The success of recognizing periodic actions in single-person-simple-background datasets, such as Weizmann and KTH, has created a need for more difficult datasets to push the performance of action recognition systems. We identify the significant weakness in systems based on popular descriptors by creating a synthetic dataset using Weizmann database. Experiments show that introducing complex backgrounds, stationary or dynamic, into the video causes a significant degradation in recognition performance. Moreover, this degradation cannot be fixed by fine-tuning the system or selecting better interest points. Instead, we show that the problem lies at the cuboid level and must be addressed by modifying cuboids.

For the above method however, one requires near-perfect localization of the action within a video sequence. To achieve this objective, we present a two stage weakly supervised probabilistic model for simultaneous localization and recognition of actions in videos. Different from previous approaches, our method is novel in that it (1) eliminates the need for manual annotations for the training procedure and (2) does not require any human detection or tracking in the classification stage. The first stage of our framework is a probabilistic action localization model which extracts the most promising sub-windows in a video sequence where an action can take place. We use a non-linear classifier in the second stage of our framework for the final classification task. We show the effectiveness of our proposed model on two well known real-world datasets: UCF Sports and UCF11 datasets.

Another application of the weakly supervised probabilistic model proposed above is in the gaming environment. An important aspect in designing interactive, action-based interfaces is reliably recognizing actions with minimal latency. High latency causes the system's feedback to lag behind user actions and thus significantly degrades the interactivity of the user experience. With slight modification to the weakly supervised probabilistic model we proposed for action localization, we show how it can be used for reducing latency when recognizing actions. This latency-aware learning formulation trains a logistic regression based classifier that automatically determines distinctive canonical poses from data and uses these to robustly recognize actions in the presence of ambiguous poses. We introduce a novel (publicly released) dataset for the purpose of our experiments. Comparisons of our method against both a Bag of Words and a Conditional Random Field (CRF) classifiers show improved recognition performance for both pre-segmented and online classification tasks. Additionally, we employ GentleBoost to reduce our feature set and further improve our results. Finally, we present experiments that explore the accuracy/latency trade-off over a varying number of actions.