Pose Filter Based Hidden-CRF Models for Activity Detection

Prithviraj Banerjee and Ram Nevatia
University of Southern California, Los Angeles, USA

Abstract. Detecting activities which involve a sequence of complex pose and motion changes in unsegmented videos is a challenging task, and common approaches use sequential graphical models to infer the human pose-state in every frame. We propose an alternative model based on detecting the key-poses in a video, where only the temporal positions of a few key-poses are inferred. We also introduce a novel pose summarization algorithm to automatically discover the key-poses of an activity. We learn a detection filter for each key-pose, which along with a bag-of-words root filter are combined in an HCRF model, whose parameters are learned using the latent-SVM optimization. We evaluate the performance of our model for detection on unsegmented videos on four human action datasets, which include challenging crowded scenes with dynamic backgrounds, inter-person occlusions, multi-human interactions and hard-to-detect daily use objects.

Keywords: Activity detection, Key-poses, CRFs, Latent-SVM.

1 Introduction

There has been considerable research in classifying segmented videos, however there has been comparatively less progress on the more challenging task of activity detection, where multiple instances of an activity are simultaneously localized and classified in un-segmented videos. Detection is an important task, as in real world applications like surveillance, the activities of interest occur only for a part of the video. We propose a novel activity detection algorithm based on automatically discovering the key-poses in the activity, and learning a key-pose filter based Hidden Conditional Random Field (HCRF) model. We focus on activities primarily defined by a sequence of complex pose and motion changes, which can involve interactions with objects or other humans in the scene.

Activity recognition algorithms can be broadly categorized based on their structure modeling capabilities. A common class of approaches [9,26] train classifiers on video-wide statistics of local features, and ignore the local temporal dynamics of the activity. To classify unsegmented videos, they typically use an inefficient sliding window approach [30], which can be sensitive to window size. A complementary approach [11,12] learns a sequential motion model, and performs classification based on state assignments inferred from every frame in the video; to keep the inference tractable, these further require a Markovian assumption.
between adjacent frames, and fail to capture long range dynamics in the activity, making them sensitive to variations in activity styles and action-durations.

We argue that for activity detection, it is sufficient to determine the presence or absence of certain key states in an observation sequence, and whether certain temporal relationships between the state detections are satisfied. Recognizing actions using a subset of the frames has been explored previously [3,15,20,21], however these do not address the problem of automatically discovering the important states/sub-sequences of an activity, and either perform exhaustive search over all possible sub-sequences [20], rely on hand annotations [3,21], or use a manually defined list of relevant poses [15], requiring separate annotated pose data for each pose-detector. There exist methods for automatic discovery of key-states [11,31], however [11] relies on hard to obtain mocap data, and [31] ignores temporal structure.

We propose a novel graphical model for activity detection, where the random variables to be inferred are the temporal locations of the key-poses. Key-poses represent the important human pose configurations in an activity, and are a natural choice for defining a key-state in our model. Our algorithm automatically discovers the relevant key-pose definitions in an activity, and learns a set of key-pose detection filters, and pools their detection responses, while satisfying the temporal relationships between them.

Our contributions are multi-fold: (1) The relevant key-poses are discovered automatically, (2) the key-pose detection filters are learned jointly in a discriminative HCRF framework, and do not require manually annotated pose-specific training data, (3) the temporal locations of the key-pose detections correspond to the active segments in the video stream, enabling activity detection in unsegmented videos, and (4) the key-poses correspond to a natural semantic interpretation. We show results on 4 datasets, which include challenging crowded scenes with dynamic backgrounds and inter-person occlusions, multi-person interactions, and actions involving hard-to-detect daily use objects.

2 Related Work

We briefly survey classification methods using subsequence based models, and methods for activity detection.

[Subsequence Models] The discriminative advantage of short snippets of video for activity classification has been recognized before [21]. There exist approaches