Rozpoznawanie obiektów trójwymiarowych na podstawie danych RGBD

Adam Kosiorek

September 22, 2013

Abstract

Abstrakt

1 Wstęp

Rozwój nowych technologii pozwala na gromadzenie oraz przetwarzanie bardzo dużych ilości danych. Aparaty i kamery cyfrowe przyczyniają się do powstawania milionów gigabajtów informacji każdego dnia. Co więcej, dobrej jakości urządzenia służące do obrazowania trójwymiarowego osiągnęły tak niskie ceny, że można je znaleźć w wielu gospodarstwach domowych. Jednym z nich jest *Microsoft Kinect* – korzystający z technologii *PrimeSense* — oparty o światło strukturalne sensor rejestrujący dane RGBD.

Nowe technologie przyczyniają się też do spadku cen robotów mobilnych. Można sobie wyobrazić, że w niedalekiej przyszłości zrobotyzowani asystenci zawitają w domach. Roboty takie będą musiały spełniać szereg wymagań związanych z wymogami bezpieczeństwa oraz wygodą użytkowania. W szczególności będą musiały być wyposażone w mechanizmy pozwalające na bezpieczną, a zarazem efektywną interakcję z otoczeniem — w tym z ludźmi.

Zapewnienie poprawnej interakcji z otoczeniem wymaga spełnienia wielu warunków. Między innymi są to:

- Interfejs człowiek-maszyna powinien pozwalać na swobodną komunikację z robotem.
- Posiadanie wiedzy na temat aktualnego położenia oraz dysponowanie mapą otoczenia - problem opisywany w literaturze jako *SLAM — Simultaneous Localization And Mapping*
- Zdolność semantycznej klasyfikacji miejsca, w którym robot się znajduje
- Zdolność semantycznej klasyfikacji pojedyńczych obiektów

Trzy ostatnie spośród wymienionych punktów są zupełnie naturalne dla ludzi. Od dnia narodzin stajemy przed zadaniem rozpoznawanie swojego otoczenia, miejsca w którym się znajdujemy i przedmiotów, z którymi mamy do czynienia. Czynności te wydają się nam bardzo łatwe — od samego początku dysponujemy bardzo dużą ilością danych dotyczących naszego otoczenia, których źródłem są nasze zmysły. Ponadto uczymy się od starszych jak zachowywać się w określonych miejscach, jak obchodzić się z konkretnymi przedmiotami. Maszyny zazwyczaj nie mają tak bogatych danych opisujących geometrię, kolor czy fakturę otaczających przedmiotów, nie mają też wiedzy nt. semantycznego znaczenia tych przedmiotów. Z tego powodu powstaje wiele algorytmów służacych do opisania otaczającego nas świata oraz takich, które uczą się rozpoznawać obiekty.

W przypadku semantycznej klasyfikacji otoczenia robota można zauważyć, że rejestracja wszystkich elementów otoczenia wykorzystując np. skaner laserowy będzie niemożliwa. Skanery takie mają ograniczony zasięg, a w przypadku przebywania w otwartej przestrzeni większość obiektów może znaleźć się poza zasięgiem skanera. Prowadzi to do utrudnienia bądź uniemożliwienia wykorzystania informacji przestrznnej do rozpoznawania otoczenia

robota. W przypadku semanycznej klasyfikacji obiektów sytuacja przedstawia się inaczej — sam fakt wystąpienia problemu dowodzi, że obiekt znalazł się w zasiegu sensorów. Możliwa jest więc rejestracja chmury punktów opisująca badany przedmiot oraz wykorzystanie informacji przestrzennej w celu klasyfikacji.

1.1 Cel pracy

Celem niniejszej pracy jest napisanie aplikacji służącej do semantycznej kategoryzacji obiektów trójwymiarowych. Kategoryzacja powinna odbywać się w oparciu o dane RGBD oraz wykorzystywać reprezentację Bag of Words. Aplikacja powinna działać w czasie rzeczywistym, a jej skuteczność powinna pozwalać na wykorzystanie jej w rzeczywistych robotach mobilnych.

1.2 Zakres pracy

W pracy zostały przyjęte następujące założenia projektowe:

- System operuje na danych RGBD
- Analizowane obrazy powinny zostać przygotowane w ten sposób, że rozpoznawany obiekt powinien wypełniać ponad połowę powierzchni zdjęcia
- Kategoryzacja obiektów odbywa się z pominięciem segmentacji obrazu.

Do realizacji projektu wykorzystano biblioteki: *OpenCV, PointCloudLibrary, Boost.* Wszystkie wykorzystywane algorytmy przetwarzania obrazu oraz uczenia maszynowego zostały zaimplementowane przez osoby trzecie, a w większości pochodzą z wymienionych bibliotek.

Zakłada się, że zostaną wykorzystane implementacje algorytmów uwzględniające wielowątkowość, w celu przyśpieszenia działania programu. Ponadto możliwe jest wykorzystanie platformy CUDA w celu porównania wydajności i dalszego przyśpieszenia obliczeń.

2 Bag of Words image representation

As portreyed in [2] the Bag of Words or BoW model has originated from the text retrieval domain. The original use enabled a vector-like representation of a text document. One of the most simple cases would be to, given a dictionary, construct a histogram depicting the incidence of words in a particular text document. Such an approach obliterates any grammatical dependencies in order to retain only the statistical information associated with each word. It is believed that the words' frequencies are connceted to the semantic meaning of the document.

The following questions emerge:

- How does this apply to computer vision?
- How can one convert an image to a text document?

In order to answer the first question one has to consider how computers process data. Digital images are nothing but streams of binary code. A simple conversion to an RGB format makes the data structured and is enough for people to understand their contents. Unforunately, such image representation is extrmelly vulnerable to translation and rotation as well as changing lighting conditions. One way to address these disadvantages would be to compute locations of some characteristic points (keypoints) and describe them somehow. Even then, however, the resulting represenation is low level and incomprehensible for humans. Defining robot's behaviour basing on such data might prove cumbersome, for there are too many details to consider. All of the above is reffered to as a semantic gap. It is defined by Tsai, with respect to the Content Based Image Retrieval or *CBIR*, as *'the gap between the extracted and indexed low-level features by computers and the high-level concepts (or semantics) of user's queries'*. Suppose a text document can be created from an image.

If this is the case, then the document can be converted into a BoW model. Since the latter contains information about the semantic significance of the image an impact of the semantic gap can be reduced.

As for the second question, there is a quite well established pipeline that enables creation of text documents from images. The steps of the BoW methodology are as follows:

- Keypoint detection — keypoints are local interest points or regions. They are usually computed in such a way so as to provide scale and location invariance. Rigid transformation and illumination invariance would be desired but it somewhat harder to achieve
- Keypoint description — each keypoint have to be described in a manner that distinguish the particular keypoint in some way.
- Vector quantization — clustering algorithms are used to find regions in the high dimensional space of keypoint descriptors. When the clusters are found, each described keypoint can be assigned to a corresponding cluster. Then, an image can be represented by the number of clusters, to which the image's keypoing were assigned. The clusters' numbers are called visual words.

3 Literature review

Scene and objet categorization, according to [1] requires employment of computer vision as well as machine learning algorithms. The authors focus on a pipline incorporating a bag of words image representation followed by a usage of a classification algorithm.

4 Proponowane podejście

W poniższej pracy wykorzystuje się podejście Bag of Words (BoW) połączone z algorytmami uczenia nadzorowanego. Model BoW ma tę zaletę w stosunku do surowych zdjęć lub chmur punktów, iż zmniejsza tzw. semantic gap - lukę znaczeniową pomiędzy zdjęciem lub niskopoziomowymi cechami charakteryzującymi obraz a wysokopoziomowy koncepcjami, nadającymi zdjęciu znaczenie semantyczne np. zachód słońca, człowiek.

5 Opis zastosowanych rozwiązań

6 Eksperymenty

7 Podsumowanie

References

[1] J. Ponce, D. Forsyth, E.-p. Willow, S. Antipolis-Méditerranée, R. d'activité RAweb, L. Inria, and I. Alumni. Computer vision: a modern approach. *Computer*, 16:11, 2011.

[2] C.-F. Tsai. Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*, 2012, 2012.