A New Evaluation Framework and Image Dataset for Key Point Extraction and Feature Descriptor Matching

Iñigo Barandiaran[1,2], Camilo Cortes[1], Marcos Nieto[1], Manuel Graña[2] and Oscar Ruiz[3]
[1] *Vicomtech-IK4 Research Alliance*
[2] *Dpto. CCIA, UPV-EHU*
[3] *CAD CAM CAE laboratory, Universidad EAFIT, Carrera 49 No 7 Sur - 50, Medellín, Colombia*
{*ibarandiaran, ccortes,mnieto*}*@vicomtech.org,ccpgrrom@gmail.com,oruiz@eafit.edu.co*

Keywords: Key point Extraction, Feature Descriptor, Key point Matching, Homography Estimation.

Abstract: Key point extraction and description mechanisms play a crucial role for image matching, where several image points must be accurately identified to robustly estimate a transformation or recognize an object or a scene. Currently several new mechanisms for key point extraction and for feature description are emerging, so normalized data and evaluation protocols are needed in order to assess them accurately. In response to these needs, we present a new evaluation framework for measuring different aspects and behaviours of the state-of-the-art feature point extraction and description mechanisms. In addition, we also propose a new image dataset and a testing image generator. This evaluation framework and dataset can be useful to help the research community improving their key point extraction and feature descriptor approaches. Also, the practitioners on computer vision applications, based on image point matching, can obtain valuable information from this contribution to select the algorithm that best suit their needs. All proposed material in this work is freely available on-line.

1 INTRODUCTION

Interest points extraction and matching is nowadays a common task in many computer vision based approaches, which are applied in many different domains, such as 3D reconstruction, object recognition, camera tracking and augmented reality. Key point extraction and description mechanisms play a crucial role during image matching processes, where several image points must be accurately identified to robustly estimate a transformation or recognize an object. Currently there is an increasing activity in the development of new approaches for key point extraction, description and matching, trying to get more robust and computationally lightweight approaches. In this way, we think that normalized data and evaluation protocols are needed in order to assess them accurately.

In this work, we present a new testing framework, an image dataset, and a testing image generator for the evaluation of the state-of-the-art key point extractors and feature point descriptors. This appraisal is done by measuring several algorithm features, such as repeatability, accuracy and invariance to affine transformations or photometric transformations. Our new proposed testing dataset comprises both a transformed image generator, that allows generating new images with geometric and photometric transformations, and a set of real images acquired with different types of sensors and conditions, showing also variations in both geometric (such as similarities or affinities) and photometric transformations. The dataset of real images, the image generator with the evaluation framework form a useful tool to help in the selection of the proper algorithm to develop computer vision applications based on image point matching, and to improve or develop new approaches for key point extraction or point matching. The paper is structured as follows: section 2 provides a brief description on key point extraction and feature descriptors and an overview of some evaluation framework and testing datasets. Section 3 describes the proposed framework for feature point descriptor evaluation. Sections 4 and 5 describe both the proposed acquired real image dataset and transformed image generatorin and finally section 6 gives final remarks and depicts the future work.

2 RELATED WORK

Several computer vision based applications rely on the identification or matching of several discrete points extracted from the images. Although this is a very common task, depending on the nature of such applications, the requirements for a specific key point extractor and descriptor may vary. For example, applications related with self-navigation or simultaneous location and mapping (SLAM) would require a fast key point extractor algorithm because of its real-time restrictions. On the other hand, an application for object or image recognition would benefit from more robust or better invariant key point extractor; even if this implies a higher computation time. In the context of point matching, a robust key point can be understood, in general, as a point of the same structure in the scene that is able to be extracted and matched even if some types of geometric or photometric transformations occur between different image acquisitions.

In (Tuytelaars and Mikolajczyk, 2008) the authors suggest that there are several parameters of a point detector and feature descriptor that can be measured; they also cite the most relevant ones. However, measuring some of them, such as the point extractor accuracy, descriptor robustness or invariance needs a normalized test protocol and test benchmark. In this way, the seminal works of (Mikolajczyk and Schmid, 2005) settled the basis for key point extractor and feature description evaluations. Since then, several new approaches for key point or region extraction (Mikolajczyk et al., 2007) and for feature descriptor (Bay et al., 2006; Heikkilä et al., 2009; Bellavia et al., 2010; Leutenegger et al., 2011) were tested against their dataset and evaluated with their corresponding scripts freely available online at `'www.robots.ox.ac.uk/~vgg/research/affine/index.html'`.

In (Fraundorfer and Bischof, 2005) the author proposed an extension of the work of (Mikolajczyk and Schmid, 2005) by analyzing key point repeatability for non-planar scenes, using tri-focal tensor geometric restriction for estimating the ground-truth data of their own dataset. They found several differences in key points repeatability scores when applied to non-planar scenes. Recently, (Gauglitz et al., 2011) proposed a dataset consisting of several videos of surfaces, with different types of textures and different light conditions, which are used to evaluate key point matching strategies oriented to camera tracking applications. The authors claim that due to restrictions in the hardware they used to move the camera for the generation of different points of view, they could not reproduce exactly the same movements every time they changed scene conditions. This implies that homographies are not the same and may bias the results of the different algorithms. They used 4 markers attached to each picture in order to compute image to image homographies.

Very recently, in (Alahi et al., 2012) the authors tested their new descriptor approach with the known dataset and evaluation framework of (Mikolajczyk and Schmid, 2002). However, they also tested their descriptor with a non-publicly accessible approach in `computer-vision-talks.com`, which is similar to our evaluation framework proposal. This framework allowed the authors to compare the robustness of their descriptor against different geometric transformation values, in the form of a ratio between correct and wrong matches. The authors affirm that this approach provides a very useful insight about the tested descriptors.

Our dataset and evaluation framework is based and inspired by the developments of (Mikolajczyk and Schmid, 2005). In comparison to Mikolajczyk's approach, our dataset comprises a higher number of images, with higher resolution and with better controlled conditions.

We also include a set of images obtained with mobile devices. We think that it is important to consider some features of these devices, such as their low dynamic range, in a testing data. This is relevant since mobile devices are becoming part of our everyday lives and computer vision applications are increasing their popularity. To the best of our knowledge, this feature lacks in the available testing datasets.

In this way, our dataset includes a set of images that can be used to evaluate the robustness of key point extractors and descriptors approaches against photometric transformations, such as luminance and chrominance noise addition.

Finally, we also propose a transformed image generator that can be used to provide more testing images to a given key point or descriptor evaluation.

All proposed material in this work, i.e. images, code and binary executables will be freely available on-line at `'www.vicomtech.tv/KeyPoints'`.

3 EVALUATION FRAMEWORK

We have implemented an evaluation framework based on the one present in the Open Source Computer Vision Library (OpenCV) (Bradski, 2000), derived from the original work of (Mikolajczyk and Schmid, 2005). This framework uses the class hierarchy implemented in OpenCV that nicely decouples key point extraction from key point description and descriptor matching.

In this way, the user can easily define experiments by mixing several point extractor with key point descriptors and matchers. Whereas Mikolajczyk's work, where the framework is written in Matlab scripting, our approach is written in C++. In the case of the mentioned Matlab-based evaluation framework, the user needs to generate both a file with detected key points in a given image, and the corresponding key points descriptors in order to evaluate them. The generation of these files can be cumbersome in some contexts, such as development of commercial computer vision based applications, because the whole solution may not be tested in the same development platform. We think that our approach helps in the evaluation of future extractor or descriptor approaches because it can be easily integrated in a development environment, without the need to export additional data to other platforms. Nevertheless, our approach also supports the reading of Mikolajczyk file format, allowing the comparison with previous approaches or studies. Figure 1 shows partial results of an evaluation conducted using the proposed dataset and evaluation framework. In addition to the precision-recall curves proposed by (Mikolajczyk and Schmid, 2002), we propose to generate more informative curves about the performance of different approaches based on the number or percentage of correct matches given specific values of the evaluated transformation. For example, Figure 1 shows the result of the number of correct matches of several feature descriptors against a dataset composed of several in-plane rotations of an image. These preliminary results suggest that, for example, BRIEF descriptors are not robust against a rotation larger than 35 degrees approximately, or how SURF approach is more sensitive to orientations like 90, 180 and 270 degrees, possibly due to discretization effects related with the use of box filters for approximating LoG filtering. In this way, a better insight of the behaviour of a given approach may be obtained.

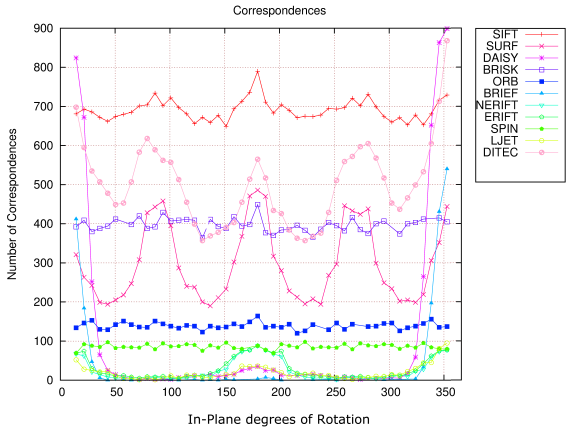

Figure 1: Results of the evaluation of several feature descriptors using the in-plane rotation.

3.1 Matching Evaluation

An image formation process is usually represented as in Equation 1 where X_w represents world point, x_i represent world points projected in the image. P represents the projection matrix, described in Equation 2, where K describes the transformation from the camera reference frame to the image reference frame, and $[R|t]$ the composition of a translation and a rotation transformation between world and camera coordinate systems.

$$x_i = PX_w \tag{1}$$

$$P = K[R|t] \tag{2}$$

When either world points X_w lie on a world plane, or the images are acquired with a rotating camera around its center of projection, the transformation between image points x_i and world points X_w are related by a 2D linear projective transformation or homography H (Hartley and Zisserman, 2004).

As in the dataset proposed in (Mikolajczyk and Schmid, 2002), in our proposed dataset all images are related by a 2D homography H_{abD}. This known transformation is used as ground truth data, allowing to know a priori where a point x_{iaD}, extracted from image a of dataset D, should be projected in image b of the same dataset, by using Equation (3).

$$x_{jbD} = H_{abD}x_{iaD} \tag{3}$$

Similarly, points extracted from image b can be projected back to image a by using the inverse of H_{abD}. Let $\tilde{x}_{jbD}$ be the estimated match of point x_{iaD} in image b obtained by the point detector algorithm. Then, the known transformation H_{abD} is used to measure the accuracy and repeatability of a point detector algorithm. This process is performed by computing the Euclidean distance d between the estimated and the ground truth points of a pair of images, as shown in Equation 4.

$$d_{ij} = d(\tilde{x}_{jbD}, H_{abD}x_{iaD})^2 + d(x_{iaD}, H_{abD}^{-1}\tilde{x}_{jbD})^2 \tag{4}$$

In order to estimate correct matches m_{ab}, as shown in Figure 2, among all potential matches or correspondences, i.e. point pairs x_{ia} and $\tilde{x}_{jb}$ extracted from images a and b respectively, we used the overlap error as proposed in (Mikolajczyk and Schmid, 2002). This error measures how well two supporting regions, usually ellipses or circles R_{ia} and R_{jb}, estimated by point extraction algorithm from key points x_{ia} and $\tilde{x}_{jb}$

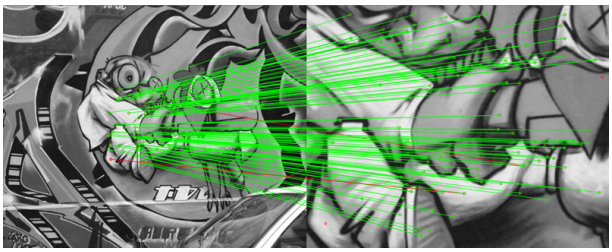

Figure 2: Correct matches(in green), wrong matches(in red) between two images.

respectively, correspond under the known geometric transformation H_{ab}. In our case, this transformation is described by an homography.

$$\varepsilon_s \leq 1 - \left(\frac{R_{ia} \cap H_{ab}^T R_{jb} H_{ab}}{R_{ia} \cup H_{ab}^T R_{jb} H_{ab}} \right) \quad (5)$$

The point pair x_{ia} and $\tilde{x}_{jb}$ that has lower error distance d_{ij} given by equation 4 and the lower overlap error given by equation 5 is considered as a true match. The overlap error reduces the probability of false positive matches. We calculate the ellipses overlap by using the software proposed in (Hughes and Chraibi, 2011) and freely distributed by the author at `'www.chraibi.de'`.

4 IMAGE DATASET

4.1 Acquisition Setup

Our image acquisition setup is composed by a DSLR Canon 7D and an iPad with a 5 Mega pixels built-in camera. In the Canon 7D scenario we used a Tamron 17-50mm f2.8 and a Canon 100mm f2.8 macro lenses. In addition to the camera, we used two Canon 580EXII flash with light diffuser, both operated wirelessly and synchronized with the acquisition. In the case of the iPad setup we can not synchronize the light with the acquisition, so we decided to use continuous light source instead of flashes.

4.2 Geometric transformations

In order to generate a set of images with perspective distortion, we carried out an approach similar to (Gauglitz et al., 2011). We used a Kuka robotic arm with a Canon 7D attached with Tamron lens in order to generate different points of view of the same target, as shown in Figure 3. The use of the robotic arm allowed us to generate known, repeatable and precise positions and trajectories around the target scene. We also used a Wacom Cintiq screen for displaying images instead of using pictures placed in a wall or in a table, as in (Gauglitz et al., 2011). Our set of displayed images covers different types of images with structured or unstructured textures, with low texture, or with repeating textures or patterns. Many authors (Tuytelaars and Mikolajczyk, 2008; Heikkilä et al., 2009; Gauglitz et al., 2011) agreed in the importance of evaluating key point extractors and descriptors in such different conditions, in order to truly evaluate the robustness of their approaches.

The robotic arm is a KUKA LWR IV+, which has 7 joints, a payload of 7 kg and a repeatability of $\pm$ 0.05 mm. The desired position and orientation of the robot's end effector can be commanded from a remote PC, using the KUKA Fast Research Interface (FRI). The FRI provides a C++ high level interface, which can be used to retrieve information of the robotic arm, such as the tool's Cartesian position/orientation, and to implement different control strategies.

We decided to generate circular trajectories (arcs) to obtain several points of view of the Wacom screen, and therefore different values of captured perspective distortion. The desired circular path is defined by three points in the Cartesian 3D space, which are used to calculate the different elements of the parametric equation of a circle. The required orientation of the camera at the initial and final points of the trajectory can be defined independently of the circular path, allowing different configurations in a flexible fashion.

The described trajectories are resampled according to a desired number of points M along them, where images are to be taken. The set $Q = \{Q_1, Q_2, \ldots Q_M\}$ constitutes the resulting discretized trajectory. Each $Q_i \in Q$ is 3x4 matrix that describes the i pose (position and orientation) of the camera, with respect to the robot's base coordinate system, where $1 \leq i \leq M$. This means that the original circular path is approximated in a piecewise linear way. Analogously, the orientation of the camera at each Q_i is determined by performing a linear interpolation of the total rotation matrix R_T, defined by $R_T = R_M (R_1)^{-1}$, where R_M and R_1 correspond to the rotation parts of Q_M and Q_1 respectively. Therefore, R_T is applied in $M-1$ steps, which can be done easily using quaternion notation. Each element of Q is used as a set point for the robot's Cartesian controller. The points in Q set are traversed in order. When the position and orientation errors with respect to a particular point Q_i are below some predefined thresholds, a signal is sent to the camera in order to take N pictures in a synchronous way. At any point Q_i the first picture to be taken corresponds to the calibration pattern image; then $N-1$ pictures of other images shown on the Wa-

Figure 4: Some images of exposure varying dataset compound of 15 different images.

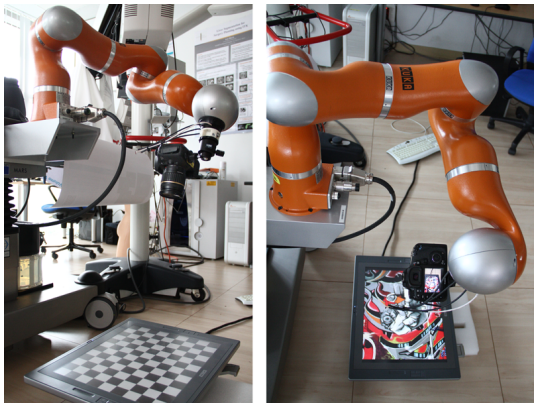

Figure 3: Image acquisition setup with Kuka robot arm and Canon 7D attached.

com Cintiq screen are taken. While pictures are being taken the robot holds its position. Figure 5 shows a 3D reconstruction of a known generated arc trajectory of the camera around the Wacom screen, from a circular sector of radius equal to 0.4m, covering a total angle of 70 degrees.

We used the calibration pattern image for calibrating the camera, i.e. estimate extrinsic and intrinsic parameters, and also for accurate estimation of the homographies between images. We used the estimated

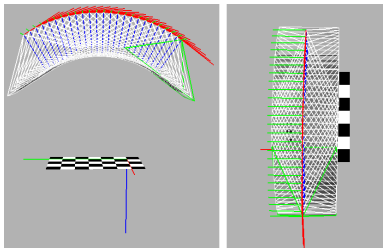

Figure 5: Recovered trajectory of a Robot driven image acquisition.

camera calibration parameters for rectifying the distortion of the images acquired with the Tamron lens, which has around a 2% of geometric barrel distortion. The Canon 100mm macro lens is able to render images with almost negligible geometric distortions. Geometric distortion can be considered as one of many types of optical aberrations. These distortions cause to the projection of incoming rays to the optical system differ from the ideal position produced by a distortion free model, such as a pinhole camera. All images of our dataset are geometrically corrected, thus neither barrel nor pincushion distortions remain.

4.2.1 Image Focus

In addition to the capability of generating unfocused images with our image testing generator, we also captured real scenes because unfocused images are not only Gaussian smoothed versions of a correctly focused image. The shape of the lens diaphragm and the value of the lens aperture, which determines depth of field, play an important role in the finally rendered image; therefore it is not easy to simulate them synthetically. We propose an image dataset where the focus point is progressively varying from a correct focus point, i.e. all objects in the scene are accurately rendered in images as sharp, to a point where all objects appear blurred or unfocused, as shown in Figure 6. In this subset of images, even if the camera was not moved along the image sequence acquisition, changes made in the camera focus required to compute the homography between images.

4.3 Photometric Transformations

Photometric transformations are also involved in the process of image formation along with geometric transformations. These transformations are related to the camera settings, light conditions and the nature of the camera hardware, mainly the camera sensor. As a photometric transformation dataset, we propose a set of images that show a variation in the light condition or light exposure, as shown in Figure 4. The purpose of this subset is to be able to evaluate the robustness of key point extractors repeatability or feature descriptors robustness against illumination changes and noise.

Image acquisition was carried out by using a protocol where no geometric transformations were applied between any of the images that form this dataset, ensuring that only photometric transformations occur between them. This implies that the homography ma-

Figure 6: Some images of focus varying dataset compound of 25 different images.

trix that relates them geometrically correspond to the identity matrix. To ensure that no geometric transformations were applied during dataset acquisition, both the illumination equipment and the camera were operated remotely. As mentioned in section 4, we used flashes to generate the illumination of the scene. The use of the flashes allow us to vary the amount of light without changing any camera acquisition parameters, i.e. setting fixed the aperture value, the exposure time, and ISO speed. In this way, neither the depth of field (DOF) is varied along the images that constitute the dataset, nor additional noise is added due to an increase of either ISO speed, or due to sensor heat because of longer exposure times. Every image in this dataset is consecutively reduced approximately an 1/3 of a f-stop, starting with a correct exposure in the first image. This dataset is composed of 15 images resulting in a difference of 4.5 f-stops between the first and last images. Figure 7 shows two images of the

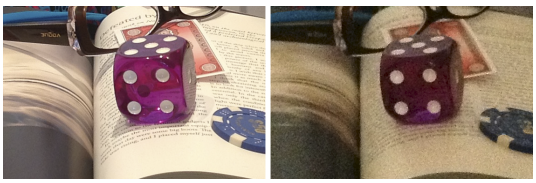

Figure 7: Images from the exposure varying dataset taken with a mobile device.

same scene taken with the iPad in controlled illumination conditions. Left image was captured with a correct value of exposure, while the right image was captured with approximately 2'5 f-stops less of exposure. As mentioned in section 4, in the mobile device setup we used a continuous light source where light intensity can be set manually. It is worth mentioning that both the focus point and exposure metering point were fixed along the capturing of all images in the dataset.

In opposite to the DSLR setup where exposure values, i.e. ISO speed, aperture, and exposure time, can be set manually, in a mobile device, such as the iPad, those values are set automatically during image acquisition. In this way, we used an application that allowed us to focus and measure exposure always in the same gray neutral part of the scene along the captures. This ensures that the exposure readings are consistent along image acquisitions, given different light conditions. As expected, in both cases, as the amount of light decreases, i.e. the signal-to-noise ratio (SNR) decreases, the amount of digital noise increases. This is clearly more noticeable in the case of the mobile device, due to the smaller size of its image sensor, and therefore a more limited dynamic range compared with the DSLR camera.

5 TRANSFORMED IMAGE DATASET GENERATOR

In addition to the proposed set of images, we implemented a set of C++ functions and Python Scripts that allow the generation of several testing images by applying either random or systematic geometric transformations, as well as photometric transformations. Through Python scripts the user can define the source image, the type of transformation, the number of images to be generated, and the minimum and maximum values for the given transformation. In this way, it is easy to generate several datasets, with different types of images, and several types of transformations and transformation ranges. Next, we describe the type of transformations implemented in the image generator.

5.1 Geometric transformations

The proposed testing image generator allows to generate transformed views of a source image by applying similarity transformations such as isotropic scaling, or in-plane rotation, as shown in Figure 8, as well as other affine transformations in one or several directions. The generation of this type of images is useful in order to evaluate the behaviour of different approaches against different values of a given transformation, as described in section 3.

5.2 Photometric Transformation

In our transformed image generator, we also incorporated a functionality that allows to generate images

Figure 8: Scale transformed views of the first image of the Graffiti dataset, proposed in (Mikolajczyk and Schmid, 2002)

contaminated with noise. Digital image noise can be split mainly in two different categories, luminance noise and chrominance noise, depending if the errors are produced in luma (intensity) or in chroma (color). There are some others types of noise such as horizontal or vertical banding (patterned noise), but it does not degrade images as luminance or chrominance noise do. Our image generator is able to create images contaminated with luminance or chrominance noise, or with both types simultaneously. Figure 9 shows,

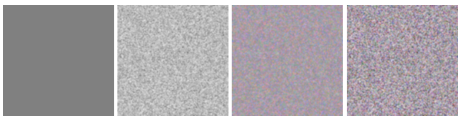

Figure 9: Types of noise

from left to right, an image patch filled with 50% gray value, contaminated with luminance noise only, with chrominance noise only and with both types of noise simultaneously. Depending on the nature of the camera and acquisition, i.e. exposure and ISO speed, these errors may vary. For example, we can check in the images of light varying dataset how noise levels increase as light decreases (SNR decreases), which is more noticeable in the case of the iPad.

6 CONCLUSIONS

We have presented a new set of images, as well as an image generator and an evaluation framework that help in the evaluation and development of new approaches related with image key point extraction, description and matching for both standard and mobile devices. Our proposed framework can be seen as an extension or an evolution of the extensively used evaluation framework of (Mikolajczyk and Schmid, 2002). Moreover, the proposed image dataset has a higher number of images, with higher resolution and with better controlled geometric and photometric conditions. The evaluation framework is entirely written in C++, and therefore easily integrable in many research environments related with the testing or development of key point extraction, description and matching mechanisms.

We are currently using and extending our proposed framework for the evaluation of state-of-the-art approaches for key point feature descriptors, such as BRIEF, ORB, RIFF, sGLOH, FREAK, NERIFT, or BRISK, among others, with real acquired images, as well as with synthetically generated ones.

REFERENCES

Alahi, A., Ortiz, R., and Vandergheynst, P. (2012). Freak: Fast retina keypoint. In *IEEE Conference on Computer Vision and Pattern Recognition (To Appear)*.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417.

Bellavia, F., Tegolo, D., and Trucco, E. (2010). Improving sift-based descriptors stability to rotations. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, pages 3460–3463. IEEE Computer Society.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Fraundorfer, F. and Bischof, H. (2005). A novel performance evaluation method of local detectors on non-planar scenes. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 33–33. IEEE.

Gauglitz, S., Höllerer, T., and Turk, M. (2011). Evaluation of interest point detectors and feature descriptors for visual tracking. *International journal of computer vision*, pages 1–26.

Hartley, R. I. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition.

Heikkilä, M., Pietikäinen, M., and Schmid, C. (2009). Description of interest regions with local binary patterns. *Pattern recognition*, 42(3):425–436.

Hughes, G. and Chraibi, M. (2011). Calculating ellipse overlap areas.

Leutenegger, S., Chli, M., and Siegwart, R. (2011). Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE.

Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector. *Computer Vision,ECCV 2002*, pages 128–142.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., et al. (2007). Affine covariant features. *Collaborative work between: the Visual Geometry Group, Katholieke Universiteit Leuven, Inria Rhone-Alpes and the Center for Machine Perception*.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280.