Оценка 6D-позы нетекстурированных объектов с помощью сверточных нейронных сетей

В.Ю. Ерхов, П.А. Лазарева, Г.Л. Дегтярев
Казанский национальный исследовательский технический университет имени А. Н. Туполева, 420111, Россия, г. Казань, ул. К. Маркса, д. 10

Аннотация

Предложен подход к оценке 6D-положения объектов без текстуры с помощью сегментации и определения угловой ориентации на основе нейросетевых алгоритмов. В ходе практической апробации обучены две нейронные сети для получения устойчивой оценки позы. Первая нейросеть устанавливает опорные точки объектов, которые затем передаются в алгоритм PnP, отвечающий за оценку положения. Вторая нейронная сеть – это сеть регрессии поворота объекта, выходные данные которой представляет кватернион ориентации. Нейронные сети были обучены на пользовательских наборах данных, содержащих как реальные, так и фотореалистичные синтезированные изображения, которые сгенерированы на основе 3D-моделей. Сами 3D-модели сформированы с помощью фотограмметрии по множеству снимков распознаваемых объектов, что и составляют массив реальных изображений в датасете. Эффективность предложенных нейронных сетей для сегментации объектов и оценки позы также была изучена и сопоставлена с использованием реальных и синтезированных изображений. Результаты экспериментов демонстрируют высокий потенциал предложенного подхода к оценке позы бестекстурных объектов, что было проверено в задаче роботизированного захвата.

Ключевые слова: компьютерное зрение, оценка 6D позы, машинное обучение, сверточные нейронные сети, YOLO.

Введение

Оценка 6D позы объектов (6 Degrees of Freedom Object Pose Estimation, 6DoF) является одной из ключевых проблем компьютерного зрения (CV), имеющей критическое значение для множества приложений в таких областях, как робототехника, дополненная и виртуальная реальность, автономные транспортные средства и системы автоматизированного производства [1]. Поиск решения заключается в определении пространственного положения (по трем декартовым координатам) и угловой ориентации (углы Эйлера) объекта относительно системы координат камеры. Установка точного положения объектов в пространстве обеспечивает для роботизированных систем возможность выполнения сложных операций, таких как точное манипулирование объектами на роботизированных производствах, а в AR/VR-приложениях – позволяет добиваться реалистичной модели взаимодействия виртуальной и реальной сред [1, 2].

В последние годы в области оценки 6D позы наблюдается значительный прогресс, связанный с широким внедрением методов глубокого обучения и появлением специализированных датасетов и бенчмарков, таких как BOP (Benchmark for 6D Object Pose Estimation) [3]. Однако несмотря на существенные достижения, тема точного позиционирования объектов в пространстве продолжает оставаться сложной из-за множества факторов, включая окклюзии, непостоянство освещения, ограничения вычислительных и аппаратных ресурсов для работы в реальном времени.

Классические методы сопоставляют ключевые точки, обнаруженные на 2D-изображениях, с соответствующими 3D-моделями; сама оценка 6D-позы выполняется по алгоритму Perspective-n-Point (PnP) [4], часто в рамках метода RANSAC [4, 5]. Но такой вид сопоставления нельзя назвать надежным для слабо текстурированных объектов, что зачастую приводит к недостоверным оценкам их положения. В работе [6] представлены новые подходы на основе шаблонов, способных справиться с задачей 6D-позы объектов без выраженных текстур. В них проводится сравнение изображения на входе с набором стандартных изображений объекта, чтобы определить его положение в пространстве. Однако методы, построенные на преобразовании пикселей изображения в координаты трехмерных объектов с целью установления 2D-3D соответствий, могут не подойти для симметричных объектов [7].

Операция распознавания является ключевой в оценке позы объекта, что требует более глубокого обсуждения и экспериментальной апробации. Современные методы оценки позы, особенно работающие со слабо текстурированными объектами, построены на обучении признакам. В Вашингтонском университете была предпринята первая попытка использовать сверточную нейронную сеть (CNN) для прямой регрессии 6DoF поз объектов [8]. Согласно исследованию, она позволяет разделить оценку смещения и оценку поворота объекта в сквозном (end-to-end) фреймворке. Фреймворк использует функции, определенные сверточными слоями сети VGG16 и их тремя различными ветвями. Две сверточные ветви выполняют семантическую сегментацию и 2D-

голосование по центру для обработки окклюзий. Третья ветвь включает в себя объединение по областям интереса (Region of Interest, RoI) и полносвязную архитектуру, отвечающую за регрессию каждого значения RoI к кватерниону. В целом, основанные на CNN подходы к оценке положения объектов, относятся либо к прямой регрессии позы в 6D по изображению [9], либо к предсказыванию размещения ключевых точек в 2D на изображении [10], из которых 6DoF может быть определена с помощью алгоритма PnP.

В передовых подходах к определению 6D-позы на основе RGB-изображений превалируют показатели точности и быстродействия [11, 12]. Существуют наборы данных в открытом доступе для сравнения производительности алгоритмов оценки 6D-позы объекта, включая OccludedLinemod [13], YCB-Video [14]. Как правило, достижение “высоких” результатов обеспечивается работой с большими данными (Big Data) в процессе обучения нейросети. В то же время, во многих реальных проектах часто возникает необходимость быстрой генерации 3D-модели конкретного объекта для проведения механических операций над ним, что особенно важно для выполнения задач, связанных с роботами-манипуляторами. Эталонная 3D-модель в качестве обратной связи может быть необходима при рендеринге синтезированных изображений для набора обучающих и тестовых данных.

В этой связи с использованием упомянутых передовых решений представлен собственный подход к сегментации и оценке 6D-позы объекта с учетом экономного сбора данных. В качестве таковых выбраны с ориентацией на производственные операции крепежные изделия - болты и шайбы (охарактеризованные ниже).

В первую очередь исследуется задача оценивания позы жесткого объекта в шести координатах на RGB-изображениях, где интересующий объект не имеет выраженной текстуры (как и в тривиальных производственных CV-задачах). Через операцию сегментации экземпляров (instance segmentation) его границы выделяются на фоне с помощью свёрточной нейронной сети YOLOv8. Далее идет обучение двух нейросетевых моделей для достижения устойчивой оценки 6D-позы объекта на RGB-изображениях. Первая нейросеть определяет опорные точки на объекте, координаты которых далее обрабатываются в алгоритме PnP, отвечающем за саму оценку 6D позы. Вторая – представляет собой сеть регрессии поворота, выдающую на выходе кватернион ориентации объекта. Нейросети обучены на пользовательских наборах данных, содержащих как реальные изображения, так и синтезированные – собранные с помощью рендеринга 3D-моделей. На завершающем этапе производительность алгоритмов проходит тестирование на изображениях реальных и смоделированных объектов. Преимущество предложенного подхода заключается в расширении датасета из небольшого набора данных. Изображения, исходно предназначенные для генерации 3D-моделей, сформировали и реальную, и синтетическую часть датасета: фотографии искомых объектов в преобразовании “фотограмметрия-модель-рендеринг” позволили более, чем вдвое умножить количество обучающих данных. В целях повышения устойчивости сегментирующей нейросети к различным условиям освещенности и окклюзий датасет был подвергнут аугментации. Далее рассматриваются ключевые этапы реализации нового подхода.

1. Нейросеть для сегментации объектов

Операция сегментации экземпляров осуществлялась с помощью нейросетевой модели YOLOv8. Архитектура этой нейронной сети представлена на рис. 1. Обучение проведено на датасете размером 420 RGB-изображений разрешения 640*480.

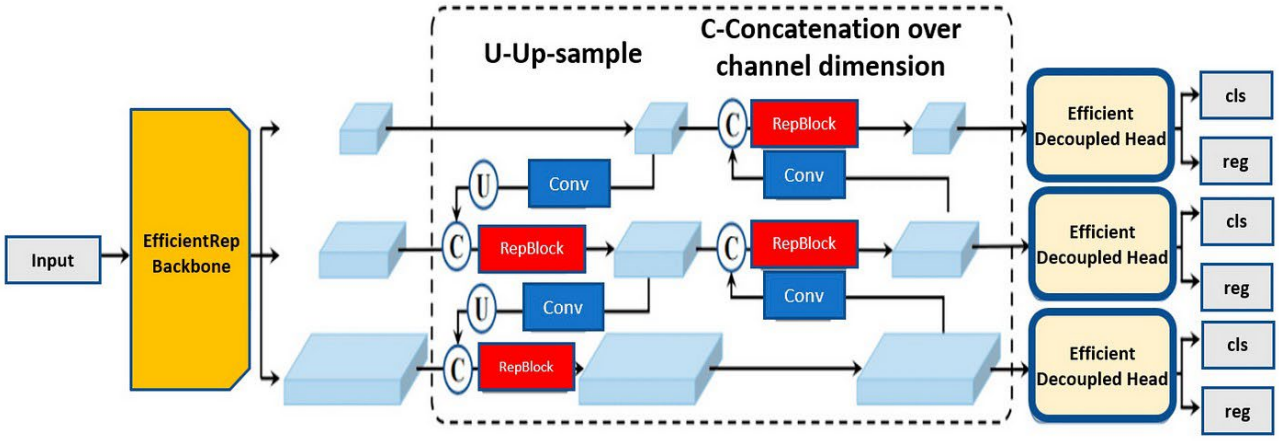

Рис. 1. Архитектура YOLOv8, примененная для сегментации объектов [15]

При обучении нейронной сети использовалась функция фокальных потерь распределения (Distribution Focal Loss, DFL) [15]:

$$L_{DFL} = -[(y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1})], \quad (1)$$

где y – истинное смещение координаты, y_i и y_{i+1} – ближайшие к y границы интервала из дискретного распределения $\{0, \ldots, y_{max}\}$, S_i и S_{i+1} – прогнозируемая вероятность логистической функции *softmax* для соответствующих границ. В полной мере *softmax* применяется для обучения на 3 и более классах объектов; в данном исследовании с 2 классами *softmax* эквивалентна логистической функции ошибки *log_loss* [16]. Приведенная формула рассматривает регрессию как классификацию по блокам, что становится предпочтительным решением для задач обнаружения и сегментации объектов.

2. 6D оценка позы объекта

Использовались две алгоритмически отличные нейросетевые архитектуры. В начале описана нейросеть и работа с ней для определения опорных точек и оценки ею 2D-положений, а также кратко охарактеризован алгоритм PnP. Далее представлена нейронная сеть для регрессионного анализа поворота объекта.

2.1. 6D оценка позы объекта с использованием опорных точек и PnP

В первом подходе оценка каждого объекта осуществляется с помощью CNN, которая устанавливает восемь опорных точек объекта на изображении. Двумерные координаты точек далее подвергаются обработке в алгоритме PnP, который вычисляет 6D-позу объекта; архитектура нейросети такого типа подробнее описана в исследовании [17]. С учетом того, что результаты, представленные в работах [18, 19] обобщенно свидетельствуют об эффективности совмещения синтетических и реальных данных, в используемом датасете их количество сопоставимо и представлено в пропорции 4:3. Для каждого объекта сгенерировано с помощью рендеринга 240 синтезированных изображений разрешения 640*480 с использованием эталонных данных, а также вручную аннотировано 180 реальных изображений. Нейронные сети обучены на масштабированных RGB-изображениях разрешения 640*640 за 600 эпох. Функция потерь представлена средней абсолютной ошибкой (Mean Absolute Error, MAE). Эталонные данные на синтезированных изображениях определены на основе трехмерных положений опорных точек на 3D-модели, которые были перепроцированы на двухмерные изображения. Следует отметить, что даже если точка скрыта, она все равно проецируется на RGB-изображение. Так, количество спроецированных опорных точек на каждом изображении равно восьми.

Согласно общей формулировке преобразования 3D в 2D, цель состоит в том, чтобы установить преобразование T_k, минимизирующее ошибку репроекции:

$$T_k = argmin \sum_i \left\| p_k^i - p_{k-1}^i \right\|^2, \quad (2)$$

где p_{k-1}^i – репроекция трехмерной точки X_{k-1}^i на изображение I_k с преобразованием T_k.

Данная задача преобразования известна как “Перспектива из N Точек” (PnP), что определяет внешние параметры камеры на основе 3D-2D соответствий. Это означает, что PnP оценивает положение откалиброванной камеры, учитывая набор из N точек в пространстве и соответствующих им двухмерных проекций. Существует множество решений этой проблемы. В данном подходе использовано EPnP [20] из библиотеки OpenCV.

2.2. Нейронная сеть для регрессии поворота объекта

Отдельно для каждого объекта была обучена нейронная сеть, предоставляющая на выходе кватернион, содержащий информацию о повороте интересующего объекта (принцип работы данной CNN развернуто изложен в [21]). Нейросеть прошла обучение на масштабированных RGB-изображениях 640*640 за 600 эпох. Как и в методе опорных точек, нейронная сеть после обучения валидирована на 240 синтезированных и 180 реальных изображениях. Эталонные данные для сгенерированных изображений были определены с помощью программных скриптов на Python. Функция потерь представлена среднеквадратической ошибкой.

3. Применение 3D-моделей

Экспериментальная оценка проводилась на крепежных изделиях (болт, гайка); данные объекты не обладают явно выраженной текстурой. CAD-модели объектов были сгенерированы с помощью фотограмметрии в Adobe Substance 3D Sampler и отредактированы в Autodesk Meshmixer, которые являются специализированными приложениями для работы с компьютерной графикой. В приведенных далее результатах в качестве оцениваемого объекта фигурирует болт с шестигранной шляпкой. Диаметр болта составляет 20 мм, сама 3D-модель состоит из 4657 вершин. Диаметр шайбы – 12 мм, 3D-модель включает 2931 вершину. Синтезированные обучающие и тестовые изображения были собраны с помощью Python-скриптов. На рис. 2 показан реальный объект под разными ракурсами и прошедшие рендеринг изображения, которые были получены на основе 3D-модели. Очевидно, что репрезентация объекта позволяет фотореалистично визуализировать его же в требуемых позициях. Камера была предварительно откалибрована с помощью библиотеки OpenCV.

Рис. 2. Реальный и синтезированный объект: первая строка содержит изображения реального объекта, вторая - изображения его 3D-модели.

Объект запечатлен с различных ракурсов (рис. 3). Для конвертации массива RGB-изображений в 3D-модель объект был размещен в интервале от 0 до 360°. На каждые десять градусов поворота было получено 5 изображений. Это означает, что для каждого объекта количество изображений, собранных в таком виде, равно 180.

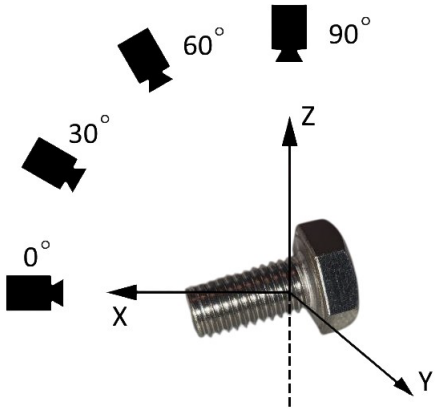

Рис. 3. Размещение камеры для сбора изображений объекта

При подготовке нейросетевой модели сегментации объекты регистрировались камерой с разных точек наблюдения. Обученные нейросети затем использовались для выделения рассматриваемых объектов на фоне, при их варьировании. В дальнейшем был собран набор данных из RGB-изображений для тестирования нейронных сетей по функциям определения 6D-положения объектов и их отслеживания в режиме реального времени.

4. Сегментация объектов

Нейронная сеть YOLOv8, рассмотренная в параграфе 1, была обучена обнаружению и выделению искомого объекта по набору сегментированных вручную изображений. На рис. 4 показаны примеры RGB-изображений с соответствующими бинарными масками, которые были получены на основе весов YOLO. Указанные изображения относятся именно к тестовой подгруппе и не включены в обучающую подгруппу датасета.

Рис. 4. Сегментация экземпляров: первая строка содержит RGB-изображения, вторая – бинарные маски, определенные по выходным данным YOLOv8

В табл. 1 представлены значения, полученные в тестовом подмножестве датасета. Коэффициент сходства для двух наборов A и B может быть выражен следующим образом [22, 23]:

$$dice(A,B) = \frac{2*|intersection(A,B)|}{(|A|+|B|)}, \quad (3)$$

где |A| и |B| означают мощность множеств A и B соответственно.

Данный коэффициент также может быть выражен в виде истинных срабатываний (True Positive, TP), ложных срабатываний (False Positive, FP) и ложноотрицательных результатов (False Negative, FN) следующей формулой:

$$dice(A,B) = \frac{2*TP}{2*TP+FP+FN}, \quad (4)$$

Тестовое подмножество содержит по тридцать изображений для каждого объекта. Как видно из табл. 1, лучшие результаты сегментации были достигнуты для YOLOv8, обученной отдельно на каждом из рассматриваемых объектов. С учетом преимущества данного сценария обучения над сегментацией нескольких классов объектов, далее обсуждаются экспериментальные результаты, полученные таким способом.

Табл. 1. Коэффициенты сходства для результатов сегментации с использованием YOLOv8 по мощности множеств на тестовой подгруппе (безразмерная величина)

Применимость	Болт	Шайба
YOLOv8 отдельно для каждого объекта	0,974	0,962
YOLOv8 для всех объектов	0,957	0,936

5. Оценочная характеристика для 6D-позы

Качественно оценка 6D-позы характеризуется погрешностью ADD (Average Distance of Model Points) [24] в 3D-пространстве. Метрика ADD вычисляет среднее евклидово расстояние между вершинами модели объекта в истинной и предсказанной позах. Расчет проводится следующим образом:

$$err_{ADD} = \frac{1}{|M|}\sum_{x\in M}\left\|(Rx+t)-\left(R_{gt}x+t_{gt}\right)\right\|, \quad (5)$$

где M – набор вершин на сетке модели объекта, R, t – квадратная матрица поворота и вектор смещения для предсказанной позы, R_{gt}, t_{gt} – аналогичные данные для истинной позы объекта.

6DoF-оценка считается достоверной, если погрешность ADD составляет менее 10% от диаметра объекта. Также включается показатель ошибки угловой ориентации на основе следующей формулы:

$$err_{rot} = \arccos\frac{T_r\left(R_{gt}R^T\right)-1}{2}, \quad (6)$$

где T_r – след матрицы (сумма всех элементов главной диагонали матрицы), R и R_{gt} – квадратные матрицы поворота, соответствующие предсказанной и истинной позе. То есть данная погрешность измеряет угловое отклонение между предсказанным поворотом R и истинным R_{gt}.

5.1. Экспериментальная оценка

В табл. 2 представлены экспериментальные результаты, которые были получены с помощью первой нейронной сети и PnP-алгоритма при оценке 6D-позы болтов и шайб по отдельным RGB-изображениям. Нейросеть была обучена только на сгенерированных изображениях и протестирована на реальных данных. Эксперименты проводились на RGB-изображениях разрешения 640*480. На основании полученных данных (табл. 2) можно утверждать, что веса нейросети, обученных только на синтезированных изображениях и примененных к реальным изображениям, показали удовлетворительные результаты. Табл. 3 отражает значения ADD, полученные нейронной сетью для определения опорных точек и оценки их 2D-положения с помощью алгоритма PnP, где сеть была обучена как на синтезированных, так и на реальных изображениях. Очевидно, что нейронные сети, обученные на разных типах изображений в практическом применении к реальным, достигают лучших результатов по сравнению с ними же, обученными только на синтетическом наборе (табл. 2).

Табл. 2. Значения ADD [%], полученные для определения опорных точек и PnP при обучении на синтезированных и тестировании на реальных изображениях

Угол обзора с камеры, град.	err_{ADD} для болта, %	err_{ADD} для шайбы, %
0°	87 ADD < 2,3 мм	86 ADD < 1,4 мм
30°	90 ADD < 2,2 мм	78 ADD < 1,6 мм
60°	88 ADD < 2,3 мм	82 ADD < 1,5 мм
90°	84 ADD < 2,4 мм	81 ADD < 1,5 мм

Табл. 3. Значения ADD [%], полученные для определения опорных точек и PnP при обучении на синтезированных и реальных изображениях и тестировании на реальных изображениях

Угол обзора с камеры, град.	err_{ADD} для болта, %	err_{ADD} для шайбы, %
0°	97 ADD < 2,1 мм	94 ADD < 1,3 мм
30°	96 ADD < 2,1 мм	89 ADD < 1,4 мм
60°	93 ADD < 2,2 мм	88 ADD < 1,4 мм
90°	88 ADD < 2,3 мм	94 ADD < 1,3 мм

В табл. 4 представлены результаты, полученные с помощью нейронной сети, проводящей регрессию поворота. Чтобы вычислить ADD, использовались данные позиций в 3D, которые были определены с помощью PnP. При таком подходе достигаются результаты, сравнимые с результатами, полученными на основе опорных точек и PnP (табл. 2). Кроме того, рассчитано пространственное положение 3D-объектов; для этого спроецированы их модели на плоское изображение, а затем сопоставлены прошедшие рендеринг объекты с сегментированными. Позы объектов определялись с использованием метода Particle Swarm Optimization (PSO) [25]. Данные кватернионов прошли преобразование в углы Эйлера, а декартовы координаты 3D-объектов были рассчитаны путем сопоставления сегментированных объектов и проецируемой на плоскость изображения 3D-модели.

Табл. 4. Значения ADD [%], полученные для регрессии поворота при обучении на синтезированных и тестировании на реальных изображениях

Угол обзора с камеры, град.	err_{ADD} для болта, %	err_{ADD} для шайбы, %
0°	92 ADD < 2,2 мм	87 ADD < 1,4 мм
30°	93 ADD < 2,2 мм	82 ADD < 1,5 мм
60°	92 ADD < 2,2 мм	88 ADD < 1,4 мм
90°	87 ADD < 2,3 мм	87 ADD < 1,4 мм

На рис. 5 показаны графики ошибок угловой ориентации в зависимости от углового положения объектов по равенству (6). На изображениях видно, что максимальная погрешность не превышает порог в 3°.

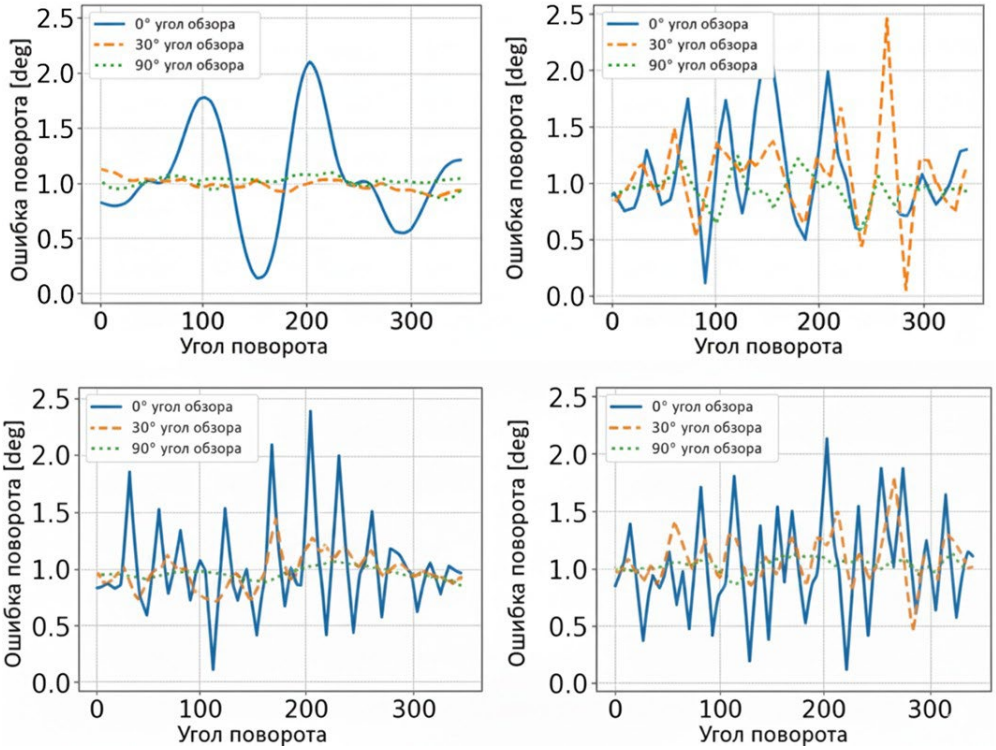

Рис. 5. Зависимость погрешности угловой ориентации от угла поворота объекта: нейронная сеть, оценивающая двумерное расположение опорных точек с алгоритмом PnP (слева); нейронная сеть, проводящая регрессию для угла поворота, выраженного кватернионом (справа); графики верхней строки относятся к классу болта, нижние – к классу шайбы

В табл. 5 приведены оценки ADD, которые были получены при использовании второго подхода с нейронной сетью, обученной как на синтезированных, так и на реальных изображениях. Сравнивая результаты, представленные в обсуждаемой таблице, и результаты, приведенные в табл. 3, можно отметить, что обсуждаемый подход позволяет достичь еще более высоких показателей ADD.

Табл. 5. Значения ADD [%], полученные для регрессии поворота при обучении на синтезированных и реальных изображениях и тестировании на реальных изображениях

Угол обзора с камеры, град.	err_{ADD} для болта, %	err_{ADD} для шайбы, %
0°	96 ADD < 2,1 мм	88 ADD < 1,4 мм
30°	85 ADD < 2,4 мм	93 ADD < 1,3 мм
60°	91 ADD < 2,2 мм	82 ADD < 1,5 мм
90°	92 ADD < 2,2 мм	94 ADD < 1,3 мм

Система оценки 6D-позы была реализована на языке Python, работает на обычном ПК с CPU/GPU. В стандартных условиях (при работе на компьютере, оснащенным процессором Intel Core i7-11700KF с частотой 3,6 ГГц), время сегментации составило около 0,20 секунд, вместе с тем интервал работы нейронной сети при определении опорных точек и регрессии поворота объекта занимает около 0,03 секунд. Экспериментальная часть исследования, проведенная с коллаборативным роботом-манипулятором Rozum Pulse 75, оснащенным RGB-D камерой Intel RealSense D455, показала жизнеспособность предложенного подхода для задачи захвата объекта (рис. 6).

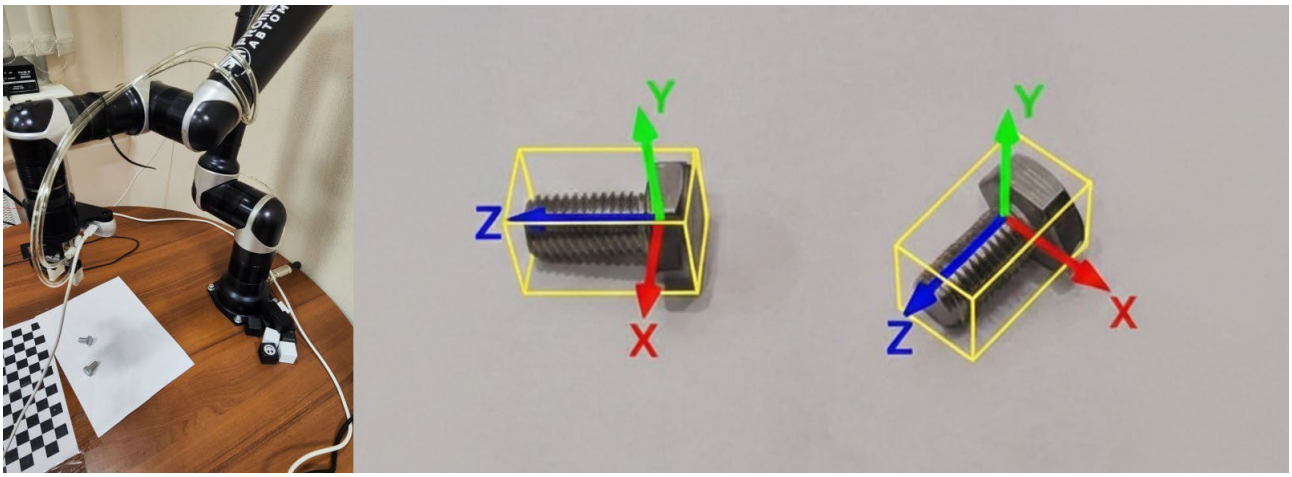

Рис. 6. Оценка и трекинг 6D-позы с помощью руки-манипулятора Rozum Pulse 75, оснащенной камерой RealSense D455

Предложенный подход по оценке 6D-позы объектов может быть внедрен на роботизированных производствах, где исполнительному устройству необходимо быстро и точно совершать механические манипуляции с мелкими объектами (например, с крепежными изделиями) без выраженной текстуры.

Заключение

Проведен сравнительный анализ двух алгоритмически разных сверточных нейронных сетей, предназначенных для оценки 6D-позы объектов на RGB-изображениях. Сами объекты были сегментированы с помощью архитектуры YOLOv8. Первая нейронная сеть определяет восемь опорных точек на объектах, которые затем передаются в алгоритм PnP, отвечающий за определение 6D-позы. Вторая нейронная сеть – это сеть регрессии поворота, выходным типом данных которой является кватернион угловой ориентации. Обе они были последовательно сопоставлены между собой по критериям ADD и погрешности поворота на нескольких наборах изображений. Алгоритмы оценивания продемонстрировали качественно близкие результаты согласно значениям из табл. 2-4, что позволяет охарактеризовать точность их расчетов как равнозначную. Генерация 3D-моделей интересующих объектов обусловлена как получением синтетической части датасета, так и самими вычислениями 6DoF. Обучение на расширенном наборе из реальных и синтезированных изображений ожидаемо привело к более надежной оценке позы, нежели при включении данных только одного типа. Также проведены тесты на основе изображений, полученных с помощью камеры, установленной на роботизированной руке. Экспериментальные результаты показали переносимость модели, обученной при участии синтетических данных, на реальные данные робототехнической сцены. Предложенный подход можно реализовать в промышленных операциях на производствах, связанных с участием роботов-манипуляторов.

Благодарности

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского национального исследовательский технического университета имени А. Н. Туполева («ПРИОРИТЕТ-2030»).

References

[1] Guan J, Hao Y, Wu Q, Li S, Fang Y. A Survey of 6DoF Object Pose Estimation Methods for Different Application Scenarios. Sensors 2024; 24(4): 1076. DOI: 10.3390/s24041076.

[2] Hugo D, Marius P, Titus Z. 6D Pose Estimation of Unseen Objects for Industrial Augmented Reality. IEEE 20th International Conference on Intelligent Computer Communication and Processing 2024; DOI: 10.1109/ICCP63557.2024.10792989.

[3] Thomas R, Julien M, Alexandre E, Antoine M. Real-time visual pose estimation: from BOP objects to custom drone — A journey. Mechatronics 2025; ISSN: 0957-4158.

[4] Zhou B, Ziqiang C, Qinghua L. An Efficient Solution to the Perspective-n-Point Problem for Camera With Unknown Focal Length. IEEE Access 2020; DOI: 10.1109/ACCESS.2020.3021313.

[5] Guixian C, Jianhao M, Salar F. RANSAC Revisited: An Improved Algorithm for Robust Subspace Recovery under Adversarial and Noisy Corruptions 2025; Source: <https://arxiv.org/abs/2504.09648>. DOI: 10.48550/arXiv.2504.09648.

[6] Hinterstoisser S, Holzer S, Cagniart C, Ilic S, Konolige K, Navab N, Lepetit V. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. International Conference on Computer Vision 2011; DOI: 10.1109/ICCV.2011.6126326.

[7] Fabian D, Sebastian K, Hanna Z, Ngo A. SyMFM6D: Symmetry-Aware Multi-Directional Fusion for Multi-View 6D Object Pose Estimation. IEEE Robotics and Automation Letters 2023; 8(9): 5315-5322. DOI: 10.1109/LRA.2023.3293317.

[8] Lei J, Xiaojuan W, Mingshu H, Jingyue W. DRNet: A Depth-Based Regression Network for 6D Object Pose Estimation. Sensors 2021; 21(5): 1692. DOI: 10.3390/s21051692.

[9] Ge G, Mikko L, Yulong W, Xiaolin H, Jianwei Z, Simone F. 6D Object Pose Regression via Supervised Learning on Point Clouds. 2020; Source: <https://arxiv.org/abs/2001.08942>. DOI: 10.48550/arXiv.2001.08942.
[10] Rad M, Lepetit V. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. IEEE International Conference on Computer Vision 2017; DOI: 10.1109/ICCV.2017.413.
[11] Ordoumpozanis K, Papakostas G. Reviewing 6D Pose Estimation: Model Strengths, Limitations, and Application Fields. Applied Sciences 2025; 15(6): 3284. DOI: 10.3390/app15063284.
[12] Wanqing X, Hao Z, Weiliang X, Xun X. Large vision-language models enabled novel objects 6D pose estimation for human-robot collaboration. Robotics and Computer-Integrated Manufacturing 2025; ISSN 0736-5845.
[13] Budi N, a Nanik S, Chastine F. Analysis of Optimization Techniques in 6D Pose Estimation Approaches using RGB Images on Multiple Objects with Occlusion. Procedia Computer Science 2024; ISSN 1877-0509.
[14] Pavel R, Thomas P. YCB-Ev 1.1: Event-vision dataset for 6DoF object pose estimation. 2023; Source: <https://arxiv.org/abs/2309.08482>. DOI: 10.48550/arXiv.2309.08482.
[15] Antunes S, Okano M, Nääs I, Lopes W, Aguiar F, Vendrametto O, Fernandes J, Fernandes M. Model Development for Identifying Aromatic Herbs Using Object Detection Algorithm. AgriEngineering 2024, 6(3): 1924-1936. DOI: 10.3390/agriengineering6030112.
[16] Richard C, Alan D, Ben C, Lucia V. Correlations of Cross-Entropy Loss in Machine Learning. Entropy 2024; 26(6): 491. DOI: 10.3390/e26060491.
[17] Mateusz M, Kwolek B. Fiducial Points-supported Object Pose Tracking on RGB Images via Particle Filtering with Heuristic Optimization. VISIGRAPP 2021; ISSN 2184-4321.
[18] Alex X, Colin R, Binh P. Blending is all you need: Data-centric ensemble synthetic data. Information Sciences 2025; ISSN 0020-0255.
[19] Goyal M, Mahmoud Q. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics 2024; 13(17): 3509. DOI: 10.3390/electronics13173509.
[20] Hansheng C, Wei T, Pichao W, Fan W, Lu X, Hao L. EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2024; 47(11): 9413-9425. DOI: 10.1109/TPAMI.2024.3354997.
[21] Mateusz M, Kwolek B. Deep Quaternion Pose Proposals for 6D Object Pose Tracking. International Conference on Computer Vision Workshops 2021; ISSN 2473-9944.
[22] Zou K, Warfield S, Bharatha A, Tempany C, Kaus M, Haker S, Wells W, Jolesz F, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index. Academic Radiology 2004; ISSN 1076-6332.
[23] Vatsal R, Nataliia M, Mara G, Andrey M, Henning M, Meritxell B, Mark G. Tackling Bias in the Dice Similarity Coefficient: Introducing nDSC for White Matter Lesion Segmentation. 2023; Source: <https://arxiv.org/abs/2302.05432>. DOI: 202310.48550/arXiv.2302.05432.
[24] Fang G, Qingyi S, Shaodong L, Wenbo L, Yong L, Jun Y, Feng S. Efficient 6D object pose estimation based on attentive multi-scale contextual information. IET Computer Vision 2022; 16(7): 596-606. DOI: 10.1049/cvi2.12101.
[25] Ahmed G. Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. Archives of Computational Methods in Engineering 2022; 29: 2531-2561. DOI: 10.1007/s11831-021-09694-4.

Сведения об авторах

Ерхов Владимир Юрьевич, 1999 года рождения, аспирант Казанского национального исследовательского технического университета имени А. Н. Туполева. Сфера научных интересов: цифровая обработка изображений, нейронные сети, робототехнические системы. E-mail: *molovlad@gmail.com*

Лазарева Полина Александровна, 1986 года рождения, к.ф.-м.н., доцент кафедры автоматики и управления Казанского национального исследовательского технического университета имени А. Н. Туполева. Сфера научных интересов: интеллектуальное управление, оптимальное управление, моделирование динамических систем, вычислительные алгоритмы. E-mail: *PALazareva@kai.ru*

Дегтярев Геннадий Лукич, 1938 года рождения, профессор, д.т.н., профессор кафедры автоматики и управления Казанского национального исследовательского технического университета имени А. Н. Туполева. Сфера научных интересов: методы оптимального управления, методы анализа и синтеза динамических систем в условиях параметрической неопределённости и внешних возмущений. E-mail: *GLDegtyarev@kai.ru*

6D Pose Estimation of texture-free objects using convolutional neural networks

V.Yu. Erkhov, P.A. Lazareva, G.L. Degtyarev
Kazan national research technical university
named after A. N. Tupolev, Karl Marx Str. 10, Kazan, 420111, Russia

Abstract

This article proposes an approach for 6D Pose Estimation for textureless objects, based on segmentation and angular orientation CNN algorithms. During practical testing, we've trained two neural networks to obtain a stable pose estimation. The first neural network establishes the reference points of objects – they are transmitted to PnP algorithm responsible for pose estimation. The second neural network is a rotated regression network – its output is a quaternion. Neural networks are trained on custom datasets containing both real and photorealistic synthesized images which were generated on 3D-models rendering. Those models are created with photogrammetry on recognizable objects multiple images – they are constituting an array of real images in dataset. An effectiveness of the proposed neural networks for instance segmentation and 6D pose estimation was also evaluated appealing to real and synthesized images. Our experimental results show valuable perspective for this Pose Estimation approach with textureless objects using camera supervision, that is mounted on the flange of a robotic arm.

Keywords: computer vision, 6D Pose Estimation, machine learning, convolutional neural networks, YOLO.

About authors

Erkhov Vladimir Yurievich, born in 1999, graduate student at Kazan National Research Technical University named after A.N. Tupolev. Area of scientific interests: digital image processing, neural networks, robotic systems. E-mail: *molovlad@gmail.com*

Lazareva Polina Alexandrovna, born in 1986, associate professor of the department of automation and control at Kazan National Research Technical University named after A.N. Tupolev. Area of scientific interests: intelligent control, optimal control, dynamic systems modeling, computational algorithms. E-mail: *PALazareva@kai.ru*

Degtyarev Gennady Lukich, born in 1938, professor of the department of automation and control at Kazan National Research Technical University named after A.N. Tupolev. Area of scientific interests: optimal control methods, methods of analysis and synthesis of dynamic systems under conditions of parametric uncertainty and external disturbances. E-mail: *GLDegtyarev@kai.ru*
