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

Iñigo Barandiaran[1,2], Camilo Cortes[1,3], Marcos Nieto[1], Manuel Graña[2] and Oscar E. Ruiz[3]
[1]*Vicomtech-IK4 Research Alliance, San Sebastián, Spain*
[2]*Dpto. CCIA, UPV-EHU, San Sebastián, Spain*
[3]*CAD CAM CAE Laboratory, Universidad EAFIT, Carrera 49 No 7 Sur - 50, Medellín, Colombia*

Keywords: Keypoint Extraction, Feature Descriptor, Keypoint Matching, Homography Estimation.

Abstract: Key point extraction and description mechanisms play a crucial role in image matching, where several image points must be accurately identified to robustly estimate a transformation or to recognize an object or a scene. New procedures for keypoint extraction and for feature description are continuously emerging. In order to assess them accurately, normalized data and evaluation protocols are required. In response to these needs, we present a (1) new evaluation framework that allow assessing the performance of the state-of-the-art feature point extraction and description mechanisms, (2) a new image dataset acquired under controlled affine and photometric transformations and (3) a testing image generator. Our evaluation framework allows generating detailed curves about the performance of different approaches, providing a valuable insight about their behavior. Also, it can be easily integrated in many research and development environments. The contributions mentioned above are available on-line for the use of the scientific community.

GLOSSARY

I	:	Set of images
I_n	:	Image n of set I
x_{in}	:	pixel position in I_n of a keypoint in $\mathbb{R}^2$
H_{ab}	:	Homography that maps pixels of I_a to I_b
$d(a,b)$	:	$\|a-b\|_2$

1 INTRODUCTION

Several computer vision-based applications rely on keypoint matching. Depending on the nature of such applications, the requirements for a specific keypoint extractor and descriptor may vary. For example, applications related with self-navigation or simultaneous location and mapping (SLAM) require a fast keypoint extractor algorithm because of their real-time restrictions. On the other hand, an application for object or image recognition benefits from a more robust keypoint descriptor; even if this implies a higher computation time.

A **keypoint** is a distinguished point in $\mathbb{R}^2$ representing the projection of a particular structure of a 3D scene. A **feature descriptor** is a vector in $\mathbb{R}^k$ that contains a set of attributes that intend to uniquely represent x_{in} .

Currently, there is an increasing activity in the development of new approaches for keypoint extraction, description and matching, pursuing robustness and low computational complexity. In order to assess these new approaches accurately, normalized data and evaluation protocols are required.

Responding to the mentioned needs, in this paper we present a new evaluation framework for the evaluation of the state-of-the-art keypoint extractors and feature point descriptors.

Formally, the framework discussed here has the following I/O specification:
INPUTS: (1) A set $I = \{I_1, I_2, \ldots, I_z\}$ captured from a particular scene. (2) A set of bijection functions $S_0 = \{f_{1,2}, f_{1,3}, \ldots, f_{i,j}, \ldots\}$, such that $f_{i,j} : I_i \rightarrow I_j$ establishes the real correspondence between pixels of I_i and I_j, so that mapped pixels actually mark the same 3D point. (3) A set of matching algorithms $A = \{A_1, A_2, \ldots, A_w\}$. Algorithm A_m is an arbitrary configuration of a keypoint extraction approach and a feature description technique. A_m produce an alternative set of functions S_m when applied on I, which match the images of I among themselves. The set S_0 is the ground-truth data of I, since it is the set of map-

Barandiaran I., Cortes C., Nieto M., Graña M. and Ruiz O..
A New Evaluation Framework and Image Dataset for Keypoint Extraction and Feature Descriptor Matching.
DOI: 10.5220/0004211502520257
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 252-257
ISBN: 978-989-8565-47-1

pings corresponding to reality. The set of functions S_m is considered imperfect, since it resembles the actual set S_0.

OUTPUTS: A set of performance evaluations for the matching algorithm A_m ($1 \leq m \leq w$). These performance evaluations obviously grade the quality of A_m against the ground-truth data. This appraisal allows measuring several algorithm's features, such as repeatability, accuracy and invariance to affine or photometric transformations.

In addition, this article reports the protocol for producing a particular set I under controlled affine and photometric transformations. The capture has been conducted using a methodology that allows to ensure that only one kind of transformation occur for a series of images. This permits to determine how sensitive is A_m to a specific factor. In order to supplement the dataset of real images, we present an image generator that allows producing images with affine or photometric transformations for testing purposes.

The research community and the practitioners on computer vision applications can obtain valuable information from the mentioned contributions to improve their approaches or to select the algorithm that best suit their needs.

2 RELATED WORK

Tuytelaars et al. suggested that there are several parameters of a point detector and feature descriptor that can be measured to assess their performance (Tuytelaars and Mikolajczyk, 2008). However, to measure some of them, such as the point extractor accuracy, descriptor robustness or invariance a normalized test protocol and test benchmark are required. In this way, the seminal works of Mikolajczyk et al. settled the basis for keypoint extractor and feature description evaluations (Mikolajczyk and Schmid, 2005). Since then, several new approaches for keypoint or region extraction (Mikolajczyk et al., 2007) and for feature description (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, which are freely available online at *www.robots.ox.ac.uk/ vgg/research/affine/index.html.*

Recently, Gauglitz et al. proposed a dataset of several videos of surfaces, with different types of textures and different light conditions, which are used to evaluate keypoint matching strategies oriented to camera tracking applications (Gauglitz et al., 2011). The authors claim that due to restrictions of 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 the evaluation of a particular factor may not be performed given the differences in the geometric transformations during the acquisition of I.

Very recently, Alahi et al. (Alahi et al., 2012) tested their descriptor approach with the dataset and evaluation framework of Mikolajczyk and Schmid (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.

Important contributions have been performed to assess the performance of the extraction and matching mechanisms using non-planar scenes. Fraundorfer and Bischof (Fraundorfer and Bischof, 2005) proposed an extension of the work of Mikolajczyk and Schmid (Mikolajczyk and Schmid, 2005) by analyzing keypoint repeatability for non-planar scenes. They used tri-focal tensor geometric restriction for estimating the ground-truth data of their own dataset. Other relevant works in this field are presented by Gil et al. (Gil et al., 2010) and Moreels et al. (Moreels and Perona, 2007).

Our dataset and evaluation framework are inspired by the developments of Mikolajczyk and Schmid (Mikolajczyk and Schmid, 2005). In comparison to that approach, our contribution comprises a higher number of images, with higher resolution and with better controlled conditions, and it is supplemented with an image generator. For the acquisition of images we used different types of sensors, including mobile devices. It is important to consider some features of these devices, such as their low dynamic range, in a testing data. To the best of our knowledge, this feature lacks in the available testing datasets. Our dataset present variations in both geometric (e.g. similarities and affinities) and photometric transformations (e.g. luminance and chrominance noise addition).

Finally, our evaluation framework is written in C++, which makes its integration in development environments straightforward, and allows generating detailed curves about the performance of different approaches.

All presented 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). It uses the class hierarchy implemented in OpenCV that decouples keypoint extraction from keypoint description and descriptor matching, allowing to try different configurations of keypoint extractor, descriptors and matchers. Whereas Mikolajczyk's work (Mikolajczyk and Schmid, 2005), where the framework is written in Matlab scripting, our approach is written in C++, allowing its easy integration in a development environment. Thus, is not necessary to export additional data to other platforms, as occurs with the mentioned Matlab-based evaluation framework, which can be can be cumbersome, especially when developing commercial software. Nevertheless, our approach also supports the reading of Mikolajczyk's file format, allowing the comparison with previous approaches or studies. Figure 1 shows partial results

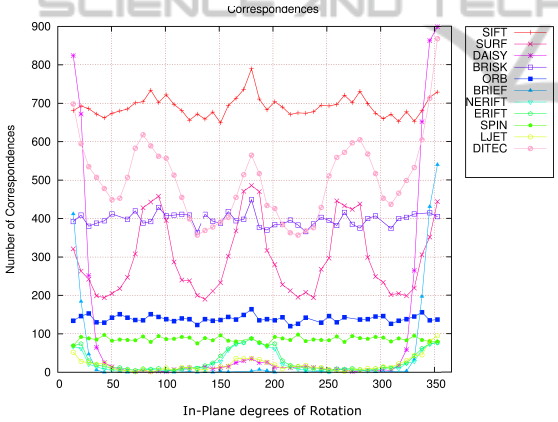

Figure 1: Results of the evaluation of several feature descriptors using the in-plane rotation.

of an evaluation conducted using our dataset and evaluation framework. In addition to the precision-recall curves proposed by Mikolajczyk and Schmid (Mikolajczyk and Schmid, 2002), our framework generates detailed performance curves based on the number 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. They suggest that, for example, BRIEF descriptors are not robust against a rotation larger than 35 degrees approximately. Also, it can be observed that 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

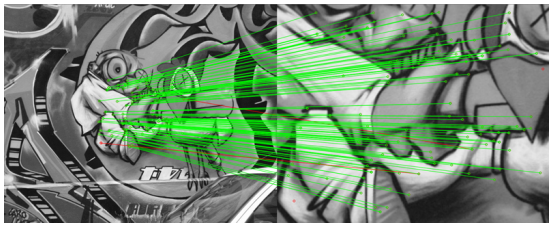

Figure 2: Correct matches (in green), wrong matches (in red) between two images.

way, a better insight of the behavior of a given approach can be obtained.

3.1 Matching Evaluation

An image formation process is usually represented as in Equation 1 where Y_w represents a point in $\mathbb{R}^3$ and Y_i corresponds to the projection of Y_w 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.

$$y_i = PY_w \tag{1}$$

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

When either points Y_w lie on a plane, or the images are acquired with a camera rotating around its center of projection, the transformation among points y_i and points Y_w corresponds to a 2D linear projective transformation or homography H (Hartley and Zisserman, 2004).

As in the dataset proposed by Mikolajczyk and Schmid (Mikolajczyk and Schmid, 2002), a 2D homograhpy relate all images in our set I. This known transformation is used as ground truth data, allowing to know a priori where x_{ia} should be projected in I_b by using Equation 3.

$$x_{jb} = H_{ab}x_{ia} \tag{3}$$

Similarly, keypoints from I_b can be projected back to I_a by using the inverse of H_{ab}. Let $\tilde{x}_{jb}$ be the estimated match of x_{ia} obtained by a given A_m. Then, H_{ab} is used to measure the accuracy and repeatability of a point detector algorithm. This process is performed by computing the error measure d_{ij} of the estimated and the ground truth keypoints of a pair of images, as shown in Equation 4.

$$d_{ij} = d(\tilde{x}_{jb}, H_{ab}x_{ia})^2 + d(x_{ia}, H_{ab}^{-1}\tilde{x}_{jb})^2 \tag{4}$$

In order to estimate correct matches m_{ab} of keypoint pairs x_{ia} and $\tilde{x}_{jb}$, as shown in Figure 2, we used the overlap error defined by equation 5 to reduce the

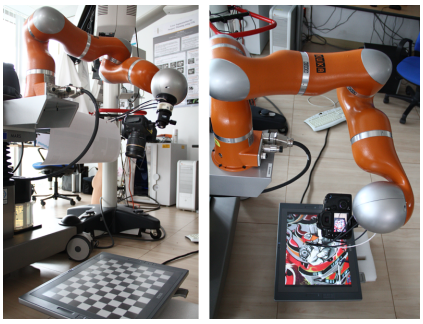

Figure 3: Image acquisition setup with Kuka robotic arm and Canon 7D attached.

probability of occurrence of false positive matches (Mikolajczyk and Schmid, 2002). Consider two ellipsoidal support regions R_{ia} and R_{jb} estimated by a point extraction algorithm. The error in equation 5 measures how well the supporting regions correspond under the geometric transformation H_{ab}.

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

We calculate the ellipses overlap by using the software proposed by Hughes and Chraibi (Hughes and Chraibi, 2011), which is available at *www.chraibi.de*.

If point pair x_{ia} and $\tilde{x}_{jb}$ present an error measure d_{ij} given by equation 4 and overlap error given by equation 5 under some predefined thresholds, then it is considered as a correspondece.

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 different values of perspective distortion, we used a Kuka robotic arm with a Canon 7D attached with Tamron lens (see figure 3) to obtain several points of view of a target scene by traversing circular trajectories (arcs), as shown in Figure 4. The robot allowed us to generate known, repeatable and precise poses and trajectories around the target scene, as an improvement to the manual acquisition described by Gauglitz et al. (Gauglitz et al., 2011). To generate and command the follow up of the desired trajectories, we developed an application in C++, which uses the Kuka Fast Research Interface to interface with the robot.

We used a Wacom Cintiq screen for displaying the target images, instead of using pictures placed in a wall or in a table as performed by Gauglitz et al. (Gauglitz et al., 2011). Our set of displayed images covers different types of images with structured, unstructured and low texture, as well as repeating patterns. Several authors (Tuytelaars and Mikolajczyk, 2008; Heikkilä et al., 2009; Gauglitz et al., 2011) agree in the importance of evaluating keypoint extractors and descriptors in different conditions to truly test their robustness.

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-th ($1 \leq i \leq M$) desired pose of the camera with respect to the robot's base coordinate system. 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.

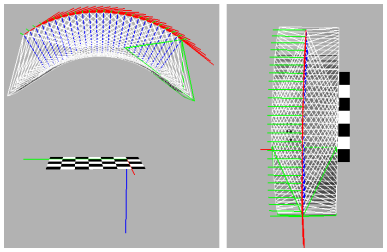

Figure 4: Recovered trajectory from a circular sector of a robot-driven image acquisition.

The set Q is traversed in order. When the camera reaches a particular Q_i, a signal is send to it in order to take N pictures in a synchronous way. At any Q_i the first picture to be taken corresponds to the calibration pattern image; then $N-1$ pictures of other images shown on the Wacom Cintiq screen are taken. While pictures are being taken the robot holds its position.

We used the calibration pattern image to calculate the extrinsic and intrinsic parameters of the camera,

Figure 5: Some images of exposure varying dataset compound of 15 different images.

and also to estimate the homographies between images accurately. This also allowed us to rectify the distortion of the images.

This novel implementation guarantees that the homography that relate the images taken at Q_i and Q_{i+1} is the same for all pictures. Thus, this allows to undertake the performance evaluations under the same geometric transformation and different image's texture patterns.

4.2.1 Image Focus

In addition to the capability of generating unfocused images with our testing image 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 final rendered image; therefore it is not easy to simulate their effect synthetically. We present 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.

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. These ones are related to the camera settings, light conditions and the nature of the camera hardware (mainly the camera sensor). Here we present a set of images that show a variation in the light condition or light exposure, as shown in Figure 5. The purpose of this subset is to be able to evaluate the robustness of keypoint extractors repeatability or feature descriptors robustness against illumination changes and noise.

Image acquisition was carried out by operating the illumination equipment and the camera remotely, ensuring that no geometric transformations were applied and only photometric transformations occur between the images that form this dataset. This implies that the homography matrix that relates them geometrically correspond to the identity matrix.

The use of flashes to generate the illumination of the scene allowed us to vary the amount of light without changing any camera acquisition parameters, 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 ISO speed or 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 6: Images from the exposure varying dataset taken with a mobile device.

Figure 6 shows two images of the same scene taken with an 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. For this setup we used a continuous light source where light intensity can be set manually. It is worth mentioning that the focus point, exposure metering point and aperture were fixed along the capturing of all images in the dataset.

As expected, in both scenarios, 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 IMAGE 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.

The proposed testing image generator allows to generate transformed views of a source image by applying similarity transformations such as isotropic scaling, as shown in Figure 7, or in-plane rotation, as well as other affine transformations in one or several directions.

Digital image noise can be classified mainly in two categories, luminance and chrominance. Our image generator is able to create images contaminated with luminance or chrominance noise, or with both types simultaneously.

Figure 7: Scale transformed views of the first image of the Graffiti dataset proposed in (Mikolajczyk and Schmid, 2002).

6 CONCLUSIONS

We have presented a new set of images, as well as an image generator and an evaluation framework that allow evaluating approaches related with image keypoint extraction, description and matching for both standard and mobile devices. Our framework can be seen as an evolution of the extensively used evaluation framework of Mikolajczyk and Schmid (Mikolajczyk and Schmid, 2002). Moreover, the presented 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 and development environments of this field.

We are currently using and extending our proposed framework for the evaluation of state-of-the-art approaches for keypoint 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.

Gil, A., Mozos, O., Ballesta, M., and Reinoso, O. (2010). A comparative evaluation of interest point detectors and local descriptors for visual slam. *Machine Vision and Applications*, 21(6):905–920.

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. *arXiv preprint arXiv:1106.3787*.

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*.

Moreels, P. and Perona, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73(3):263–284.

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