論文の内容の要旨

Modeling and Recognizing Human Activities from Video
（映像にもとづく人物行動のモデリングと認識）

キタニ クリス マコト

This thesis presents a complete computational framework for discovering human actions and modeling human activities from video, to enable intelligent computer systems to effectively recognize human activities. This work is motivated by a desire to create an intelligent computer system that can understand high-level activities of people, thus allowing computer systems to efficiently interact with people. A bottom-up computational framework for learning and modeling human activities is presented in three parts. First, a method for learning primitive actions units is presented. It is shown that by utilizing local motion features and visual context (the appearance of the actor, interactive objects and related background features), the proposed method can effectively discover action categories from a video database without supervision. Second, an algorithm for recovering the basic structure of human activities from a noisy video sequence of actions is presented. The basic structure of an activity is represented by a stochastic context-free grammar, which is obtained by finding the best set of relevant action units in a way that minimizes the description length of a video database of human activities. Experiments with synthetic data examine the validity of the algorithm, while experiments with real data reveals the robustness of the algorithm to action sequences corrupted with action noise. Third, a computational methodology for recognizing human activities from a video sequence of actions is presented. The method uses a Bayesian network, encoded by a stochastic context-free grammar, to parse an input video sequence and compute the posterior probability over all activities. It is shown how the use of deleted interpolation with the posterior probability of activities can be used to recognize overlapping activities. While the theoretical justification and experimental validation of each algorithm is given independently, this work taken as a whole lays the necessary groundwork for designing intelligent systems to automatically learn, model and recognize human activities from a video sequence of actions.