Invariant color descriptors for efficient object recognition

van de Sande, K.E.A.

Citation for published version (APA):

General rights
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), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Finding the objects is a natural first step for humans in their processing of images. The position and identity of an object comes so natural to us that it is hard to imagine that the brain splits the pathways for ‘what is it’ from ‘where is it’ [54][84], though they are not completely separate [37]. The distinction between ‘what’ and ‘where’ is illustrated in Figure 1.1. We learn how to recognize objects while we are still infants even before we can speak. This is one indication why the understanding of the content of images does not have to be a rational process: there is no need to explain the understanding of an image in words if we do not use the explanation in action.

Figure 1.1: Given an image, the goal of object recognition is to determine ‘what’ objects are present and ‘where’ they are in the image.
We argue that there is even no possibility to explain all understanding of an image in words, as the informative content of images is so much bigger than what words can express. This contrasts with textual information and learning how to read: children are taught how to read in school step by step. Compared to recognizing the content of pictures, the process is much better understood: roughly it goes from recognizing letters to words to phrases. To recognize objects, a computer generates the certainty that the pixel values that make up a digital image map to ‘what it is’ (a label of the objects category) and ‘where it is’ (here defined as a bounding box around the object). The approach of a computer to do the recognition does not necessarily correspond to how a human would do it. Therefore, for this thesis we focus on object recognition by computer, ignoring the human approach.

Note that in human life ‘where’ is much less relevant than ‘what’. The ‘where’ is only needed to avoid things (whatever they are) or to grab them. For the latter it is good to know ‘what’ it is first. Knowing ‘what’ it is (a dog, a cactus or a teddy bear) is much more operationally useful. Still, the ‘where’ can be used to attach semantic meaning to an image. For example, when looking for a human hugging a dog, if human and animal are not close to each other, it is highly unlikely that the human is hugging the dog nor has the intent to do so. Thus, finding the ‘where’ besides the ‘what’ allows for a deeper understanding of a scene. Therefore, apart from the ‘what’ we also investigate the ‘where’ of objects.

When describing what something is, there is a gap between literal phrases and their meaning. If we assign this semantic meaning based on the text associated with a picture or movie, the sorting order can be surprising. For example, when sorting a movie collection textually by the word dogs, the movie Reservoir dogs will be in the top results. However, despite its title, there are no dogs in this movie. The movie Beethoven will not be in the top, as Beethoven is typically associated with music and not the movie about a St. Bernard dog. Had the semantic meaning been assigned based on ‘what’ objects are visually present, these sorting mistakes would not have been made. Here, we consider specific object categories with a visual manifestation and no ambiguity.

### 1.1 Object appearance in the world

When recognizing objects, humans are able to deliver even when there are large fluctuations between object appearances and imaging conditions. The three most important causes of fluctuations in this thesis are as illustrated in Figure 1.2:

- Illumination changes
- Viewpoint changes
- Different instances

To automatically recognize objects from real-world images, the approach used by a computer needs to be robust to these kind of changes in object appearance. This thesis makes two contributions: (1) we accurately find the ‘what’ even under illumination changes and (2) we show how to find the ‘where’ in a reasonable computation time.
1.2 ‘What’

In object classification, the visual presence of an object category is determined, i.e., we are only concerned with finding the ‘what’ and not the ‘where’. An object classification system consists of two main steps: training and classifying. To train an object classification system, supervised machine learning algorithms such as Support Vector Machines [15,121] are typically used. These algorithms learn how to separate an object category from other categories based on positive example images and negative example images. The resulting classifier can be used to determine the likelihood the object is present in a new, unseen image.

In both steps, representation plays an essential role. To represent the image, a well-known and robust method is the bag-of-words model [95]. The stages in extracting this representation are shown in Figure 1.3. The bag-of-words model computes image descriptors at specific points in the image. These points are often selected using salient point methods [75,81,108], which address viewpoint changes: they find regions in the image which can be robustly detected under translation, rotation and scaling transformations. The resulting points are translation- and rotation-invariant and are covariant with scale changes. In the bag-of-words model, the descrip-
Figure 1.3: Given an image, the bag-of-words model creates a representation of the image. Tors extracted around these points are quantized against a codebook. The occurrence frequency in the image of each element from the codebook forms the representation of the image.

Common descriptors are the SIFT descriptor [75], the SURF descriptor [6] and the HOG descriptor [24]. However, these descriptors include intensity information only and do not include color information. To increase photometric invariance and discriminative power, color descriptors have been proposed which are robust against certain photometric changes [1, 9, 10, 50, 117]. As there are many different methods to obtain color descriptors, however, it is unclear what similarities these methods have and how they are different. To arrange color invariant descriptors in the context of object classification, a taxonomy is required based on principles of photometric changes. Therefore, we pose the following question:

**Question 1.** How do viewpoint and illumination changes affect existing color descriptors and subsequently visual object classification? Which invariance properties are desirable for object classification?

The next logical step is to look for new color descriptors which possess these desirable invariance properties but have higher discriminative power:

**Question 2.** Can we design new color descriptors which effectively improve object classification?

The illumination-invariant descriptors sampled at salient points give a powerful contribution to the bag-of-words model for accurate object classification. Where accuracy increases, it comes at a price of much higher computational cost. This cost is a severe drawback of the bag-of-words model on standard computers. In recent years, tremendous new opportunities for computational speedup have arrived in both CPUs and GPUs by increasing parallelism: the number of processing units has increased, instead of the speed of the processing units. Knowing how to exploit parallelism is important to speedup the bag-of-words model:

**Question 3.** Can we exploit parallelism in CPU and GPU architectures, to handle the computational cost of the bag-of-words model?
1.3 ‘Where’

When looking for ‘where’ in the image an object is, the most successful approaches so far are based on exhaustive search over the image to find the best object positions [24,32,38,58,124,131]. However, as the total number of images and the number of potential windows therein is huge, it is necessary to constrain the computation per location and the number of locations considered. The computation is currently reduced by using a weak classifier with simple-to-compute features [24,38,58,124,131], and by reducing the number of locations on a coarse grid and with fixed window sizes [24,38,123], as illustrated in Figure 1.4. This comes at the expense of overlooking some object locations and misclassifying others. The bag-of-words model is too computationally expensive to use it in combination with an exhaustive search. A more selective search which considers fewer locations enables the use of more expensive bag-of-words features. This leads to:

**Question 4.** Can we design a selective search strategy which visits the locations where it is probable that the object is actually present? Will this improve object recognition?

1.4 Organization of the Thesis

The answer to question 1 is in Chapter 2 which has appeared as [112]. Question 2 is investigated in Chapter 3 which has been submitted as [113]. Question 3 forms the main topic of Chapter 4 which is published as [115]. Finally, in answer to question 4, we present a selective search strategy in Chapter 5 which comes from [116]. For a brief overview of the work done and achievements made in this thesis, we provide a summary and conclusions in Chapter 6.