Robust visual scene categorization in context
van Gemert, J.C.

Citation for published version (APA):
van Gemert, J. C. (2010). Robust visual scene categorization in context

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: http://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.
Chapter 7

Summary and Conclusions

7.1 Summary

In this thesis we explore robust and practical methods for visual scene categorization. To this end, we scrutinize and improve the bag-of-visual-words, a.k.a. codebook model. In the codebook model, image features are represented by discrete prototypes, describing an image as a histogram of prototype counts. Prototype-histograms are subsequently used by a classifier to separate images of visual scene categories. In this thesis we focus on the codebook model, and identify four core parts where robust methods may improve the practical application of the model:

1. Prototype vocabulary size and eloquence: the more compact, i.e. smaller, the vocabulary, the larger the image collections that can be indexed. Moreover, a prototype vocabulary may be tuned to the image domain at hand.

2. Image feature sampling: the contextual surroundings of an object may be more informative than the object itself.

3. Prototype to feature assignment: representing an image feature by multiple prototype candidates over merely the best prototype.

4. Classification parameter tuning: careful classification performance estimation may allow more accurate parameter tuning.

In the following paragraphs we summarize the contributions of this thesis per chapter:

Chapter 2, Episode-Constrained Cross-Validation in Video Concept Retrieval. In this Chapter we propose an episode-constrained cross-validation method for estimating scene classification performance in video. The traditional method of cross-validation is based on shots, whereas we propose a method based on episodes. Our episode-constrained method prevents the leaking of nearly identical shots to the rotating hold-out set. Consequently, episode-constrained cross-validation produces sets with an unbalanced number of relevant items. Such unbalances sets apriori have a better Average Precision (AP) score, since AP is not normalized for the number of relevant items. To remedy this bias, we introduce a new performance measure: Balanced Average Precision (BAP). We experimentally compare BAP with AP, and episode-constrained cross-validation with shot-based cross-validation for two classifiers on a large video collection. The results show that the bias of AP for unbalanced data does occur. However, in our dataset, BAP performs equal to AP because the effect does not occur frequently enough in this set. Further experimental evaluation shows that the episode-constrained method yields a more accurate estimate of the classifier performance than the shot-based method. Moreover, when cross-validation is used for parameter optimization, the episode-constrained method is better able to estimate the optimal classifier parameters, resulting in higher performance on validation data compared to the traditional shot-based cross-validation.
Summary and Conclusions

Chapter 3, Visual Scene Categorization by Learning Image Statistics in Context. We present a scene category classification method by learning the contextual occurrence of proto-concepts like sky, water, vegetation, etc., in images. We compactly represent these proto-concepts by using color invariance and natural image statistics properties. We exploit similarity responses as opposed to strict selection of a codebook vocabulary, and we have been able to generalize these proto-concepts to be applicable in general image collections. We demonstrated the applicability of our approach in a) learning 50 scene categories from a large collection of news video data; b) a collection of 101 categories of images; c) two instances of the Pascal VOC object recognition challenge and d) two large collections of photo-stock images, comprising 89 categories, where categories are learned from one and categorized from the other. An important contribution is scalability, showing that the proposed scheme is effective in capturing visual characteristics for a large class of concepts, over a wide variety of image sets. Where specific methods may have better performance for specific datasets, we have shown a method which is neither tuned nor optimized in parameters for each collection. Hence, the method has proven to robustly categorize scenes from learned context.

Chapter 4, Comparing Compact Codebooks for Visual Categorization. In this Chapter we focus on compact, and thus efficient, models for visual concept categorization. We use the codebook scene classification algorithm where model complexity is determined by the size of the vocabulary. We structurally compared four approaches that lead to compact and expressive codebooks. Specifically, we compared three methods to create a compact vocabulary: 1) global clustering, 2) concept-specific clustering and 3) a semantic vocabulary. The fourth approach increases expressive power by soft-assignment of codewords to image features. We experimentally compared these four methods on a large and standard video collection. The results show that soft-assignment improves the expressive power of the vocabulary, leading to increased categorization performance without sacrificing vocabulary compactness. Further experiments showed that a semantic vocabulary leads to compact vocabularies, while retaining reasonable categorization performance. A concept-specific vocabulary leads to reasonable compact vocabularies, while providing fair visual categorization performance. Given these results, the best method depends at the application at hand. In this Chapter we presented a guideline for selecting a method given the size of the video dataset, the desirability of manual annotation, the amount of available computing power and the desired categorization performance.

Chapter 5, Visual Word Ambiguity. With visual word ambiguity we refer to modeling soft-assignment in the codebook model. One inherent component of the codebook model is the assignment of discrete visual words to continuous image features. Despite the clear mismatch of this hard assignment with the nature of continuous features, the approach has been applied successfully for some years. In this Chapter we investigate four types of soft-assignment of visual words to image features. We demonstrate that explicitly modeling visual word assignment ambiguity improves classification performance compared to the hard-assignment of the traditional codebook model. The traditional codebook model is compared against our method for five well-known datasets: 15 natural scenes, Caltech-101, Caltech-256, and Pascal VOC 2007/2008. The results of all experiments show that soft-assignment outperforms the traditional hard assignment over all dimensions, all vocabulary sizes, and over all datasets. We demonstrate that large codebook vocabulary sizes completely deteriorate the performance of the traditional model, whereas the proposed model performs consistently. Moreover, we show that our method profits in high-dimensional feature spaces and reaps higher benefits when increasing the number of image categories.

Chapter 6, Color Invariant Object Recognition using Entropic Graphs. In this Chapter we combine an unparameterized entropy estimator with color invariant features for object recognition. We use color invariant features that keep image measurements constant under varying intensity, viewpoint and shading. For similarity matching we employ a measure based on entropic spanning graphs. Entropic graphs provide an alternative to traditional approaches of image matching such as assuming a fixed probability distribution or histogram binning. The parameters required are the number of nearest neighbors and the value for $\alpha$ in the $\alpha$-entropy. The number $k$ of the $k$-nearest neighbors is not critical, however a higher $k$ adds more robustness. The value of $\alpha$
7.2 Conclusions and Discussion

This thesis contributes to practical automatic scene classification by endowing the bag-of-visual-words model with more robust properties. The proposed properties increase the classification performance or allow indexing of large, real-world image and video collections. This work allows us to draw the following conclusions.

From chapter 2 we conclude that respecting the contextual narrative structure in video data leads to accurate estimation of classification performance. More accurate estimation, in turn, leads to better parameter selection yielding improved performance. We retain narrative structure by treating a video episode as an atomic element during cross-validation. We imagine that more advanced techniques such as automatic story-segmentation allow more fine-grained atomic story elements. Smaller story units may improve performance estimation by achieving a more diverse spread of stories over the rotating hold-out sets.

Chapters 3 and 4 allow us to conclude that scene context is capable of capturing the global essence of an image. Moreover, a vocabulary of semantic prototypes like sky, water, vegetation, etc., is suitable for many datasets. Such a semantic vocabulary is related to a recently proposed method [31] that describes an image by its attributes such as has wheel, has head, is furry, is shiny, etc. Such attributes allow a compositional approach to scene classification, where the meaning of an image is made up of the meaning of its parts. In some sense this is a Homunculus argument, where the problem of image classification is simply postponed to another level. Nevertheless, classification of semantic prototypes or image attributes is intended to be simpler and limited in options for atomic compositional elements. These elements fully determine what the model can ‘see’. What compositional elements to choose, remains an open question [82].

Our third conclusion, drawn from chapters 3, 4 and in particular chapter 5, states that in the codebook model the soft-assignment of image features to vocabulary elements is always beneficial when compared to hard-assignment. We have observed this benefit for all vocabulary types: semantic, concept-specific, and generic, for all vocabulary sizes: ranging from extremely small to extremely large, for many datasets: Scene-15, Caltech-101, Caltech-256, the Mediamill challenge and Trecvid collections, and for several image features: SIFT, Wiccest and Gabor features. Soft-assignment reflects ambiguity in the visual word vocabulary. We model this ambiguity with a global similarity function fitting to the image features at hand. However, more advanced methods may readily be applied. We can imagine a classifier’s posterior probability, or a learned distance metric that may change depending on its position in feature space.

On the matter of visual vocabulary compactness and large scale image indexing as studied in chapter 4, we conclude that generally there is a compactness vs. performance tradeoff. This balance may be tipped somewhat towards better performance by using soft-assignment and a visual word vocabulary that is tuned to the problem at hand. Nevertheless, higher performance comes at a price of less compact models. The choice of how much performance is ‘good enough’ depends on the application at hand, or alternatively, determined by the size of the dataset and the choice of available hardware.

Our final conclusion from chapter 6 states that entropy-based similarity measures can outperform histogram-based methods. Since a histogram is also used in the codebook model, an obvious extension would be to use an entropic similarity measure instead of a visual word histogram. A somewhat similar approach based on flexible image-to-image matching has shown excellent performance [155]. Removing the histogram would eliminate the need for a visual word vocabulary altogether; and with it the need for ambiguity modeling.