別紙４－１（課程博士（英文））

Date of Submission (month day, year) : January 6th, 2023

Department of Computer Science and Engineering	Student ID Number	D179304		Supervisors	Masaki Aono Shigeru Kuriyama
Applicant's name	Yuri Yudhaswana Joefrie				

Abstract (Doctor)

Title of Thesis	On Spatio-Temporal and Motion Feature Representation Learning with Deep Neural Network for Accurate Video Action Recognition

Accurate action recognition from digital videos is one of core and challenging computer vision problems that has been studied for decades. The potential applications of action recognition include video surveillance, human-computer interface, visual information retrieval, and unmanned driving. How to perform effective and efficient analysis of the video footage is vital given the recent exponential growth of surveillance data on the Internet. More and more individuals are becoming accustomed to posting photographs and phrases on social media sites like Instagram, Facebook, Twitter, and Flickr to express their feelings and ideas on almost any occasion or topic. As a result, studying the vast quantity of videos on social media has become increasingly important. Traditional machine learning techniques that only extract computable characteristics have limitations and need to function better with large amounts of visual input. However, convolutional neural networks (CNNs) and other deep learning techniques have significantly improved in this area.

Even though deep learning models exhibit such prominent results, the proposed work for action recognition still needs to improve, despite the numerous solutions that have been developed thus far. The model's ability to precisely detect changes in action variety, in addition to its ability to effectively recognize targets and actions from the background, presents a significant challenge. The conventional model will be hampered as a result of this issue.

Meanwhile, action recognition in videos can be divided into two tasks. Recognizing various activities consecutively or simultaneously in each video at an arbitrary time is the first task. Each activity may have a unique set of features. Another task is to recognize a single action inside a video Thus, different approaches are needed to deal with any issues that may develop. In this dissertation, we propose two approaches to tackle the challenges of action recognition and validate them in challenging datasets.

We introduce deep fusion schemes for multi-label multi-class classification for the first task. This approach extends the previous work by introducing a fusion of spatial and temporal branches to provide superior action recognition capability toward multi-label multi-class classification problems. We propose three fusion models with different fusion strategies. We first build several efficient Temporal

Gaussian Mixture (TGM) layers to form spatial and temporal branches to learn a set of features. In addition to these branches, we introduce a new deep spatio-temporal branch consisting of a series of TGM layers to learn the features that emerged from the existing branches. Each branch produces a temporal-aware feature that assists the model in understanding the underlying action in a video. We verify the performance of our proposed models on the well-known MultiTHUMOS benchmarking dataset.

We tackle several issues regarding spatio-temporal and motion feature learning for the second approach. As these aspects are the key to action recognition, we build a novel building block for a 2D CNN-based network for efficient and effective learning of the abovementioned features. Typical previous approaches utilize 3D CNNs to cope with spatial and temporal features while they suffer from massive computations. Other approaches are to utilize (1+2)D CNNs to learn spatial and temporal features efficiently while they need to pay more attention to the importance of motion representations. To overcome problems with previous approaches, we propose a novel block that can capture spatial and temporal features more faithfully and efficiently learn motion features. This proposed block includes Motion Excitation (ME), Multi-view Excitation (MvE), and Densely Connected Temporal Aggregation (DCTA). The purpose of ME is to encode feature-level frames difference; MvE is designed to enrich spatiotemporal features with multiple view representations adaptively; DCTA is to model long-range temporal dependencies. We inject the proposed building block, which we refer to as the META block (or simply "META"), into 2D ResNet-50. We verify the performance of our proposed model on three large-scale benchmarking datasets, including the Something-something v1, Jester, and Moments in Time Mini datasets.

Extensive experimental studies prove that both approaches demonstrate competitive results on several benchmarking datasets compared to the baseline and state-of-the-art methods.