Исследование алгоритма детектирования позы человека на изображении и в видеопотоке

Е.В. Авдонина[1], П.Ю. Якимов[1]

[1]Самарский национальный исследовательский университет имени академика С.П. Королёва, Московское шоссе 34А, Самара, Россия, 34443086

Аннотация. Данная статья посвящена исследованию различных алгоритмов детектирования человека на изображении и в видеопотоке. Сравнение алгоритмов на основе точности вычисления, используя датасеты. Выбор алгоритма с наиболее точными результатами и потреблением наименьшего количества ресурсов на устройствах с различными техническими характеристиками.

1. Введение

Распознаванию позы человека на изображениях и в видеопотоке уделяется все большее и большее значение. Необходимость в решении такой задачи возникает в самых разных областях - от военного дела и систем безопасности, до сфер развлечений. Наиболее часто свое применение распознавание объектов находит в системах беспилотного управления, виртуальной реальности, робототехнике, трехмерной визуализации, в задачах управления жестами, и так же в интерактивных инсталляциях.

Множество современных исследований направлены на улучшение алгоритма распознавания поз (например, на увеличение точности и производительности). Рассмотрим существующие подходы к решению задачи и алгоритмы, которые реализуют детектирование позы человека.

2. Обзор существующих алгоритмов детектированию человека на изображении и в видеопотоке

2.1. Классический подход

Основная идея подхода заключается в представлении объекта набором "частей", расположенных в деформируемой конфигурации (не жесткой). "Часть" - это шаблон внешнего вида, который соответствует изображению. Пружины показывают пространственные связи между деталями.

Авторы метода модели деформируемых деталей [1] используют смешанную модель деталей, которая выражает сложные совместные отношения. Деформируемые модели деталей представляют собой набор шаблонов, расположенных в деформируемой конфигурации, и каждая модель имеет глобальный шаблон и шаблоны деталей.

2.2. Подходы, основанные на глубоком обучении

Данные подходы позволяют классифицировать поданное на вход изображение (или сигнал) в соответствии с предварительной настройкой (обучением) сети. Наиболее популярной

нейронной сетью, которая используется в большинстве методов является сверточная нейронная сеть CNN [2], которая ищет определенные паттерны на изображении.

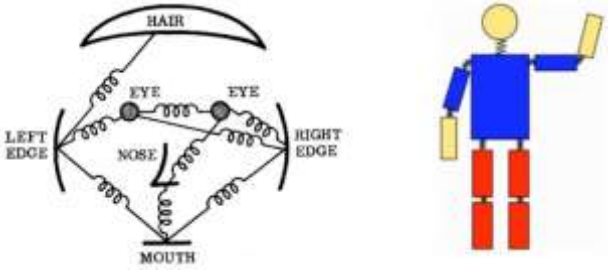

Рисунок 1. Схематичное изображение модели деформируемых деталей.

С появлением “DeepPose”[3], исследования по детектированию позы начали смещаться от классических подходов к нейронным сетям. Метод состоит в том, чтобы сначала установить примерное положение всех суставов, а после уточнить исходные координаты суставов независимо друг от друга.

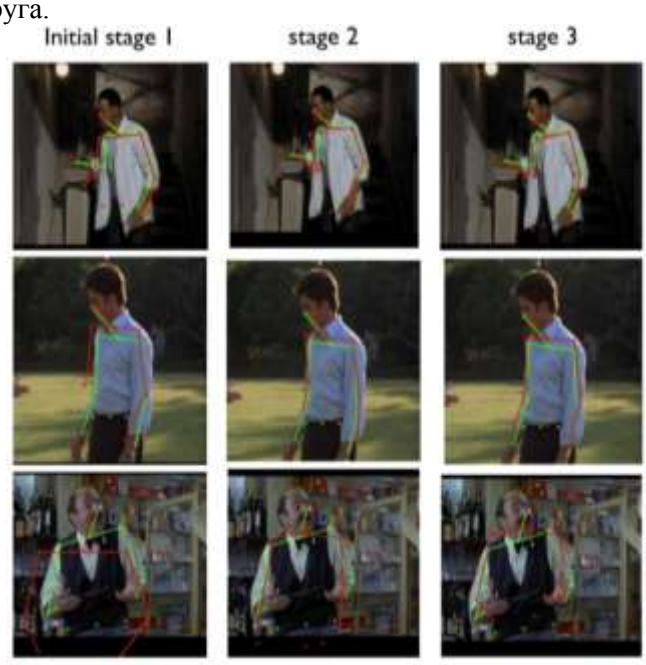

Рисунок 2. На изображениях зеленым цветом нарисована истинная поза, а красным — поза, предсказанная моделью на каждом этапе.

В “ OpenPose ”[4] описывается подход, использующий непараметрическое представление, которое авторы называли полями сходства частей (PAF). Как и во многих подходах снизу-вверх, OpenPose сначала обнаруживает части (характерные точки), не связывая к какому именно человеку они относятся, а затем присваивает части отдельным людям.

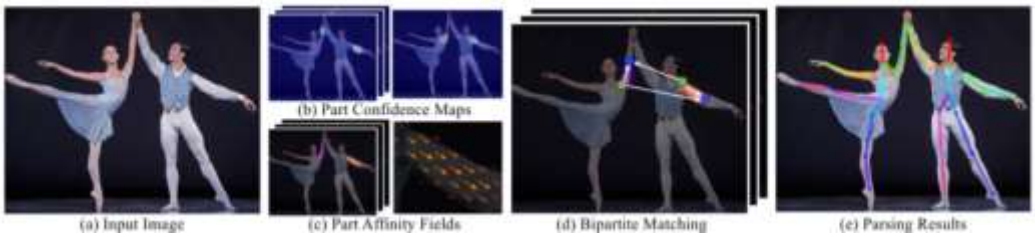

Рисунок 3. Иллюстрирует общую схему метода.

На вход подается изображение любого размера (a) для сверточная нейронной сети (CNN) с двумя ветвями. Метод предсказывает карты достоверности для обнаружения частей тела,

показанные на (b), и поля сходства частей для ассоциации частей, показанные на (c). Этап синтаксического анализа выполняет набор двудольных сопоставлений, чтобы связать кандидатов частей тела (d). Наконец, сопоставления собираются в позы для всех людей на изображении (e).

В своей статье [5] авторы “DeepCut” подошли к поставленной задаче, определив следующие задачи:

1. Производится набор D кандидатов на часть тела. Этот набор представляет все возможные местоположения частей тела для каждого человека на изображении. Выбирается подмножество частей тела из приведенного выше набора кандидатов частей тела.
2. Обозначается каждая выбранная часть тела одним из C классов деталей тела. Классы частей тела представляют собой типы частей, таких как “рука”, “нога”, “туловище” и т.д.
3. Части тела разделяются на группы по принадлежности одному и тому же человеку.

Рисунок 4. Обзор метода: (a) начальные обнаружения (кандидаты в детали) и попарные термины (графы) между всеми обнаружениями, которые (b) объединены в кластеры, принадлежащие одному человеку (один цветной подграф - это один человек), и каждая часть помечена как соответствующая классу его части (разные цвета и символы соответствуют разным частям тела); (в) показывает предсказанные позы линии.

В [6] описывается сеть порождающих разделов (GPN). Авторы предлагают метод, который отличается от существующих моделей, которые являются либо полностью сверху вниз, либо снизу-вверх. GPN предлагает новый подход - он генерирует рамки(боксы) для всех людей на изображении и выводит конфигурации поз одновременно для всех.

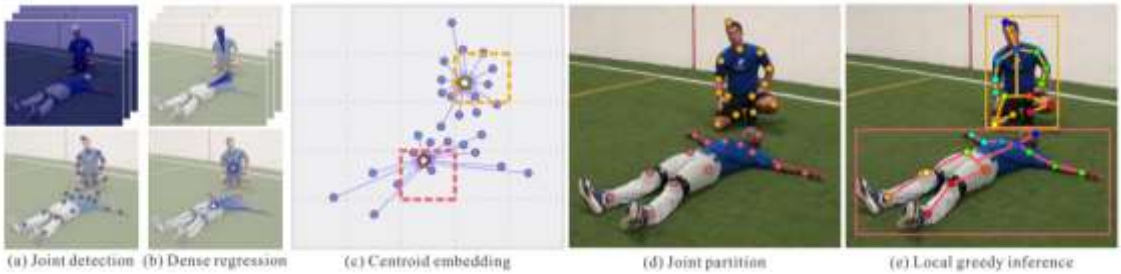

Рисунок 5. Сеть порождающих разделов.

GPN сначала использует CNN, чтобы предсказать (a) совместные карты достоверности и (b) плотные регрессионные карты суставов (модуль плотной регрессии для эффективного и надежного разделения суставов тела нескольких человек, что является ключом к ускорению оценки позы для нескольких человек). Затем GPN выполняет (c) центроидное внедрение для всех совместных кандидатов в пространстве посредством плотной регрессии, чтобы создать (d) совместные разбиения при обнаружении человека. Наконец, GPN проводит (e) локальный вывод для локального создания конфигураций соединений для каждой совместной секции, предоставляя результаты оценки позы для нескольких человек.

3. Постановка задачи

Задачей является сравнение алгоритмов на устройствах с различными техническими характеристиками по потреблению вычислительной мощности и выбор наилучшего алгоритма по отношению потребляемой мощности к точности детектирования позы человека. Есть множество успешных попыток решить данную задачу, помимо рассмотренных ранее [7, 8, 9]. Однако все методы не одинаково производительны. Поэтому для анализа рассмотрим несколько наиболее актуальных и добившихся лучших результатов в состязании на основе датасете MPII[10]. Ниже приведены таблицы результатов для отдельных алгоритмов:

GPU

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Total	Time [s]
Iqbal and Gall [13]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Insafutdinov *et al.* [12]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Levinkov *et al.* [16]	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6	-
Insafutdinov *et al.* [11]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3	-
Cao *et al.* [3]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6	1.24
Fang *et al.* [8]	88.4	86.5	78.6	70.4	74.4	73.0	65.8	76.7	1.5
Newell and Deng [18]	92.1	89.3	78.9	69.8	76.2	71.6	64.7	77.5	-
GPN (Ours)	**92.2**	**89.7**	**82.1**	**74.4**	**78.6**	**76.4**	**69.3**	**80.4**	**0.77**

Рисунок 6. Сравнение с современным уровнем техники по полному тестовому набору данных MPII Human Pose Multi-Person.

OpenPose

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	**mAP**	s/image
Subset of 288 images as in [1]									
Deepcut [1]	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
Iqbal et al. [41]	70.0	65.2	56.4	46.1	52.7	47.9	44.5	54.7	10
DeeperCut [2]	87.9	84.0	71.9	63.9	68.8	63.8	58.1	71.2	230
Newell et al. [48]	91.5	87.2	75.9	65.4	72.2	67.0	62.1	74.5	-
ArtTrack [47]	92.2	**91.3**	80.8	71.4	**79.1**	72.6	67.8	**79.3**	**0.005**
Fang et al. [6]	89.3	88.1	80.7	75.5	73.7	**76.7**	**70.0**	79.1	-
Ours	**92.9**	**91.3**	**82.3**	72.6	76.0	70.9	66.8	79.0	**0.005**
Full testing set									
DeeperCut [2]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Iqbal et al. [41]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Levinko et al. [71]	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6	-
ArtTrack [47]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3	**0.005**
Fang et al. [6]	88.4	86.5	78.6	**70.4**	74.4	**73.0**	**65.8**	76.7	-
Newell et al. [48]	**92.1**	89.3	78.9	69.8	76.2	71.6	64.7	77.5	-
Fieraru et al. [72]	91.8	**89.5**	**80.4**	69.6	**77.3**	71.7	65.5	**78.0**	-
Ours (one scale)	89.0	84.9	74.9	64.2	71.0	65.6	58.1	72.5	0.005
Ours	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6	**0.005**

Рисунок 7. Результаты в наборе данных MPII.

Вверху: результаты сравнения на тестовом подмножестве, определенном в [1]. Посередине: результаты сравнения на всем наборе испытаний. Тестирование без поиска по шкале обозначается как «(одна шкала)».

DeepCut

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
oracle 2000	98.8	98.8	97.4	96.4	97.4	98.3	97.7	97.8	84.0
DPM scale 1	48.8	25.1	14.4	10.2	13.6	21.8	27.1	23.0	13.6
AlexNet scale 1	82.2	67.0	49.6	45.4	53.1	52.9	48.2	56.9	35.9
AlexNet scale 4	85.7	74.4	61.3	53.2	64.1	63.1	53.8	65.1	39.0
+ optimal params	88.1	79.3	68.9	62.6	73.5	69.3	64.7	72.4	44.6
VGG scale 4 optimal params	91.0	84.2	74.6	67.7	77.4	77.3	72.8	77.9	50.0
+ finetune LSP	**95.4**	**86.5**	**77.8**	**74.0**	**84.5**	**78.8**	**82.6**	**82.8**	**57.0**

Рисунок 8. Результаты метода на наборе MPII.

4. Постановка эксперимента

Основываясь на таблицах результатов, выберем сильнейший из них, это OpenPose, посмотрим потребление вычислительной мощности на разных компьютерах:

Таблица 1. Характеристики компьютеров.

Машина	CPU	GPU	GPU (INTEGR.)	RAM(Гб)
A	Intel(R) Core(TM) i5-6200U	NVIDIA GeForce 920MX	Intel(R) HD Graphics 520	12
B	Intel(R) Core(TM) i5-7200U	NVIDIA GeForce MX150	Intel(R) HD Graphics 620	8

Таблица 2. Потребляемые ресурсы. (Отношение потребляемого к общему объему).

Машина	CPU	GPU	GPU (INTEGR.)	RAM
A	47%	60%	1%	39%
B	36%	47%	1%	55%

5. Заключение

Был проведен анализ методов детектирования позы человека на изображении и в видеопотоке, выбраны и сравнены алгоритмы. Лучшим алгоритмом на данных момент является OpenPose, самый точный среди своих конкурентов. Для выбранного алгоритма были проведены исследования на количество потребляемых ресурсов для устройств с различными техническими характеристиками.

6. Благодарности

Работа выполнена в рамках государственного задания по теме FSSS-2020-0017 при частичной поддержке РФФИ: проект № 17-29-03112 офи_м и проект № 19-29-01235 мк».

7. Литература

[1] Yang, Y. Articulated Human Detection with Flexible Mixtures of Parts / Y. Yang, D. Ramanan // Proceedings of the IEEE conference on computer vision and pattern recognition, 2013. – P. 2878-2890.

[2] Escontrela, A. Convolutional Neural Networks from the ground up, 2018 [Electronic resource]. – Access mode: https://towardsdatascience.com/convolutional-neural-networks-from-the-ground-up-c67bb41454e1 (22/12/2019).

[3] Toshev, A. DeepPose: Human Pose Estimation via Deep Neural Networks / A. Toshev, C. Szegedy // Proceedings of the IEEE conference on computer vision and pattern recognition, 2014. – P. 1653-1660.

[4] Cao, Z. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh // arXiv: 1812. 08008v2, 2019.

[5] Pishchulin, L. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation / L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele // arXiv:1511.06645v2, 2016.

[6] Nie, X. Generative Partition Networks for Multi-Person Pose Estimation / X. Nie, J. Feng, J. Xing, S. Yan // arXiv:1705.07422v2, 2017.

[7] Hui, J. Image segmentation with Mask R-CNN, 2018 [Electronic resource]. – Access mode: https://medium.com/@jonathan_hui/image-segmentation-with-mask-r-cnn-ebe6d793272 (22.12.2019).

[8] Oved, D. Real-time Human Pose Estimation in the Browser with TensorFlow, 2018 [Electronic resource]. – Access mode: https://medium.com/tensorflow/real-time-human-pose-estimation-in-the-browser-with-tensorflow-js-7dd0bc881cd5 (22.12.2019).

[9] Fang, H. RMPE: Regional Multi-person Pose Estimation / H. Fang, S. Xie, Y. Tai, C. Lu // arXiv: 1612.00137v5, 2018.

[10] Andriluka, M. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis / M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele // Proceedings of the IEEE conference on computer vision and pattern recognition, 2014. DOI: 10.1109/CVPR.2014.471.

Research of algorithm of detection of a pose of the person on the image and in a video stream

E.V. Avdonina[1], P.U. Yakimov[1]

[1]Samara National Research University, Moskovskoye shosse 34A, Samara, Russia, 443086

Abstract. This article is devoted to the study of various algorithms for detecting a person in an image and in a video stream. Comparison of algorithms based on calculation accuracy using datasets. The choice of the algorithm with the most accurate results and the consumption of the least amount of resources on devices with different technical characteristics.