Research in human action recognition strives to develop increasingly generalized methods that are robust to intra-class variability and inter-class ambiguity. Training such models need thousands of precise spatio-temporal manual annotations, which require many human annotators, hundreds of hours, and are subject to human biases.

In the first part of this dissertation, we explore the reasons for poor classifier performance when tested on novel datasets, and quantify the effect of scene background on action representations and recognition. We propose a new method to obtain a measure of confidence in each pixel of the video being a foreground region using motion, appearance, and saliency together in a 3D "Markov Random Field (MRF) based framework. We also propose multiple ways to exploit the foreground confidence: to improve bag-of-words vocabulary, histogram representation of a video, and a novel histogram decomposition based representation and kernel.

The above-mentioned method does not provide the precise spatio-temporal location of the actor and needs manual spatio-temporal annotations to train an action detector. Therefore, in the second part of this dissertation, we propose a weakly labeled approach to automatically obtain spatio-temporal annotations of actors in action videos. We first obtain a large number of action proposals in each video. To capture a few most representative action proposals in each video, we rank them using motion and saliency cues and select a few proposals using MAP based proposal subset selection method. Our next challenge is to iteratively select one proposal from each video so that all proposals are globally consistent. We formulate this as Generalized Maximum Clique Graph problem. The output of our method is the most action representative proposals from each video which are used to train action detector.

The above-mentioned annotation method uses multiple videos of the same action. Therefore, in the third part of this dissertation, we tackle the problem of spatio-temporal action localization in a video, without assuming the availability of multiple videos or any prior annotations. We present a novel approach for action localization in videos using web images. Given a video, first we generate multiple spatio-temporal action proposals. To obtain the most action representative proposals, we reconstruct action proposals in the video by leveraging the action proposals in images. We solve this optimization problem using the variant of two-metric projection algorithm.

Finally, we propose a framework to generate a few better action proposals that are ranked properly. We first divide each action proposal into sub-proposals and then use Dynamic Programming based graph optimization scheme to select the optimal combinations of sub-proposals from different proposals and assign each new proposal a score. We propose a new unsupervised image-based actionness detector that leverages web images and employs it as one of the node scores in our graph formulation. We demonstrate that properly ranked proposals produce significantly better action detection as compared to state-of-the-art proposal based methods.

Our extensive experimental results on different challenging and realistic action datasets, comparisons with several competitive baselines and detailed analysis of each step of proposed methods validate the proposed ideas and frameworks.

Major: Computer Science

Educational Career:
Bachelor's of Electrical Engineering, BS, 2006, University of Engineering and Technology
Master's of Electrical Engineering, MS, 2010, Seoul National University