氏　　名	DIPANKAR DAS
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 813 号
学位授与年月日	平成 22 年 9 月 17 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	Multiple Object Detection and Localization Using Generative and Discriminative Models （生成的モデルと識別的モデルを用いた複数物体の検出と位置決め）
論文審査委員	委員長　教　授　久野　義徳 委　員　教　授　池口　　徹 委　員　教　授　前川　　仁 委　員　准教授　内田　淳史

論文の内容の要旨

Multiple object detection and localization with large appearance variation is a fundamental problem in computer vision. The appearance of object categories can change due to intra-class variability, viewpoint changes, and illumination variations. In this thesis, we propose an integrated approach of generative and discriminative models to explore the problem of multiple object detection and localization task by introducing an efficient hypothesis generation method and using an appropriate combination of features.

We first investigate whether the integrated approach of generative and discriminative models is beneficial for the task of multiple object detection and localization. In the generative stage, the probabilistic latent semantic analysis (pLSA) model is fitted to the training data without knowledge of labels of bounding boxes, and topics are assigned based on the object specific topic probability under each category. In the testing stage, an algorithm is proposed and implemented to efficiently generate promising hypotheses for multiple object categories with their positions and scales. For this purpose, our algorithm considers the bag-of-visual-words (BOVW) extracted from the image and the number of topics generated during the learning stage. Then an initial region of interest (ROI) containing all visual words belonging to a topic is defined based on the maximum topic specific word probability and is searched for final probable object's locations (promising hypotheses). The initial ROIs reduce the search space and speed up the hypothesis generation stage. Once hypotheses have been generated, a discriminatively trained SVM (Support Vector Machine) classifier verifies these hypotheses using merging features. Since the hypothesis generation stage effectively acts as a pre-filter, the discriminant power is applied only where it is needed. Thus, our system is able to detect and localize multiple objects with a large number of categories. In the post-processing stage, environmental context information along with the probabilistic output of the SVM classifier is used to improve the overall performance of the system.

We then propose a new sub-category optimization approach that automatically divides an object category into an

appropriate number of sub-categories based on appearance variation. In our flexible learning strategy, a single object category can be represented with multiple topics (sub-categories) and the model can be adapted to diverse object categories with large appearance variations. Instead of using a predefined intra-category sub-categorization based on domain knowledge, we divide the sample space by unsupervised clustering based on discriminative image features. Then the clustering performance is verified using a sub-category discriminant analysis. Based on the clustering performance of the unsupervised approach and sub-category discriminant analysis results, we determine the optimal number of sub-categories per object category. Furthermore, we employ the optimal sub-category representation as the basis and a supervised multi-category detection system with X^2 merging kernel function to efficiently detect and localize object categories within an image.

Finally, we provide a new scale invariant feature descriptor to locate multiple object categories in 3D range images with depth information. For this purpose, the fragmented local edgels (*key-edgel*) are efficiently extracted from a 3D edge map by separating them at their corner points. The 3D edge maps are reliably constructed by combining both boundary and fold edges of 3D range images. Each key-edgel is described using our scale invariant descriptors that encode local geometric configuration by joining the edgel to adjacent edgels at its start and end points. Using key-edgels and their descriptors, our model generates promising hypothetical locations in the image. These hypotheses are then verified using more discriminative features.

Our proposed models and methods for object detection and localization can be applied for service robot applications. In particular, we focused on detecting multiple objects in an image of a scene in different environment (office, kitchen etc.) and finding their exact locations within an image with depth information. We present an extensive experimental evaluation involving authors' own and standard databases. The system has shown the ability to accurately detect and localize many objects even in the presence of cluttered background, substantial occlusion, and large appearance changes. The experimental results demonstrate that the hypothesis generation algorithm is able to generate nearly accurate hypotheses for all objects. The SVM verification stage, on the other hand, uses the merging features and category specific weighted merging features to enrich the performance of the system. Finally, the environmental context information in the post-processing stage compensates for ambiguity in an object's visual appearance.

From experimental result on 15 specific and 10 categories of objects on our database, we can conclude that the merging features improve the average detection and localization rate by approximately 8% more than the single feature. The weighted merging feature increases the detection rate by 7% more than the merging features on both specific and categories of objects. Finally, the context information increases the detection results by 5% and 4% more than weighted merging features on specific and categories of objects, respectively. The performance of our system is compared to other three methods on four categories of standard databases (car, motorbike, cow, and horse). In all cases except car category our method performs better than the other three methods. We have also compared the performance on MIT-CSAIL database on three object categories (computer screen, keyboard, and bookshelf). In this case, we have achieved superior performance on *computer screen* and *bookshelf* categories than the state-of-the-art. Our better performance compared to others could be due to the integration of both generative and discriminative classifiers with feature combination instead of using only the generative model.

For object categories with large appearance changes, our current system automatically sub-categorizes an object

category to the appropriate number to generate more accurate hypotheses within an image. The system is able to discriminate among diverse object categories using feature specific X^2 merging kernel with both shape and appearance features. The sub-category optimization technique increases the detection and localization performance by 17% on our database. Adding the background features as an additional category, the system is able to improve more 5.3%. In all cases, our feature specific X^2 merging kernel improves the detection and localization accuracy by approximately 2% more than the rbf kernel of the SVM classifier. We have compared the performance of our automatic sub-categorization method to other two related approaches on ETH-80 multi-view and ETHZ shape databases. With sub-category optimization, the system produces average classification result of 84.7% on 8-category ETH-80 multi-view database. Our average classification result is better than the other two approaches. Averaged over all 5 categories on ETHZ shape database, we improve the detection and localization performance by 12.3% and 4.3% more than the other two approaches, respectively. The superior performance compared to others could be due to the use of the optimized topic model with the sub-categorization technique and the use of shape and appearance features with X^2 merging kernel.

Finally, we have shown the usefulness of local edgel geometry for multiple object category detection and localization from range images. Our scale invariant local edgel descriptors are robust for partial occlusion, background clutter, and significant scale changes. The detection rate and computational efficiency suggest that our technique is suitable for real time use. The method is useful for service robot applications because it can use 3D information to know the exact position and pose of the target objects.

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 21 年 7 月 22 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の分野における物体認識に関するものである。物体認識の研究は、近年、インターネットにおける画像検索の応用を想定して発展してきているが、この研究はロボットの視覚としての物体認識を目指すものである。ネット検索の場合は、検索対象の物体が画像中にあるかないかが分かればよいが、ロボット視覚の場合は、ロボットの周りにある多くの物体を認識し、その位置を求める必要がある。また、ネット検索の場合は、対象物はだいたいの場合、一般的な視点から見た画像であり、見え方の変化は比較的小さい。それに対し、ロボット視覚の場合は対象物を様々な方向から見る可能性がある。したがって、ネット検索の場合と同様に同一カテゴリー内の多様な物体の認識に対応するとともに、視点変化による見え方の変化にも対処する必要がある。本論文は、以上のような問題を解決し、介護ロボットなどのサービスロボットに有効な物体認識法を提案するものである。

本論文は 5 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。

第 2 章では、提案する物体認識の枠組みについて述べている。提案の物体認識法は、ネット検索用の物体認識で注目されている自然言語処理とのアナロジーから生まれた生成的な物体認識法と、パターン認識の分野で中心的に研究されてきた識別的な物体認識法を組み合わせたものである。前者は、物体の画像特徴を「単語」、そして物体を「話題」と考えて、こういう複数の「単語」（画像特徴）が現れるのは、こういう「話題」（物体）の文書であろうということを学習しておくものである。多様な物体の認識に対応できるが、物体がないのに誤って物体を検出してしまうことが多い。後者は指定の物体を他の物体と識別するモデルを学習するもので、認識性能は高いが、多様な物体に対応しようとすると、認識に非常な時間を要する。そこで、本論文では前者を基にした方法により、画像中に存在する可能性のある多くの物体の候補を検出し、その候補に対して後者の方法に基づく方法を適用することにより、高速に、多数の物体を、その画像上の位置情報も含めて認識する方法を提案している。物体認識研究で用いられる標準的なデータセットによる実験で、既存の他の手法と比べて同等以上の認識性能が得られることを示している。さらに、標準的なデータセットにはサービスロボットが対象とするようなシーンの画像が少ないため、複数の日常用品を複雑に配置した標準的なデータセットの画像より認識が難しいと思われるシーンの画像を準備し、実験により有効性を実証している。

第 3 章では、前章の枠組みの中で、さらに良好な認識を行うために、物体を適切な数のサブカテゴリーに分けたモデルを自動生成する方法を提案している。物体認識で難しいのは見え方の変化にどう対応するかである。同一の物体でも見る方向により見え方は変わるし、同じ名前の物体でも、様々な形のものがある場合がある。これらを一つの物体のモデルで扱おうとしても、認識性能が上がらない。そこで同一カテゴリーの物体をいくつかのサブカテゴリーに分けて認識モデルを作ることが考えられる。このようなサブカテゴリーへの分割につては、これまでにも研究はあるが、サブカテゴリー数をあらかじめ与える場合が多く、いくつのサブカテゴリーに分けるのがよいかまで進んだ研究はなかった。本論文では、この問題に対し、物体認識はカテゴリー名称がついた画像からの教師あり学習だが、そこに教師なし学習を取りこんで、最適なサブカテゴリー分割を行う方法を提案している。これについても標準的な画像と自分で集めた画像に対して実験を行い、有効性を確認している。

第 4 章では、第 2 章の枠組みをレーザレンジファインダの距離画像に適応する方法を提案している。最近、リアルタイムで距離画像の得られるセンサが利用できるようになってきているが、３次元の絶対的な距離も

得られるので、実際に物体に対して作業を行うロボットには都合がよい。そこで、これまでは濃淡画像を対象にしていたが、距離画像を対象とする場合についても検討を行っている。距離画像の場合に有効な画像特徴としてエッジ要素の接続関係を記述したものを用いる方法を提案し、実験で有効性を示している。

最後に第５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。