CS231A Course Project Proposal: Unsupervised Depth Mask Learning for Object Detection in RGBZ Images

Andrew Duchi
Stanford University
aduchi@cs.stanford.edu

1. Overview

I will be finalizing an action classification pipeline proposed by Pawan Kumar, Ben Packer, and Tim Tang that makes use of latent variables for object and human detection and poselet features for human action evaluation.

The focus of this project is to classify the actions of people in various images based on either human pose or the objects that the people are interacting with. This task is important because the semantic information of an image often goes far beyond simply the image contents; for example, there is a distinction between pictures of people kicking or throwing balls and the images give very different information (perhaps imply we are viewing a different sport).

1.1. Proposed Datasets

The dataset to be used will be the set of images that has been made available for training for the Pascal challenge, which is available at http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/actionexamples/index.html. These images are labeled with bounding boxes around all humans, but not any object detections. This dataset has also been augmented by using Google image search.

1.2. Proposed Methods

In order to perform action classification, we will use precomputed poselets and object detection probabilities for each image, these will be done with existing packages and will not be changed. The main coding for this project will be to implement a Mean Average Precision struct SVM in C++ and finalize a pipeline for processing and classifying images, as well as performing evaluation. The use of latent variables for an iterative classification procedure (where we instantiate, update our beliefs, then reevaluate instantiation of objects) is the novel aspect of this method – contrasting with most prior work that simply used raw object detection probability as a feature. Other changes will likely be made as I get past the point of implementing the base model (changing weightings in the model, adding features) but it is difficult to foresee what the error analysis will point to as the most important aspects to update.

1.3. Background Material

There are a number of papers on action classification that I will reference for the general task[3, 4]. I will read papers on deformable parts models for interfacing with our object detection framework and on poselets as I will use those features as well[1]. Additionally, I also will read about the Mean Average Precision SVM to understand how to appropriately implement and optimize that framework[5] and refine my understanding of latent variable models [2]. For full paper titles, please refer to the bibliography at the end of this paper.

1.4. Evaluation and Reporting

As this is a classification task, the primary evaluation metric will be looking at precision and recall displayed on an ROC plot. Results will also be displayed and analyzed in images with bounding boxes around detected pepople and actions.

References

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1365 –1372, 29 2009-oct. 2 2009.

[2] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models, 2010. NIPS 2010.

[3] B. Liu. Human action recognition using diverse data and self-pace learning, 2011. Unpublished senior honors thesis.

[4] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance, 2010. CVPR 2010.

[5] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 271–278, New York, NY, USA, 2007. ACM.