages 36.1–36.10, sep 2002.

[9] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005.

[10] Jean-Michel Morel and Guoshen Yu. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[11] Luc Vincent and Pierre Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations, 1991.

[12] Zhengyou Zhang Zhengyou Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 0–7, 1999.