Form 3

Abstract of Thesis

Name （Buntueng Yana）	
Title	A Study on Fusion Framework for Air-writing Recognition Based on Spatial and Temporal Hand Trajectory Modeling (手先軌道の静的・動的モデリングに基づく空中手書き文字認識に関する研究)

Abstract of Thesis

Air-writing refers to writing alphabet or numeric gestures by hand or finger movement in free space. It has attracted attention since it can offer verbal communication. Even the air-writing recognition has been studied more than three decades, creating a robust system is still challenging. The primary objective of this work is studying a fusion framework for air-writing recognition from a vision-based sensor. We employ a fusion scheme by modeling the air-writing with the temporal feature augmented with the spatial feature. We address the air-writing recognition in two categories: motion character recognition and motion word recognition. The underlying assumption of motion character is that the gesture is correctly spotted; therefore, a segmentation process is not necessary for the air-writing recognition. In contrast, the motion word gesture was captured from a user in a motion stream. It does not have a sign to indicate the writing and non-writing part. Moreover, there are ligatures between the characters in motion word.

For learning the motion character, we model the air-writing by using spatial features augmented with an image-like feature. The proposed structure comprises three main parts: a CNN part, an RNN part, and a Fusion part. The CNN part consists of three convolution layers and two subsampling layers. The convolution layers in the CNN part are employed to extract information from the image-like feature. In the RNN part, there are two types of structures that were considered. The first structure is a Bidirectional Long-Short Term Memory (BLSTM), and the other one is a simplified Bidirectional Recurrent Neural Network (simplified BRNN). To obtain useful information from the temporal features, the BLSTM was deployed in the RNN part. The output of the CNN and the RNN parts were combined before feeding into the Fusion part. In the first experiment, the performance of the proposed structure was compared with three baseline references: the CNN, the BLSTM, and Yang's work. The result confirms the fusion scheme outperforms all of the references. In the second experiment, the effects of the recurrent units were examined by varying the number of BLSTM units in the RNN part. The optimum number of the BLSTM units are 15 and 25 for the numeric gesture and alphabet gesture, respectively. From the experimental results, we confined that the execution time of the fusion structure is high due to the complexity of the BLSTM unit. In the third experiment, the simplified BRNN was considered. When comparing the results with the previous experiment, the execution time of simplified BRNN unit reduces in half while the accuracy drops insignificantly. In the last experiment, we demonstrated that using hand position feature (RNN part) and image-like feature (CNN part) are adequate for the fusion network.

For leaning the motion word, a deep recurrent neural network was studied. In the output layer of the proposed structure, the Connectionist Temporal Classification (CTC) loss was considered. The main advantage of using the CTC loss is removing a predefined alignment to create the training set. The features that we studied are the hand position feature and the path signature feature. In the preprocessing stage, we employ a sliding window technique to segment a long sequence of motion gesture into small pieces. Then, each piece of motion was used to generate the hand position feature and the path signature feature. When using the sliding window technique, the most critical parameter is the size of the sliding window. The output of the fusion structure attempts to predict characters in a word; therefore, the size of the sliding window should be set to capture the data no more than one character at a time. For examining the performance of the proposed structure, two public datasets were studied, i.e. a palm-writing dataset and a finger-writing dataset. Each dataset was analyzed to

obtain a writing duration per character, which could be used to set the maximum size of the sliding window. The shortest of writing duration per one character in the palm-writing dataset and finger writing dataset are 0.88 seconds and 1.38 seconds, respectively. From the experiments, the appropriate window size of the palm-writing and the finger-writing dataset are determined as 0.5 seconds and 0.25 seconds, respectively. The best recognition accuracy on the palm-writing dataset and the finger-writing dataset are 86.90% and 75.81%, respectively. We also confirmed that the required prediction time per word on the palm-writing dataset and the finger-writing dataset are 3.91 milliseconds and 6.37 milliseconds, respectively. These results confirm the proposed algorithm can be executed in a real-time.

様式 7

論文審査の結果の要旨及び担当者

氏　名　（　Buntueng Yana　）		
	（職）	氏　名
論文審査担当者	主 査　教授	尾上 孝雄
	副 査　教授	橋本 昌宜
	副 査　准教授	谷口 一徹
	副 査　客員教授	岸野 文郎（関西学院大学）

論文審査の結果の要旨

本論文は、空中手書き文字（ハンド・ジェスチャー）認識に関する研究の成果をまとめたものであり、以下の主要な結果を得ている。

1. 手先軌道の静的・動的混合モデリング手法の提案

空中手書き文字認識に用いるモデリング手法として、従来は、手先軌道の静的データ(画像データ)と動的データ(時系列データ)のいずれかもしくは両方について、個々に畳み込みニューラルネットワークなどの機械学習手法を適用することで文字を識別し、その結果を組み合わせる手法が採られていた。本研究では、静的データと動的データを混合し、多層ニューラルネットワークを用いた深層学習により識別する手法を提案している。具体的には、静的データならびに動的データに対し、畳み込みニューラルネットワークと再帰型ニューラルネットワークをそれぞれ適用し、全結合層で統合・識別する構成となっている。アラビア数字ならびに英語アルファベットの手書き文字ベンチマークデータセットに適用した結果、それぞれ 99.83% および 99.25% という高い認識率を達成している。また、長短期記憶(LSTM; Long Short-Term Memory) ネットワークの構造を簡素化することで、認識精度劣化を抑えつつ処理時間の高速化も達成している。

2. コネクショニスト時系列分類損失を用いる単語認識への拡張

空中手書き文字認識では、個別文字の認識手法を時系列に接続する形で単語認識への拡張が期待できるが、空中手書き文字の場合、個別文字の時間的なばらつきが課題となっていた。本研究では、同様の課題を持つ音声認識分野で有効に用いられているコネクショニスト時系列分類損失を計算しつつ、スライディングウィンドウを用いて単語認識範囲を調整することで、時間的なばらつきを許容した単語認識を実現している。本手法を、スライディングウィンドウサイズを変化させつつ、掌軌道および指軌道のデータセットに対して適用した結果、掌軌道データセットについては、ウィンドウサイズが0.5秒の際に86.90%、指軌道データセットについては、ウィンドウサイズが0.25秒の際に75.81%の認識精度を達成した。認識に必要となる処理時間を確認した結果、いずれのデータセットに対しても、リアルタイム処理が達成可能であることを検証できている。

以上のように、本研究による空中手書き文字認識に関する一連の研究成果は、Society5.0におけるサイバーフィジカルシステム等で今後ますます重要となる人間とコンピュータとのインタラクションの機能性を増大させるという側面からも非常に有用である。認識精度のみならず処理時間等に関しても議論されており、本論文は次世代ユーザインタフェースの実用化にも寄与するものと期待できる。従って、博士（情報科学）の学位論文として価値あるものと認める。