As imaging systems become ubiquitous, the ability to recognize human actions is becoming increasingly important. Just as in the object detection and recognition literature, action recognition can be roughly divided into classification tasks, where the goal is to classify a video according to the action depicted in the video, and detection tasks, where the goal is to detect and localize a human performing a particular action. A growing literature is demonstrating the benefits of localizing discriminative sub-regions of images and videos when performing recognition tasks. In this thesis, we address the action detection and recognition problems.

Action detection in video is a particularly difficult problem because actions must not only be recognized correctly, but must also be localized in the 3D spatio-temporal volume. We introduce a technique that transforms the 3D localization problem into a series of 2D detection tasks. This is accomplished by dividing the video into overlapping segments, then representing each segment with a 2D video projection. The advantage of the 2D projection is that it makes it convenient to apply the best techniques from object detection to the action detection problem. We also introduce a novel, straightforward method for searching the 2D projections to localize actions, termed Two Point Subwindow Search (TPSS). Finally, we show how to connect the local detections in time using a chaining algorithm to identify the entire extent of the action. Our experiments show that video projection outperforms the latest results on action detection in a direct comparison.

Second, we present a probabilistic model learning to identify discriminative regions in video from weakly-supervised data where video clips are only assigned a label describing what action is present in the frame or clip. While our first system requires every action to be manually outlined in every frame of the video, this second system only requires that the video be given a single high-level tag. From this data, the system is able to identify discriminative regions that correspond well to the regions containing the actual actions. Our experiments on both the MSR Action Dataset II and UCF Sports dataset show that the localizations produced by this weakly supervised system are comparable in quality to localizations produced by systems that require each frame to be manually annotated. This system is able to detect actions in both 1) non-temporally segmented action videos and 2) recognition tasks where a single label is assigned to the clip. We also demonstrate the action recognition performance of our method on two complex datasets, i.e. HMDB and UCF101.

Third, we extend our weakly-supervised framework by replacing the recognition stage with a two-stage neural network and apply dropout for preventing over-fitting of the parameters on the training data. Dropout technique has been recently introduced to prevent over-fitting of the parameters in deep neural networks and it has been applied successfully to object recognition problem. To our knowledge, this is the first system using dropout for action recognition problem. We demonstrate that using dropout improves the action recognition accuracies on HMDB and UCF101 datasets.

Major: Electrical Engineering

Educational Career:
Bachelor’s of Electrical and Electronics Engineering, BS, 2002, Hacettepe University
Master’s of Electrical and Electronics Engineering, MS, 2005, Bilkent University

Committee in Charge:
Marshall Tappen, Chair, Electrical Engineering and Computer Science
Hassan Foroosh, Electrical Engineering and Computer Science
Mingjie Lin, Electrical Engineering and Computer Science
The public is welcome to attend.