Nuances in visual recognition

Gavves, E.

Publication date
2014

1

INTRODUCTION

"Of things said without any combination, each signifies either substance or quantity or qualification or a relative or where or when or being-in-a-position or having or doing or being-affected. To give a rough idea, examples of substance are man, horse; of quantity: four-foot, five-foot; of qualification: white, grammatical; of a relative: double, half, larger; of where: in the Lyceum, in the market-place; of when: yesterday, last-year; of being-in-a-position: is-lying, is-sitting; of having: has-shoes-on, has-armour-on; of doing: cutting, burning; of being-affected: being-cut, being-burned."

– Aristotle (384-322 BC), *Categories* [6]

Before continuing with the reading, please close the book and have a look at the front page for a moment. The cover depicts the *Penrose Triangle*. You quickly realize there is something wrong with it. It appears that the upper acme is supported by the vertical one. However, at the same time, the upper acme indirectly supports the vertical acme via the lower acme, thus leading to a paradox. This paradox of the Penrose triangle forms a showcase of the perils that lie within vision, visual understanding and computer vision; although it feels trivially simple to see, in reality it is an extraordinarily complex mechanism. For understanding what is being seen, one needs to detect the fine details that make all the difference. Hence, the fundamental question naturally arises for the Penrose triangle, and any other object, action or scene for that matter: *what are the fine details, the nuances, we need to discover to make sense of a visually complex image?* This thesis is dedicated to studying fine nuances for visual recognition as the crucial step towards visual understanding.

Understanding the important nuances in a visual composition is primarily a philosophical endeavor. Consider an animal with its view of the world. A dog can identify objects, such as a "bowl with food" or a "pool of water". After years of training some breeds even recognize more complicated objects, such as "traffic lights" or "door handles". Still, the dog cannot achieve a higher level of understanding of the visual world. For example, although the dog can see the wheel, it cannot understand why the wheel spins. In fact, a dog cannot recognize itself in front of a mirror, a capability that only a handful of animals possess.

In vision there is a layer between the low level information that resides in an image, such as the colors, the textures, or the edges, and the high level information that the image conveys. From a Kantian perspective [62], computer vision aims at combining experiences-in the form of training data- and pre-existing concepts to reconstruct the intermediate layer. The nuances, which render this combination fruitful, shall, therefore, uncover some of the mechanisms of reasoning, *e.g.* the rules that make a cat look like a cat [73], or a chair look like a chair [50].

Except for philosophical reasons, there are also practical motivations for searching for the nuances in visual understanding. Imagine you are about to cross a road. Certainly, sounds can inform you about the general direction a car approaching. Still, you look at the car before crossing the road to estimate the danger precisely. For a robot, GPS coordinates hints its whereabouts, a gyroscope hints whether it is standing or it is lying down and a compass hints in what direction the robot faces. Yet, none of these senses gives a clue for the precise movements a robot should make to grab the cup on the table. So, vision has a unique place among the senses, in that it provides this fine level of precision required for high-level cognitive tasks.

As a result, autonomous vehicles, human-friendly industrial robots, security systems, semantic search engines, human emotion analyzers, social networks, augmented reality, smartphones are products that actually use computer vision today. The better we capture the nuances present in an image, the more precise, and therefore more practical in everyday's activities, computer vision becomes.

In this thesis we study automatic recognition of nuances, without requiring any human assistance at runtime. Image search looks for images that are similar to a query picture, without necessarily requiring to understand the underlying semantics. Image classification focuses more on the underlying semantics, attempting to assign one or more labels to an image given a pre-defined set of labels.

As a starting point for looking for visual nuances, we research whether a relation exists between what humans and computers perceive as identical. This brings us to the first question discussed in the thesis.

Is the machine representation of physically identical elements constant?

Figure 1: ***Detecting the nuances*** *is important for recognizing visually similar landmarks, such as the Amsterdam Centraal from the Rijksmuseum.*

This research question is considered in Chapter 2. To answer this question we need to address two issues: first, how to represent the appearance of visual patterns and second, how to reliably identify the visual patterns that belong to physically identical elements. We resort to the constant geometry of rigid, richly-textured objects, like the Dam Square palace in Amsterdam, see Figure 1, or the Parthenon in Athens. Comparing then the visual patterns in geometrically coherent locations allows for inferring the conditions under which local appearance is constant.

Although geometry allows for a respectable decrease in the image signature, the detected nuances are not necessarily optimal for recognition, especially if some prior knowledge of the distribution of images is provided. This brings us to the second question of the thesis.

What set of nuances matters most for image search?

This research question is discussed in Chapter 3. We show that for image search directly discretizing the visual appearance into fine nuances and selecting the informative ones is better than opting for more generic, and therefore less specific nuances. We approach the search of the informative nuances from an importance sampling perspective [20]. Since explicitly visiting the vast number of all possible combinations of nuances would be a colossal effort, we phrase the optimal set of nuances as a rarity to be simulated by an appropriately learnt sampling distribution.

Even if the detected nuances are capable of describing accurately the objects of interest, we do not know what nuances would be interesting from an object classification point of view. Consequently, we pose the third question of this thesis.

What set of nuances matters most for image classification?

This research question is discussed in Chapter 4. Traditionally, image classification is built upon similarities, often expressed in terms of the kernel trick [116]. Loosely speaking, the kernel trick allows to use only the pair-wise similarities in a high-dimensional space formed by the image representations, without the need to explicitly define these representations. Starting from linear kernel representations, we use common, sparsity-driven selection methods such as regularized least squares, to discover the nuances capable of retaining the kernel comparisons close to the original ones.

In the next chapter of the thesis we elaborate further on situations, where discovering distinctive visual nuances are not just important, but vital for successful classification and visual understanding. We, therefore, phrase the forth question of the thesis.

Which visual nuances discriminate fine-grained object categories?

This research question is discussed in Chapter 5. To answer the question, we consider the difference between fine-grained and generic object categories by using birds as an example. There

Figure 2: ***Classification of fine-grained bird categories****, like the warblers in this picture, depends and does not just benefit from detecting the important nuances.*

are thousands of bird species, such as the "Bobolink", the "Vermilion Flycatcher" or the various warblers in Figure 2. Fine-grained classification rests on categories that look visually similar, as they belong to a common *super-category*. We start from the observation that fine-grained categories share similar shapes and poses, inherent to their common super-category. Inspired by cognitive psychology literature [111], we propose to use the common shape of fine-grained objects to align and classify them by precisely examining their corresponding body regions.

In order to align fine-grained objects one needs to segment them from the background. As segmentation, and locality in general, appears to be important for the accurate recognition of categories, we arrive at the final question of this thesis.

Can we decompose the interpretation of image nuances into local, reusable components, allowing for exact region-level nuance reconstruction?

This research question is discussed in Chapter 6. We start from the observation that both state-of-the-art image representations, as well as object classification machines are composed of summations of constituing terms. Exploiting the homomorphic properties of the summation operations, we derive an image representation that separates the classification scores of individual elements in the image from their normalization coefficients. Therefore, we are able to exploit heavy, state-of-the-art feature encodings and reconstruct the classification score of arbitrary regions including normalization, which is vital for successful recognition. As a result, the decomposition facilitates state-of-the-art semantic segmentation and classification as the one in Figure 3.

Figure 3: ***The decomposition of the interpretation of visual nuances*** *leads to highly accurate segmentation and classification, as in the examples above.*