Chapter 8

Conclusions

In this thesis, we proposed a novel semantic image representation based on co-occurrence of semantic concepts. The proposed modeling is quite simple, and builds upon the bag-of-features appearance representation and the availability of robust appearance classifiers. Images are represented by their posterior probabilities with respect to a set of semantic concepts. This results in mapping of the images from the space of appearance feature to that of semantic features. Denoted as the semantic space, each dimension of this space encodes an appearance-based posterior probability with respect to a semantic concept. Semantic image representation is shown to have a higher level of abstraction than bag-of-features appearance representation. Three novel visual recognition systems; for the task of image retrieval, scene classification and cross-modal multimedia retrieval, were proposed. All three recognition systems build upon the proposed semantic image representation.

First, the design of a content based retrieval system, query-by-semantic-example (QBSE), was introduced, where the retrieval operation was carried on the semantic space. QBSE system, apart from yielding state of the art retrieval performance, was instrumental in evaluating the intrinsic benefit of semantic representation for image retrieval. The results above provide *strong* support in favor of the argument, that is *semantic representations have an intrinsic benefit for image retrieval*. While this could be dismissed as a trivial conclusion, we believe that doing so would be unwise, for two main reasons. First, it had not been previously shown that query-by-text systems can generalize beyond the restricted vocabulary on which they are trained. This is certainly not the case for the current standard text based query paradigm. Second, the results above suggest some interesting hypotheses for future research, which could lead to long-term gains that are more significant than simple out-of-vocabulary generalization. For example, given that the higher abstraction of the semantic representation enables better performance than visual matching, it appears likely that semantic spaces constructed with better abstraction or that exploits the structure of natural language, can lead to better retrieval systems. The QBSE paradigm now proposed could be easily extended to the multi-resolution semantic spaces that are likely to result from a hierarchical

concept representation. Furthermore, it would allow an objective characterization of the gains achievable at the different levels of the taxonomy. We intend to explore these questions in future work.

Second, the design of a scene classification system based on semantic image representation was presented. Inspired from the recent works on scene classification, where a low-level intermediate "theme" space is introduced, a framework based on semantic space – which serves as a proxy for the intermediate space, was proposed. All classification decisions were performed on this space. An implementation of the proposed framework was presented and compared to various existing algorithms, on benchmark datasets. The results allow a number of conclusions. First, while low dimensional semantic representations are desirable for the reasons discussed in Section 4.1, previous approaches based on latent-space models have failed to match the performance of the flat bag-of-words model, which has high dimensionality. We have shown that this is indeed possible, with methods that have much lower complexity than the latent-space approaches previously proposed, but make better use of the available labeling information. Next, a study of the effect of dimensionality on the classification performance was presented, which indicates that the dimensionality would grow sub-linearly with the number of scene categories. This could be a significant advantage over the flat bag-of-words models which, although successful for the limited datasets in current use, will likely not scale well when the class vocabulary increases.

Third, the design of cross-modal multimedia retrieval system was proposed. This entails the retrieval of database entries from one content modality in response to queries from another. While the emphasis was on cross-modal retrieval of images and text, the proposed models support many other content modalities. By requiring representations that can generalize across modalities, cross-modal retrieval establishes a suitable context for the objective investigation of fundamental hypotheses in multimedia modeling. We have considered two such hypotheses, regarding the importance of low-level cross-modal correlations and semantic abstraction in multi-modal content modeling. The hypotheses were objectively tested by comparing the performance of three new approaches to cross-modal retrieval:

1) CM, based on the correlation hypothesis, 2) SM, based on the abstraction hypothesis, and 3) SCM, based on the combination of the two. All of these map objects from different native spaces (*e.g.*, text and images) to a pair of isomorphic spaces, where a natural correspondence can be established for cross-modal retrieval purposes. The retrieval performance of the three solutions was extensively tested on two datasets, "Wikipedia" and "TVGraz", containing documents that combine images and text. While the two fundamental hypotheses were shown to hold for the two datasets, where both CM and SM achieved significant improvements over chance retrieval, SM achieved overall better performance than CM. This implies stronger evidence for the abstraction than for the correlation hypothesis. The two hypotheses were also found to be complementary, with SCM achieving the best results of all methods considered.

Finally, the design of a two-layer holistic context modeling system based on the probability of co-occurrence of objects and scenes was proposed. The first layer represents the images in a semantic space, which has a higher level of abstraction, but suffers from a certain amount of contextual noise, due to the inherent ambiguity of classifying image patches. The second layer enables robust inference in the presence of this noise, by modeling the distribution of each concept in the semantic space. An image is then represented by its posterior probabilities with respect to these *contextual* distributions. This was shown to produce posterior distributions that emphasize concept co-occurrences due to true contextual relationships and inhibit accidental co-occurrences due to ambiguity. Interestingly, we found a weak correlation between the quality of the appearance classification and the corresponding quality at the contextual level. In fact, some variations of the representation with weak appearance-level performance were top-performers at the contextual level. It appears that, while supervision is critical to bridging the semantic gap during learning, soft appearance-level decisions are more effective during inference. This is an interesting finding, given the emphasis on highly accurate appearance classification in the literature. Recognition systems that operates on the clean contextual representation were shown to outperform both noisy semantic representation and the appearance representation in the tasks of scene

classification and image retrieval. In both cases, it was also shown that, despite its simplicity, the proposed contextual models are superior to various previous proposals in the literature. The gains with respect to appearance modeling were shown to hold irrespectively of the choice and accuracy of the underlying appearance models.

The overall representation is similar to a topic model, but where topics are learned in a supervised manner. Supervised learning is a necessary condition for overcoming the semantic gap between the low-level patch representation and the higher-level contextual relationships. While multiple instance learning is required to cope with the uncertainty of the appearance representation, multiple instance inference was shown ineffective. Best results are obtained with weaker, patch-based, inference that leads to an appearance representation of higher entropy. This prevents a greedy commitment to premature image explanations that, while consistent with appearance statistics, do not take context into account. The latter goal is better served by inference procedures that assign non-zero probability to multiple plausible classes, at the appearance level. We proposed topic supervised topic models that address the limitations of the existing topic models, enabling them to achieve better classification accuracies.

It should be noted that our current implementation does not incorporate spatial information of any form. Current evidence [74, 83] suggests that integration of weak spatial information, by diving an image in a 2×2 or 4×4 grid of spatial bins, can improve the accuracy of visual recognition systems. Furthermore, in this thesis the proposed semantic and contextual image representations, were tested on datasets composed of ten to a few hundred concepts. The benefits of the proposed representation in recognition tasks with much higher number of semantic concepts, remains to be tested. We intend to explore these issues as a part of future work.