Selecting and robustifying local image descriptors

Everts, I.

Citation for published version (APA):
Everts, I. (2014). Selecting and robustifying local image descriptors
Introduction

“Learn how to see”

Try to imagine to learn how to see, as demanded by this quote of Leonardo da Vinci. Are you seeing any pictures?

As infants, most people learn how to see so well that it comes fully natural to us for the rest of our lives. From motion detection for predator evasion, to semantic visual interpretation for recognition: the ability of sight is of vital importance. Look, it is so firmly grounded in human language that it is common to appeal to the reader’s or listener’s ability to see.

The ease and flexibility with which humans see is accomplished by the highly complex visual processing system. In rough terms, this constitutes the transduction of photons that hit the retina in the back of the eyeball into signals sent through the optic nerve to the brain. Information at any layer of abstraction is extracted by an extremely large, complex and interrelated network of neurons. The system responsible for the processing of visual input (the visual cortex) is the largest functional system in the brain, which in turn is the largest energy consumer in the human body. Without such a vast and complex processing capacity, human vision is not possible.

When one tries to imagine to learn how to see, or to convey the ability of sight to someone who has never seen, one should start by accessing photoreceptors for determining the intensity and the wavelength (i.e. color) of the light entering the eye. Subsequently these signals should be filtered to detect visual phenomena such as edges and textures. Then, segments must be grouped together when
being similar in terms of location, intensity, color, texture and shape, before fin-
ally being interpreted on a semantic level. However, the mental experiment in
which it is attempted to learn how to see reveals that it appears impossible to
imagine to actively ‘use’ our cells, nerves and neurons that enable sight. Also,
it appears difficult to communicate about the process without referring to any
kind of visual phenomenon. In the development of computer vision this has
been taken to the extreme, as machines have to be instructed exactly what to
do. The question is how to convey the ability of sight step by step.

Because of the enormous amount of image data, online and in digital archives,
security databases, in medical domains, numerous industrial settings and for
human-computer interaction applications, the usefulness of sight granted to
machines is evident. Structuring these large amounts of complex image data
is possible only with the aid of effective and efficient computer vision, which is
reflected at the time of writing by the recent installment of computer vision de-
partments at Google and Facebook Inc. Computer vision is used to perform face
detection and recognition to reveal links in social networks, to perform multi-
media search based on the media content, and for the automatic categorization
of photo collections made up of ordinary images like the ones in figure 1.1.

![Figure 1.1: Examples of ordinary images from a typical photo collection. All images may be categorized as containing ‘cars’, whereas image (A) and (B) are taken ‘outdoors’, (B) and (C) are both images of a ‘Saab’ and image (A) also contains a ‘van’ and ‘buildings’.](image)

The images in figure 1.1 reveal a few things. All images contain ‘cars’, which is easily observed despite many large variations in car manufacturer, car model, position, scale, viewpoint, illumination, color and camera hardware. Image (C) was photographed ‘indoors’, which renders illumination effects different to the ‘outdoor’ settings of (A) and (B). Image (A) contains several ‘buildings’, ‘cars’ and a ‘van’, which is a different instance in the class of ‘vehicles’ just like a ‘car’.
Also, it is worth noting that the cars in images (B) and (C) are both a ‘Saab’. The examples illustrate many aspects of the problem faced by the field of computer vision: there may be large differences between instances of the same visual category, whereas instances across categories might appear very similar.

In order to be able to explain why objects and scenes appear in images the way they do it is important to understand the process of image formation. Similar to the first steps of biological vision, images are formed by capturing the projection of the continuous 3D-world on a discrete 2D-lattice. The process involves many complex factors, which produce incidental variations in the appearance of objects and scenes. Geometric variations result from camera viewpoint, the object location, the object pose and the scene [36]: the images in figure 1.1 exhibit cars in the side-, rear- and frontal view, in close-up and far away, and from different makings. Illumination conditions will cause photometric variations [31], illustrated in figure 1.1 by the indoor and outdoor settings, highlights and shadows. The appearance of an object and scene is best understood as a specific geometric and photometric instantiation.

The information in a digital image is a list of pixels consisting of a location \((x, y)\) and color \((\text{Red, Green, Blue})\). Such a representation is affected by the variations in image formation. For computer vision, it is advised to formulate the problem in terms of invariance against factors of image formation and class variation that are disturbing the task at hand. For example, one way of correctly recognizing a ‘Saab’ in images 1.1 (B) and (C) is the detection of the Saab logo, despite viewpoint variations and the presence of strong highlights. Simultaneously, the representation should adequately capture the visual properties that are informative for the content: the highlights should somehow be ignored while preserving the structure and color of the Saab logo. The ability of doing so results in image representations exhibiting discriminative power. Computer vision has to consider trade-offs between invariance and discriminative power because informative content might be ignored due to invariant image measurement [93]. This trade-off is referred to as robustness [91].

The current mainstream approach to robust computer vision considers large amounts of measurements of local image features, known as local image descriptors, which are either learned [45] or modeled [55, 31]. Deep learning is more and more popular as a result of recent top performances in recognition benchmarks. A drawback is the large amount of parameters in the neural network, the robust estimation of which requires enormous amounts of data. Also, due to the complexity of the network, it is difficult to interpret the representations
Introduction

at the various layers and neurons.

This thesis is concerned with physical models of local image features and representations thereof [76]. Local image measurements are relatively insensitive to global geometric variations as well as to a range of photometric variations, and are used for tasks as diverse as object and scene recognition [76, 89], object tracking [81], motion estimation [53], panoramic stitching [8], depth estimation [62] and 3D reconstruction [36]. Thus, representations based on local image descriptors are suitable for modeling image content. The robustness of a computer vision system then depends on the robustness of the local descriptors themselves.

Currently, there exist two main paradigms for achieving descriptor robustness. First, invariance can be imposed by design [55, 31]. Invariant properties are derived analytically from physical models of image features conforming to a model of image formation, or the algorithm for feature extraction and aggregation may be engineered such that certain variations are ignored. Second, the building blocks of the descriptor extraction algorithm (and parameter settings thereof) may be globally optimized in order to identify the trade-off between invariance and discriminative power [7]. Both approaches result in a single descriptor exhibiting a fixed level of robustness. However, due to the enormous variety of local image content resulting from variations in image formation, there exists no single local image descriptor that is optimal under all circumstances [93, 21]. For example, textured image regions may be represented best by measurements of structure, where color could be most informative for homogeneous regions. Such a diversification of descriptors also applies to the discrimination of materials. Shiny materials may be best represented with a highlight-invariant, whereas the best representation of matte materials is in general a shading-invariant.

This thesis aims to identify the issues related to the diversification of local descriptors arising from the variations in image formation and object properties. In general, this requires an adaptation of the extraction and modeling of local descriptors based on properties of the underlying image signal itself. By doing so, a descriptor is selected dynamically. The dynamic selection of local image descriptors results in novel image representations with different robustness characteristics.
1.1 Objectives

The objectives of this thesis are illustrated by analyzing the stylized pictures in figure 1.2. The images 1.2(A) through 1.2(D) show a sequence of vinyl recorded under varying illumination conditions. Here, in addition to the ambient light, an illumination source is positioned at various locations relative to the vinyl, which renders photometric variations in the image data. The images 1.2(E) through 1.2(H) in the bottom row show a sequence of the vinyl record in rotation, recorded at 0°, 90°, 180° and 270°. These pictures exhibit one type of geometric variation.

![Figure 1.2: Illustration of the chapters in this thesis. A vinyl record is photographed under varying illumination conditions (A-D) and in different poses (E-H), producing photometric resp. geometric variations of the object. Photometric variations are most apparent on the homogeneous, shiny vinyl, whereas geometric variations are best observed on structured and distinctively colored object regions. The way in which the object appearance responds to a variation in the recording conditions is dependent on properties of the object itself. This may impose conflicting demands on the way in which the image is best represented for a computer vision task.](image)

Images 1.2(A) and 1.2(B) show that a photometric transformation has most effect on unstructured image regions such as those exhibiting the black vinyl, whereas structured regions containing e.g. the apple stem appear to not change at all. This illustrates that the degree to which a photometric change in the recording conditions affects the object appearance is in fact dependent on structure properties of the object itself. The relation is denoted as stability, which
Introduction

propagates to any attempt to represent the image data for accomplishing some visual recognition task. As a consequence, the descriptors that are used for image representation may be very unstable on regions exhibiting little or no structure and could thus negatively impact recognition performance. Therefore, it is aimed in this thesis to explicitly account for the stability of an image descriptor such that the associated performance loss is prevented.

**Objective 1.** *Formulate and exploit measurements of photometric descriptor instability in standard image representations to improve robustness of visual recognition systems* (chapter 2).

In contrast with the observed effect of varying illumination conditions in images 1.2(A)-(B), it is shown in 1.2(E)-(F) that the appearance of the same image regions behave differently under a geometric transformation of the imaged content such as rotation. Here, the visual effect of the transformation is observable on structured regions, whereas no variation is observed on the black vinyl due its lack of structure (at this scale of observation). Thus, the observed effect of the transformation on the image depends on the imaged content itself, as is also the case for photometric transformations, however reversed in relation to image structure. Knowledge of the effect of transformations is valuable as it identifies what regions remain stable. This, in turn, permits an image matching scheme to be adaptable on-the-fly such that more importance is granted to stable image regions and as a consequence matching performance might be improved.

**Objective 2.** *Formulate and exploit the effect of local geometric image transformations for improved robustness of image matching* (chapter 3).

In addition to properties related to structure, other aspects of object appearance are material and albedo. The images 1.2(A)-(D) show that vinyl is a shiny material. A colorful object region is indicated in image 1.2(G). Structure, material and color together impose complex and potentially conflicting demands on the way images are being represented. The demands move beyond globally optimal, task-specific image representations. Instead, they focus on descriptor diversification at the level of image regions. The aim is to select the most representative and discriminative descriptor for a given image region, based on structure, material and color properties of that image region.

**Objective 3.** *Select from a diversified set of image descriptors the best descriptor based on properties of the underlying image content for robust matching* (chapter 4).
Finally, it is observed from images 1.2(C)-(D) and 1.2(G)-(H) that there are multiple sources of apparent motion in an image sequence. On object regions without structure such as the black vinyl in figure 1.2(C)-(D), the only observed variations are not due to object motion but to a change in position of an illumination source. Structured and colorful regions such as the apple stem in figures 1.2(G)-(H) exhibit variations that comply with the actual object motion. Thus, for motion detection and description it is important to discriminate between deceptive motion of illumination effects and variations due to actual (object and camera) motion. Moreover, better motion estimates may result from considering color in addition to structure in the representation.

**Objective 4.** Reformulate local video descriptors in order to enable invariant color image measurement for improved robustness of visual recognition in the video domain (chapter 5).

In summary, the central question of this thesis is how to select (parameters of) local image descriptors, such that the robustness of visual recognition and matching is improved. The objectives 1-3 approach the problem by learning how to select or adapt a descriptor per image region based on visual properties of that region. Objective 4 aims to select the single best descriptor based on batch processing of video datasets.