A Novel Global Image Description Approach for Long Term Vehicle Localization

Fabien Bonardi*, Samia Ainouz*, Rémi Boutteau†, Yohan Dupuis‡, Xavier Savatier† and Pascal Vasseur*
*Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes
Normandie Univ, UNIROUEN, UNIHAVRE, INSA Rouen, LITIS, 76000 Rouen, France {firstname.lastname}@litislab.fr
†Institut de Recherche en Systèmes Électroniques Embarqués
Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, 76000 Rouen, France; {firstname.lastname}@esigelec.fr
‡Centre d'Études et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement
CEREMA, 76000 Rouen, France; yohan.dupuis@cerema.fr

***Abstract*—Long-term *place recognition* for vehicles or robots in outdoor environment is still a tackling issue: numerous changes occur in appearance due to illumination variations or weather phenomena for instance, when using visual sensors. Few methods from the literature try to manage different visual sources while it could favor data interoperability across variable sensors.**

In this paper, we emphasis our works on cases where there is a need to associate data from different imaging sources (optics, sensors size and even spectral ranges). We developed a method with a first camera which composes the *visual memory*. Afterwards, we consider another camera which partially covers the same journey. Our goal is to associate live images to the prior visual memory thanks to visual features invariant to sensors changes, with the help of a probabilistic approach for the implementation part.

I. Introduction

Instead of visual sensors, recent prototypes in vehicular automation field reached milestones with strong use of LIDAR scanners to perceive their surrounding environment, to create a dynamic 3D reconstruction [1] and for localization [2]. Nevertheless, imaging sensors have widespread applications thanks to cheaper cameras, embeddability ease and plurality of informations provided by them. Like a human eye, ideal system design involves a camera as a single sensor for varied processing and tasks.

SLAM (Simultaneous Localization And Mapping) techniques consist of composing a map with sensors' data of the surrounding environment (whenever it is LIDAR or visual sensors) while the robot is evolving in its environment and needs localization in the map. In this way, place recognition (or visual localization) task permits the robot to simplify its map and eventually to minimize some drift. This task, also known as *loop closure* remains a tricky issue if a camera is used alone. Perceptual-aliasing in outdoor environments easily tricks even human perception. Indeed, when you are looking for long-term visual memory, changes from the surrounding environment could be considerable.

Outdoor environments suffer from various kind of changes: illumination, weather and seasonal variations and consequences on vegetation, but also changes due to human hands. Most of the time, images from a first experience, called *memory* or *visual memory*, are compared with *live* acquisition. A metric is defined on chosen features of interest and the best score determines if current position and place in the memory snapshot are the same. As this scheme generally yield errors and false matchings, methods have been improved to filter results and gain robustness, notably thanks to *temporal consistency* [3].

Our main contribution concerns a global image description for outdoor localisation, robust enough to sensors characteristics changes as well as perception changes of the environment. This image signature is a part of an overall application we have made with a particle filter implementation.

In section II, we sum up related works and methods from both robotics and image retrieval framework which have inspired this contribution. Section III gives details on the method we developed. Section IV recaps experiments we made so far with this approach, section V and section VI draw up conclusion and potential future works.

II. Related works and motivation

A. Image retrieval framework

Classical visual only approaches for localization deal with the same outline than the *image retrieval* framework: extraction (or *sampling*) of the features of interest in an image, choosing the most discriminant data, which should be invariant to changes (illumination for example), and condensing it for fast comparison. A recent survey on vision-based mapping and localization methods [4] divides approaches according to four image retrieval categories as well, making use of global descriptors, local features, "bag-of-words schemes" or combined approaches.

Most of the recent image retrieval advances are bound to *mid-level features* techniques as a unified overview that make use of local image descriptors like image patches or feature keypoints (SIFT or SURF for example [5]). For the most part, mid-level techniques algorithms place emphasis on grabbing significant and distinctive pieces from the huge amount of information contained in images: for instance, a common approach as in [6] consists in building up a *codebook* of the most relevant vis-terms included in a corpus. Later, methods evaluate images by *quantifying* visual words with this codebook [7].

Mid-level techniques are globally efficient, but their complexity can be a burden for fast computation needs. Several

ISBN 978-0-9928626-7-1

recently emerging methods rely straight on raw data from the camera. Visual sensors are particularly faced to high dynamic appearance changes in outdoor. This changes are inherent to the illumination of the scene (sun visibility, shadows, *etc*). That is why [8] works place emphasis on a transform on raw images called illumination invariant transform.

Another source of problems raises from the diversity of sensors which lead to various images of the real scene. Different visual sensors have been used for the SLAM task in the literature. For instance, [9] developed a method close to monoSLAM but with an infrared monocular camera. Some works deal with multimodality thanks to *visual servoing* like [10]. Some others rather start from standard descriptors used in computer vision and made them more robust to multimodal matching after further modifications [11].

B. Probabilistic filtering

For a few years, localization and SLAM became more and more efficient and robust thanks to probabilistic approaches applied to the estimated state of the system. [3] is an example involving LIDAR sensors and SLAM techniques using probabilistic approaches. We concentrate our work with Probabilistic filtering, namely particles filtering implemented from general Bayes filter formalism as described in [12].

III. Proposed method

Globally, our method can be separated into two main steps as summed up in fig. 1: firstly, we compose the visual memory. Secondly, we localize on-line with another camera as the only input.

A. The memory: Visual map creation

A visual map is created from data provided by a first run with our instrumented vehicle. It has a differential GPS and a roof-mounted video camera looking in front of the car. The GPS receiver allows us to precisely associate each image with the vehicle position at the same moment. We call this video sequence the *memory*. This memory could be compared to a metrical map with distinctive positions (or *places*) where we have a recorded view. We extract from each frame of the memory an *image signature*, that is to say a distinctive feature of the whole image. The way we extract images signatures for memory and online sequence is exactly the same. We will further compare signatures from the memory with live signatures.

B. Images signatures computation

1) How to slice images: We chose to describe and compare images with global descriptors aggregating local patches descriptions. We downsample images and then divide them according to a regular grid (*grid sampling*). Resulting patches size is around thirty pixels for most of the papers. If we associate different kind of sensors, we need to be sure that data in patches care approximately the same information from the physical world. Contrary to methods from the state of the art which define an arbitrary grid, we propose here to use a grid linked to the geometry of the optic.

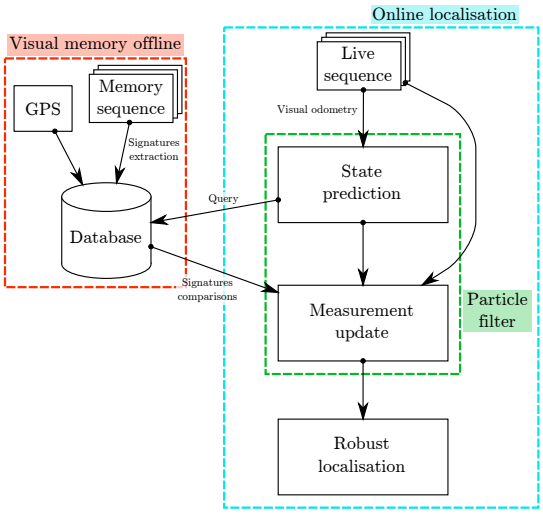


Figure 1: Global diagram of the method

We consider the well-known pinhole model [13] with f_x and f_z focal length in terms of pixels for x and y axis, (u, v) coordinates of the principal point in pixels, $X = [x, y, z, 1]^T$ homogeneous coordinates of a 3D-point of the environment relative to the camera and $\tilde{x}$ its projection to image coordinates. We define then a sphere centered on the optical center of the pinhole model as schematized in the fig. 2.

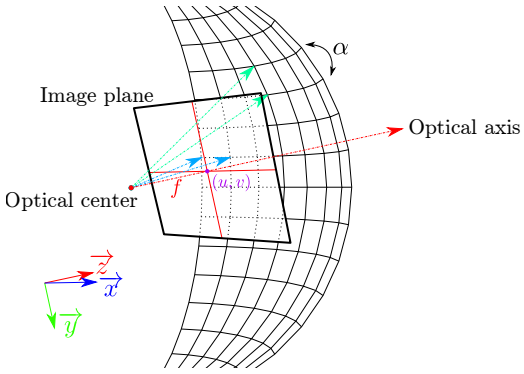


Figure 2: Image slicing with geometric constraints

We name α the *opening angle*.

Projections of directions spaced out by α on $\overrightarrow{x}$ and $\overrightarrow{y}$ axis give us the patches bound coordinates according to the following formula (we note w_s the width of the image sensor in pixels, h_s its height):

$$\tilde{x}_{m,n} = \begin{pmatrix} f_x & 0 & u \\ 0 & f_y & v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -\sin(m\alpha) \\ \sin(n\alpha) \\ (\cos(m\alpha) + \cos(n\alpha)) \end{pmatrix} \quad (1)$$

$$\text{with all } m \in \left[\!\!\left[\frac{\arctan\left(\frac{u-w_s}{f_x}\right)}{\alpha}; \frac{\arctan\left(\frac{u}{f_x}\right)}{\alpha} \right]\!\!\right]$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

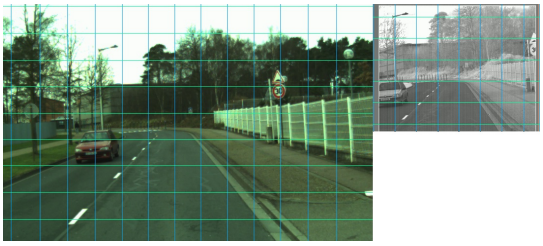


Figure 3: Slicing applied to different sensors

$$\text{and all } n \in \left[\!\!\left[\frac{\arctan\left(\frac{-v}{f_y}\right)}{\alpha}; \frac{\arctan\left(\frac{h_s - v}{f_y}\right)}{\alpha} \right]\!\!\right]$$

$\tilde{x}_{m,n}$ values are rounded to get image patches with plain pixels. This way, we get several image patches side by side. Dimensions of the patches bounds should be fixed carefully: large size would indeed weaken consequences of little viewpoint changes (as explained in [14]), that is why we determined the best angle α on our dataset in a range from $0°$ to $10°$ (see section IV-A). It is clear that patches dimensions, once projected on sensor plane, are bigger at the periphery of the image than at its centre (an example is given in fig. 3 with $\alpha = 1°$). As shown in fig. 3, depending on characteristics of optics, we don't have necessarily the same number of subdivisions in our images.

2) Modified histogram of gradient: From each sliced tile, we compute a eight bins *Histogram Of Gradient* (HOG) [15]. For that, we use the gradient routine implemented in OpenCV library. We compute angles of the gradients for each pixel and aggregate them in each tile according to eight directions as usual HOG features (seen as 8-dimensions vectors). The main reason for using HOG descriptor is its quite robust invariance to modality change [11].

We add a process step inspired by [11] and represented in fig. 4: empirical analysis easily suggest that some objects or material appears mainly black in visible spectra whereas they are bright

in infrared spectra and *vice versa*. As a consequence, gradient orientations are sometimes inverted across different spectral ranges. To make Histograms of Oriented Gradient invariant to gradient way, we divide 8-bins histograms by two and sum up them together (fig. 4).

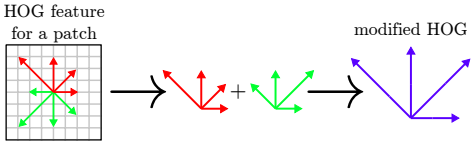


Figure 4: Traditional HOG feature & way-invariant HOG

3) HOG features pooling: Resulting 4-bins histograms are normalized in accordance with the number of pixels included in the subregion and then stored together in a 3-dimensional array to obtain an image signature.

C. Live sequence

We refer to *live sequence* or *on-line* experience a new run. This journey may happen several days or months later so that environment has severe changes. The only one input considered is image stream from another camera: optics and sensor size may change, even spectral range. We compute images signatures with the same opening angle parameter in our grid sampling method and according to new sensor calibration characteristics.

D. Comparing images signatures

In order to compare two different images, we compute the previously described signature for each one and use cosine similarity to compute a matching score. Cosine similarity is a common similarity measurement used in information retrieval, particularly in text mining [16]. Cosine similarity has been used by [14] for visual place recognition and [17] for multimodal stereo correspondence. Given two image signatures σ_a and σ_b, computing matching score is given by the following formula:

$$\text{score}(\sigma_a, \sigma_b) = \cos(\theta) = \frac{\sigma_a \cdot \sigma_b}{\|\sigma_a\| \|\sigma_b\|} \tag{2}$$

The cosine similarity by definition is always defined in the range $[0, 1]$. In the case we have image signatures of different sizes, we try all the possibilities on the bigger one from left to right and up to down in the field of view and keep the best score. Considering σ_a and σ_b, with respectively (m, n) and (k, l) sizes, $k \leq m$ and $l \leq n$:

E. Temporal consistency

As explained previously, we consider that following frames cannot represent places far away from each other: image sequences have a temporal consistency. This hypothesis allows us to consider the system as a *Hidden Markov Model* and to adapt a particle filter to our method. We implemented a particle filter as described in [12]. State transitions are estimated thanks to a traditional visual odometry method and measurement update is made with signature comparison.

1) State space: The particle filter evaluates position of the vehicle: its coordinates on a 2D-map and its heading. An additional parameter represents the scale factor of the motion computed by the odometry algorithm [18].

2) State transition estimation: We use 600 particles for the application. Each particle represents a weighted possible state of the vehicle. States are updated with the visual odometry computation. Gaussian noise with $0.5°$ standard deviation is applied on heading values. As odometry measure returns a unit vector at each step,

another Gaussian noise with 5 meters standard deviation is applied on scale value to encompass variations of car's speed.

IV. Experimental results

We divided our experimentation following two axis: first we searched for an optimal opening angle α value. Secondly, we made experiments to check if our feature is distinctive enough on real data and sufficiently invariant to spectral range changes.

ISBN 978-0-9928626-7-1 © EURASIP 2017

A. *Optimal opening angle for patches slicing*

We acquired a first dataset in order to check if the purposed global feature is itself sufficiently discriminant and well-conditioned at the same time to associate data from camera radically different. This dataset condenses a journey across both urban areas and highway and has been done with three cameras: two identical visible cameras with a 30 cm baseline and a third, between the two previous ones, a SWIR (Short Wavelength Infrared) camera. A sample of visible and SWIR images has been given in fig. 3. Each sequence is composed of 200 images.

We synchronized our three cameras with a trigger. We then took several video sequences and compute similarity matrix between sequences from both visible cameras. This experiment permits to verify if a small variation of the point of view infers on the similarity measure. Computation on the first record is displayed in fig. 5. Higher scores remain on similarity matrix diagonal, chosen feature is discriminant enough and ensures very few false matching. Precision-Recall curves for visible to visible matching according to α value are given in fig. 6.

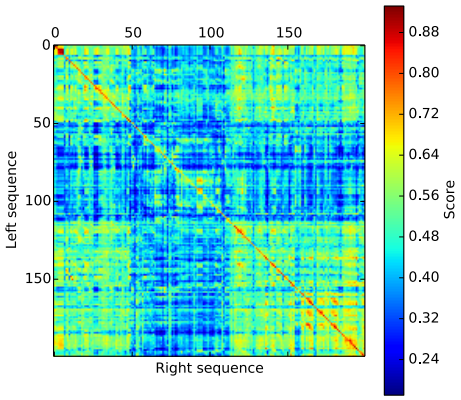


Figure 5: Similarity matrix of two synchronized visible sequences

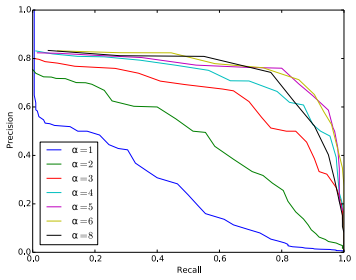


Figure 6: Precision-Recall curve for visible to visible association

We applied the same checking to a pair of visible and SWIR cameras (fig. 7 and fig. 8). This task is much more harder as expected but similarity measure seems nevertheless a good assumption prior. Subsequently, we try on our datasets several values for tiling the images. Tested values go from 1° to 10° for the opening angle. We added in fig. 8 a comparison with a Bag-of-Words retrieval method using SIFT feature and a codebook of 1000 visual words.

A slicing defined by a 2° opening angle revealed to be the best compromise for both visible/visible and visible/SWIR association. With such parameters, the computation of association between both sequences lasts 1 minutes and 30 seconds on a desktop computer equipped with an *Intel core i5* processor and $8Gio$ of ram.

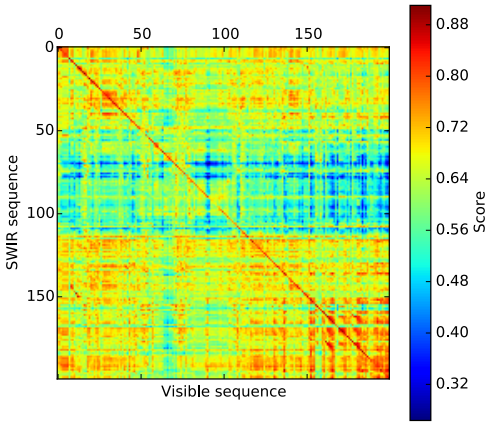


Figure 7: Similarity matrix of visible and SWIR sequences

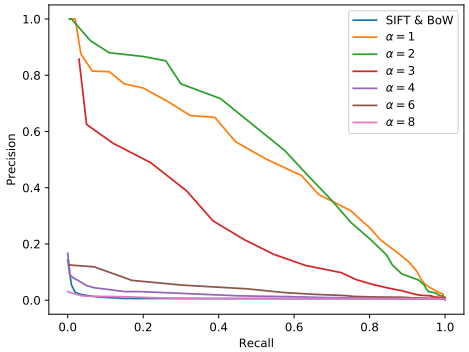
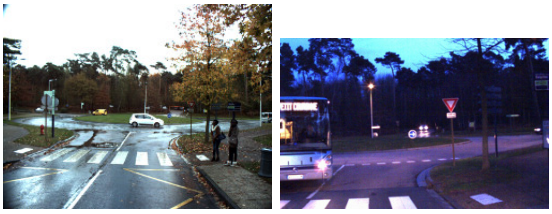


Figure 8: Precision-Recall curve for SWIR to visible association

B. *Tests on our dataset with visual odometry*

In this part, we apply our whole method with particle filter on our own dataset. First run with GPS registration has been done with a 1624×1234 pixels size camera (fig. 9b). Second run has been made on evening several weeks later with another camera (752×480 pixels) (fig. 9a). On the fig. 10, we give positions of the images in memory, estimated position of the

ISBN 978-0-9928626-7-1 © EURASIP 2017



(a) Memory sequence (b) Live sequence

Figure 9: Live and memory sequences samples

vehicle by averaging particles position and an example of a particles cloud computed during a step of our algorithm.

We notice that the live estimated positions based only on vision sensors are generally close to the ground truth given by the GPS data of the map. Sometimes, some successive faulty odometry estimations can make the estimated path diverging from the ground truth (like the bottom right path in fig. 10) but are compensated several steps further thanks to a coherent image retrieval matching.

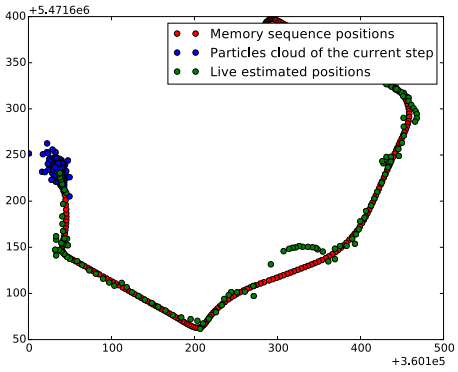


Figure 10: Visual localization test with particle filter

V. Conclusion

We developed in this paper a global descriptor for visual localization. This approach use geometric parameters given by usual calibration matrix in order to compare data provided by different cameras (optics, sensor size). We then compare live signatures and signatures in memory, balancing the scores thanks to a probabilistic filter.

The specificity of our work relies on its multi-sensors approach. Our main contribution consists of using two different cameras for mapping task on one hand and localization task on the other hand. We moreover use cameras having far different spectral range sensitivity: visible and SWIR spectra. Such technical choices aim at bringing interoperability between highly different sensors with a view to future mass market systems sharing the same visual map.

VI. Future works

We hope to develop further this approach by testing other and more complex mid-level encoding techniques as a first step, as well as improving our probabilistic model for the filtering step. Our choice will probably focus on particle filters more efficient on scale factor computation. Another avenue would be to define and generate a more evolved database structure, to implement graph models modeling a more realistic neighborhood of a place with several other nearest places. Furthermore, we would like to tackle deeply the issue concerning high perceptual changes.

References

[1] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt *et al.*, "Towards fully autonomous driving: Systems and algorithms," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 163–168.

[2] R. W. Wolcott and R. M. Eustice, "Visual localization within lidar maps for automated urban driving," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 176–183.

[3] H. Strasdat, J. Montiel, and A. J. Davison, "Real-time monocular slam: Why filter?" in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2657–2664.

[4] E. Garcia-Fidalgo and A. Ortiz, "Vision-based topological mapping and localization methods: A survey," *Robotics and Autonomous Systems*, vol. 64, pp. 1–20, 2015.

[5] P. Koniusz, F. Yan, and K. Mikolajczyk, "Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 479–492, 2013.

[6] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 696–709.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[8] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," in *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China*, 2014.

[9] F. Abrate, B. Bona, and M. Indri, "Experimental ekf-based slam for mini-rovers with ir sensors only." in *EMCR*, 2007.

[10] G. Caron, A. Dame, and E. Marchand, "Direct model based visual tracking and pose estimation using mutual information," *Image and Vision Computing*, vol. 32, no. 1, pp. 54–63, jan 2014. [Online]. Available: http://hal.inria.fr/hal-00879104

[11] D. Firmenichy, M. Brown, and S. Susstrunk, "Multispectral interest points for rgb-nir image registration," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 181–184.

[12] S. Thrun, W. Burgard, D. Fox *et al.*, *Probabilistic robotics*. MIT press Cambridge, 2005, vol. 1.

[13] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[14] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," 2014.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[16] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[17] T. Mouats and N. Aouf, "Multimodal stereo correspondence based on phase congruency and edge histogram descriptor," in *Information Fusion (FUSION), 2013 16th International Conference on*. IEEE, 2013, pp. 1981–1987.

[18] D. Nistér, "An efficient solution to the five-point relative pose problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 6, pp. 756–770, 2004.

ISBN 978-0-9928626-7-1 © EURASIP 2017