氏　　名	Sulfayanti
授与した学位	博　士
専攻分野の名称	工学
学位授与番号	博甲第１３６号
学位授与の日付	令和２年３月２４日
学位論文の題目	Development of Hand Posture Classification and Food Constituent Estimation as Welfare Technology Using Convolutional Neural Network
学位審査委員会	主査　金川　明弘　副査　菊井　弦一郎　副査　國島　丈生 副査　山内　仁

学位論文内容の要旨

This dissertation is motivated to be able to realize a support system for welfare technology.

The number of aging people is increasing across the world while the number of young people entering the workforce is decreasing. Therefore, human resources will be faced with more difficult to provide assistance to people in need if they merely depend on the same services and technologies. Another similar situation is caring for people with a sickness or disability. Welfare technology is the technology used to enhance human welfare in daily life, especially for the welfare society (aging people, disability, and people with chronic disease).

A monitoring system as a part of welfare technology works to monitor and provide information to the caregiver. Monitoring technology mostly used sensors or cameras or the combination of both. Therefore, many applications based on computer vision techniques have been widely developed. However, in the research field of computer vision that uses only RGB images captured by a visible light camera, it is difficult to execute tasks stably under various conditions or different environments such as lighting, weather, and background. Robustness with respect to these conditions can be achieved by integrating with data obtained from external sensors. Multi-modal sensor (depth and thermal sensors) and computer vision approaches are leveraged in a system to detect and monitor people's activity continuously. The system was developed for direct monitoring and analyzing patterns of people's activity in daily life to improve the caregiver's ability to assist. Sign language recognition system and food constituent estimation system are two

representatives of the support system for human communication and healthcare in welfare technology.

Many of the recognition tasks in computer vision are conventionally solved using the handcrafted feature-based approach. The experts specifically design the approach for feature detectors and descriptors, and subsequently, the classification task is usually followed by trained classifiers. However, generally in image recognition, environmental circumstances such as background and illumination affect the object appearances in the image. Therefore, it is difficult to manually construct a powerful feature descriptor that entirely illustrates all kinds of objects. Especially, in food recognition and food constituent estimation task, accuracies of handcrafted feature-based approaches are relatively low since appearance of some food has a high intra-class and low inter-class variance. The main factor is that food consists of typically deformable objects.

Nevertheless, the problem in handcrafted feature can be avoided by using deep learning, which is another approach in computer vision. Deep learning represents learning in a computational model using multiple layer processing. These layers automatically extract features and determine them as data input features. During the past years, the appearance of the ImageNet dataset increasingly supported the convolutional neural networks (CNN), which is one of the deep learning techniques, become the most effective architecture to perform visual recognition. CNN automatically extracts relevant features and shapes from a large-scale training dataset for classification. However, CNN required a sufficient dataset to optimize a large number of parameters. Generally, data augmentation and transfer learning methods are frequently applied to resolve the issue of small datasets. Although existing data augmentation focused on two-dimensional data, augmentation data in three-dimensional data (which can be obtained from the RGB-D sensor) has not been proposed.

This dissertation focuses on sign language recognition and nutritional estimation based on deep learning to realize sophisticated welfare technology. In particular, two methods by employing CNN for classification and estimation are proposed. First, a data augmentation for effective hand posture classification has been proposed. The proposed data augmentation strategy generates a large number of hand images with various appearances based on the three-dimensional rotation for depth image data. Second, the

Xception model, which is the state-of-the-art model of CNN, has been applied for feature extraction and classification.

On the other hand, automatic food category classification and food constituent estimation method based on a multi-task CNN has been proposed in this dissertation. In particular, to achieve lifestyle disease prevention, the focus is on the recognition of the food category and the estimation of calories and salinity. The effective estimation of calories and salinity using multi-task learning with food category classification have been realized by defining both calorie and salinity estimation as a regression problem.

Chapter 1 describes the background and purpose of the research. This explanation begins with the importance of fulfilling welfare technology for aging people, disabilities, and people with chronic diseases. This is accompanied by examples of welfare technologies that have been implemented, both assistive technology and monitoring technology. Then, the equipment such as sensor and camera that is widely used for the realization of welfare technology for monitoring, and the involvement of computer vision in offering a digital solution by analyzing the captured image/video are described. Next, the discussion of the computer vision approaches consisting of handcrafted features and deep learning, as well as the consideration in the application of each approach. Finally, I describe the focus of the problem that will be resolved and define the objectives to be achieved in this study.

Chapter 2 describes the theory of CNN based on deep learning. This discussion covers the stages of preprocessing input for CNN, CNN as feature extraction, and the modification of the last CNN layer that is utilized not simply for the classification, but also for the regression. Here, I also discuss the approaches to increase the effectiveness of using CNN, such as fine-tuning and CNN architectures.

Chapter 3 presents the proposed hand posture classification method with CNN in sign language recognition. To achieve effective and efficient hand posture classification using depth data, a hand posture classification based on the Xception model and data augmentation for hand depth data has been

proposed. The proposed data augmentation method using the 3D rotation on depth data is effective in generating various appearances of the hand posture and increasing classification accuracy on the manually obtained dataset. On the other hand, the Xception model, which is one of the state-of-the-art CNN models, is applied to hand posture classification. Furthermore, the proposed method has been evaluated and compared with state-of-the-art researches.

Chapter 4 presents the development of an automatic food constituent estimation method from food images using multi-task CNN. The research focuses on the recognition of food categories and the estimation of calories and salinity. First, a new food image dataset has been constructed by using public images from several recipe-gathering websites because there is no large food image dataset with detail information on calorie and salinity. However, the number of food images with calorie and salinity obtained from the Internet is not sufficient for the effective learning of CNN. In order to address this issue, two-stage transfer learning using a large number of food categories recognition was proposed. The effectiveness of calories and salinity estimation using multi-task learning with food category classification by defining both calorie and salinity estimation as a regression problem is demonstrated. Here, the relationship between the food category and salinity is also experimentally shown by using multi-task CNN.

Chapter 5 provides the conclusions of the overall system and comments on some future possibilities to improve the system.

主業績

No.1	
論文題目	Food Constituent Estimation for Lifestyle Disease Prevention by Multi-task CNN
著者名	Sulfayanti F. Situju, H. Takimoto, S. Sato, H. Yamauchi, A. Kanagawa, and A. Lawi
発表誌名	Applied Artificial Intelligence, Vol. 33, Issue 8, pp. 732-746, (2019)
No.2	
論文題目	Hand Posture Classification of Augmented Depth Image Data Using a Convolutional Neural Network
著者名	Sulfayanti F. Situju, H. Takimoto,H. Yamauchi, and A. Kanagawa
発表誌名	Journal of The Japan Society for Welfare Engineering, Vol. 21, No. 2, pp. 38-46 (2019)

副業績

No.1	
論文題目	Hand Posture Classification on RGB-D Images with Convolutional Neural Network
著者名	Sulfayanti F. Situju, H. Takimoto, H. Yamauchi, and A. Kanagawa
発表誌名	Proc of the 18th Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS2017), ID165, C1-32, (2017)
No.2	
論文題目	Visual Saliency Estimation Based on Multi-task CNN
著者名	H. Takimoto, S. Katsumata, Sulfayanti F. Situju, A. Kanagawa, and A. Lawi
発表誌名	Proc. of The Fourteenth International Conference on Industrial Management (ICIM2018), (2018)

論文審査結果の要旨

本論文は，社会問題化している福祉・介護の現場における慢性的な労働者不足対策のひとつとして，自宅や施設等において高齢者や要介護者の様々な活動を支援するスマートホームを実現することを最終目標とし，深層学習に基づくモニタリング技術の提案を行っている．深層学習に基づく畳み込みニューラルネットワーク（CNN）は，多くのコンピュータビジョンに関する課題を解決する最も効果的なアーキテクチャである．一方で，複雑な構造を持つ CNN が高精度な認識を達成するためには，訓練用データセットとして膨大な量の画像データを必要とする．しかし，この膨大な量の訓練用データの収集に要する莫大なコストが CNN 活用における大きな課題である．

本論文では，上記の深刻な問題を解決するため，限られた小規模サイズの訓練用データを用いた場合であっても，手形状認識と料理画像からの成分推定を効果的に実現する技術を提案している．具体的な本論文の成果は以下の 2 点に要約される．

（1）手形状認識は，手話・ジェスチャー認識の重要な要素技術の１つとして古くから注目されている．本論文では，RGB-D センサにより三次元情報として取得した小規模な手形状データセットに対して三次元回転を適用することにより，様々な外観を有する多くの手形状データを仮想的に生成する手法を提案している．また，最先端の CNN アーキテクチャに対して仮想的に生成された大規模手形状データセットを適用することにより，高精度な手形状認識が可能であることを示している．加えて，この三次元手形状データに特化したデータ拡張法を実現するため，三次元深度画像から手領域のみを自動的に検出・セグメンテーションする手法も提案している．評価実験により，各種手法の有効性を示している．

（2）撮影された料理画像から手軽に栄養情報を取得できる環境の実現に向けて，マルチタスク CNN を用いて料理画像から複数の食品成分を自動的に推定する方法を提案している．特に，過剰なカロリーと塩分の摂取が重大な健康リスクを引き起こすことから，料理カテゴリ認識とカロリー量と塩分量の推定を対象としている．具体的には，Xception アーキテクチャに基づくマルチタスク CNN を提案しており，高精度な 2 種類の食品成分の同時推定を実現している．また，料理カテゴリ認識のため大規模な画像データベースに基づく 2 段階転移学習を提案しており，カロリー量と塩分量情報が紐づいた小規模なデータセットのみを用いた場合であっても高精度なカロリー量と塩分量の推定が可能であることを示している．

第 1 章では，現在の高齢者・要介護者の社会環境，各種センサに基づくモニタリング技術の現状を説明し，本研究の意義と各章の内容について概説している．第 2 章では，深層学習に基づく CNN とその関連技術について詳細が述べられている．第 3 章と第 4 章では，それぞれ手形状の自動認識システムと料理画像に対する成分推定システムの詳細がまとめられている．第 5 章では，本論文の結論がまとめられている．なお，予備審査で指摘された博士論文題目や今後の課題について適切に修正されていることを確認した．

以上の結果より，本論文の内容は，学術的，実用的価値が極めて高いものと判断し，本学位論文審査委員会は博士（工学）の学位論文として価値あるものと認める．