

# Computational commonalities of image understanding algorithms

**Author(s):**
Boix Bosch, Xavier

**Publication Date:**
2015

**Permanent Link:**

https://doi.org/10.3929/ethz-a-010470357 →

DISS. ETH NO. 22432

# *COMPUTATIONAL COMMONALITIES OF IMAGE UNDERSTANDING ALGORITHMS*

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

*XAVIER BOIX BOSCH*

*M.S. Computer Vision and Artificial Intelligence,*
*Universitat Autònoma de Barcelona*
*M.S. Telecommunications Engineer, Universitat Ramon Llull*
*M.S. Electronics Engineer, Universitat Ramon Llull*

born on *17.03.1985*

citizen of Catalonia, Spain

accepted on the recommendation of

*Prof. Dr. Luc Van Gool, examiner*
*Prof. Dr. Cristian Sminchisescu, co-examiner*

2015

# Abstract

A fundamental goal of Computer Vision is the automatic understanding of images. Semantic segmentation is a modality of image understanding, that aims at labeling each pixel of an image with a semantic class. In the literature, there is a plethora of semantic segmentation algorithms. Often these algorithms are engineered at the algorithmic and implementation levels, and they appear to be disconnected from prior art at the computational level. This has led to algorithms designed for specific applications and datasets.

Many semantic segmentation algorithms can be divided into $4$ fundamental processes: perceptual organization, object recognition, hypotheses generation, and attentional feedback. These processes highly depend on each other, and do not necessarily are used in all semantic segmentation algorithms. In this thesis, we analyze the computational commonalities of some of these processes.

In the first part of the thesis, we analyze the computational commonalities of object recognition algorithms. We first introduce a new, more principled formulation for encoding approaches based on vector quantization (VQ) into a set of templates. We take advantage of the capabilities of our formulation to design novel, more discriminative and computationally efficient (binary) descriptors. Afterwards, we analyze a method that was recently introduced, called Second-order Pooling (O2P). This method apparently does not use templates as in VQ. We show that in fact O2P automatically adapts the templates to the input features at testing phase, rather than using templates learned during the training phase. This formulation of O2P as a template-based methods allows for significant accuracy improvements of O2P in standard benchmarks of image classification.

In the second part of the thesis, we analyze the processes of attentional feedback and hypotheses generation. Most algorithms use the hypotheses generation to incorporate contextual information of the image to the results of perceptual organization and object recognition. We introduce a hypotheses gener-

ation algorithm, the harmony potential, that generalizes some of the previous algorithms, and allows incorporating consistency among object classes in the image. Then, we introduce a hypotheses generation algorithm that infers the semantic labeling in all the image by only attending to a reduced set of regions in the image. The algorithms guides the attentional feedback by selecting the regions of the image to extract new information. In doing so, it dramatically reduces the complexity of computing the image descriptors and applying classifiers in all the image regions, without a significant performance drop.

# Zusammenfassung

Ein wesentliches Ziel des Computer Vision ist die automatische Bildverstndnis. Semantische Segmentierung ist eine Modalitt der Bildverstehen, die an die Kennzeichnung jedes Pixel eines Bildes mit einer semantischen Klasse zielt. In der Literatur gibt es eine Vielzahl von semantischen Segmentierungsalgorithmen. Oft sind diese Algorithmen sind an den algorithmischen und Umsetzung Ebenen entwickelt, und sie scheinen aus dem Stand der Technik in der Rechenstufe getrennt werden. Dies hat zu Algorithmen fr spezifische Anwendungen und Datenmengen konzipiert gefhrt.

Wahrnehmungsorganisation, Objekterkennung, Hypothesen Generation und Aufmerksamkeits Feedback: Viele semantische Segmentierungsalgorithmen knnen in $4$ grundlegenden Prozesse aufgeteilt werden. Diese Prozesse stark voneinander abhngig, und nicht notwendigerweise in allen semantischen Segmentierungsalgorithmen verwendet. In dieser Arbeit analysieren wir die Rechen Gemeinsamkeiten einiger dieser Prozesse.

Im ersten Teil der Arbeit analysieren wir die Rechen Gemeinsamkeiten der Objekterkennungs-Algorithmen. Wir zunchst eine neue, grundstzliche Formulierung fr die Codierung Anstze, die auf Vektorquantisierung (VQ) in eine Reihe von Vorlagen. Wir nutzen die Fhigkeiten unserer Formulierung zu entwerfen Roman, mehr diskriminierend und rechnerisch effizient (binr) Deskriptoren. Danach analysieren wir eine Methode, die vor kurzem eingefhrt wurde, genannt zweiter Ordnung Pooling (O2P). Diese Methode offensichtlich nicht verwenden Vorlagen wie in VQ. Wir zeigen, dass in der Tat O2P passt sich automatisch die Vorlagen, um die Eingabe-Features in Testphase, und nicht mit Hilfe von Vorlagen whrend der Trainingsphase gelernt. Diese Formulierung O2P als Vorlage basierenden Verfahren ermglicht eine erhebliche Verbesserung der Genauigkeit O2P in Standard-Benchmarks der Bildklassifizierung.

Im zweiten Teil der Arbeit analysieren wir die Prozesse der Aufmerksamkeits Feedback und Hypothesen Generation. Die meisten Algorithmen die Hypo-

thesen Generation zu Kontextinformationen des Bildes, um die Ergebnisse der Wahrnehmungsorganisation und Objekterkennung zu integrieren. Wir fhren eine Hypothesen Erzeugungsalgorithmus, die Harmonie Potenzial, dass einige der frheren Algorithmen verallgemeinert und ermglicht Einbeziehung Konsistenz zwischen Objektklassen im Bild. Dann fhren wir eine Hypothese Generation Algorithmus, der die semantische Kennzeichnung schliet in all dem Bild durch die Teilnahme an nur einem reduzierten Satz von Bereichen im Bild. Die Algorithmen fhrt die beobachteten Rckkopplung durch Auswhlen der Bereiche des Bildes, um neue Informationen zu extrahieren. Damit es reduziert die Komplexitt der Berechnung der Bilddeskriptoren und Aufbringen Klassifikatoren in allen Bildbereichen ohne wesentliche Leistungsabfall.