	ならんちめぐ ぼると
氏　　名	**Naranchimeg Bold**
本籍（国籍）	モンゴル国
学位の種類	博士(工学)
学位記番号	工博　第３１８号
学位授与年月日	令和２年３月２３日
学位授与の要件	学位規則第５条第１項該当　課程博士
研究科及び専攻	工学研究科デザイン・メディア工学専攻
学位論文　題目	**Study on Efficient Multimodal Fusion Strategy with Deep Neural Network using Audio-Visual Data for Fine-Grained Recognition** **(細粒度認識のための視聴覚データを用いたディープニューラルネットワークを用いた効率的なマルチモーダル融合戦略に関する研究)**
学位審査委員	主査　教授　　藤本　忠博 副査　教授　　今野　晃市 副査　准教授　　明石　卓也

論　文　内　容　の　要　旨

In recent decade, many state-of-the-art algorithms on object recognition as well as audio recognition have achieved noticeable successes with the development of deep convolutional neural network (CNN). However, most of the studies only exploit a single type of data. For that reason, the multimodal research field brings some unique challenges for researchers to make progress in understanding the things (i.e., the world around us) by processing relate information from multiple modalities. Learning from multimodal sources offers the possibility of capturing correspondences between modalities and gaining an in-depth understanding of an object, event, or activity of interest. Furthermore, multimodal fusion is one of the original topics in multimodal research with different fusion approaches: early, late, and hybrid approaches. Multimodal fusion has a very broad range of applications, including image and sentence detection, multimodal emotion recognition, action detection, and especially audio-visual speech recognition (AVSR). The results have shown that one modality can enhance the performance of the other by providing relevant information. However, extracting robust features from different modalities

and fusing them in an effective way is crucial for attaining high recognition performance. Moreover, it is practically challenging for a learning model to fuse different modalities while learning features or to learn joint features while fusing different modalities.

On the other hand, fine-grained recognition (i.e., categorization, classification) is one of the latest and most challenging object recognition task which aims to distinguish similar object categories from another, e.g., species of birds, models of cars, breed of dogs, species of flowers, etc. Fine-grained recognition have been used for widespread applications such as analyzing biodiversity and scene understanding because it has the ability to describe the things in the world in more detail. However, compared to generic object recognition, fine-grained recognition is quite challenging due to the small difference among categories and can be easily overwhelmed by other factors, such as pose, viewpoint, or location of the object in the image. Due to the recent advances in deep learning lead to remarkable progress on fine-grained recognition. Nevertheless, most of the current state-of-the-art methods employee part/pose detection approach, which is also a challenging task to achieve good performances.

This dissertation presents a study on multimodal fusion strategies with deep neural network and audio-visual data for fine-grained recognition. Since sound also provides us important information about the world around us, the goal of this study is to enhance the performance of fine-grained recognition by exploiting the combination of both visual and audio data using CNN, which has been sparsely treated so far. For this purpose, this study aims to answer three research considerations: (1) what to fuse, i.e., what feature representations to use for audio and visual modality, (2) when to fuse, i.e., which fusion strategy performs best when fusing both modalities, (3) how to fuse, i.e., propose or utilize practically efficient methods under different fusion strategies while learning CNN features with both modalities.

Specifically, the kernel-based fusion approach which fuses audio and visual features at the kernel level is studied. First, audio and visual CNN are trained separately by adapting the weights of the pre-trained model. After the training, deep neural features from both modalities are extracted based on the activation values of an inner layer of the trained CNN and combined by multiple kernel learning (MKL) to perform the final classification. Since generating deep neural features based on visual representations of audio recordings have proven to be very effective,

spectrogram representation of audio data is used to extract CNN features. To train Audio and Visual CNN, the most suitable and large enough dataset is required. Such data were not available for audio modality, the audio dataset corresponding to the popular fine-grained image dataset has been collected with proper matching. Unlike support vector machine (SVM) based on a single kernel, MKL uses multiple kernels and learns optimal composite kernels by combining those kernels constructed from different modalities. To automatically determine the kernel weights, an lp-norm MKL algorithm that produces non-sparse kernel combinations is employed. The experimental results indicate that proposed CNN+MKL method which utilizes the combination of audio-visual data outperforms single-modality kernel methods, some simple kernel combination methods, and the conventional early fusion method.

Furthermore, CNN-based multimodal learning models with three types of fusion strategies (early, middle, late) are proposed to settle the issues of combining training data of both modalities. The advantage of the proposed method lies in the fact that CNN is utilized not only to extract features from visual and audio data but also to combine the features across modalities. Experiments are conducted to evaluate the multimodal learning models as well as different fusion strategies with respect to the classification accuracy on the integrated dataset. In qualitative and quantitative results, a model that utilizes the combination of both data outperforms models trained with only either type of data, and fusing them at the late stage performs better with a significant margin. It is also shown that transfer learning can significantly increase the classification performance.

論 文 審 査 結 果 の 要 旨

本論文は，Fine-Grained Recognition（詳細認識）のために視聴覚データを使用したディープニューラルネットワークによる効率的なマルチモーダルフュージョン戦略に関する研究について述べている．詳細認識は，カテゴリ間の特徴量のわずかな違いのために非常に困難であり，画像内のオブジェクトのポーズ，視点，位置などの他の要因によって容易に認識精度が低下する可能性が高い．ディープラーニングの最近の進歩により，詳細な認識が著しく進歩してきている．それにもかかわらず，現在の最先端の方法である生物の部位やポーズ検出アプローチのほとんどは，良好なパフォーマンスを達成しておらず，新規性の高い研究課題で

ある．

提案手法では，上記の課題を解決するために，マルチモーダルな情報を含むデータベースを用いて学習することで，モダリティ間の対応を把握し，対象のオブジェクト，イベント，またはアクティビティの詳細な理解を得ることを可能としている．さらに，マルチモーダルフュージョンにおける，異なるフュージョンアプローチ（初期，後期，およびハイブリッドアプローチ）を分析・研究した前例はほとんどない．

本研究の目標は，convolutional neural network（CNN）を使用して視覚データと音声データの両方を組み合わせ，詳細認識のパフォーマンスを向上させることである．この目的のために，この研究は以下の 3 つの研究上の課題を扱っている．（1）融合する対象，すなわち，オーディオおよびビジュアルモダリティに使用する特徴表現を検討．（2）融合するタイミング，つまり，両方を融合する際の最も効率的な融合戦略を検討．（3）異なる融合戦略の下で実際に効率的な方法を提案または利用する融合方法．以上の結果，オーディオおよびビジュアルモダリティの最適な融合方法を提案し，従来よりも高精度な分類精度を達成している．

本論文の構成は以下の通りである．

第 1 章は序論であり，研究の背景，論文で扱う検討事項，提案手法について概説している．

第 2 章では，詳細認識だけではなく，鳥の種類の認識，ディープニューラルネットワーク，マルチモーダルフュージョン，カーネルベースのフュージョン法における既存の研究の概要について述べられている．

第 3 章では，カーネルレベルでオーディオとビジュアルの両モダリティの深いニューラルネットワークの神経機能を組み合わせて，詳細な鳥の種類を分類するためのカーネルベースの融合に関する研究について述べられている．最初に，事前にトレーニングされたモデルの重みを調整することにより，オーディオとビジュアルの CNN を別々にトレーニングしている．トレーニング後，両方のモダリティからの深い神経機能が，トレーニングされた CNN の内層の活性化値に基づいて抽出され，multiple kernel learning（MKL）によって結合され，最終的な分類が実行される．音声記録の視覚的表現に基づいて深い神経特徴を生成することは非常に効果的であることが証明されているため，音声データのスペクトログラム表現を使用して CNN 特徴を抽出している．また，オーディオとビジュアルのデータセットを用いた学習において，単一のカーネルに基づく support vector machine とは異なり，MKL は複数のカーネルを使用し，さまざまなモダリティから構築されたカーネルを組み合わせて最適な複合カーネルを学習する．実験結果として，視聴覚データの組み合わせを利用するために提案した CNN + MKL 手法が，単一モダリティカーネル手法，いくつかの単純なカーネル組み合わせ手法，および従来の初期融合手法よりも優れていることを示している．

第 4 章では，3 つのタイプの融合戦略（初期，中期，後期）を使用した CNN ベー

スのマルチモーダル学習モデルについて述べ，両方のモダリティのトレーニングデータを組み合わせる問題を解決している．提案手法の利点は，視覚および音声データから特徴を抽出するだけでなく，モダリティ全体で特徴を結合するためにも CNN が利用されるという点にある．実験では，統合データセットの分類精度に関して，マルチモーダル学習モデルとさまざまな融合戦略を評価している．定性的および定量的結果では，両方のデータの組み合わせを利用するモデルは，いずれかのタイプのデータのみでトレーニングされたモデルよりも優れており，後期でそれらを融合することで精度が向上している．

第 5 章では，オーディオ・ビジュアルモダリティを使用した詳細認識のための CNN とのマルチモーダル融合戦略に関する研究の成果をまとめている．

以上，本論文は詳細認識のために音声と視覚のデータを融合させて処理することにより，ディープニューラルネットワークを使用したマルチモーダル融合戦略の研究について述べ，その有効性と有用性を示したものであり，メディア工学分野やコンピュータビジョンの発展に寄与するところが大である．

よって，本論文は博士（工学）の学位論文として合格と認める．

原著論文名（３編を記載．ただし，単位取得満期退学後１年以内の申請の場合は，１編を記載）

Naranchimeg Bold, Chao Zhang, Takuya Akashi: Cross-domain Deep Feature Combination for Bird Species Classification with Audio-visual Data, IEICE TRANSACTIONS on Information and Systems, Vol.E102-D, No.10, pp.2033-2042, 2019.