Abstract

In computer vision, one is interested in extracting useful information from a given image or video. However, depending upon the sensor involved, a typical camera system may only capture a subset of the available information about a scene. For example, an RGB camera does not capture the depth information. Finite wavelength response, spatial and temporal resolution, as well as occlusion and noise during the capturing process, further reduce the available information. This results in inherent ambiguity in the inference for many perception tasks in computer vision. A common way to get around the ambiguities in an inference problem is to use auxiliary information in the form of priors and/or evidence. The focus of this thesis is to propose novel techniques and algorithms for exploiting such auxiliary information to improve prediction performance in common computer vision problems.

In an MRF-MAP framework, one formulates a computer vision problem as a Markov Random Field and then computes the labeling configuration with maximum a posteriori probability. The framework captures the prior conditional dependencies between the pixel/node labels through the use of clique potential, with larger cliques allowing for wider and more complex priors. However, the inference in such higher order MRF-MAP problems is NP-hard. This has led researchers to focus on various subsets of the problem, for which efficient algorithms for optimal inference may be proposed. One such subset, which is also the focus of this thesis, is the MRF-MAP problems with submodular clique potentials. In the first part of the thesis, we give an efficient algorithm for provably optimal inference even when the clique size grows to 25. Here we make use of an important observation that only an extremely small subset of constraints defined by a clique potential are typically useful for defining the optimal MAP configuration. We exploit the observation and perform inference with a small number of such constraints in a particular iteration. As the iterations progress, we bring the constraints as per the need, ultimately converging to the optimal configuration using a fraction of memory and orders of magnitude improvement in inference time. The algorithm is called the Lazy Generic Cuts, due to its use of Generic Cuts as a baseline, and characteristic behavior of lazily bringing in the required constraints.

The effective use of complex higher order priors requires not only the efficient inference algorithm, but the capability to learn such clique potentials from the training data. In the second part of the thesis, we propose a novel algorithm for learning higher order submodular clique potentials. We use the standard max-margin framework within a structured SVM formulation. However, imposing additional submodularity constraints on parameters creates scalability issues, since the number of submodularity constraints increases exponentially with the clique size. It becomes difficult for existing algorithms to scale to large problems arising out of learning higher order potential in an MRF-MAP formulation. As an important contri-

bution, we propose a relaxation of submodularity constraints and give an efficient subgradient based method for the learning. The proposed algorithm, also exploits the Lazy Generic Cuts for optimal inference in the inner loop. Whereas the earlier state of the art could learn clique potentials of size only upto 9, using our algorithm one can learn the clique potentials even upto size 16.

In the last decade, fuelled by the improvement in the computation capabilities and the availability of large annotated datasets, deep neural networks have become a de-facto technique for many computer vision problems. Though our earlier work on MRF-MAP problems can be used as post-processing with many such deep neural network techniques, the lack of end-to-end training and inference of deep neural network with the MRF-MAP module, hampers the overall prediction performance. In the third and last part of the thesis, we explore the direct use of auxiliary information in the form of evidence, to improve predictions of deep neural networks. Unlike priors, which are dataset level statistics, evidence is much more targeted, and represent additional sample level information at the test time. We see the abundance of such auxiliary information around us. For example, while the primary inference task of inference may be semantic or instance level segmentation of an image, we often have image level tags or associated image captions. Although one can design an appropriate deep neural network which takes multi-modal input, the approach requires training such models from scratch, requiring a large amount of annotated data with labels for the primary task as well as auxiliary information. On the other hand, we propose a novel multi-task learning (MTL) framework for exploiting the auxiliary information. We propose to model the task of interest as the primary task in an MTL framework. The auxiliary information in the form of evidence is incorporated by modeling it as the output of a secondary task. Given a particular sample, we first perform inference on the secondary task and back-propagate the loss on the secondary task based upon the given auxiliary information. This allows the model to adapt the weights as per the given sample, which are then used to perform inference on the primary task. The advantages of the proposed framework include the possibility of using standard pre-trained models for both primary and secondary tasks, as well as the ability to train even when no jointly annotated data is available. The second advantage flows directly from the use of the MTL framework.

In summary, this thesis proposes novel techniques to exploit the auxiliary information, either in the form of prior or evidence, to improve the predictive accuracy of underlying machine learning models. In each case, we demonstrate the efficacy of our proposed techniques by experimenting on real-world data and improving the performance of the state-of-the-art.