氏　　名	Cao Lu
博士の専攻分野の名称	博士（学術）
学位記号番号	博理工甲第 919 号
学位授与年月日	平成 25 年 9 月 20 日
学位授与の条件	学位規則第 4 条第 1 項該当
学位論文題目	Comprehending Spatial Relations for Interactive Object Recognition （対話物体認識のための空間位置関係理解）
論文審査委員	委員長　教　授　久野　義徳
	委　員　教　授　前川　　仁
	委　員　教　授　島村　徹也
	委　員　准教授　小室　　孝

論文の内容の要旨

Human can effortlessly express their spatial experience and talk about *where* objects are located in relation to some underlying objects. Since it is impossible for us to learn all objects, such information has been critical to explore the visual world. Intuitively, if we know where the objects are, recognizing them will become easier. In computer vision, in order to mimic human's ability, an important and open problem is to endow robotic systems the ability to comprehend spatial relations as humans do. This is somewhat like a school child does when learning to write a descriptive sentence, such as the CD is to the left of the book.

The primary goal of this work is to design and demonstrate spatial recognition methods to bridge the gap between visual information and human cognition. Towards this goal, we treat spatial relations as a kind of feature as well as other visual features, such as *color*, *size*, etc and have developed computational templates to represent spatial relations. We propose a novel model to encode linguistic spatial expressions.

We first investigate how humans manipulate space by the action of natural language and classified basic class of relation. We then extend the observations from cognitive systems to computer vision applications. We propose templates for recognizing spatial relation, translating linguistic expression into visual information, and representing spatial terms in an angular fashion. The templates have been tested over 720 scenarios where 1-3 *unknown* objects within.

Comprehending spatial relations is beyond simply distinguishing them. It is noticeable that spatial relation needs a pair of objects. In determination of different classes of relation, the underlying objects which are named as reference objects play a decisive role. Concretely, objects like *humans, animals* and *computer displays* are somewhat different from objects like *balls, boxes, cups* in that they have intrinsic *front* side. The former's *front* is independent from interlocutors' viewpoint whereas the latter's is not. It turns out that the front orientation adjacent to the frontal-side of those objects are transformed accordingly if they are rotated from the frontal view. We then focus on introducing an estimation model for those objects, from estimating pose transformations, to adjusting *intrinsic-front* orientation. The first step studies one prominent type of pose variation given viewpoint transformation in supervised fashion. Naïve Bayesian classifier is followed for prediction. The estimator performs highly competitively with the state of the arts on

the ETH-80 database, and an everyday-object database that we collected on our own.

The models profit from an interactive interface, which is developed to understand some simple English words and grammatical structures. Our models can make the interaction closer to the way of human-human interaction. Finally, we conduct experiments integrally within the system, which consists of an object detector, a spatial recognition model, a pose estimator and a user interface. The goal is recognizing *unknown* objects via comprehending spatial relations by interactive means. The simple yet effective models outperform in recognition tasks in the author's database.

論文の審査結果の要旨

当論文審査委員会は、当該論文の発表会を平成 25 年 7 月 25 日に公開で開催し、詳細な質問を行い論文内容の審査を行った。その論文発表を含む学位論文の審査の結果、本提出論文を博士（学術）の学位論文として合格と判定した。以下に審査結果の要約を示す。

本提出論文はコンピュータビジョン分野における物体認識に関するものである。物体認識は頑健な局所特徴量と統計的機械学習の利用により進展してきているが、まだ、どのような環境でも確実に多くの物体を認識できるというような状況ではない。そこで、自動での認識に失敗したら人間に関連情報を教えてもらい認識を試みるという対話物体認識が検討されている。対話物体認識では、物体の色や形といった属性を使うものが検討されているが、本論文は空間位置関係について検討している。人間同士でも、「その本の左にあるCD」というように、目的の物体を他の物体との位置関係で表現することは多い。しかし、人間の言語による空間位置関係の表現は、物体に応じて変わるなど複雑である。空間位置関係に基づき対話物体認識を行うためには、この複雑な人間の言語表現を正しく解釈し、さらに、その表現に対応する情報を画像から得る方法を検討しなければならない。本論文は以上の問題を検討し、対話物体認識のための空間位置関係の理解法を提案するものである。

本論文は 6 章からなるが、まず、第 1 章では、上述のような研究の背景と目的を述べている。そして、心理学、言語学、哲学における空間位置関係についての関連研究を調査してまとめている。さらに、ロボット分野での関連研究について述べている。

第 2 章では、言語学等の知見に基づき、対話物体認識のための空間位置関係表現のモデルを提案している。人間の自然言語による空間位置関係表現は多様であり、また、言語により異なる部分もあるが、主要なものとして、intrinsic frame of reference を用いるものと relative frame of reference を用いるものがある。前者は人間やテレビのように固有の方向が定まった物体を基準にして、他の物体の位置を表現するものである。例えば、テレビの画面の前方にあるものに対しては、そのシーンを撮影した画像上で、その物体がテレビの右に写っていても、左に写っていても「テレビの前」と表現する。後者は、そのような固有の方向のない物体、あるいは固有の方向がある物体でも、その固有の方向以外の方向について言及する場合に用いられる。例えば、「そのコップの右」というような表現である。この場合は、物理的に同じ空間位置関係でも、発話者の視点が変われば、言語表現が変わる場合がある。本論文では、最初に、簡単なモデルとして、画像上での位置関係から、空間位置関係の表現を定める 2 次元計算モデルを提案し、それに基づく自動判定と人間の被験者による表現の比較実験を行った。その結果、物体間の距離等を考慮していないので、自動判定に誤りが起こる場合があることが分かった。そこで、物体の 3 次元位置関係を考慮した 3 次元計算モデルを提案し、実験により有効性を確認した。さらに、論文では複数の物体がグループになっている場合の空間位置関係の表現についても検討している。これには、グループの中の物体をグループ全体の中の位置で表現する場合とグループ全体を基準物体として他の物体の位置を表現する場合があることを見出し、モデルを提案している。

第 3 章では、物体の姿勢推定を単眼画像から行う方法を検討している。これは、intrinsic frame of reference を用いた表現を解釈する場合には、基準となる物体の方向を画像から求めることが必要なためである。PHOG(Pyramid Histogram of Oriented Gradients) 特徴を用いた方法を提案し、実験により既存の方法と同程度の識別結果をより高速に得られることを確認している。

第 4 章では、実験に必要なデータベースの構築について述べている。物体認識の研究では性能の比較のために実験用の画像を集めたデータベースがいくつか公開されている。しかし、それらは単独の物体認識用なので、画像中に位置関係がポイントになるような複数の物体が写っていない。そこで、空間位置関係の実験

用のデータベースを開発した。まず、姿勢検出等の基礎実験用に、24 種類の 122 個の物体について各 42 方向からの画像、計 5,124 枚を集めた。次に、それらの物体から 5 個程度までを選んで、いろいろに配置した、400 程度の画像を集めてデータベースを構築した。

第 5 章では、これまでの成果をもとに、実際に空間位置関係について対話を行い物体を認識するシステムを提案している。そして、第 4 章で構築したデータベースを用いて、対話により目標物体が検出できることを検証している。

最後に第 6 章で、論文の内容をまとめ、今後に残された課題を議論している。

以上のように、本論文の内容は、学術的に意義のある研究であると評価できる。よって、当学位論文審査委員会は、本論文を博士（学術）の学位論文として合格と判定した。