Chapter 10

Conclusions

10.1 Contributions of the thesis

This thesis introduced a new decision-theoretical formulation for the visual information retrieval problem. This formulation was shown to lead to 1) new insights on the retrieval problem, and 2) new guidelines for the design of practical systems. In particular, we have shown that the decision-theoretic formulation has the following appealing properties.

- Provides a unified solution to the problems of visual recognition and learning, that is optimal in the sense of minimizing the probability of retrieval error.
- Establishes a universal probabilistic language for the retrieval problem which enables the design of systems that can seamlessly integrate information from multiple content modalities.
- Establishes objective guidelines for the design of the three main components of a visual recognition architecture: feature transformation, feature representation, and similarity function.
- Unifies a significant number of recognition approaches that have been previously proposed both in terms of image representation and similarity function.
- Enables the design of systems that learn across multiple temporal scales.

These theoretical properties were shown to be of important practical consequence. In particular, we have presented new solutions to the following challenging problems.

- Joint modeling of image color and texture.
- Precise characterization of the trade-off between feature transformation and feature representation, and guidelines for the design of each of these modules.
- Unified support for local and global queries without requiring image segmentation.
- Integration of textual and visual queries.
- Decision-theoretically optimal design of short-term learning (relevance feedback) algorithms that allow fast convergence to the desired images.
- Decision-theoretically optimal design of long-term learning (concept learning) algorithms that, over time, allow personalization of the retrieval system to the preferences of the user.

These solutions were combined into a new visual recognition architecture that was experimentally shown to 1) perform well on object, texture, and generic image databases, 2) provide a good trade-off between retrieval accuracy, invariance, and complexity, 3) lead to perceptually relevant judgments of similarity, and 4) support learning through belief propagation algorithms that, although optimal in a decision-theoretic sense, are extremely simple, intuitive, and easy to implement. This recognition architecture is the basis of the RaBI image retrieval system that was designed according to the theoretical principles laid out by the thesis.

10.2 Directions for future work

Obviously, there are several interesting questions in retrieval that we could not solve, or even address, in the course of the thesis research. Some of these questions were ignored simply because of temporal constraints. Two good examples are 1) how to extend the models now used for static images to other content types, such as video or audio, and 2) how to create indexing structures compatible with Bayesian retrieval?

We would like to emphasize that one of the added, and not thoroughly discussed in the thesis, advantages of Bayesian retrieval is exactly the fact that it provides a unified solution to these questions. On one hand, probabilistic representations from the class of mixture models are among the best known for modeling speech (where hidden Markov models are predominant [140]) and there is good reason to believe that they will be equally successful for video [69]. On the other, we have already shown that probabilistic representations are amenable to the design of hierarchical descriptors that exploit the structure of the database to efficiently build indexing structures [188]. In fact, because they maintain a complete description of the conditional density of each image class at each step of the hierarchy, we have strong reason to believe that they will outperform many of the standard indexing techniques that keep only a representative vector. Hence, while indexing and extensions to other data types remain topics for future work, we believe that they will not pose major problems to Bayesian retrieval.

A more challenging question is how to incorporate spatial relationships in Bayesian retrieval. Ideally, one would like to allow not only local queries, but also queries of the type "a region similar to x *above and to the left* of a region similar to y". Theoretically, there is no fundamental difference between these and the local queries currently supported, one simply has to rely on a more sophisticated model, capable of capturing spatial dependencies *between* regions, e.g. a Markov random field. In practice, however, it usually turns out that this is more complex than predicted by the theory since inference is much more difficult, sometimes even intractable, in such models. In spite of this difficulty and the fact that, so far, we do not have a completely satisfying solution to the problem, we are convinced that Bayesian retrieval is the best framework in which to address it. The key question is how to develop models that achieve a good trade-off between the expressive power required to account for spatial relationships and complexity. Most alternative solutions that we are aware of (e.g. [93, 164]), tend to be based on heuristics that are not always easy to justify, rely on assumptions that are usually not made explicit, and lead to representations that cannot be easily extended to deal with other components of the retrieval problem.

Establishing a language to deal with spatial relationships would bring us one step closer to the holy grail of image retrieval: the automatic extraction of semantic content descriptors. This is, without question, the main challenge for the next generation of retrieval systems.

While the retrieval by *visual similarity* presented in this thesis is sufficient in some domains, a substantial number of applications require instead *semantic* retrieval, e.g. support for queries such as "pictures of a child pointing to a bird on the sky", or the "the scene where the murder takes place".

In [186, 191], we have shown that it is possible to extend the Bayesian retrieval framework introduced in this thesis to the extraction of semantic-level descriptors, and introduced a semantic classifier based on attributes such as action, type of set (man-made vs natural), presence of close-ups (commonly associated with dialog), and crowds (scenes containing a large number of people). While this classifier demonstrates the feasibility of extracting semantic information from images and video, its practical value is somewhat limited by the fact that it requires expert knowledge about the content domain where the characterization takes place. This is an expected limitation for semantic characterization since some form of regularization will always be required to disambiguate between the multiple interpretations that a given scene may have. Better understanding of the semantic content characterization problem can only be attained though a substantial amount of research in questions such as 1) which semantic attributes can and cannot be modeled and detected and 2) how generic can semantic classifiers be?

In the absence of semantic classifiers, or as a complement to these, it is imperative that retrieval systems can learn from user-interaction and become better matched to the interests of their users as time progresses. We believe we have presented convincing evidence supporting the claim that Bayesian retrieval is a natural solution to the learning problem. However, we relied on several assumptions that may not always hold. It remains to be seen how much is lost by relying on these assumptions, what would be the complexity of learning algorithms that did not rely on them, and what other forms of learning could be implemented.