Proseminar Objektposenschätzung

Robert Jeutter

3. November 2021

Damit ein Roboter einen Gegenstand greifen kann, ist es meist notwendig die genaue Lage des Objektes zu kennen. Dies kann sowohl über klassische Verfahren als auch über Deep-Learning-Verfahren erreicht werden. Ziel dieses Seminars ist es den Stand der Technik für die Objektposenschätzung aufzuarbeiten und vorzustellen. Der Fokus sollte dabei auf Verfahren liegen, bei denen zuvor kein Objektmodell benötig wird, sodass auch die Lage von unbekannten Objekten geschätzt werden kann.

1 Motivation

Die Erkennung von Objekten und die Schätzung ihrer Lage in 3D hat eine Vielzahl von Anwendungen in der Robotik. So ist beispielsweise die Erkennung der 3D-Lage und Ausrichtung von Objekten wichtig für die Robotermanipulation. Sie ist auch bei Aufgaben der Mensch-Roboter-Interaktion nützlich, z. B. beim Lernen aus Demonstrationen. In einigen Fällen kann dies durch Vorwärtskinematik erreicht werden, wobei angenommen wird, dass die Bewegung des Objekts der Bewegung des Endeffektors entspricht. Häufig reicht die Vorwärtskinematik jedoch nicht aus, um die Lage des Objekts genau zu bestimmen. Dies kann durch Schlupf beim Greifen oder bei der Manipulation mit der Hand, bei der Übergabe oder durch die Nachgiebigkeit eines Saugnapfes bedingt sein. In diesen Fällen ist eine dynamische Schätzung der Objektposition aus visuellen Daten wünschenswert. Methoden zur 6D-Positionsschätzung aus Einzelbildern wurden ausgiebig untersucht. Einige von ihnen sind schnell und können die Pose für jedes neue Bild von Grund auf neu schätzen. Dies ist jedoch redundant, weniger effizient und führt zu weniger kohärenten Schätzungen für aufeinanderfolgende Bilder. Andererseits kann die Verfolgung von 6D-Objektposen über Bildsequenzen bei einer anfänglichen Posenschätzung die Schätzgeschwindigkeit verbessern und gleichzeitig kohärente und genaue Posen liefern, indem die zeitliche Konsistenz genutzt wird.
Traditionell wird das Problem der Objektposenschätzung durch den Abgleich von Merkmalspunkten zwischen 3D-Modellen und Bildern angegangen. Diese Methoden setzen jedoch voraus, dass die Objekte reichhaltig texturiert sind, um Merkmalspunkte für den Abgleich zu erkennen. Daher sind sie nicht in der Lage, mit Objekten ohne Textur umzugehen. Die meisten bestehenden Ansätze zur Objektposenschätzung setzen den Zugriff auf das 3D-Modell einer Objektinstanz voraus. Der Zugang zu solchen 3D-Modellen erschwert die Verallgemeinerung auf neue, unbekannte Instanzen. Darüber hinaus erfordern 3D-Modelldatenbanken oft einen nicht unerheblichen manuellen Aufwand und Expertenwissen, um sie zu erstellen, wobei Schritte wie Scannen, Netzverfeinerung oder CAD-Design erforderlich sind. Zusätzliche Komplexität einer Szene, die durch Unordnung und Verdeckungen zwischen den Objekten verursacht wird, senkt die korrekte Erkennung bei modellbasierten Verfahren deutlich. Bei schablonenbasierten Methoden wird eine starre Schablone konstruiert und verwendet, um verschiedene Stellen im Eingabebild zu scannen. An jeder Stelle wird ein Ähnlichkeitswert berechnet, und die beste Übereinstimmung wird durch den Vergleich dieser Ähnlichkeitswerte ermittelt. Schablonenbasierte Methoden sind nützlich für die Erkennung texturloser Objekte. Sie können jedoch nicht sehr gut mit Verdeckungen zwischen Objekten umgehen, da die Vorlage einen niedrigen Ähnlichkeitswert hat, wenn das Objekt verdeckt ist. Alternativ dazu können Methoden, die eine Regression von Bildpixeln auf 3D-Objektkoordinaten erlernen, um 2D-3D-Korrespondenzen für die 6D-Positionsschätzung herzustellen, nicht mit symmetrischen Objekten umgehen. Zudem können bei der Verfolgung durch solche dynamische on-the-fly Rekonstruktion von Objekten Fehler entstehen, wenn Beobachtungen mit fehlerhaften Posenschätzungen in das globale Modell einfließen. Diese Fehler wirken sich nachteilig auf die Modellverfolgung in nachfolgenden Bildern aus.
Motiviert durch die oben genannten Einschränkungen, zielt diese Arbeit auf eine genaue, robuste 6D-Objekterkennung ab, die auf neuartige Objekte ohne 3D-Modelle verallgemeinert werden kann.

2 Anforderungen

Für verschiedene Anwendungszenarien bestehen unterschiedliche Anforderungen und Möglichkeiten um ein bestimmtes Verfahren verwenden zu können. Um eine schnelle Übersicht über die Verfahren geben zu können wird jedes in mehreren Kategorien eingeteilt und verglichen.
Die Kategorisierung aller Verfahren erfolgt nach folgendem Schemata

Objektmodelle müssen für das Training oder Nutzung Objektmodelle (2D, 3D, CAD), Schablonen-Modelle oder merkmalsbasierte Modelle vorhanden sein?

Video-Input verarbeitet das Verfahren 2D Bilder, 3D Bilder mit Tiefenwahrnehmung und kann die Position der Kamera verändern?

genutzte Datensätze mit welchen Datensätzen wurde das Verfahren trainiert oder getestet?

Genauigkeit Wie akkurat ist die Objektposenschätzung im Vergleich?

Ressourcenintensivität Wie Ressourcenintensiv ist das Verfahren und werden spezielle Hardware benötigt?

Laufzeit mit welcher Geschwindigkeit ist die Verarbeitung von Eingabedaten möglich und stabil?

3 Verschiedene Verfahren

3.1 BundleTrack[WB21]

Objektmodelle
Video-Input
genutzte Datensätze
Genauigkeit
Ressourcenintensivität
Laufzeit

BundleTrack ist ein Framework für die 6D-Positionsverfolgung neuartiger Objekte, das nicht von 3D-Modellen auf Instanz- oder Kategorieebene abhängt. Es nutzt komplementären Eigenschaften für die Segmentierung und robuste Merkmalsextraktion sowie die speichererweiterte Pose-Graph-Optimierung für die räumlich-zeitliche Konsistenz. Dies ermöglicht eine langfristige, abdriftarme Verfolgung in verschiedenen anspruchsvollen Szenarien, einschließlich erheblicher Verdeckungen und Objektbewegungen.
Im Vergleich zu modernen Methoden, die auf einem CAD-Modell der Objektinstanz basieren, wird eine vergleichbare Leistung erzielt, obwohl die vorgeschlagene Methode weniger Informationen benötigt. Eine effiziente Implementierung in CUDA ermöglicht eine Echtzeitleistung von 10 Hz für das gesamte System. Der Code ist verfügbar unter: https://github.com/wenbowen123/BundleTrack

3.2 DeepIM[Li+18]

Objektmodelle
Video-Input
genutzte Datensätze
Genauigkeit
Ressourcenintensivität
Laufzeit

3.3 MaskFusion[RBA18]

Objektmodelle
Video-Input
genutzte Datensätze
Genauigkeit
Ressourcenintensivität
Laufzeit

3.4 Neural Analysis-by-Synthesis[Che+20]

Objektmodelle
Video-Input
genutzte Datensätze
Genauigkeit
Ressourcenintensivität
Laufzeit

3.5 6-PACK[Wan+19]

Objektmodelle
Video-Input
genutzte Datensätze
Genauigkeit
Ressourcenintensivität
Laufzeit

3.6 PoseCNN[Xia+17]

Objektmodelle
Video-Input
genutzte Datensätze
Genauigkeit
Ressourcenintensivität
Laufzeit

Ein neues Convolutional Neural Network für die 6D-Objektposenschätzung. PoseCNN schätzt die 3D-Verschiebung eines Objekts, indem es sein Zentrum im

Bild lokalisiert und seinen Abstand zur Kamera vorhersagt. Die 3D-Rotation des Objekts wird durch Regression auf eine Quaternion-Darstellung geschätzt. Dabei führt man eine neue Verlustfunktion ein, die es PoseCNN ermöglicht, symmetrische Objekte zu behandeln. Erreicht Ende-zu-Ende 6D Posenschätzung und ist sehr robust gegenüber Verdeckungen zwischen Objekten.
PoseCNN entkoppelt die Schätzung von 3D-Rotation und 3D-Translation. Es schätzt die 3D-Verschiebung durch Lokalisierung des Objektzentrums und Vorhersage des Zentrumsabstands. Durch Regression jedes Pixels auf einen Einheitsvektor in Richtung des Objektzentrums kann das Zentrum unabhängig vom Maßstab robust geschätzt werden. Noch wichtiger ist, dass die Pixel das Objektzentrum auch dann wählen, wenn es von anderen Objekten verdeckt wird. Die 3D-Drehung wird durch Regression auf eine Quaternion-Darstellung vorhergesagt. Es werden zwei neue Verlustfunktionen für die Rotationsschätzung eingeführt, wobei der ShapeMatch-Verlust für symmetrische Objekte entwickelt wurde. Dadurch ist PoseCNN in der Lage, Okklusion und symmetrische Objekte in unübersichtlichen Szenen zu verarbeiten. Dies eröffnet den Weg zur Verwendung von Kameras mit einer Auflösung und einem Sichtfeld, die weit über die derzeit verwendeten Tiefenkamerasysteme hinausgehen. Wir stellen fest, dass SLOSS manchmal zu lokalen Minimums im Pose-Raum führt, ähnlich wie ICP. Es wäre interessant, in Zukunft einen effizienteren Umgang mit symmetrischen Objekten in der 6D-Positionsschätzung zu erforschen.

3.7 Robust Gaussian Filter[Iss+16]

Objektmodelle
Video-Input
genutzte Datensätze
Genauigkeit
Ressourcenintensivität
Laufzeit

4 Vergleich verschiedener Verfahren

5 Fazit

Literatur

[Che+20] Xu Chen u. a. *Category Level Object Pose Estimation via Neural Analysis-by-Synthesis*. Aufgerufen 27.10.2021. 2020. arXiv: 2008.08145. URL: arxiv.org/abs/2008.08145.

[Iss+16] Jan Issac u. a. „Depth-based object tracking using a Robust Gaussian Filter“. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)* (März 2016). Aufgerufen 27.10.2021. DOI: 10.1109/icra.2016.7487184. URL: dx.doi.org/10.1109/ICRA.2016.7487184.

[Li+18] Yi Li u. a. „DeepIM: Deep Iterative Matching for 6D Pose Estimation“. In: *International Journal of Computer Vision* 128.3 (Nov. 2018). Aufgerufen 16.10.2021, S. 657–678. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01250-9. URL: arxiv.org/abs/1804.00175.

[RBA18] Martin Rünz, Maud Buffier und Lourdes Agapito. *MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects*. Aufgerufen 27.10.2021. 2018. arXiv: 1804.09194. URL: arxiv.org/abs/1804.09194.

[Wan+19] Chen Wang u. a. *6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints*. Aufgerufen 27.10.2021. 2019. arXiv: 1910.10750. URL: arxiv.org/abs/1910.10750.

[WB21] Bowen Wen und Kostas Bekris. *BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models*. Website. Aufgerufen 23.10.2021. 2021. arXiv: 2108.00516. URL: arxiv.org/abs/2108.00516.

[Xia+17] Yu Xiang u. a. *PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes*. Website. Aufgerufen 16.10.2021. 2017. arXiv: 1711.00199. URL: arxiv.org/abs/1711.00199.