Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 421 – 428

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Information System for Automation of Counterfeited Documents Images Correlation

Rafael Vieira[1], Catarina Silva[1,2*], Mário Antunes[1,3], Ana Assis[4]

[1]*School of Technology and Management, Polytechnic Institute of Leiria, Portugal*
[2]*Center for Informatics and Systems of the University of Coimbra, Portugal*
[3]*Center for Research in Advanced Computing Systems, INESC-TEC, University of Porto, Portugal*
[4]*Scientific Police Laboratory - Judiciary Police, Portugal*

Abstract

Forgery detection of official documents is a continuous challenge encountered by documents' forensic experts. Among the most common counterfeited documents we may find citizen cards, passports and driving licenses. Forgers are increasingly resorting to more sophisticated techniques to produce fake documents, trying to deceive criminal polices and hamper their work. Having an updated past counterfeited documents image catalogue enables forensic experts to determine if a similar technique or material was already used to forge a document. Thus, through the modus operandi characterization is possible to obtain more information about the source of the counterfeited document.

In this paper we present an information system to manage counterfeited documents images that includes a two-fold approach: (i) the storage of images of past counterfeited documents seized by questioned documents forensic experts of the Portuguese Scientific Laboratory in a structured database; and (ii) the automation of the counterfeit identification by comparing a given fraudulent document image with the database images of previously catalogued counterfeited documents. In general, the proposed information system aims to smooth the counterfeit identification and to overcome the error prone, manual and time consuming tasks carried on by forensic experts. Hence, we have used a scalable algorithm under the OpenCV framework, to compare images, match patterns and analyse textures and colours. The algorithm was tested on a subset of counterfeited Portuguese citizen cards, presenting very promising results.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the organizing committee of CENTERIS 2016

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: catarina@ipleiria.pt

1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.178

Keywords: Information systems; Fraud detection; Pattern Matching; Image Processing.

1. Introduction

Digital image processing and manipulation areas are increasingly used by individuals or criminal organizations engaging in illegal activities related to documents forgery, mainly security documents like citizen cards, driving licenses and passport. The counterfeiting of security documents by its illegal reproduction are high profile crimes involved in terrorism, financial transactions or drug trafficking.

There are a lot of different digital techniques to produce fake documents, for example, resampling, copy-paste and splicing[1]. These techniques are described in Section 2.1.

Forensic laboratories have a well-defined set of protocols to detect counterfeited and forged documents and the correspondent fraudulent sources. For a counterfeited document detection, a set of illumination techniques is applied, to enhance/detect printing processes, holograms, watermarks and other security features. Furthermore, a verification of old cases similar to the current one being analyzed is carried out, usually manually. If a match is found, the counterfeit is identified and a correlation with all the identical cases already detected in past will be successfully pursued. Otherwise, it will be created a new counterfeit number for future correlations.

The whole process of comparing a fake document with a list of previously catalogued counterfeited ones is usually made manually by the questioned forensic experts of the Scientific Police Laboratory. Having in mind that the catalogue of documents, even for a specific document type is potentially overwhelming, the time involved in such manual analysis may thus be prohibitive and certainly inefficient for a fast criminal investigation response. Hence, an information system based on image detection algorithms that could automate, or semi-automate such process could bring numerous advantages.

The main challenges faced by such an image detection algorithm are its ability to distinguish different types of features/printing techniques in specific regions of the image, that is to identify the processing area of the document to be processed and the type of tests and comparisons that should be applied. The algorithm should also cope with the content of the data source, like the type of document and illumination that was used or if it is the front or versus of the document. If the data source has a lot of images but they are not uniformed, the output of the system will be affected. For instance, one may consider the following cases which will affect negatively the output of the system:

- Images with ultraviolet filters;
- Existence of zoomed zones;
- Existence of a scale rule.

In this paper we propose an integrated information system that can tackle with this issue by storing the images related to each former fraudulent documents' case in a normalized and structured database, where one can easily search country, type of document, and main printing technique. The proposed system also provides an automated process to compare and eventually match the patterns in a new counterfeiting case with the existing questioned documents related with the same document type and country.

The aim of this Information System (IS) is not to replace the human in the loop, as such manual verification should always be carried out in any case, but to implement an algorithm that ranks the compare image of questioned documents by a level of similitude. That is, to discard automatically the documents that have less or no similitude with the document being analyzed, thus directing forensic expert's attentions to the most relevant documents.

The rest of the paper is organized as follows. Section 2 presents the fundamentals regarding fraud in security documents and image processing algorithms. Section 3 details the proposed architecture and the experimental approach. Section 4 describes the results obtained with the use of the integrated information system and the image processing algorithm with real testing data provided by Scientific Police Laboratory of the Portuguese Judiciary Police. Section 5 concludes the paper and presents future lines of research.

2. Background

In this section we present the main background topics that are needed to understand this work, namely the meaning of counterfeit in security documents and an introduction to OpenCV and its most interesting algorithms for digital image processing.

2.1. Counterfeited documents

Counterfeited documents are reproductions or imitations of the originals ones. Nowadays this process starts with a digital creation/reproduction of the original document[2,3,4]. To produce the physical document it is necessary to use materials and printing techniques from available technologies.

The forensic analyses of all the constituent elements of the questioned document is carried out through different techniques and methodologies (physical and chemical examinations). Those elements may include printing process, watermarks, fluorescent fibers and planchettes, guilloche pattern, fluorescent and magnetic inks, optically variable inks, rainbow printing, microprinting, latent images, scrambled indicia, laser printing, photos, signatures, embossing stamps, optically variable devices, protective films, perforations, machine readable security, retro-reflective pattern, among others. This analysis provides information that may lead to the classification of the documents as genuine, false or forged. All technical observations are intelligence information that may conduct to the discovery of the counterfeiting operation, i.e. associate the counterfeit with the components of its production.

The database registration of a counterfeited document depends on its uniqueness individualization characteristics.

2.2. Image Processing Algorithms

There is a wide set of developed algorithms to cope with image comparison and fraud detection. A well-known and very popular open source library for computer vision is OpenCV[†] that has several tools for digital image processing. It includes a set of widely used algorithms for pattern detection and image comparison that are briefly explained below.

The Harris Corner Detection algorithm[5] was developed by Chris Harris and Mike Stephens. It uses a mathematic model to detect corners and edges using a function that calculates a correlation. This function considers window in the image and determine the average changes of image intensity that result from shifting the window by a small amount of pixels in various directions. It is used to detect corners in digital images, by comparing the same area in both fake and genuine documents.

Another corner detection algorithm was proposed by Lowe[6]: the Scale-Invariant Feature Transform algorithm (SIFT). The rationale was that all the algorithms were based on corner detections and that if the image was zoomed out or zoomed in, the corners would be affected. The SIFT was created with the purpose, as the authors describe, to be invariant to image scale and rotation.

More recently, Bay et al. presented a new variation of SIFT, named Speeded-Up Robust Features (SURF), with the goal of obtaining a version more optimized to computer vision processing[7]. At the same time, Rosten and Drummond developed the Fast Algorithm for Corner Detection (FAST)[8], aiming to have better performance in real time applications. Although these two new algorithms are patent protected, which means that they cannot be used commercially without royalties' payment.

In 2011, Rublee et al. implemented Oriented Fast and Rotated Brief algorithm (ORB)9 that can be seen as a mix of SIFT and FAST algorithms. Given that ORB free for commercial use, the OpenCV framework includes an implementation along with other interesting functionalities to deal with patterns' detection in digital images, e.g. homography and matching template. The former relates to the detection of an image inside another, even with rotation or perspective. The later improves homography by providing six different mathematic formulas to detect one image into another.

[†] http://www.opencv.org

With the release of OpenCV version 2, an API it became available to help the implementation of *intelligent* algorithms through libraries for K-Nearest Neighbor, SVM (Support Vector Machines), K-Means Clustering, Neural Networks and Decision Trees.

3. Integrated Information System for Automation of Counterfeited Documents Images Correlation

In this section we will present the proposed architecture and describe the main features related to the counterfeit documents images correlation automation system

3.1. Proposed Architecture

Current processes use desktop applications that access the images, manually organized by type of document and country. These pictures are uploaded to a directory in a local server through dedicated hardware that scans the documents with high level of quality. The procedure used to analyse the pictures and identify the corresponding falsification is thus fully manual and consists in analysing each picture individually.

The proposed architecture, depicted in Figure 1 is web-based. It is composed of a web client that sends HTTP requests to a web server by taking advantage of an embedded REST API functions set. The overall procedure is as follows: i) the web server validates the HTTP request; 2) the web server queries the database to find the images related with the type of document and country; 3) the web server then executes the scripts that implement the image processing modules; 4) finally, the web server returns to the client the similitude level of each image with the document that is being analysed.

The structured database is able to store the images and their general information, in order to provide a better performance to the matching images' algorithm. During the transition phase both databases will be used simultaneously, being the database of images fed with the data stored in the current one. By using a REST API we enable the use of any web-based client, not only a web browser, giving flexibility to the overall solution.

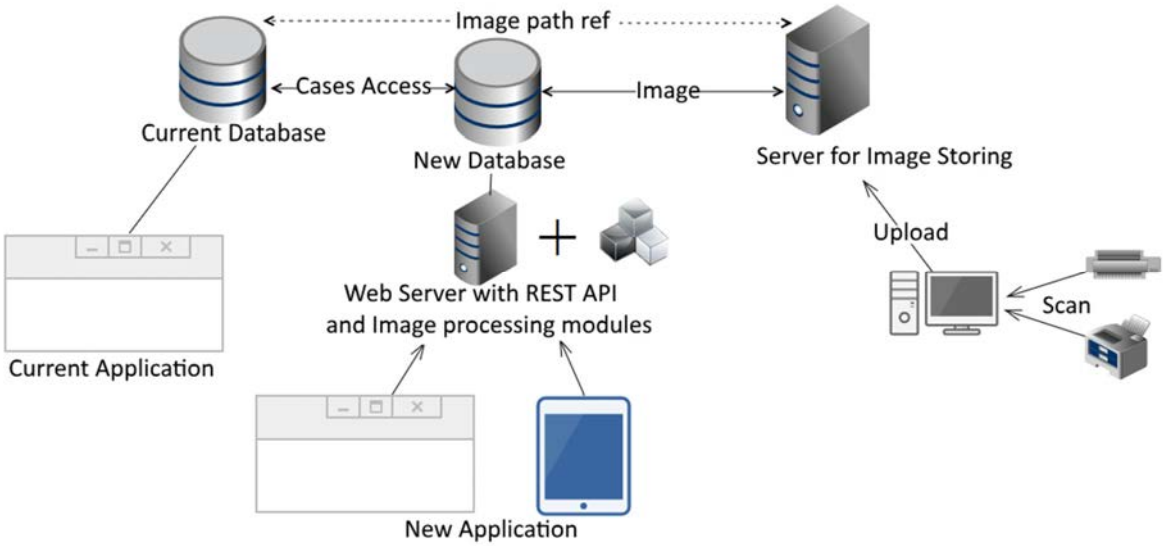

Figure 1. General Architecture

3.2. Technologies

Table 1 presents the technologies used to implement the solution described below. The implementation was carried out with free open source software.

Table 1. Adopted Technologies.

Module	Client	Web Server	Image Processing	Database
Technologies	HTML5, CSS3, JavaScript	Apache2 v2.4.7 PHP v5.5.9	C++ with OpenCV2.4.9	PostgreSQL v9.5

3.3. Automation of Counterfeited Documents Images Correlation

A comprehensive set of automatic digital image processing techniques and algorithms must be applied in order to achieve better accuracy and therefore, to reduce the group of documents that needs to be manually analysed.

As represented in Figure 2, the image processing algorithm used to automate the analysis of counterfeited documents comprises different modules, each one responsible for a specific analysis task: 1) texture analysis; 2) comparison of image areas; 3) detection of similar imperfections in text areas.

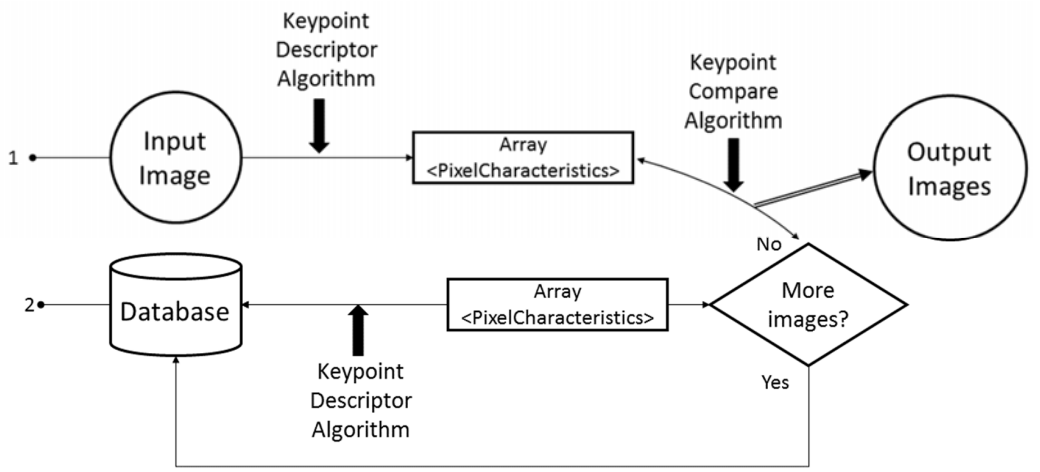

Figure 2. Image processing algorithm

Figure 2 depicts the image processing algorithm, including the flow of data throughout the process using OpenCV features. Firstly (1) a given image will be processed with a "keypoint descriptor" algorithm, that is OpenCV SURF, SIFT or ORB implementations. In this step the image is processed and we obtain an array with the following relevant information: angles of corners, edges, pixel's intensity and directions of the most pronounced intensity changes. Next (2) a "descriptor" algorithm is applied to the images of known fraudulent documents stored in the database that meet the same characteristics (type of document and country) of the input document. For instance, if the input document is a Portuguese citizen's card, only documents of this type will be processed. The information retrieved for each processed document is then compared with the array obtained in step (1), using a "descriptor compare" algorithm. The output obtained is an index of similitude between the input document and each one of the evaluated documents. Regarding texture area, the texture descriptor processing algorithms, like OpenCV native algorithms (e.g. HogDescriptor) extracts the necessary parameters to describe it.

For *"keypoint descriptors"* processing we used OpenCV SURF implementation, which is not OpenCV native. By giving an image to a keypoint descriptor for each pixel, the algorithm computes an *"interestingness"* function, which measures the likelihood and uniqueness of each point in another similar image. To that effect, the keypoint

descriptor analyzes the area around each pixel (the corners) and calculates statistical values and hashes that will be retrieved for future comparison. The next step consists on applying the same *"keypoint descriptor"* algorithm to the images stored on database and evaluate whether they contain the same information.

4. Tests and Results Analysis

4.1. Experimental setup

Forgery and counterfeits in security identity documents can be of several types and may use different resources. We have conducted two distinct tests, both with images from the Portuguese citizen card, as depicted in Figure 3:

A. Image of the chip area of a Portuguese identification card (Figure 3A);
B. Image of the optically variable ink security feature (Figure 3B).

Figure 3. (A) Input Image of Test A; (B) Input Image of Test B

For each test, the input image was compared with images of the same type from the counterfeited images database. For the test A the dataset contains 6 false documents and 1 valid document. For the test B the dataset contains 8 false documents and also 1 genuine document. By using the SURF algorithm, it is possible to how closely the images in a dataset are from the input image by detecting the most important keypoints. The tool provides a first output where the input image is displayed side by side with each compared images along with the representation of the common keypoints that were analyzed and their location.

Figure 4 presents two samples for the image of the Portuguese citizen card chip (Figure 3A) being compared in their keypoints. Notice that neither the documents nor the taken images have a perfect cut around the area of interest to analyze, which must be accepted by the algorithm. Since a high number of keypoints are considered, in some cases the image might be confusing. To overcome this situation, the tool provides an alternative output in which the keypoints are ranked, being the best and the worst keypoints highlighted.

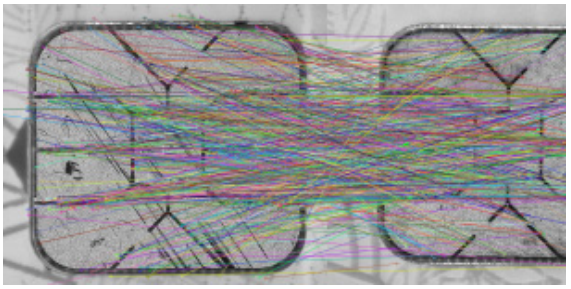

Figure 4. Visual output representing keypoints matching.

Keypoint analysis emphasizes the fact that an image with more similar keypoints is the best candidate to have the same *modus operandi* of the given fake document.

Table 2 presents the distance between keypoints from the testing image and elected documents stored in the database. The columns (from left to right respectively) represents the medium value for the 10 keypoints with more

similitude; the medium value for the 10 keypoints with less similitude; and the percentage of keypoints with more than 80% of similitude. That is, the lower the values for keypoints, more similar are the documents and thus higher the percentage of similitude.

In Table 2 the degrees of similitude for each document in the dataset are presented. The results of the Test A show that the valid (positive) document is the one with keypoints with higher difference in the worse keypoints (40.18%) and the number of keypoints with 80% or more similitude (898). Although the false document #6 has similar results to the valid document, this can be justified by the fact that not all the area of the chip suffered edition processing. For this same reason, the valid document has similar values in the better keypoints. In this text, the candidates to have a similar fraud to the input document are the documents #4 and #5 since they have the higher degree of similitude in the better and worse keypoints, as well as the higher percentage of keypoints over 80% of similitude.

Table 2. Degrees of similitude

Test A	**Better Keypoints**	**Worse Keypoints**	**Keypoints<0.2 (>80%)**
Distance to false document #1	0.0856 (91.44%)	0.4964 (50.36%)	1101 (41.02%)
Distance to false document #2	0.0873 (91.27%)	0.5203 (47.97%)	1267 (47.20%)
Distance to false document #3	0.0774 (92.26%)	0.5069 (49.31%)	1154 (42.96%)
Distance to false document #4	0.0500 (95.00%)	0.4865 (51.35%)	1697 (63.22%)
Distance to false document #5	0.0470 (95.30%)	0.4590 (54.10%)	1745 (65.01%)
Distance to false document #6	0.1052 (89.48%)	0.4891 (51.09%)	786 (29.28%)
Distance to false document #7	-	-	-
Distance to false document #8	-	-	-
Distance to a genuine document	0.0791 (92.09%)	0.5982 (40.18%)	898 (33.46%)

Test B	**Better Keypoints**	**Worse Keypoints**	**Keypoints<0.2 (>80%)**
Distance to false document #1	0.1173 (88.27%)	0.4810 (51.90%)	41 (15.30%)
Distance to false document #2	0.1427 (85.73%)	0.5118 (48.82%)	35 (13.06%)
Distance to false document #3	0.1086 (89.14%)	0.6137 (38.63%)	48 (17.91%)
Distance to false document #4	0.1116 (88.84%)	0.5377 (46.23%)	44 (16.42%)
Distance to false document #5	0.1227 (87.73%)	0.6696 (33.04%)	32 (11.94%)
Distance to false document #6	0.1149 (88.51%)	0.6126 (38.74%)	50 (18.66%)
Distance to false document #7	0.1105 (88.95%)	0.5791 (42.09%)	34 (12.69%)
Distance to false document #8	0.1094 (89.06%)	0.5208 (47.92%)	58 (21.64%)
Distance to a genuine document	0.1973 (80.27%)	0.5682 (43.18%)	4 (1.49%)

Regarding Test B, the valid document is undoubtedly the one that presents results with less similitude (1.49%) of keypoints over 80% and the lower percentage in the better keypoints (80.49%). In this test we have not a false document with similar results to the valid document because the area is a hologram, and normally when a hologram is forged it can be detected at any point of the hologram area. However, according to the results obtained, the better candidate is the document #8 since it has 89% (best of all along with #3) on the better keypoints and more keypoints over 80% of similitude (21.64% against 18.66% of the second best).

Table 3 presents the medium values for all the false documents comparing to the values of the valid document.

Table 3. Medium of degrees of similitude between false documents and valid documents

Test A	**Better Keypoints**	**Worse Keypoints**	**Keypoints<0.2 (>80%)**
Distance false documents	0.0754 (92.46%)	0.4930 (50.07%)	1296 (48.29%)
Distance genuine document	0.0791 (91.27%)	0.5982 (40.18%)	898 (33.46%)

Test B	**Better Keypoints**	**Worse Keypoints**	**Keypoints<0.2 (>80%)**
Distance false documents	0.1172 88.28%)	0.5658 (43.42%)	43 (16.04%)
Distance genuine document	0.1973 (80.27%)	0.5682 (43.18%)	4 (1.49%)

In general, the valid document has lower similitude, which was the expected result. Although in both tests the better and worse keypoints values are similar between the valid and the false documents, this may be explained by

the fact that in a faked area not all the points/pixels suffer distortion or any kind of image change. The number of similar keypoints over a threshold (in this case 80%) can thus give us better confidence about the documents that should be further manually analyzed. In our tests the valid documents have much less keypoints comparing to the false ones.

5. Conclusions and Future Work

The purpose of this paper was to approach a systematic solution to enhance forensic analysis in correlating *modus operandi* between successive counterfeiting cases. For a criminal investigation in document fraud scenes it is of utmost importance to detect the source to avoid future falsifications from the same source. A way to determine the source is retrieving information from old cases, making a similitude link between those previous cases and the new ones. Nowadays, such operation is carried out manually which consumes too much human resources and time.

To fulfill this gap an integrated information system is presented in this paper that uses visual computing algorithms form OpenCV to aid in the semi-automatic comparison of the areas of the document when a counterfeit is identified. Tests were carried out using source images and letting the algorithm retrieve the degree of similitude to each image in the dataset, using different areas of documents that had went through a fraudulent process. The results achieved assigned a higher percentage to the false document in both tests which was the expected output. With these results it is possible to assume that this can be a solid base for a more complex algorithm able to respond positively for different documents, cases and types of falsification.

As future work we aim to test other algorithms besides SURF, like ORB or SIFT. But for that we have to have a larger dataset and improve the relevance of the obtained results. The dataset must be as uniform as possible, in order to provide less chances to the algorithm to compare wrong keypoints. And finally, to test the algorithm in the complete data source of a criminal authority, where there are hundreds or thousands of different cases with several images each, and where experts can confirm that the output results are reliable.

Acknowledgements

The authors would like to thank the collaboration of the Questioned Documents Forensic Experts of the Scientific Police Laboratory of the Portuguese Judiciary Police.

References

1. H. Farid, “Image Forgery Detection”, in *IEEE Signal Processing Magazine*, pp. 16-25, March 2009.
2. A. Kaur and A. K. G. Vaibhav Saran, “Digital Image Processing for Forensic Analysis of Fabricated Documents”, in *International Journal of Advanced Research in Science, Engineering and Techonology*, pp. 84-89, September 2014.
3. R. Bertrand, P. Gomez-Kramer, O. R. Terrades, P. Franco and J.-M. Ogier, “A System Based On Intrisic Features for Fraudulent Document Detection”, in 12th Internation Conference on Document Analysis and Recognition, pp. 106-110, August 2013.
4. J. Fridrich, D. Soukal and J. Lukás, “Detection of Copy-Move Forgery in Digital Images”, in Forensic Science International, vol. 231, pp. 284-295, September 2013.
5. C. Harris and M. Stephens, “A Combined Corner and Edge Detector”, in 4th Alvey Vision Conference, pp 147-151, 1998.
6. David G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, in International Journal of Computer Vision, pp. 91-110, 2004.
7. H. Bay, A. Ess, T. Tuytelaars and Luc Van Gool, “Speeded-Up Robust Features (SURF)”, in Computer Vision and Image Understanding, vol. 110, pp. 346-359, June 2008.
8. Edward Rosten and Tom Drummond, “Machine Learning for High-Speed Corner Detection”, in 9th European Conference on Computer Vision, May 2006.
9. E. Rublee, V. Rabaud, K. Kunolige and G. Bradski, “ORB: An Efficient Alternative to SIRF or SURF”, in IEEE International Conference on Computer Vision, pp. 2564-2571, November 2011.