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

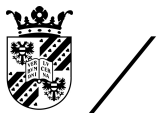
ISBN: 978-94-034-2323-4 (ISBN ebook: 978-94-034-2322-7)

UNIVERSIDAD DE MÁLAGA

AUTOR: Manuel López Antequera

http://orcid.org/0000-0002-9930-8106

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



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

Mapillary

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 neuraal 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 netwerkarchitecturen. 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 netwerkarchitecturen. 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

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 they 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
January 17, 2021

Chapter 1

Introduction

Robotics and artificial intelligence are, at the time of publication of this thesis, very popular topics. If developed to to their ultimate potential, they may ultimately free us from performing tasks that are dangerous, soul-crushing or simply uninspiring. However, these technologies could be disruptive enough to threaten critical aspects of our economy as it raises some interesting questions: Who controls the means of production when all that is needed to produce are the means themselves? What will we occupy our time with in a world where no one needs to work and abundance is simply present? These questions reveal the relevance of the field, but are not for this author to answer. There is a lot of work to be done in order to actually enable these technologies to be disruptive. In particular, this thesis deals with the development of computer vision techniques applied to different tasks that are relevant for robotics and other fields.

Computer vision is a subset of artificial intelligence that computes or extracts information from images. The field is currently in a period of fast exponential expansion, as evidenced by the number of participants and publications in top conferences and the funding volume being invested in computer vision projects and companies through private and public sources over the last five years. This expansion is largely due to the fact that computer vision techniques have recently matured enough to become useful in many commercial applications.

This maturity is partly due to the development of geometric computer vision during the last decade, enabling applications such as 3D reconstruction, camera pose estimation and basic augmented and virtual reality. However, the recent explosion of the field is mostly due to the advent of convolutional neural networks, a technique that has brought problems like general object detection, face recognition and human pose detection to commercially viable performance levels. The combination of these two branches of computer vision where an analytical understanding of multiple view geometry is combined with convolutional neural networks is currently a very relevant theme in the research community. This combination of geometry and learning is present in most chapters of this thesis, as we deal with the problems of single image camera calibration, place recognition and visual localization.

1.1 Thesis Organization

This thesis is organized in two blocks: The first block, covering chapters 2 to 5, deals with a series of works related to single-image camera calibration and visual localization. In the second block (chapters 6 and 7) we introduce two general-purpose, biologically-inspired modules for convolutional neural networks.

Chapter 2 deals with the problem of camera calibration, which is the first step in many computer vision applications, particularly those dealing with three-dimensional geometry. Target-based calibration is a well-understood problem. It is performed by capturing sets of images of a calibration target from different angles and optimizing the parameters of a camera model so that the observations fit the calibration target's known geometry. We deal instead with the problem of single-image calibration, that is, the prediction of camera parameters based on a single image. This is an ill-posed or even unsolvable problem in most cases from a purely geometric viewpoint. However, a semantic interpretation of the information in the image opens up the possibility of performing robust single-image calibration, as the real world dimensions and orientations of many objects are tightly coupled to their semantic class. For example, man-made structures are dominated by lines parallel and perpendicular to the gravity vector, the sky is up and the ground is down, trees mostly grow vertically, and so on. These relationships are difficult to include in a hand-crafted system, but can be exploited by learning-based methods when trained to perform this task. We discuss the training of a convolutional neural network to effectively and efficiently perform single-image camera calibration in chapter 2

Chapter 3 describes a learning-based solution for the problem of visual place recognition. Visual place recognition deals with the classification of image pairs as being taken at the same location or not: Given two images, the system must produce a positive label if they are taken from a similar viewpoint, regardless other factors that might change the actual pixels in the image, such as illumination or seasonal changes. Within the context of robotics, it is of critical importance as part of SLAM (Simultaneous Localization And Mapping) systems, as it allows a robot to successfully detect previously visited locations. This in turn enables the correction of errors on the internal map maintained by the robot as it navigates a new environment.

A general approach to visual place recognition is to process each input image into a image-wide representation, also known as whole-image or holistic descriptor, that is compact and robust to perturbations, such that the result of comparing these representations is not affected by changes in imaging conditions. These representations are then stored in place of the images and used as a database of previously

visited locations. The method described in chapter 3 is a learning-based approach for the generation of such representations.

Chapter 4: Localization and mapping systems in robotics are usually built on top of local keypoint and descriptor matching: small image regions are tracked over multiple frames and form the basis for the generation of a three-dimensional representation of the world where the tracked image regions correspond to 3D points known as landmarks. These systems are precise, but their robustness is limited by the matching local descriptors, as they are not robust to changes in appearance (due to illumination, point of view or seasonal changes). Chapter 4 deals with a localization system that foregoes any use of local descriptors. Instead, whole-image (holistic) representations such as those developed in chapter 3 are used as part of an observation model for a particle filter. The resulting localization system achieves robust localization without the use of local keypoints and descriptors.

Chapter 5 builds upon the system developed in chapter 4. In it, two modifications to the framework are proposed. First, an approximated method for the observation model enables the system to perform on large on large scale scenarios, such as a large area in the city of Málaga spanning 8 km^2 and 172.000 images. Also, an appearance-based resampling method for the particle filter allows the system to recover from degenerate situations (such as when the system is first started and the location of the camera is completely unknown, or when the filter converges to a wrong location)

Chapter 6 describes a general-purpose module for image classification and place recognition. It is an extension of the COSFIRE method by **?**, a brain-inspired computer vision technique that uses the relative arrangement of local patterns in an image to perform detection and classification. The COSFIRE method generally makes use of traditional non-learned image filters as the basis for detection of local patterns to be arranged. We extend the method to be able to work with learnable filters instead, such as those computed internally by convolutional neural networks.

Chapter 7 develops a new module for convolutional neural networks to improve performance when there is noise present in the input images. It is inspired by inhibition mechanisms in the human visual system that enable correct processing of images when contaminated by noise. The standard way for dealing with noisy images in convolutional neural network pipelines is to augment the training set with artificially generated noisy versions of the images. Instead, our new module encodes, by design, prior knowledge about noise suppression mechanisms that have

been proven to be useful in non-learned image processing techniques. Our module is a so-called push-pull layer, as it models the inhibition mechanism with the same name. Using the push-pull layer in CNN architectures achieves better performance on standard classification tasks when dealing with noisy images, with no decrease in performance on the original noise-free images. The use of our module does not increase the number of learnable parameters and comes with a negligible increase in computation.

Published as:

Manuel Lopez-Antequera, Roger Marí Molas, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, Gloria Haro, "*Deep Single Image Camera Calibration with Radial Distortion,*" The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, June 2019, 10.1109/CVPR.2019.01209

Chapter 2
Single-image camera calibration

Abstract

Single image calibration is the problem of predicting the camera parameters from one image. This problem is of importance when dealing with images collected in uncontrolled conditions by non-calibrated cameras, such as crowd-sourced applications. In this work we propose a method to predict extrinsic (tilt and roll) and intrinsic (focal length and radial distortion) parameters from a single image. We propose a parameterization for radial distortion that is better suited for learning than directly predicting the distortion parameters. Moreover, predicting additional heterogeneous variables exacerbates the problem of loss balancing. We propose a new loss function based on point projections to avoid having to balance heterogeneous loss terms. Our method is, to our knowledge, the first to jointly estimate the tilt, roll, focal length, and radial distortion parameters from a single image. We thoroughly analyze the performance of the proposed method and the impact of the improvements and compare with previous approaches for single image radial distortion correction.

Preprint:

Ruben Gomez-Ojeda, Manuel Lopez-Antequera, Nicolai Petkov, Javier Gonzalez-Jimenez, *"Training a Convolutional Neural Network for Appearance-Invariant Place Recognition,"*, 27 May 2015, arXiv: 1505.07428

Published as:

Manuel Lopez-Antequera, Ruben Gomez-Ojeda, Nicolai Petkov, Javier Gonzalez-Jimenez, *"Appearance-invariant place recognition by discriminatively training a convolutional neural network,"* Pattern Recognition Letters, Volume 92, 1 June 2017, Pages 89-95, ISSN 0167-8655, 10.1016/j.patrec.2017.04.017

Chapter 3

Trainable image descriptors for place recognition

Abstract

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. The difficulty of recognizing a revisited location increases with appearance variations caused by weather, illumination or point of view changes. In this paper we present a convolutional neural network (CNN) embedding to perform place recognition, even under severe appearance changes. The network maps images to a low dimensional space where images from nearby locations map to points close to each other, despite differences in visual appearance caused by the aforementioned phenomena. In order for the network to learn the desired invariances, we train it with triplets of images selected from datasets which present a challenging variability in visual appearance. Our proposal is validated through extensive experimentation that reveals better performance than state-of-the-art methods. Importantly, though the training phase is computationally demanding, its online application is very efficient.

Published as:

Manuel Lopez-Antequera, Nicolai Petkov, Javier Gonzalez-Jimenez, *"Image-based localization using Gaussian processes,"* International Conference on Indoor Positioning and Indoor Navigation (IPIN), *best paper award*, 4-7 October 2016, ISSN 2471-917X, 10.1109/IPIN.2016.7743697

Chapter 4
Visual localization using Gaussian Processes

Abstract

Visual localization is the process of finding the location of a camera from the appearance of the images it captures. In this work, we propose an observation model that allows the use of images for particle filter localization. To achieve this, we exploit the capabilities of Gaussian Processes to calculate the likelihood of the observation for any given pose, in contrast to methods which restrict the camera to a graph or a set of discrete poses.

We evaluate this framework using different visual features as input and test its performance against laser-based localization in an indoor dataset, showing that our method requires smaller particle filter sizes while having better initialization performance.

Published as:
Manuel Lopez-Antequera, Nicolai Petkov, Javier Gonzalez-Jimenez, *"City-scale continuous visual localization,"* European Conference on Mobile Robots (ECMR), 6-8 September 2017, 10.1109/ECMR.2017.8098692

Chapter 5
City-scale continuous visual localization

Abstract

Visual or image-based self-localization refers to the recovery of a camera's position and orientation in the world based on the images it records. In this paper, we deal with the problem of self-localization using a sequence of images. This application is of interest in settings where GPS-based systems are unavailable or imprecise, such as indoors or in dense cities.

Unlike typical approaches, we do not restrict the problem to that of sequence-to-sequence or sequence-to-graph localization. Instead, the image sequences are localized in an image database consisting on images taken at known locations, but with no explicit ordering. We build upon the Gaussian Process Particle Filter framework, proposing two improvements that enable localization when using databases covering large areas: 1) an approximation to Gaussian Process regression is applied, allowing execution on large databases. 2) we introduce appearance-based particle sampling as a way to combat particle deprivation and bad initialization of the particle filter. Extensive experimental validation is performed using two new datasets which are made available as part of this publication.

Published as:
Manuel Lopez-Antequera, María Leyva-Vallina, Nicola Strisciuglio, Nicolai Petkov,
"Place and Object Recognition by CNN-based COSFIRE filters," IEEE Access, Volume 7, 22 May 2019, Pages 66157-66166, ISSN 2169-3536, 10.1109/ACCESS.2019.2918267

Chapter 6
CNN-based COSFIRE filters

Abstract

COSFIRE filters are effective means for detecting and localizing visual patterns. In contrast to a Convolutional Neural Network (CNN), such a filter can be configured by presenting a single training example and it can be applied on images of any size. The main limitation of COSFIRE filters so far was the use of only Gabor and DoGs contributing filters for the configuration of a COSFIRE filter.

In this paper we propose to use a much broader class of contributing filters, namely filters defined by intermediate CNN representations. We apply our proposed method on the MNIST data set, on the butterfly data set, and on a garden data set for place recognition, obtaining accuracies of 99.49%, 96.57%, *and* 89.84%, *respectively.*

Our method outperforms a CNN-baseline method in which the full CNN representation at a certain layer is used as input to a SVM classifier. It also outperforms traditional non-CNN methods for the studied applications. In the case of place recognition our method outperforms NetVLAD when only one reference image is used per scene and the two methods perform similarly when many reference images are used.

Published as:
Nicola Strisciuglio, Manuel Lopez-Antequera, Nicolai Petkov,
"Enhanced Robustness of Convolutional Networks with a Push-Pull Inhibition Layer," Neural Computing & Applications, Volume 32, 5 February 2020, Pages 17957–17971, 10.1007/s00521-020-04751-8

Chapter 7
Push-Pull networks

Abstract

Convolutional Neural Networks (CNNs) lack robustness to test image corruptions that are not seen during training. In this paper, we propose a new layer for CNNs that increases their robustness to several types of corruptions of the input images. We call it a 'push-pull' layer and compute its response as the combination of two half-wave rectified convolutions, with kernels of different size and opposite polarity. Its implementation is based on a biologically-motivated model of certain neurons in the visual system that exhibit response suppression, known as push-pull inhibition. We validate our method by replacing the first convolutional layer of the LeNet, ResNet and DenseNet architectures with our push-pull layer. We train the networks on original training images from the MNIST and CIFAR data sets, and test them on images with several corruptions, of different types and severities, that are unseen by the training process. We experiment with various configurations of the ResNet and DenseNet models on a benchmark test set with typical image corruptions contructed on the CIFAR test images. We demonstrate that our push-pull layer contributes to a considerable improvement in robustness of classification of corrupted images, while maintaining state-of-the-art performance on the original image classification task. We released the code and trained models at the url `http://github.com/nicstrisc/Push-Pull-CNN-layer`.

Chapter 8
Summary and Outlook

In this thesis we proposed advances in computer vision for several applications, as well as some general-purpose methods. Chapters 2 to 5 detail the development of solutions for applications related to camera calibration and visual localization. Chapters 6 and 7 introduce two general-purpose, biologically-inspired modules for convolutional neural networks.

In chapter 2 we dealt with the problem of **camera calibration and orientation estimation**, developing a method to predict intrinsic (focal length and radial distortion) and extrinsic (tilt and roll angles) camera parameters using a single image. Although this is considered an ill-posed problem from a geometric point of view when only a single image is available, we noticed that it is not the case when semantics are involved and proposed to use a learning-based approach. Our method is not a replacement for intrinsic camera calibration in laboratory conditions, but produces useful results in applications where the camera capture is not controlled, such as crowd-sourced scenarios. The work described in chapter 2 involves training a convolutional neural network using a fully supervised scheme where panoramas are cropped to simulate images taken with cameras of arbitrary orientation, focal length and radial distortion. This line of work is progressing further at Mapillary, where we are exploring ways to train the network without direct supervision, possibly enabling training with arbitrary non-annotated images of the desired domain.

In chapter 3 we explore the problem of **visual place recognition**, that is, the task of finding the location of a query image given a database of images with known locations. The problem is similar to that of content-based image retrieval or image-based search. At the time of publication of the related research paper, bag-of-words models were the state-of-the-art solution for this problem. We developed a learning-based approach using convolutional neural networks trained on datasets of images taken at known locations with challenging illumination and weather conditions in order to produce a per-image feature vectors. The resulting descriptors are compact and enable efficient image-based querying that is robust to weather and illumination changes. Since the publication of this work, the state of the art in trainable descriptors for place recognition has advanced. At the time of publication of this thesis, the best performing methods (**?**) integrate translation-invariant aggregation

of features (much like the state of the art before the advent of convolutional neural networks) in the network architecture itself.

A **localization** system based on such features was developed in chapter 4. We use these descriptors in a Gaussian Process Particle Filter framework in order to accumulate evidence over time as the camera moves in the environment, enabling localization in cases where single-shot systems would fail. As our framework encodes each image in a single low-dimensional feature vector, this solution is compact, efficient and scalable. We successfully validated our method on an indoor localization presenting hard cases such as lack of textured surfaces and repetitive environments. We continued this line of work in chapter 5, adding two modifications to the framework that enable the system to perform on large on very large scale scenarios, such as an area in the city of Málaga spanning 8 km^2 and 172.000 images.

The precision achieved by the framework described in chapters 4 and 5 is limited, as images are described by a single descriptor and fine-grained geometric positioning is infeasible without point-based correspondences. This work could be extended by utilizing the intermediate activations of the convolutional neural network that extracts the descriptor as a local features. Work along these lines is currently being proposed at localization workshops in computer vision conferences.[1]

The last chapters of the thesis dealt with general-purpose modules for convolutional neural networks. In chapter 6 we developed **CNN-COSFIRE**, an extension of the COSFIRE method by **?**. COSFIRE traditionally uses non-learned image filters as the basis for detection of local patterns to be arranged. We extended the method to be able to work with learnable filters instead, such as those computed internally by convolutional neural networks. We validated the method in classification and place recognition tasks.

Finally, in chapter 7 we developed the **push-pull** layer, a new module for convolutional neural networks to improve performance when there is noise present in the input images. It was inspired by inhibition mechanisms in the human visual system. The module encodes, by design, prior knowledge about noise suppression mechanisms that have been proven to be useful in non-learned image processing techniques. We validated this module on standard classification tasks where the images are contaminated with noise, achieving better performance in these cases with no decrease in performance on the original noise-free images. The module is a drop-in replacement for the convolution layer used as a basic building block in all convolutional neural networks, facilitating its inclusion in existing architectures.

[1] Workshop on Long-Term Visual Localization under Changing Conditions, CVPR, 2019

Research Activities

Published work

- Manuel López-Antequera, Ruben Gómez-Ojeda, Nicolai Petkov and Javier González-Jiménez, *"Appearance-invariant place recognition by discriminatively training a convolutional neural network,"* Pattern Recognition Letters, Volume 92, 1 June 2017, Pages 89-95, ISSN 0167-8655, 10.1016/j.patrec.2017.04.017
  Original draft available as: *"Training a Convolutional Neural Network for Appearance-Invariant Place Recognition"*, arXiv, 1505.07428, 2015.

- Manuel López-Antequera, Nicolai Petkov, and Javier González-Jiménez, *"Image-based localization using Gaussian processes,"* International Conference on Indoor Positioning and Indoor Navigation (IPIN), 4-7 October 2016, Madrid, 10.1109/IPIN.2016.7743697

- Manuel López-Antequera, Nicolai Petkov, and Javier González-Jiménez, *"City-scale continuous visual localization,"* European Conference on Mobile Robots (ECMR), 6-8 September 2017, Paris, 10.1109/ECMR.2017.8098692

- Manuel López-Antequera, Javier González-Jiménez and Nicolai Petkov, *"Evaluation of Whole-Image Descriptors for Metric Localization,"* in International Conference on Computer Aided Systems Theory (EUROCAST), 2017, Las Palmas de Gran Canaria, 10.1007/978-3-319-74727-9_33

- Mariano Jaimez, Javier G. Monroy, Manuel López-Antequera, and Javier González-Jiménez, *"Robust Planar Odometry Based on Symmetric Range Flow and Multiscan Alignment,"* IEEE Transactions on Robotics, vol. 34, no. 6, pp. 1623–1635, 2018. 10.1109/TRO.2018.2861911

- Manuel López-Antequera, Roger Marí Molas, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, Gloria Haro, *" Deep Single Image Camera Calibration*

with Radial Distortion" The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, June 2019

- María Leyva-Vallina, Nicola Strisciuglio, Manuel López-Antequera, Michael Blaich, Radim Tylecek, Nicolai Petkov, *"Tb-Places: A Data Set for Benchmarking Place Recognition in Garden Environments"* IEEE Access, Volume 7, 24 April 2019, Pages 52277-52287, ISSN 2169-3536, 10.1109/ACCESS.2019.2910150

- Manuel López-Antequera, María Leyva-Vallina, Nicola Strisciuglio, Nicolai Petkov,
"Place and Object Recognition by CNN-based COSFIRE filters," IEEE Access, Volume 7, 22 May 2019, Pages 66157-66166, ISSN 2169-3536, 10.1109/ACCESS.2019.2918267

- Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, Peter Kontschieder, *"Disentangling Monocular 3D Object Detection,"* The IEEE International Conference on Computer Vision (ICCV), 2019

- Nicola Strisciuglio, Manuel Lopez-Antequera, Nicolai Petkov,
"Enhanced Robustness of Convolutional Networks with a Push-Pull Inhibition Layer," Neural Computing & Applications, Volume 32, 5 February 2020, Pages 17957–17971, 10.1007/s00521-020-04751-8

Awards

- **Best Paper Award**: Manuel López-Antequera, Nicolai Petkov, and Javier González-Jiménez, *"Image-based localization using Gaussian processes,"* International Conference on Indoor Positioning and Indoor Navigation (IPIN), 4-7 October 2016, Madrid, 10.1109/IPIN.2016.7743697

Attended Conferences

- International Conference on Indoor Positioning and Indoor Navigation (IPIN), 4-7 October 2016, Alcalá de Henares, Spain.

- European Conference on Mobile Robots (ECMR), 6-8 September 2017, Paris, France.

- International Conference on Computer Aided Systems Theory (EUROCAST), 2017, Las Palmas de Gran Canaria, Spain.

- European Conference on Computer Vision (ECCV), 2018, Munich, Germany.
- Computer Vision and Pattern Recognition (CVPR), 2019, Long Beach, California.
- International Conference on Computer Vision(CVPR), 2019, Seoul, South Korea.

Other activities

- ICVSS, International Computer Vision Summer School, Ragusa, Sicily, July 2014.
- Reviewer for the IEEE International Conference on Intelligent Robots and Systems (IROS) (2016, 2017, 2018)