Multimodal Visual Perception for Real-World Scene Parsing

中嶋, 一斗

（別紙様式 2）

氏　　名　：中嶋　一斗

論 文 名　：Multimodal Visual Perception for Real-World Scene Parsing
（実世界シーン認識のためのマルチモーダル視覚に関する研究）

区　　分　：甲

論　文　内　容　の　要　旨

知能ロボットが人を取り巻く一般環境に適応するには，カメラ等の視覚センサからシーンを撮像し，得られた画像からその特徴や実世界の構成物体について記述する画像認識技術が極めて重要である．近年は，ウェブ上の大規模データを活用した深層学習の台頭により，精度の大幅向上やタスクの多様化等，当分野の進展は著しい．一方，人間は視覚だけではなく触覚・味覚・聴覚・嗅覚等のマルチモーダルな共起感覚情報を取り入れることで，モダリティごとの曖昧性を解消し，より正確で抽象度の高い認知処理を実現している．画像認識分野においても人間の知覚機能に倣い，異種センサを組み合わせることで推論の頑健性を高める試みが進められてきた．本論文では，特に視覚センサの計測物理量と視点に伴う 2 種類のモダリティに着目して，マルチモーダル視覚情報に基づく新たなシーン認識問題と深層学習の枠組みによるその解法を提案する．

まず,視覚センサの計測物理量に関して,マルチモーダル視覚情報を利活用する手法を提案する．近年，Microsoft 社 Kinect に代表される低価格な RGB-D カメラが登場し，シーンの 3D 幾何情報を手軽に計測できるようになった．それに伴い，RGB 画像と距離画像を融合するマルチモーダル手法も数多く提案されてきた．しかし，屋外のような照明条件が大きく変化するシーンを対象とした場合，夜間や雨天時において輝度特徴の消失やパターン多様化による精度低下が課題となる．そこで，本論文では，照明条件に頑健な 3D LiDAR センサから得られるマルチモーダル画像情報を用いた屋外環境識別手法を提案する．3D LiDAR センサは，数万点のレーザ測距により全方位の 3D 幾何情報を計測するアクティブセンサであり，各測距点に対して対象物体までの距離値とレーザ反射強度が得られる．本手法ではこれらをパノラマ画像組として表現し，多層ニューラルネットで屋外環境シーンの識別モデルを構築する．実験では，入力モダリティや内部融合方法を変えた種々のモデルを構築し，大規模データセットを用いた精度検証およびモデルの内部解析により提案手法の有効性を示す．

次に，視覚センサの視点に関する手法を提案する．日常生活の状態・体験をタグ付きの画像で記録したものをビジュアルライフログと呼び，生活リズムの把握や視覚記憶の補助に応用できる．ウェアラブルカメラを使って一人称視点画像を記録するアプローチが主流であるが，カメラ装着者自身の状態やコンテキストを記録できない課題があった．一方，外部固定した客観的視点を用いると記述対象の全容を撮影できるが，ライフログ用途では頻繁なオクルージョンや解像度低下が問題となる．そこで．撮影視点に伴う可視領域の相補性に着目し，複数人称視点の融合によるタグ生成の高精度化を試みる．特に，分散センサを配備した人・ロボット共生空間を想定し，人の一人称視点・ロボットの二人称視点・環境固定カメラの三人称視点を用いたシーン説明文生成を検証タスクとする．本論文では，映像組と参照説明文のデータセット構築および複数画像を統合するモデルの新規

提案を行い，データセットを用いた人称視点のアブレーション評価により提案手法の有効性を示す．

本論文は 4 章から構成される．第 1 章は序論であり，本研究の背景と目的について述べる．第 2 章では，3D LiDAR センサから得られる距離値・反射強度に基づく屋外環境識別手法について述べる．第 3 章では，人・ロボット共生空間の複数人称視点に基づく説明文生成手法について述べる．第 4 章では，本研究で得られた結果を総括し，今後の展望について述べる．