Selecting and robustifying local image descriptors

Everts, I.

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

The individual chapters in this thesis are summarized before conclusion.

6.1 Summary

Chapter 2: The current most popular local image descriptors represent the gradient structure of the underlying image [55]. This family of descriptors is unstable on image regions with low signal strength (i.e. without structure). This instability arises from the fact that perturbations in the image signal due to photometric variations, noise and compression have a relatively large effect on regions with little structure. As a consequence, region descriptors also suffer from these instabilities. The standard approach to dealing with unstable descriptors is to ignore them all together [59, 94]. This, however, is an ad hoc treatment of which the optimal setting is data dependent and may lead to performance loss. Therefore, it is proposed in this chapter to explicitly model the instability in the descriptor coding scheme for image representation. This is accomplished by designing a measure of descriptor instability which can be interpreted as the variance of the observed descriptor and can be computed on-the-fly. The descriptor variance is incorporated in the Fisher Vector encoding framework, which changes the learning problem because the observations are now represented as high dimensional densities instead of vectors. These proposed Fisher Vectors are shown to increase recognition performance in image matching, retrieval and object recognition tasks.

Chapter 3: Local image descriptors are generally designed and optimized to
be robust against small geometric variations. To this end, it is common to simulate the sets of relevant transformations, and to analyze and exploit the resulting descriptor distributions [60, 10]. While this generally leads to a global optimum of descriptor settings, it is shown in this chapter that the quality of descriptor matching is dependent on the distributions per patch, as the specific effect of the transformations is determined by the image content itself. However, the computational complexity involved in such a simulation strategy at run time is prohibitively large. Therefore, it is proposed in this chapter to formulate the effect of a given image transformation in closed form. In essence, it is aimed to compute the covariance of an image sample using prior knowledge about the expected image transformations. A metric is derived from the thus obtained covariance and used in an image matching task, which performs comparably to the case in which the image transformations are actually repetitively applied to the image for covariance estimation.

Chapter 4: Whereas the chapters 2 and 3 focus on photometric and/or geometric disturbances, additional sources of appearance variation are due to object material and albedo. All these sources of variation are relevant to the formation of images in practice. Therefore, many image descriptors have been designed to ideally be robust against all disturbances [55, 31, 89]. Despite advances in this respect, the optimal descriptor characteristics vary per patch and as a consequence there exists no single best descriptor under all circumstances. This statement is supported by the experimental results in chapter 4. Here, it is shown in an ‘oracle’ experiment that the best image descriptor varies on a per-patch basis, and as a consequence that selecting the best descriptor per patch for matching performs better than using the globally best single descriptor for all patches. It is further proposed to perform the selection automatically by a classifier that predicts the best descriptor for a given image patch based on visual properties related to structure, material and color of the imaged content.

Chapter 5: It is well known from studies in the image domain that descriptors extracted from invariant image representations contribute substantially to the robustness of many computer vision systems. Another well known result is that the best performing descriptor varies per dataset [89], and even per image or image region, as shown in this thesis. In the video domain, however, such studies have not yet been performed. Here, local descriptors have been designed based on measurements of the luminance signal only [98]. As a consequence, representations based on these descriptors are sensitive to photometric variations and ignore potentially discriminative features from the color channels. Therefore, in chapter 5, existing approaches for (human action) recogni-
tion in video based on spatio-temporal interest points (STIPs) are reformulated in order to operate on multi-channel video representations. This essentially involves extending the existing formulations to handle vector observations instead of scalars only. It is shown that this may result in compact and powerful representations that substantially outperform their single-channel counterparts.

### 6.2 Conclusion

This thesis is based on the observation that different types of image content require different representations. Variations in image content apply globally as well as locally, ranging from image level annotations to the appearance of local structures. Because of the wide applicability of local image descriptors in computer vision, this thesis contributes by focusing on variations in local image content. The central question is whether local measurement types can be dynamically selected or adapted based on properties of the image signal such that the robustness of visual recognition methods is improved. This thesis concludes with three flavors of the central answer: concept, practice and future.

From a conceptual point of view, it is concluded from the results in this thesis that it is indeed possible to select or adapt a local descriptor based on properties of the image content itself. The thesis proposes to 1) incorporate a measure of descriptor instability in the image encoding scheme based on measurements of the underlying signal itself, 2) adapt the local image matching scheme based on the expected covariance per descriptor under a given image transformation and 3) dynamically select an image descriptor based on visual properties of the underlying image content. The implementations of these ideas are shown to improve the robustness of image matching, retrieval and recognition methods.

The practical applicability of the concepts in this thesis depends on the degree to which they are intrusive for the image representation or recognition method. In chapter 2, the measure of descriptor instability can be computed from the descriptor itself and can be modeled as the descriptor variance in the image encoding scheme. This contrasts with the proposed methods in chapters 3 and 4, in which the metric that is derived from the estimated descriptor covariance (chapter 3) resp. the selected descriptor (chapter 4) has to be applied to the to-be-matched descriptors as well. This is not viable for large scale application. Note that chapter 5 deviates from the central question in the sense that it does not propose to dynamically select or adapt a local descriptor. Instead, a set of novel video descriptors is introduced from which the best one is selected for practical usage.
Future work with respect to the specific content of this thesis should focus, where possible, on the efficiency of the proposed methods in chapters 3 and 4 in particular. The aim should be to improve the trade-off between the merits of per-patch analysis regarding matching performance and the advantage of global image analysis regarding processing efficiency.

In a broader sense, the exploitation of image descriptors that are engineered based on physical models as advocated in this thesis will remain a very important aspect of computer vision in the future. There are two main reasons for this, both of which relate to the current upcoming popularity of deep learning. Consider the case of deep learning resulting in the best performance for accomplishing some visual recognition task according to some measure. First, while the learned representation is then apparently very robust and powerful, it can not be the case that this representation subsumes the set of all engineerable descriptors. Thus, there can always be image descriptors engineered that add to the information as contained in some learned representation. Second, and more importantly, in our mission to enrich industry and society with computer vision and machine learning capability, it is of critical importance that the techniques are conveyed in ways that are easy to understand. In practice, it is seldomly the case that a company or individual has large amounts of relevant and clean data available to train models for solving exactly their well defined problem. Instead, it is essential to infiltrate the various domains in which ‘data science’ appears important to actively cooperate and discover the value of data together. In the case of visual data, this means being able to clearly express the type of image measurements that are being used, i.e. ‘shape’ and ‘color’, while ignoring ‘shadows’ and ‘highlights’. Without such communication it is not possible to do trials, make errors, try again and accomplish joint success.