報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　A Robust Approach to Hough-based Action Detection
（ハフ変換に基づく行動検出の頑健化に関する研究）

氏　名　HARA Kensho

論　文　内　容　の　要　旨

There has been rapid growth in computers' understanding of the real world through images and videos. One of the most important aspects of real-world information is human actions. To support and communicate with humans, computers must understand their actions. In addition, a huge number of videos exist in the world; to manage so much data, automatic understanding of these videos by computers is required. Finally, understanding actions is especially important because the main subjects of most videos are humans.

Our goal is to develop an automated approach to understanding actions in videos. In this study, we focus on action detection in videos; action detection finds where, when, and what actions occur within videos. Target actions in this study, such as *kicking a ball* and *running*, are constructed by primitive elements such as *lifting a leg* and *swinging an arm*.

We focus on Hough-based action detection methods. Hough-based methods extract local spatiotemporal features from an entire video, then cast votes for action class labels, positions and scales. Here, an action position is usually defined as a spatiotemporal center position. Voting scores are calculated by accumulating the votes at each position based on all local features for each related action class. The local maxima of the accumulated voting scores indicate candidate detected actions. These methods manage the actions that have scores over a threshold as detected actions. Hough-based methods can detect actions robustly with partial observations; the votes based on observed local features are not affected by unobserved local features, because the voting process for each local feature is performed independently. When actions are spatially occluded by other

objects and humans, only partial observations are available. Robustness is also useful to early action detection, which can use only early observations of actions.

Various factors, such as occlusion, human orientation variety, motion similarity, temporal variations, and action manners variety, make accurate action detection from videos difficult. To detect actions accurately, detection methods should be robust to these factors. Hough-based methods are already robust to occlusions. In this study, we propose methods that increase robustness to four additional factors: *human orientation variety, motion similarity, temporal variations*, and *action manners variety*. This thesis is organized with the following chapters.

Chapter 1 presents background, problems, and purpose of this thesis. Contributions of this study are also provided in this chapter.

Chapter 2 reviews related work; we provide an overview of feature representation methods for action recognition and detection and describe related approaches to action detection.

Chapter 3 describes the basic Hough-based action detection algorithm and the implementation of our baseline method.

Chapter 4 presents a Hough-based method that uses multiview videos to provide additional robustness to variety in human orientations. The appearances of actions change relative to a person ' s orientation to cameras. Multiview videos are synchronous videos captured from multiple cameras. Capturing actions with multiview videos gives observations from various viewpoints. These observations include different relative orientations of human subjects to the cameras. Therefore, these observations reduce the differences in relative orientation between training and test data and contribute the robustness to human orientation variety. Our proposed method casts independent votes in each view. Here, we assume that human feet touch the ground plane when they start an action. We then integrate votes in global coordinates based on assumptions using homographic transformations. The proposed method uses multiview information effectively and detects actions robustly.

Chapter 5 describes a novel Hough-based action detection method to overcome the problem of motion similarity. Discriminating between similar local motions that exist in different action classes is difficult; in such cases, conventional Hough-based methods often cast votes for false action classes. The false votes do not occur randomly such that they depend on relevant action classes. We introduce vote distributions, which are distributions of the voting scores for each action class. We assume that the distribution of false votes includes important infor-

mation necessary for improving action detection. These distributions are used to build a model that represents the characteristics of Hough voting, including false votes. This method estimates likelihood using this model and reduces the influence of false votes, which leads to robustness to motion similarities across action classes.

Chapter 6 presents a method for achieving robustness to temporal variations that can exist in the same action class. Conventional Hough-based methods perform poorly for actions with temporal variations because such variations change the temporal relation between local feature positions and action positions. Some votes may be scattered in the temporal dimension because of such variations. We propose a method for concentrating scattered votes through time warping. The proposed method estimates the offsets between scattered and concentrated voting positions based on conventional Hough votes. The offsets warp the scattered votes to concentrate them, providing a method for robustness even in the presence of temporal variations.

Chapter 7 describes a method that focuses on the number of local features. Various factors, such as occlusions, human orientation variety, temporal variations, and action manners variety, change not only the feature descriptors of actions but also the number of local features. Conventional Hough-based methods perform poorly with variations in the number of local features extracted from actions. Changes in voting scores that depend on the number of local features produce difficulties in determining a voting score detection threshold. Our proposed method improves two parts of the Hough-based method. The first is the extraction of local features; the proposed method reduces the method ' s dependency on the number of local features based on spatial scales. It adjusts the number of local features for each spatial scale using a sampling method. The second part is detection thresholding. The proposed method determines appropriate thresholds for voting scores based on the number of local features by learning the relation between the number of local features and voting scores. These changes reduce the influence of the number of local features (i.e. improving robustness in different aspects from previous chapters).

Finally, Chapter 8 concludes this thesis and presents directions for future work.