

Research Statement

Discriminative Graphical Models for Image Classification

Sanjiv Kumar (skumar@cs.cmu.edu)

A large number of real-world classification problems in the domain of text, speech, image, biology, etc. involve labeling of data items that have multiple components, each requiring a classification label. These problems are challenging because the interactions among the components can be rich and complex. I am interested in sound theoretical formulations that can *model* such complex interactions, and robust algorithms for *efficient parameter learning and inference* over such models in real-world applications.

Computer Vision provides a challenging arena for exploring intricate interactions among various image components to realize the long standing goal of *semantic scene interpretation*. An image component may refer to a single pixel, a patch of pixels, an object or the whole image itself. Existing interaction models are exclusively developed in a generative paradigm needing various simplifying assumptions. This severely restricts their classification power in a variety of applications. During my doctoral research, I introduced a new family of models in computer vision that *discriminatively* model interactions among different components, leading to significant improvements in image denoising, texture classification, semantic segmentation and object detection [1,11,12,19].

The Curse of Ambiguity: Why should we care to model the interactions at all? The answer is that classifying individual components independently is extremely difficult since data from different classes often appear similar. This is due to the physics of imaging (illumination, pose, noise), or the intrinsic nature of the data itself (some objects look similar). Incorporating spatial interactions can help disambiguate the classification significantly.

The next question is: Can we solve the ambiguity problem without modeling spatial interactions? My early work on scene classification introduced a new technique for combining supervised learning on the training data with unsupervised learning on the test data to alleviate the ambiguity problem [2]. Even though the ambiguity was reduced in the specific application, use of spatial interactions further improved the results significantly [9]. In fact, the data interaction models are much more general and several applications can be addressed through a single model.

Modeling Spatial Interactions: The causal models for image interactions suffer from the problem of non-stationarity yielding inaccurate classification, as shown in my work on building detection [13]. Markov Random Fields (MRFs) are the traditional noncausal interaction models used in vision, but they exclusively use generative models for classification. On the other hand, a wealth of powerful discriminative classifiers including kernel machines has been proposed in machine learning literature recently, but these classifiers are limited to independent data, allowing no interactions among them. To overcome these limitations, I introduced discriminative probabilistic models named Discriminative Random Fields (DRFs) [11,12] in computer vision, which extended the idea of Conditional Random Fields from 1D graphs to arbitrary graphs with loops, essential for vision problems. The DRFs allow use of arbitrary task-dependent discriminative classifiers for data interactions.

Other advantages of the DRFs are: First, they capture arbitrarily complex spatial dependencies in the labels and as well as the observed data simultaneously, while the MRFs generally use the simplifying assumption of conditional independence of the observed data. Second, by allowing data-dependent label interactions, the DRFs can model relationships between parts of an object (geometric configurations) or between different objects themselves, leading to robust object detection [19], and possibly scene or event recognition. This was not possible using the conventional MRFs.

The present DRF models use only pairwise interactions between the components. For several applications, it is essential to incorporate higher order interactions, e.g., affine invariance can be achieved in object detection by considering the interactions between component triples. I intend to explore the formulation and applicability of such models in the future.

Parameter Learning and Inference: Even if we are able to create powerful models, they will remain useless unless efficient parameter learning and inference is possible in them. Since the underlying graph in DRFs has loops, in general, exact methods cannot be used for learning. In my work on approximate parameter learning in discriminative fields, I have investigated efficient ways of achieving learning over such fields [18] and demonstrated their efficacy by learning a large number of parameters for object detection [19]. A major advantage of these techniques is that they can also be used to train conventional MRFs. Efficient training of the DRFs when using kernels is still an open question, especially for large problems. My work on DRF based building detection is currently being used for real-time robotic landmark detection and video retrieval.

In many vision applications, a-priori deterministic information may be given. For example, in a scene, all the SKY components *must* lie above all the WATER components. A straightforward inclusion of these constraints into standard inference techniques may be hard due to the loss of locality in the graph structure. I have developed a constrained inference technique for one such case where no two image components were allowed to take the same label [10]. In the future, I plan to systematically investigate the inclusion of local or global constraints in inference, which remains virtually unexplored in computer vision.

Image Synthesis: The spatial interaction models can be used not only for image analysis but also for image synthesis. As an intern at Microsoft Research, I applied a random field formulation to obtain a visual summary (*digital tapestry*) of a collection of consumer images, which can be used either as a virtual thumbnail for the whole collection or as an image retrieval engine [10]. In the future, I would like to explore image synthesis tasks to verify the modeling power of the richer models that I plan to develop.

Future Goals: To solve the most challenging problems in computer vision such as generic object recognition, event or activity recognition, and image and video retrieval, a principled yet practical approach is needed. Clearly, this demands expertise in both computer vision and machine learning. My research goal is to harness the synergy at the interface of these two areas. In this direction, I have successfully addressed several sub-topics during my PhD at CMU, and described future research directions as stated above. Another important problem I plan to address in the future is to develop formal ways to automatically learn task-dependent data representation (features) in images. I would also like to explore the applications of this research in medical image analysis and robotics, areas in which I have done previous research [3-8,15-17], cognitive vision and computational biology.