УДК 004.93'1

АНАЛИЗ И КЛАССИФИКАЦИЯ МЕТОДОВ ОЦЕНИВАНИЯ «ПОЗЫ ОБЪЕКТОВ», ПРИМЕНЯЕМЫХ ПРИ РЕШЕНИИ ЗАДАЧ МАШИННОГО ЗРЕНИЯ

Д.С. Гаджиев
Институт проблем управления им. В.А. Трапезникова
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: danialgad2002@gmail.com

А.В. Макаренко
Институт проблем управления им. В.А. Трапезникова
Россия, 117997, Москва, Профсоюзная ул., 65

Ключевые слова: Оценка позы, нейронные сети, машинное зрение, фотоизображение.

Аннотация: Проведен анализ работ, опубликованных с 1969 г. по первую половину 2023 г., посвященных методам оценивания «позы объектов» по фотоизображению и видеопоследовательности. Выявлены и классифицированы три основных периода развития алгоритмов: до 1999 г. – на основе классических методов машинного зрения; 1999–2014 г. – переход от классических методов к подходам на основе машинного обучения, сопровождающийся развитием аннотированных наборов данных; после 2014 г. – подходы на основе нейросетевых, и, в том числе, трансформерных архитектур. Установлено, что наиболее актуальным направлением дальнейших исследований в области развития методов оценивания «позы объектов» является устранение критических недостатков и модернизация трансформерных нейросетевых архитектур.

1. Введение

В настоящее время задача оценивания «позы объектов» является ключевым алгоритмическим компонентом во многих системах компьютерного зрения. Как показано на рисунке 1 суть данной задачи заключается в распознавании экземпляров целевого объекта, с последующей локализацией предварительно заданных ключевых точек объекта, по которым возможно построить его «шарнирную модель».

Задача оценки позы является, с одной стороны, самостоятельным элементом систем машинного зрения, с другой – обеспечивает базис для решения множества других проблем компьютерного зрения более высокого порядка, в том числе: уточнение ориентации и локализация компонентов объектов, построение их 3D-моделей, распознавание и прогнозирование действий, генерация сцен для моделей виртуальной и дополненной реальности и т.д.

Анализ исторических аспектов данной научной области свидетельствуют о том, что первые конструктивные методы и алгоритмы оценки позы, по всей видимости, были разработаны в 70-х годах прошлого столетия [1, 2], тем не менее, в силу ряда причин, они активно не применялись. Однако, с развитием компьютерных технологий и появлением новых методов машинного обучения, таких как, глубокие нейронные сети,

с 2014-го года методы оценки «позы объектов» вновь стали востребованными на практике, и достигли новых высот в точности и скорости работы [3, 4].

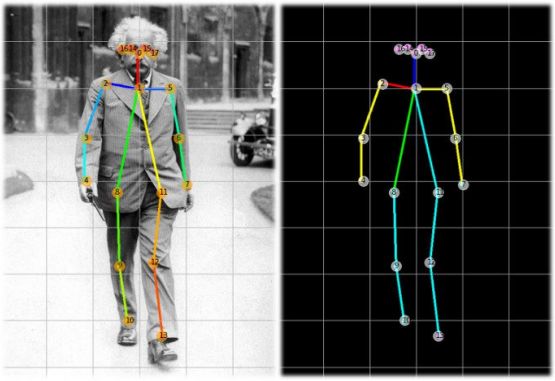

Рис 1. Выделение шарнирной модели целевого объекта на примере изображения человека.

Для того чтобы разобраться в том, как проблема pose estimation решалась ранее и какую прошла эволюцию до наших дней, нами было собрано и проанализировано около сотни научно-технических работ, опубликованные в период с 1969 г. по первую половину 2023 г. Первичный анализ показал, что на данном отрезке времени явно выделяются три основных периода развития алгоритмов:

- 1969–1999 г.г. – традиционные методы, базирующиеся на классических алгоритмах компьютерного зрения;
- 1999–2014 г.г. – переход от традиционных алгоритмических подходов к нейросетевым за счет активного развития методов машинного обучения и сопровождающийся развитием аннотированных наборов данных;
- 2014–2023 г.г. – нейросетевые подходы, получившие активное применение с появлением разнообразных аннотированных наборов данных.

2. Традиционные методы оценки позы

В период 1969-1999 гг. – задача оценки позы решалась классическими алгоритмами компьютерного зрения, без применения машинного и тем более глубокого обучения по двум основным причинам: отсутствие должных размеченных датасетов для обучения и тестирования моделей, и отсутствие эффективных технологий обучения глубоких нейросетей (хотя первые нейронные сети для анализа и обработки изображений появились еще в 1989 году [5], но их функциональные возможности ограничивались классификацией «простых» визуальных объектов). Таким образом, используемые на тот момент времени алгоритмы возможно обобщенно разделить на 5 основных групп:

- Модели структур изображений [6, 7].
- Упругие композиции компонент [8].
- Цветовые гистограммы [9, 10].
- Анализ краев и контуров [11, 12].
- Гистограммы ориентированных градиентов [13].

3. Методы на основе машинного обучения

В период 1999-2014 гг. – задача оценки позы стала привлекать все большее внимание исследователей в области компьютерного зрения и машинного обучения. В

этот период произошел значительный прогресс в разработке новых алгоритмов, что привело к более точным и эффективным решениям для оценки позы.

Одной из основных причин этого прогресса стала доступность размеченных наборов данных для обучения и тестирования моделей. С развитием технологий, собирать и организовывать такие наборы данных стало проще, и исследователи стали активно использовать их для обучения моделей оценки позы.

Кроме того, в рассматриваемый период оказались доступными более эффективные технологии для обучения и тестирования моделей машинного обучения, первоначально предложенных задолго до обозначенного периода, что позволило перейти к решению более сложных задач компьютерного зрения. Применение машинного обучения позволило повысить точность оценки позы и улучшить обобщающую способность моделей. В целом большинство алгоритмов машинного обучения, применяемых в решении данной задачи, возможно разделить на 6 основных категорий:

- Методы на основе опорных векторов [14];
- Методы на основе релевантных векторов [15];
- Модели смеси экспертов [16, 17];
- Методы пространственного обучения [18];
- Леса хафа [19];
- Деревья решений [20].

4. Нейросетевые подходы

Начиная с 2014 г. начали появляться первые иерархические и каскадные модели на основе сверточных нейронных сетей [21, 22], представляющие собой последовательности многоклассовых предикторов, а также модули вычислений характеристик изображений и предсказаний, позволяющих изучать и изображения, и контекстные представления признаков на них. В дальнейшем алгоритмы были усовершенствованы таким образом, чтобы последовательная структура прогнозирования изучала неявные пространственные признаки с помощью больших рецептивных полей на картах уверенности (представленных в виде тепловых вероятностных карт) полученных с предыдущих этапов обработки изображений [23]. Это позволило изучать пространственные связи между деталями объекта на больших расстояниях и повысить точность за счет более точных оценок расположения деталей на последующих этапах.

Следует отметить, что со времени создания и в процессе эволюции нейросетевых архитектур вплоть до 2020 г. появлявшиеся подходы не претерпевали существенных изменений. Росли размеры и менялись архитектуры моделей (включая применение новых слоев). Однако большинство вышедших в тот период работ все также были основаны на сверточных слоях и использовали принцип генерации вероятностных тепловых карт для предсказания расположения ключевых точек на изображениях.

В 2020 г. была представлена новая архитектура, основанная на идеях и концепциях применения трансформеров к обработке визуальных данных [24]. Основным преимуществом визуальных трансформеров является их способность моделировать долгосрочные зависимости и взаимодействия между пикселями изображения. В отличие от сверточных нейронных сетей, которые оперируют локальными областями изображения, трансформеры способны учитывать глобальную структуру и контекст, а также масштабируются по отношению к обработке больших изображений и/или видео. Однако трансформеры требуют существенных выборок данных для обучения и существенно более вычислительно затратны, как на этапе обучения, так и в режиме эксплуатационного функционирования, нежели сверточные нейронные сети. Кроме

того, они могут быть более чувствительны к шуму и неустойчивы к изменениям входных данных. Более того, текущая парадигма конструкции их входных структур разрушает топографию изображений, как локальную, так и глобальную. Это связано с тем, что основная идея применения визуальных трансформеров в оценке позы (см., например, работу [25]) заключается в том, чтобы использовать их для моделирования зависимостей между различными частями изображения и позой объекта. Визуальные трансформеры обрабатывают изображение, как правило, как последовательность 1D патчей, а затем моделируют взаимодействие между этими патчами, чтобы получить более точное представление о позе объекта.

Также с появлением первых трансформерных архитектур началась разработка гибридных архитектур, сочетающих визуальные трансформеры с другими типами нейронных сетей. Например, в работе [26] предложена гибридная архитектура, которая комбинирует визуальные трансформеры со сверточными нейронными сетями для решения задачи оценки позы. Гибридные архитектуры используют преимущества обоих типов нейронных сетей: сверточные слои могут эффективно извлекать локальные признаки, а также адаптироваться к сравнительно небольшим наборам данных, при этом визуальные трансформеры могут моделировать глобальные зависимости и взаимодействия между различными частями изображения.

5. Заключение

В данной работе за период с 1969 г. по первую половину 2023 г. были проанализированы подходы, методы и алгоритмы конструктивного решения задачи оценки позы объекта по фотоизображению и видеопоследовательности.

Классические методы оценивания позы, проанализированные в данной работе, отличаются относительной простотой реализации, малой вычислительной сложностью, а также высокой степенью интерпретируемости. Однако описанная группа методов не способна адекватно обобщаться на изменчивость в данных (например вариации поз объектов или различные условия освещения), а также ограничена в части обработки сложных зависимостей между признаками и объектами.

Методы на основе машинного обучения, начавшие активно применяться с появлением и развитием аннотированных наборов данных, являются компромиссными с точки зрения вычислительной сложности и способности к обобщению. Интерпретируемость этих алгоритмов варьируется внутри категорий (выделенных в 3-ем разделе) и прослеживается явная тенденция к ее последовательному уменьшению с развитием технологий. Можно сказать, что методы машинного обучения являются переходной ступенью развития области, от классических алгоритмов к нейросетевым.

Нейросетевые подходы (включая современные трансформерные архитектуры) напротив, отличаются способностью самостоятельного изучения признаков в режиме метаобучения, а также обработки сложных зависимостей между признаками и объектами, что приводит к более точным и емким результатам. Однако данные алгоритмы весьма требовательны к объемам данных (что особенно справедливо для трансформерных архитектур) и намного более ресурсоемкие (особенно на этапе обучения). Еще одним недостатком является гораздо меньшая интерпретируемость результатов.

В целом, возможно заключить, что нейросетевые подходы являются наиболее перспективными и активно развивающимися в данной области. С развитием технологий, доступностью вычислительных ресурсов и объемных наборов данных, крупные модели становятся все более привлекательными, для решения задач оценки позы, за их производительность и способность к обработке сложных зависимостей.

Следовательно, исследования необходимо сосредоточить на устранении критических недостатков и модернизации наиболее перспективных архитектур, например, таких как трансформеры.

Список литературы

1. Chaffin D.B. A computerized biomechanical model—development of and use in studying gross body actions // Journal of biomechanics. 1969. Vol. 2, No. 4. P. 429-441.
2. Nashner L.M. Adapting reflexes controlling the human posture // Experimental brain research. 1976. Vol. 26, No. 1. P. 59-72.
3. Jangade J., Babulal K. S. Study on Deep Learning Models for Human Pose Estimation and its Real Time Application // 2023 6th International Conference on Information Systems and Computer Networks (ISCON). IEEE, 2023. P. 1-6.
4. Zheng C., et al. Deep learning-based human pose estimation: A survey // ACM Computing Surveys. 2023. Vol. 56, No. 1. P. 1-37.
5. Макаренко А.В. Глубокие нейронные сети: зарождение, становление, современное состояние // Проблемы управления. 2020. №. 2. С. 3-19.
6. Lee H.J., Chen Z. Determination of 3D human body postures from a single view // Computer Vision, Graphics, and Image Processing. 1985. Vol. 30, No. 2. P. 148-168.
7. Fischler M.A., Elschlager R.A. The representation and matching of pictorial structures // IEEE Transactions on Computers. 1973. Vol. C-100, No. 1. P. 67-92.
8. Yang Y., Ramanan D. Articulated human detection with flexible mixtures of parts //IEEE transactions on pattern analysis and machine intelligence. 2012. Vol. 35, No. 12. P. 2878-2890.
9. Ekvall S., Hoffmann F., Kragic D. Object recognition and pose estimation for robotic manipulation using color cooccurrence histograms // Proc. 2003 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2003) (Cat. No. 03CH37453). IEEE, 2003. Vol. 2. P. 1284-1289.
10. Roberts T.J., McKenna S.J., Ricketts I.W. Human pose estimation using learnt probabilistic region similarities and partial configurations // ECCV (4). 2004. P. 291-303.
11. Jebara T.S. 3D pose estimation and normalization for face recognition. Centre for Intelligent Machines, McGill University. 1995.
12. Lee M.W., Cohen I. Human upper body pose estimation in static images // Computer Vision-ECCV 2004: 8th European Conference on Computer Vision. Prague, Czech Republic, May 11-14, 2004. Proceedings, Part II 8. Berlin, Heidelberg: Springer, 2004. P. 126-138.
13. Kollnig H., Nagel H.H. 3D pose estimation by directly matching polyhedral models to gray value gradients // International Journal of Computer Vision. 1997. Vol. 23, No. 3. P. 283.
14. Hearst M.A., et al. Support vector machines // IEEE Intelligent Systems and their applications. 1998. Vol. 13, No. 4. P. 18-28.
15. Tipping M. The relevance vector machine // Advances in neural information processing systems. 1999. Vol. 12.
16. Jacobs R.A., Jordan M.I., Nowlan S.J., Hinton G.E. Adaptive mixtures of local experts // Neural computation. 1991. Vol. 3. P. 79-87.
17. Waterhouse S., MacKay D., Robinson A. Bayesian methods for mixtures of experts // Advances in neural information processing systems. 1995. Vol. 8.
18. Gong W., et al. A literature review: Geometric methods and their applications in human-related analysis // Sensors. 2019. Vol. 19, No. 12. P. 2809.
19. Schulter S., et al. On-line Hough Forests // BMVC. 2011. P. 1-11.
20. Quinlan J.R. Induction of decision trees // Machine learning. 1986. Vol. 1. P. 81-106.
21. Ma C., et al. Hierarchical convolutional features for visual tracking // Proceedings of the IEEE international conference on computer vision. 2015. P. 3074-3082.
22. Toshev A., Szegedy C. Deeppose: Human pose estimation via deep neural networks // Proc. of the IEEE conference on computer vision and pattern recognition. 2014. P. 1653-1660.

23. Wei S.E., et al. Convolutional pose machines // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016. P. 4724-4732.
24. Dosovitskiy A., et al. An image is worth 16x16 words: Transformers for image recognition at scale // arXiv preprint arXiv:2010.11929. 2020.
25. Xu Y., et al. Vitpose: Simple vision transformer baselines for human pose estimation // arXiv preprint arXiv:2204.12484. 2022.
26. Aidoo E., et al. Cofopose: Conditional 2D Pose Estimation with Transformers // Sensors. 2022. Vol. 22, No. 18. P. 6821.