Chapter 1

Introduction

Humans have an intriguing ability to process visual information amazingly fast and with nearly perfect recognition rates. However, with the proliferation of the Internet, availability of cheap digital cameras, and the ubiquity of cell-phone cameras, the amount of accessible visual information has increased astronomically. Websites such as Flickr alone boast of over 5 billion images, not counting the hundreds of other such websites and countless other images that are not published online. With such enormous collections of available visual content, manual processing becomes prohibitive and it is therefore of great practical importance to build "visual recognition systems".

Visual recognition is a fundamental problem in computer vision. It subsumes the problems of scene classification [74, 77, 17, 114, 120], image annotation [21, 41, 72, 35, 12], image retrieval [28, 133, 119, 156], object recognition/localization [140, 128, 47], object detection [165, 122, 42] etc. In recent years the application of machine learning technologies — that allow computers to make intelligent decisions based on empirical data — for tackling visual recognition is becoming increasingly popular with advancements being made by both the research communities. While the last decade has produced significant progress towards the solution of the visual recognition problems, the basic strategy has remained the same: 1) identify a number of visual classes of interest, 2) design a set of "appearance" features that are discriminative for those classes, 3) postulate an architecture for their recognition, and 4) rely on sophisticated statistical tools to learn optimal recognizers from training data. We refer to this strategy as *appearance-based* visual recognition, because the associated recognizers rely on image representations which are either image pixels, features, or parts, derived by simple deterministic mappings of those pixels. The main innovations of the last decade have been associated with better appearance-based features e.g. the ubiquitous scale-invariant feature transform (SIFT) descriptor [85], the widespread adoption of statistical modeling e.g. generative graphical models (such as Gaussian Mixture Models (GMM) [157, 21], Latent Dirichlet Allocation [12, 77], etc.), sophisticated families of discriminants (such as support vector machines(SVMs) with various kernels tuned for vision [51, 23, 17, 177, 20] etc.), the application of

powerful machine learning techniques (such as variational learning [14], Markov chain Monte Carlo [50]) etc.) to the design of the recognizers themselves etc.

While there is no question that appearance based classifiers will retain a predominant role in the future of recognition, it is not as clear that they will be *sufficient* to solve the recognition problem. In fact, there is little evidence so far that they can solve all but a small class of problems (such as face detection) with accuracies comparable to those of biological vision. One striking property of the latter, at least in what concerns humans, is that it rarely seems to ground decisions exclusively on low-level visual features. This has been well documented in psychophysics, through unambiguous evidence that scene interpretation depends on *context* [8, 100]. By this, it is usually meant that the detection of an object of interest (e.g. a locomotive) is facilitated by the presence, in the scene, of other objects (e.g. railroad tracks or trains) which may not themselves be of interest. The presence of these *contextual cues* (e.g. that locomotives are usually on tracks and pull trains) increases the detection rate for the object of interest. This is illustrated in Figure 1.1, which shows the posterior probabilities of a locomotive image belonging to a number of visual concept classes, according to a number of appearance based visual detectors trained on those classes. The presence of an 'arch-like structure in the locomotive's rooftop makes the weight of the "bridge" concept slightly higher than that of the "locomotive" concept, for the adopted appearance based recognizer. However, by noting that the contextual cues "railroad", and "train" also have high posterior probability, a context-sensitive recognizer could still assign the image to the "locomotive" class.

Another striking property of human vision, which suggests that raw appearance is not the whole story for recognition, is unveiled by a set of relatively recent findings on the neural structure of the recognition process. In a series of now extensively replicated seminal experiments, Thorpe and collaborators [144] demonstrated an intriguing ability of humans to perform *decent* scene classification with *very small computation*. More precisely, EEG recordings have shown that humans are capable of solving visual recognition problems such as the detection of "food", "animals", and so forth, with $90 - 95\%$ accuracy in close to $150\,ms$, i.e. only

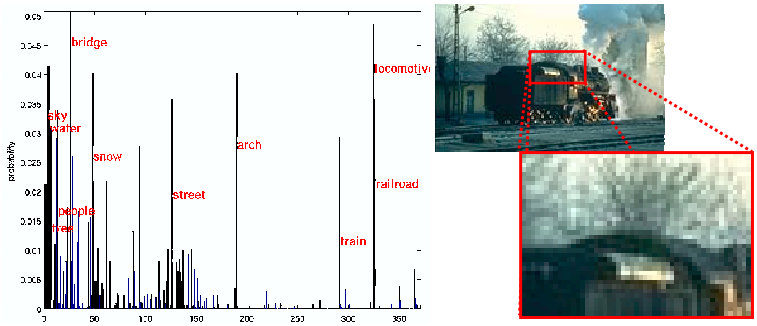

Figure 1.1: Probability of a locomotive image belonging to a number of visual concept classes according to appearance based visual classifiers. Note that, while most of the concepts of largest probability are present in the image, the SMN assigns significant probability to "bridge" and "arch". This is due to the presence of a geometric structure similar to that of "bridge" and "arch", shown on the image close-up.

enough time to propagate the visual stimulus (in a feed-forward manner) through a small number of neural layers. Given that 95% is nowhere near the recognition rates that the visual system can achieve for these classes, this raises the question of what these low-grade, but fast, classifiers could be useful for. While we do not profess to know the answer to this question, one possibility is that they could be *contextual classifiers*, whose goal is not to solve the vision problem per se, but detect the contextual cues that could make the solution easier.

In this thesis, image representation and visual recognition form the core body of work, where we address the problem of incorporating contextual cues in the image representation to tackle visual recognition problems. More precisely, the aims of this thesis are twofold. First, the design of a representation that accounts for the contextual cues present in an image. Second, the design of visual recognition systems that build upon the proposed image representation and achieve state of the art visual recognition performance.

1.1 Contributions of the thesis

This thesis provides a novel framework for visual recognition which is based on incorporation of contextual cues. To this end, first a *semantic image representation* is introduced which builds upon recent developments in visual recognition, namely the availability of robust appearance classifiers and image databases annotated with respect to a sizable concept vocabulary. This representation besides being well correlated with the human understanding of the images, is very useful in the design of visual recognition systems that yield state of the art recognition performances. Next, building upon the semantic image representation, we present three frameworks for three different visual recognition tasks, viz. image retrieval, scene classification and cross-modal multimedia retrieval. Under image retrieval, the task is to *retrieve* images from a given image repository in response to a *query* provided by the user. Under scene classification, the task is to assign one of several class labels from a given vocabulary of concepts to a user specified image. Both image retrieval and scene classification are well studied problems in computer vision [133, 28, 119, 77, 74, 105, 120]. Cross-modal multimedia retrieval on the other hand is a relatively recent problem in computer vision, where the retrieval operation is performed across different data modalities e.g. to retrieve text documents in response to an image query.

Next, we show that the semantic image representation, although effective at solving visual recognition problems, suffers from certain drawbacks, in particular the issue of *contextual noise*. In the latter part of this thesis, we introduce the framework of *holistic context modeling*, that addresses these drawbacks. Holistic context models also build upon the semantic image representation and are able to explicitly learn *true* contextual relationship between different concepts directly from the data. Holistic context models are shown to further improve the performance of visual recognition systems. Finally, a formal analysis of the holistic context models in the form of a *generative graphical model* is presented and connections to the existing work in the literature are drawn. In the remainder of this section we briefly discuss the significant contributions of this thesis.

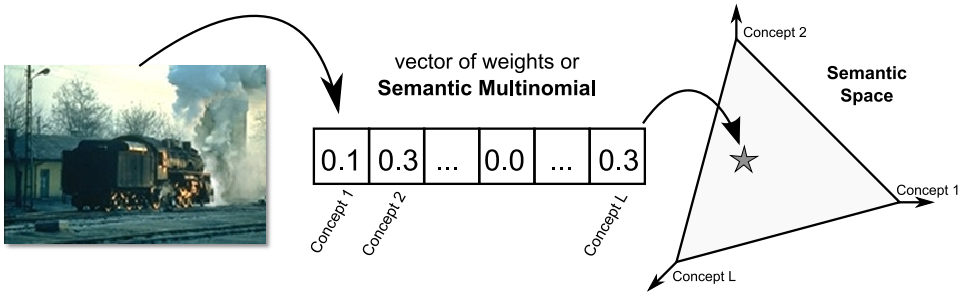

Figure 1.2: An illustration of image representation on the *semantic space*. An image is represented as a *semantic multinomial* which is a weight vector obtained using an array of appearance based classifiers.

1.1.1 Semantic Image Representation

Semantic image representation is a novel image representation, that brings a paradigm shift in the way image are represented. Under semantic image representation, instead representing the images on the space of low level appearance features derived from the image, a *semantic space* — a space where each dimension represents a meaningful visual concept — is introduced, upon which the images are represented and all recognition decisions are performed. To obtain the semantic representation of an image, first a vocabulary of visual concepts is defined and statistical models are learned for all concepts in the vocabulary with existing appearance modeling techniques [21, 74, 77]. Next, the outputs of these appearance classifiers are then interpreted as the dimensions of the semantic space. This is illustrated in 1.2, where an image is represented by the vector of its posterior probabilities under each of the appearance models. This vector is denoted as a *semantic multinomial* (SMN) distribution as the image features themselves define a multinomial distribution over the semantic concepts. An example SMN for a natural image is the probability vector shown in 1.1(left).

1.1.2 Visual Recognition Systems

A significant contribution of this thesis is the design of visual recognition systems based on the proposed semantic image representation. Below we discuss three different visual recognition systems that build upon the semantic image rep-

(a) (b)

Figure 1.3: An illustration of "semantic gap" — two images which are similar for humans as they depict the semantic concept of "beach". However they have different low-level visual properties of color, shape, etc.

resentation.

Image Retrieval: Query by Semantic Example

Current image search engines tend to rely on information extracted from image filenames or neighboring text in the webpage to retrieve the images that best satisfy a given query. This approach is fruitful only if a meticulous and complete textual description of the image is available, but this is rarely the case. It ignores the wealth of information available in the visual information stream itself, i.e. the image content. The image retrieval community studies content-based solutions to the design of retrieval systems. One popular retrieval paradigm is that of *query-by-example* — the user provides a query image, and retrieval consists of finding the closest visual match in an image collection, to this query. However, this paradigm restricts the definition of similarity to a *strict visual form*, declaring images as similar as long as they exhibit identical patterns of color, texture, shape, etc. In most cases, this narrow definition of similarity is weakly correlated with those adopted by humans for image comparison. For example, 1.3 shows two images which are similar for humans as they depict the semantic concept of "beach", however they have different low-level visual properties of color, shape, etc. This is commonly known as the "semantic gap" between *low-level processing* and the *higher level semantic abstraction* adopted by humans [133, 119].

In this thesis we propose a novel image retrieval framework, Query-by-

semantic-example (QBSE), that addresses the semantic gap. QBSE leverages on the semantic image representation by extending the query-by-example paradigm into the semantic domain, whereby the nearest neighbor retrieval operation is performed directly on the semantic space. This is shown to have two main properties of interest, one mostly practical and the other philosophical. From a practical standpoint, because QBSE has a higher level of abstraction, it enables retrieval systems with higher *generalization ability* that are more accurate than what was previously possible. Philosophically, because it allows a direct comparison of visual and semantic representations under a common query paradigm, QBSE enables the design of experiments that explicitly test the value of semantic representations for image retrieval.

Scene Classification

Scene classification is an important problem for computer vision, and has received considerable attention in the recent past. *Scene* classification differs from *object* classification, in that a scene is composed of several entities often organized in an unpredictable layout[113]. For a given scene, it is virtually impossible to define a set of properties that would be inclusive of all its possible visual manifestations. Early efforts at scene classification targeted binary problems, such as distinguishing indoor from outdoor scenes [142], city views from landscape etc. More recently, there has been an effort to solve the problem in greater generality, through design of techniques capable of classifying a relatively large number of scene categories [166, 77, 113, 74, 16, 83], and a dataset of 15 categories has been used to compare the performance of various systems[74, 83]. Several of these approaches aim to provide a compact lower dimensional representation using some intermediate characterization on a latent space, commonly known as the intermediate "theme" or "topic" representation [77]. The rationale for this strategy is that images which share frequently co-occurring visual features have similar representation in the latent space, even if they have no features in common.

In this thesis we propose an alternative solution using the semantic image representation, where the semantic space serves as the intermediary for the low di-

mensional "theme" representation. However instead of the themes being learned in an unsupervised manner, as is the case with existing approaches, they are explicitly defined. The number of semantic classes or themes used, defines the dimensionality of the intermediate semantic space. Experiments show that scene classification based on semantic image representation outperforms the unsupervised latent-space approaches, and achieves performance close to the state of the art, using a much lower dimensional image representation.

Cross Modal Multimedia Retrieval

Classical approaches to information retrieval are of a *uni-modal* nature [125, 133, 84]. Text repositories are searched with text queries, image databases with image queries, and so forth. This paradigm is of limited use in the modern information landscape, where multimedia content is ubiquitous. Recently, there has been a surge of interest in *multi-modal* modeling, representation, and retrieval [106, 148, 132, 138, 28, 60, 31]. Multi-modal retrieval relies on queries combining multiple content modalities (*e.g.* the images and sound of a music video-clip) to retrieve database entries with the same combination of modalities (*e.g.* other music video-clips). However, much of this work has focused on the straightforward extension of methods shown successful in the uni-modal scenario which limits the applicability of the resulting multimedia models and retrieval systems. For example, these systems are inadequate when the task is to query with objects that do not share the same modality as the retrieval set e.g. using images to find similar documents in a text corpus.

In this thesis, a richer interaction paradigm is considered, which is denoted *cross-modal* retrieval. The goal is to build multi-modal content models that enable interactivity with content *across* modalities. Such models can then be used to design *cross-modal retrieval systems,* where queries from one modality (*e.g.* video) can be matched to database entries from another (*e.g.*, the best accompanying audio-track). The central problem in the design of cross-modal retrieval systems is the inherent inconsistency between the representations of different modalities. To address this, a mathematical formulation is proposed, equating the design of cross-

modal retrieval systems to that of designing isomorphic feature spaces for different content modalities. Semantic image representation naturally lends itself as an effective solution to the design of the isomorphic feature spaces. By extending the semantic image representation to other modalities, all modalities are represented at a higher level of abstraction which establishes a common semantic language between them. This is referred to as the *abstraction hypothesis*. Another solution, based on maximizing correlations between different modalities, denoted as *correlation hypothesis*, is also proposed. By means of extensive experimental evaluation it is concluded that both hypotheses enable design of effective cross-modal retrieval systems and are complementary to each other, although the evidence in favor of the abstraction hypothesis is stronger than that for correlation.

1.1.3 Holistic Context Modeling

While the semantic image representation captures co-occurrences of the semantic concepts present in an image, not all these correspond to *true* contextual relationships. This is usually not due to poor statistical estimation, but due to the inherent *ambiguity* of the underlying features representation. Since appearance based features typically have small spatial support, it is frequently difficult to assign them to a single visual concept e.g. just looking at the close up of the "arch like feature" in 1.1 its is not possible to assert that this feature is from a "locomotive" image and not a "bridge". Hence, the semantic image representation extracted from an image usually assigns some probability to concepts unrelated to it e.g. "arch" and "bridge" concepts for the "locomotive" image in 1.1. We term this ambiguity as *contextual noise* i.e. casual coincidences due to the ambiguity of the underlying appearance representation (image patches that could belong to either a "locomotive" or an "arch").

Rather than attempting to eliminate contextual noise by further processing of appearance features, we propose a procedure for *robust* inference of contextual relationships *in the presence of contextual noise*. This is achieved by introducing a second level of representation, that operates on the semantic space. Each visual concept is modeled by the distribution of the posterior probabilities extracted from

all its training images. This *distribution of distributions* is referred as the *contextual model* for the concept. For large enough and diverse enough training sets, these models are dominated by the probabilities of true contextual relationships which can be found by identifying peaks of probability in semantic space. An implementation of contextual modeling is proposed, where concepts are modeled as mixtures of Gaussian distribution on appearance space, and mixtures of Dirichlet distributions on semantic space. It is shown that the contextual descriptions observed in contextual space are substantially less noisy than those characteristic of semantic space, and frequently *remarkably clean*. It is also argued that these probabilities capture the *contextual co-occurrences* of concepts and constitute the global context representation of an image. The effectiveness of the proposed approach to context modeling is further demonstrated through a comparison to existing approaches on scene classification and image retrieval, on benchmark datasets. In all cases, the proposed approach is superior to various contextual modeling procedures in the literature.

We also present a comparison of holistic context models with the existing work on "topic models", in particular latent Dirichlet allocation (LDA). It is shown that although both these models share a common generative modeling framework, one key property of the holistic context models, that enables it to achieve higher classification accuracy, is that of *supervision*. However, since the holistic context models and LDA use different image representations, its difficult to assess the true gains achieved by supervision in these model. To enable a systematic study of the benefits of supervision, we present a family of topic models, denoted as *topic-supervised* LDA, where supervision is introduced in the LDA framework. All other attributes of the LDA model are kept constant. It is shown that topic-supervised LDA models are able to outperform their unsupervised counterparts, for the task of scene classification.

1.2 Organization of the thesis

The organization of the thesis is as follows. In Chapter 2 we first review two problems of visual recognition viz. retrieval and scene classification, and two popular low-level appearance based image representations viz. discrete cosine transform and scale invariant feature transform. Next we introduce the proposed semantic image representation. In Chapter 3 we highlight the problems of existing image retrieval solutions based on appearance features and introduce query-by-semantic-example retrieval paradigm. Next, in Chapter 4 we present a scene classification system using the semantic features as an intermediate representation. The problem of cross-modal multimedia retrieval is presented in Chapter 5, where we also discuss two novel solutions for retrieval across different content modalities; first based on the semantic image representation and second based on maximizing correlation between different modalities. The issues of contextual noise are discussed in Chapter 6, where we propose holistic context modeling that addresses it. In Chapter 7 we compare the holistic context model to existing the "topic model". Conclusions are provided in Chapter 8. Finally, a brief discussion on the implementation details of various recognition systems proposed in this work is provided in Appendix G.