University of Groningen

Computer vision techniques for calibration, localization and recognition

Lopez Antequera, Manuel

DOI:
[10.33612/diss.112968625](https://doi.org/10.33612/diss.112968625)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Lopez Antequera, M. (2020). *Computer vision techniques for calibration, localization and recognition*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.
<https://doi.org/10.33612/diss.112968625>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

rijksuniversiteit
 groningen

UNIVERSIDAD
 DE MÁLAGA

UNIVERSITY OF GRONINGEN

BERNOULLI INSTITUTE FOR MATHEMATICS, COMPUTER SCIENCE
 AND ARTIFICIAL INTELLIGENCE

UNIVERSITY OF MÁLAGA

DEPARTMENT OF SYSTEMS ENGINEERING AND AUTOMATION

COMPUTER VISION TECHNIQUES FOR CALIBRATION, LOCALIZATION AND RECOGNITION

A dissertation supervised by promotors

PROF. DR. SC. TECHN. NICOLAI PETKOV

PROF. DR. JAVIER GONZÁLEZ JIMÉNEZ

and submitted by

MANUEL LÓPEZ ANTEQUERA

in fulfillment of the requirements for the Degree of

PHILOSOPHIÆDOCTOR (PH.D.)

Sept 2019

ISBN: 978-94-034-2323-4 (ISBN ebook: 978-94-034-2322-7)

university of
 groningen

UNIVERSIDAD
 DE MÁLAGA

Computer Vision Techniques for Calibration, Localization and Recognition

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.
and

to obtain the degree of PhD of the
University of Málaga
on the authority of the
Rector J.A. Narváez Bueno
and in accordance with
the decision by Doctoral Academic Committee.

This thesis will be defended in public on
Friday 7 February 2020 at 12.45 hours

by

Manuel López Antequera

born on 30 March 1988
in Caracas, Venezuela

Supervisors

Prof. J. González Jiménez

Prof. N. Petkov

Assessment committee

Prof. E. Alba

Prof. F. Torres

Prof. M. Biehl

Prof. X. Jiang

This research has been conducted at the Intelligent Systems group of Johann Bernoulli Institute for Mathematics and Computer Science of University of Groningen, the MAPIR research group of the University of Málaga and Mapillary.

This research has been supported by the University of Groningen through an “Ubbo Emmius” scholarship for international sandwich PhD programs, the Spanish Government (DPI2014-55826-R), the European Horizon H2020 program (projects MOVECARE and TrimBot2020), and Mapillary.

**rijksuniversiteit
 groningen**

UNIVERSIDAD
DE MÁLAGA

Computer Vision Techniques for Calibration, Localization and Recognition

Manuel López Antequera

ISBN: 978-94-034-2323-4 (printed version)

ISBN: 978-94-034-2322-7 (electronic version)

To my mother

Abstract

In this thesis we explore several practical applications of computer vision, with the use of learning based techniques, in particular convolutional neural networks (CNNs), as a common thread.

We begin by exploring the task of **single image camera calibration**. That is, the prediction of both intrinsic (focal length and radial distortion) and extrinsic (rotation with respect to the gravity vector) parameters from single images. We advance beyond the state of the art by proposing a novel parameterization for the camera model that facilitates the learning task. Additionally, we introduce a reprojection-based loss function to combine heterogeneous loss components into a single metric. Our solution is more robust than approaches that solve the problem by relying on geometric primitives such as vanishing points, as the learning-based solution can harness subtle but important cues available in the images.

Later on we tackle the problems of visual place recognition and visual localization in three independent studies. **Visual place recognition** is the task of automatically recognizing a previously visited location through its appearance, and plays a key role in mobile robotics and autonomous driving applications. Correctly recognizing a location even when its visual appearance has changed (for example, due to weather conditions) is a very challenging problem. We propose a learning-based solution where we train a convolutional neural network to produce image-level representations that are invariant to conditions such as lighting and weather. In order for the network to learn the desired invariances, we train it with triplets of images selected from datasets containing images from the same locations presenting challenging variability in appearance.

Visual localization is the task of recovering the pose (position and orientation) of a camera using only the appearance of the images captured by the camera and a map consisting of known image and pose pairs. In this work we refer to visual

localization when more than one image is used to perform localization once the system is deployed. The technique can complement or replace GPS in situations where it is not precise or robust enough, such as indoors. We propose a system that performs visual localization using only image-level representations computed from a sequence of images captured by a moving camera. Our approach does not rely on patch-level (local) features. Unlike contemporary approaches, we do not restrict the problem to that of sequence-to-sequence or sequence-to-graph localization. Instead, the sequence is localized in a database consisting of images taken at known locations, but with no explicit spatial structure. We build upon the Gaussian Process Particle Filter framework, proposing two improvements that enable localization when using databases covering large areas as well as robustifying the behavior when dealing with particle deprivation or incorrect initialization of the filter.

Finally, we develop two novel general-purpose modules for convolutional neural architectures. First we propose the **CNN-COSFIRE** module for the task of image recognition. CNN-COSFIRE adapts and extends the COSFIRE framework for its inclusion in convolutional neural network architectures. It explicitly models the relative in-plane arrangement of convolutional neural network responses, and can be used in detection or classification tasks. We validate our proposal on several challenging place and object recognition datasets. In the final chapter of this thesis we introduce a drop-in replacement for convolutional layers in CNN architectures to increase their robustness to several types of noise perturbations of the input images. We call this a **push-pull layer** and compute its response as the combination of two half-wave rectified convolutions, with kernels of opposite polarity. It is based on a biological phenomenon known as push-pull inhibition. The proposed layer is composed of a pair of push and pull convolutions that implement a non-linear model of inhibition as exhibited by some neurons in the visual system of the brain. The layer's parameters can be trained by gradient backpropagation, similarly to those of convolutional layers.

Samenvatting

In dit proefschrift onderzoeken we verscheidene praktische beeldherkenningsapplicaties door middel van machine leertechnieken, en convolutionele neurale netwerken (CNN) in het bijzonder.

We beginnen bij het onderzoeken van **enkel beeld camerakalibratie** (single image camera calibration), oftewel de voorspelling van zowel intrinsieke parameters (brandpuntafstand en radiale vertekening) als extrinsieke parameters (oriëntatie ten opzichte van de zwaartekracht vector) op basis van individuele beelden. We streven de huidige stand van techniek voorbij door een nieuwe leertaakfaciliterende parametrisatie voor het camera model voor te stellen. Daarnaast introduceren we een op reprojectie gebaseerde verliesfunctie om heterogene verliescomponenten te combineren in één metriek. Onze oplossing is robuuster dan oplossingen gebaseerd op geometrische primitieven zoals verdwijnpunten, omdat de op machine learning gebaseerde oplossing subtiele, belangrijke aanwijzingen in de afbeeldingen kan bundelen.

Later pakken we de problemen omtrent visuele plaatsherkenning (visual place recognition) en visuele lokalisatie (visual localization) aan in drie onafhankelijke studies. **Visuele plaatsherkenning** betreft de automatische herkenning van een eerder bezochte plaats middels de visuele kenmerken van die plaats en deze taak speelt een sleutelrol in mobiele robotica en zelfbesturingsapplicaties. Het correct herkennen van een locatie, zelfs wanneer de visuele kenmerken ervan zijn veranderd door bijvoorbeeld weersomstandigheden, is een zeer uitdagende opgave. We leggen een op machine learning gebaseerde oplossing voor waarin we een convolutioneel neurale netwerk trainen om representaties op beeldniveau te presenteren die invariant zijn voor omstandigheden zoals licht en weer. Om het netwerk de gewenste invarianties aan te leren, trainen we het met drietallen van beelden. De drietallen worden geselecteerd uit datasets die beelden van dezelfde locaties met

lastige variabiliteit in beeldkenmerken bevatten.

Visuele lokalisatie betreft het hervinden van de pose (positie en oriëntatie) van een camera middels de kenmerken van de beelden die het vastlegt en een kaart bestaande uit bekende beeld-pose sets. In dit proefschrift refereren we aan visuele lokalisatie als er meer dan één beeld wordt gebruikt om lokalisatie uit te voeren wanneer het systeem in werking is gezet. De techniek kan GPS complementeren of vervangen in situaties waar GPS niet precies of robuust genoeg is, bijvoorbeeld binnen. We stellen een systeem voor dat visuele lokalisatie uitvoert enkel op basis van representaties op beeldniveau, welke berekend zijn uit een reeks beelden die door een bewegende camera zijn vastgelegd. Onze benadering steunt niet op (lokale) kenmerken op patch-niveau. In tegenstelling tot hedendaagse benaderingen, beperken wij het probleem niet tot reeks-tot-reeks of reeks-tot-graaf lokalisatie. In plaats daarvan wordt de reeks gelokaliseerd in een database bestaande uit beelden waarvan de opnamelocatie bekend is, hoewel de locaties geen expliciete ruimtelijke structuur hebben. We bouwen voort op het Gaussian Process Particle Filter-kader en stellen twee verbeteringen voor die het mogelijk maken om zowel lokalisatie met databases van grote oppervlakten uit te voeren, als ook de prestatie bij deeltjes deprivatie of incorrecte filterinitialisatie te verbeteren.

Tenslotte ontwikkelen we twee nieuwe, algemene modules voor convolutionele netwerkachitecturen. Ten eerste stellen we de **CNN-COSFIRE** module voor beeldherkenning voor. CNN-COSFIRE past het COSFIRE-kader aan en breidt het uit voor inclusie in convolutionele neurale netwerkachitecturen. Het modelleert expliciet de relatieve tweedimensionale opstelling van convolutionele neurale netwerkreacties en kan gebruikt worden in detectie- of classificatietaken. We valideren ons voorstel middels verscheidene datasets voor plaats- en objectherkenning. In het laatste hoofdstuk van dit proefschrift introduceren we een drop-in vervanging voor convolutionele lagen in CNN-architecturen om hun robuustheid tegen verschillende soorten ruis in de inputbeelden te vergroten. We noemen dit een **'push-pull layer'** en berekenen de respons ervan als de combinatie van twee ReLu-geactiveerde convoluties, met kernen van tegengestelde polariteit. Het is gebaseerd op een biologisch fenomeen: 'push-pull' inhibitie. De voorgestelde laag bestaat uit een set van push en pull convoluties die een non-lineair model van inhibities implementeren; een proces dat ook vertoond wordt door sommige neuronen in het visuele systeem van het brein. De parameters van de laag kunnen getraind worden door terugpropagatie, vergelijkbaar met die van convolutionele lagen.

Resumen

En esta tesis exploramos varias aplicaciones prácticas de la visión por computador, con un hilo común: el uso de técnicas basadas en aprendizaje, en particular las redes neuronales convolucionales –*Convolutional Neural Networks (CNN)*–.

Comenzamos explorando la tarea de **calibración de cámara con una única imagen** –*single-image camera calibration*–, que consiste en la predicción de los parámetros de calibración de una cámara a partir de una única imagen: Tanto los intrínsecos, que modelan la proyección de la luz sobre el sensor de la cámara como los extrínsecos, que describen la posición y orientación de la cámara con respecto a un eje de coordenadas del entorno. Avanzamos el estado del arte proponiendo una nueva parametrización del modelo de proyección que facilita la tarea de aprendizaje. Proponemos además una nueva función de coste basada en la reproyección de puntos para reducir la función de coste a un único término, solventando la problemática del balanceo de sus componentes y simplificando la dinámica del entrenamiento. Nuestra solución es más robusta que los métodos basados en primitivas geométricas como los puntos de fuga y las líneas, ya que al tratarse de un método basado en aprendizaje puede aprovechar sutiles pero importantes elementos visuales que son difíciles de modelar explícitamente.

A continuación, nos enfrentamos a los problemas de reconocimiento visual de lugares –*Visual place recognition*– y de localización visual –*Visual localization*– en tres estudios diferenciados. El **reconocimiento visual de lugares** consiste en reconocer de forma automática un lugar previamente visitado, utilizando únicamente la apariencia visual, a pesar de posibles cambios en la apariencia de las imágenes (ya sea por cambios de iluminación, el clima o la estación del año). Juega un papel fundamental en la robótica móvil y en aplicaciones de conducción autónoma. Proponemos la utilización de un algoritmo basado en aprendizaje: Entrenamos una red neuronal convolucional para producir una representación de imágenes compacta y *holística*

(representando la totalidad de la imagen, en lugar puntos característicos). El algoritmo se entrena con juegos de imágenes obtenidas con apariencias diferentes (en distintas épocas del año, con distintos niveles de iluminación, etc), con el objetivo de obtener representaciones invariantes a dichos cambios de apariencia.

La **localización visual** consiste en recuperar la pose (posición y orientación en el espacio) de una cámara a partir de las imágenes capturadas por la misma, dada una base de datos (mapa) de imágenes previamente capturadas en el mismo entorno con poses conocidas. En este trabajo nos referimos a localización visual cuando se utiliza más de una imagen para obtener la posición de la cámara (por ejemplo, una secuencia). La localización visual puede sustituir o complementar a los sistemas de posicionamiento global cuando estos no son suficientemente precisos o robustos (por ejemplo, en interiores). Proponemos un sistema que utiliza como entrada representaciones holísticas (un vector por imagen) de una secuencia de imágenes obtenidas por una cámara en movimiento para obtener la pose de la misma. Al contrario que otras técnicas contemporáneas, no nos limitamos al problema de localización entre dos secuencias o al problema de localización en un grafo: Nuestro mapa consiste en una colección desordenada de pares imagen-pose sin estructura explícita. Para ello utilizamos un filtro de partículas con un modelo de observación basado en procesos Gaussianos.

Finalmente, desarrollamos dos módulos de propósito general para arquitecturas de redes neuronales convolucionales. En primer lugar proponemos **CNN-COSFIRE**, un módulo para la tarea de clasificación y detección de objetos. CNN-COSFIRE extiende y adapta el método COSFIRE para ser incluido en arquitecturas basadas en redes neuronales. Modela de forma explícita las relaciones geométricas de las activaciones de la red neuronal en el plano de la imagen y puede ser utilizado tanto para detección como para clasificación.

En el último capítulo de la tesis introducimos un módulo bio-inspirado que puede utilizarse en arquitecturas de redes neuronales obteniendo mejoras en robustez con respecto al ruido en las imágenes de entrada. Su funcionamiento está inspirado en un fenómeno biológico conocido como inhibición **push-pull**, donde neuronas espacialmente adyacentes modulan y compensan sus activaciones recíprocamente. Los parámetros del módulo se pueden entrenar junto con el resto de la arquitectura, de forma que se puede sustituir cualquier capa convolucional por el módulo propuesto con facilidad. Validamos de forma exhaustiva el módulo, demostrando su efectividad en la clasificación de imágenes perturbadas por distintos modelos de ruido con un incremento en el coste computacional despreciable al sustituir las capas convolucionales tradicionales por el módulo propuesto.

Contents

Acknowledgements	3
1 Introduction	5
1.1 Thesis Organization	6
2 Single-image camera calibration	9
2.1 Introduction	9
2.2 Related Work	11
2.3 Method	12
2.3.1 Camera Model	12
2.3.2 Parameterization	13
2.3.3 Bearing Loss	18
2.3.4 Dataset	20
2.4 Experiments	21
2.4.1 Evaluation of the Loss Functions	22
2.4.2 Effect of Distortion Parameterization	23
2.4.3 Error Distributions	23
2.4.4 Comparison with geometric-based undistortion	25
2.4.5 Qualitative results	26
2.5 Conclusions	26
3 Trainable image descriptors for place recognition	33
3.1 Introduction	33
3.2 Related Work	35
3.3 Methodology	37
3.3.1 Architecture of the CNN	37

Contents

3.3.2	Triplet similarity embedding	39
3.3.3	Triplet selection	40
3.3.4	Training	42
3.4	Experimental Evaluation	42
3.4.1	Results	44
3.4.2	Computational performance	47
3.5	Conclusions	49
4	Visual localization using Gaussian Processes	51
4.1	Introduction	51
4.2	Related work	54
4.3	Gaussian processes for modelling visual observations	55
4.3.1	Gaussian Processes	55
4.3.2	Using poses as the input variables in a GP	56
4.4	Observation model for particle filter localization	58
4.4.1	Observation model with GPs	59
4.5	Experiments	60
4.5.1	Descriptor selection	62
4.5.2	Comparison with a laser-based observation model	64
4.6	Conclusions and future work	65
5	City-scale continuous visual localization	69
5.1	Introduction	69
5.2	Related work	71
5.3	Gaussian Process Particle Filters	73
5.3.1	Fast GP regression	74
5.3.2	Appearance-based particle sampling	76
5.4	Experimental evaluation	77
5.5	Conclusions	82
6	CNN-based COSFIRE filters	83
6.1	Introduction and related work	83
6.2	Method	86
6.2.1	Overview	86
6.2.2	Convolutional Neural Networks (CNNs)	87
6.2.3	CNN-based contributing filters	88
6.2.4	Combining the contributing filter responses	89
6.2.5	Classification using CNN-COSFIRE filters	91
6.3	Results	91
6.3.1	MNIST	92

6.3.2	Butterfly data set	94
6.3.3	Garden place recognition data set	96
6.4	Discussion	98
6.5	Conclusions	101
7	Push-Pull networks	103
7.1	Introduction	103
7.2	Related works	105
7.3	Method - CNN augmentation with a push-pull layer	108
7.3.1	Implementation	108
7.3.2	Use of the push-pull layer	111
7.4	Experiments and results	111
7.4.1	LeNet on MNIST	112
7.4.2	ResNet and DenseNet on CIFAR	116
7.4.3	Sensitivity to push-pull parameters	123
7.4.4	AlexNet on CIFAR-C: baseline results	124
7.5	Discussion	125
7.5.1	Brain-inspired design	126
7.6	Conclusions	127
8	Summary and Outlook	129
	Bibliography	131
	Research Activities	143

Acknowledgements

When one looks back at the sequence of events that lead to any given day, it is easy to start believing in something like fate. Any step could have been different, but life-changing events play out just the way they do, for reasons sometimes purely related to chance. Retroactively, it is easy for me to point out at some people that were fundamental for everything being the way it is today:

I'd like to thank my uncle Enrique for teaching me my first words – "*Pink Floyd*" – and for designing the cover for this thesis.

To my mother: Thanks for your immense dedication in order to provide for us, for your constant encouragement that has given me the confidence to always believe in myself and for supporting me during the times when I was abroad, never letting me know that you missed me and instead pushing me to continue developing.

To my brother, Jose: thanks for taking care of dad when I wasn't there.

To my childhood friend Adrián Ruiz Sánchez, who already at a young age was a very dedicated student and taught me how to have responsible attitude towards studying. My first year of university would have been very different without those long hours at the library. To Carlos Sánchez Garrido, for being an excellent study partner through most of my university life before the PhD.

To professors Francisco Sánchez Pacheco and Pedro Sotorrío Ruiz for noticing me during my early years in the University of Málaga, and allowing me to participate in internships and research opportunities that developed my problem solving skills and practical experience way faster than studying for exams could ever do.

To Professor Fernando de la Torre for the opportunity of spending a year at his lab at Carnegie Mellon University in Pittsburgh, where my interests expanded from electrical engineering into the fields of robotics and computer vision.

To my doctoral supervisors, Javier González Jiménez and Nicolai Petkov, for entrusting me with the position as a sandwich PhD student in two excellent labs at the universities of Málaga and Groningen. Thanks for giving me the trust and freedom to explore research topics that were novel to both labs.

Thanks to my colleagues at the MAPIR lab in Málaga, *Jesús, Javi, Raúl, Carlos, Andy, Curro, Rubén, Mariano* for sharing their passion about our work and for mak-

ing so many days at the lab a bliss thanks to a healthy dose of humor. To my friends and colleagues from inside and around the Intelligent Systems lab at the University of Groningen. *Nicola, Ugo, Laura, Estefanía, Astone, Jiapan, Daniel and Renata, Andreas, George*: thanks for the barbecues, the dinner parties and the roadtrips. The rain and cold was much easier to deal with with such a warm group of friends, now scattered all over the world. In particular, I'd like to thank *Rubén, Jesús* and *Nicola* for the intense research discussions and direct collaborations. Thanks to the master students that worked with me: *Leonardo* and *Alberto* at MAPIR, and *Roger* at Mapillary.

To my colleagues at Mapillary, and particularly to *Pau* and *Yubin*: Thanks for trusting in a PhD student with a modest CV to join your team, and thanks for the incredible level of support, autonomy, encouragement and trust that I get daily.

Kitty, thanks for your selfless companionship as I focused on my PhD during our first period in Spain, for helping me develop in areas that I wasn't paying attention to, and for showing me my own home through your eyes. Also, thanks for the translation of the abstract to Dutch. I look forward to the rest of our story.

Manuel López Antequera
December 18, 2019