ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2023, № 4, с. 121–124

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.93.

ПРИВИЛЕГИРОВАННОЕ ОБУЧЕНИЕ С ПОМОЩЬЮ РЕГУЛЯРИЗАЦИИ В ЗАДАЧЕ ОЦЕНКИ ПОЗЫ ЧЕЛОВЕКА

© 2023 г. М. С. Каприелова[a,*], Р. Г. Нейчев[b,], А. Д. Тихонова[b,***]**

[a]*Федеральный исследовательский центр "Информатика и управление" РАН, Москва, Россия*

[b]*Московский физико-технический институт, Москва, Россия*

**e-mail: kaprielova.ms@phystech.edu*

***e-mail: neychev@phystech.edu*

****e-mail: tikhonova.ad@phystech.edu*

Поступила в редакцию 03.01.2022 г.
После доработки 08.01.2023 г.
Принята к публикации 06.02.2023 г.

Решается задача оценки позы человека по видеоданным. Производится анализ различных ключевых точек тела человека. Исследуется изменение точности фиксированной модели при использовании различных пропорций в регуляризационном слагаемом функции потерь. Показано, что при фиксированном количестве тренировочных эпох точность модели отличается в зависимости от выбранных пропорций. Кроме того, продемонстрировано, что линейная корреляция между траекториями ключевых точек, входящих в состав регуляризационного слагаемого, не является основным критерием при прогнозировании эффективности применения регуляризационного слагаемого функции потерь.

DOI: 10.31857/S000233882303006X, **EDN:** EULHZW

Введение. Оценка позы человека является одной из активно исследуемых задач в компьютерном зрении. Интерес к решению данной задачи обусловлен как прямыми прикладными результатами, так и возможностью получения простого и информативного описания тела и поведения человека. В качестве иллюстрации прямого применения могут выступать такие прикладные задачи, как оценка корректности определенных позиций в технически сложных видах спорта (например, в гимнастике, фигурном катании, гольфе), отслеживание движений в виртуальной реальности или же управление различными устройствами (такими, как дроны или персональные компьютеры) без использования дополнительных физических устройств. Эти приложения получили широкое развитие в последнее десятилетие, в первую очередь благодаря интенсивному продвижению области глубокого обучения и значительным результатам в построении искусственных нейронных сетей. Промежуточные, или скрытые, латентные представления данных внутри нейронных сетей также широко применяются в решении других задач. Например, для распознавания языка жестов (определенных последовательностей движений рук) могут использоваться представления из нейронных сетей для оценки позы с предпоследних слоев. Обнаружение аномалий в поведении также может производиться как на основе итоговой оценки позы, так и на базе латентных представлений, порождаемых моделью. Еще одной задачей, решаемой с помощью предварительной оценки позы, является распознавание движений человека. Аналогично задаче распознавания изображений обучение моделей оценки позы человека позволяет построить достаточно простое и информативное признаковое описание тела человека, которое потом может быть использовано в других областях: безопасности, медицине и т.д. Учитывая возможные условия съемки, решения задачи оценки позы человека должны быть устойчивы к частичному или полному перекрытию ключевых точек, различиям в телосложении людей и силуэтов одежды и изменению количества людей в кадре. Для этого могут применяться априорные предположения [1] о структуре тела человека и характере его движений. Наличие информации о некоторых инвариантах в наблюдаемых данных часто используется в задачах компьютерного зрения [2]. В случае оценки позы в качестве априорной информации могут быть применены пропорции тела человека: они не зависят от угла съемки и расстояния до объекта. Один из способов учета априорной информации непосредственно на этапе обучения модели – добавление регуля-

ризационного слагаемого к оптимизирующей функции потерь. В данной работе в качестве априорной рассматривается информация о различных пропорциях человека. Предшествующие этой работе исследования показали, что некоторые пропорции могут выступать в роли регуляризатора более эффективно, чем остальные. Данная работа представляет анализ различных ключевых точек тела человека. Проводится исследование процесса обучения модели оценки позы человека при использовании различных пропорций в регуляризационном слагаемом. Экспериментальная проверка осуществляется на реальных данных о движениях человека, представленных в датасете Human3.6m [3].

1. Обзор литературы. Задача распознавания активности человека требует анализа временных рядов и построения информативных признаковых представлений. В работе [4] приводится подход распознавания активности человека с помощью временных данных, полученных с IMU-датчиков, которые включают в себя показания акселерометров и гироскопов. Эти данные позволяют оценить траекторию движения частей тела человека в пространстве. Для упрощения работы с временными рядами высокой сложности предлагается построение матрицы фазовых траекторий. Но активность человека описывается информацией о движении множества точек, которая часто может быть избыточной. Для построения более простого признакового пространства [5] могут использоваться методы сравнения временных рядов и оценки связи между ними, описываемые в [6]. Также при анализе сложных зависимостей крайне полезными будут априорные предположения [2], включающие в себя как информацию о физических свойствах рассматриваемых процессов, так и некоторые ограничения на решение задачи. В анализе одновременно могут использоваться данные различной природы и частоты, что позволяет получить избыточное описание процесса и построить более точное решение, устойчивое к шумам и возможным пробелам в данных (например, из-за перекрытия частей тела на этапе оценки позы человека). Также дополнительные данные выступают в роли привилегированной информации [1]. Кроме того, в литературе встречаются работы, где знания о физике взаимодействий применяют в качестве предпосылки для разработки новых архитектур, для использования в виде дополнительного слагаемого в функции ошибки или жестких ограничений на выходные данные нейросети [8]. Однако количество ключевых точек, взаимодействия которых требуется описать для реализации такого подхода, значительно повышает вычислительную сложность решения [9]. Анализ данных разметки может сократить размерность признакового пространства и снизить вычислительную сложность решений на основе согласованности с физическими принципами.

2. Определение наиболее эффективных пропорций. В общем случае задачу оценки позы по видеоданным можно сформулировать следующим образом. Существует последовательность кадров $Q = \{q_1,\ldots,q_t\}$. Также существует набор ключевых точек $J = \{j_1,\ldots,j_k\}, j_i \in \mathbb{R}^3, \forall i : j_i = \{x_i, y_i, z_i\}$ на теле человека для каждого кадра. Количество ключевых точек может меняться в зависимости от специфики конкретного практического приложения задачи, но обычно ключевые точки выбираются на запястьях, локтях, плечах, голове, корпусе, бедрах, коленях и лодыжках. Моделью оценки позы человека назовем отображение $f : Q \mapsto J$. В этой части работы мы проанализировали датасет Human3.6m [9]. Пусть J – множество ключевых точек на теле человека. Каждой ключевой точке j_i соответствует три координаты $\{x_i, y_i, z_i\}$. Временным рядом ключевой точки для видеоряда будем называть последовательность значений, принимаемых $x_i = \{x_1,\ldots,x_t\}$, $y_i = \{y_1,\ldots,y_t\}$ или $z_i = \{z_1,\ldots,z_t\}$ на отрезке времени, соответствующем длине видеоряда. Гипотеза: временные ряды некоторых ключевых точек имеют линейную корреляцию. Обработка временных рядов производится по следующему алгоритму:

1) гауссово сглаживание,

2) дифференцирование.

Затем для временных рядов считается линейная корреляция. Используется линейная корреляция Пирсона:

$$r = \frac{\sum_{k=1}^{t}(x_k - \overline{x})^2(y_k - \overline{y})^2}{\sqrt{\sum_{k=1}^{t}(x_k - \overline{x})^2 \sum_{k=1}^{t}(y_k - \overline{y})^2}},$$

где $\overline{x}$ и $\overline{y}$ – средние x и y соответственно. Результаты анализа датасета представлены в табл. 1.

Таблица 1. Результат анализа данных

Группа	Ноги	
	Колено–лодыжка	Бедро–лодыжка
Среднее значение	0.43	0.27
Группа	Руки	
	Локоть–кисть	Плечо–кисть
Среднее значение	0.76	0.56
Группа	Корпус	
	Шея–таз	Голова–таз
Среднее значение	0.73	0.66
Группа	Бедра	Плечи
Среднее значение	0.89	0.79

Таким образом, все рассмотренные группы точек имеют линейно скоррелированные ряды, но средние значения линейной корреляции отличаются в зависимости от группировки точек. Так, руки в группировке локоть–кисть в среднем сильнее скоррелированы, чем в группировке плечо–кисть. То же самое можно сказать о группировке колено–лодыжка и бедро–лодыжка: первый вариант группировки показывает более сильную корреляцию, чем второй. На основании проведенного анализа данных можно предположить, что наиболее скоррелированные группы точек могут быть использованы в качестве регуляризационного слагаемого функции потерь, накладывающего ограничения на пропорции человека. Применение именно этих пропорций в составе регуляризатора предположительно наиболее эффективно для улучшения сходимости модели.

3. Сравнение регуляризаторов. Проанализируем, как использование в регуляризационных слагаемых точек из разных групп влияет на точность решения. В качестве исследуемых групп точек рассматриваем руки (локоть–кисть), корпус (голова–таз), корпус (шея–таз), ноги (колено–лодыжка), бедра и плечи. Мы обучали модель Poseformer [10] на датасете Human36M [3] с оригинальной функцией ошибки (MPJPE) и модифицированной (MPJPE + ProportionLoss) для каждой группы точек. Для предсказания положения человека в 3D используем реальные 2D-позы людей с 81 кадра. Для моделей с регуляризатором функция ошибки строится следующим образом:

$$L = L_{mpjpe} + L_{prop} = \frac{1}{J}\sum_{k=1}^{J}\|a_k - \widehat{a_k}\|_2 + \frac{1}{M}\sum_{n=1}^{M}(p_n - \widehat{p_n})^2,$$

где a_k и $\widehat{a_k}$ – реальное и предсказанное положение k-й ключевой точки в 3D, p_n и $\widehat{p_n}$ – реальное и предсказанное евклидово расстояние между двумя точками n-й группы. Другими словами, к стандартной MPJPE добавим регуляризатор на расстояние между крайними точками внутри каждой группы, который можно интерпретировать как информацию о пропорциях человеческого тела. Обучение всех моделей проводилось с дефолтными параметрами, как в оригинальной имплементации Poseformer [10], оптимизатором Adam с начальными гиперпараметрами $lr = 0.0001$ и $weightdecay = 0.1$. Разбиение датасета на обучающую и тестовую выборку проводится по аналогии [10]. Мы оценивали точность модели с помощью метрик, использованных в оригинальной статье Poseformer [10]. MPJPE – стандартная метрика, которая является усредненным Евклидовым расстоянием между реальными и предсказанными ключевыми точками в миллиметрах. P-MPJPE – это MPJPE после ригидного выравнивания предсказаний относительно реальных точек.

4. Результаты эксперимента. Для каждой группы точек результаты усреднялись по пяти запускам. Результаты представлены в табл. 2. Из результатов эксперимента видно, что вне зависимости от величины корреляции использование регуляризационного слагаемого положительно влияет на обучение модели. Интересно, что некоторые менее скоррелированные группы эффективнее, чем более скоррелированые. Примером таких групп являются ноги (колено–лодыжка) и корпус (голова–таз). Скорее всего, это объясняется тем, что некоторые точки, входящие в более скоррелированные группы, часто подвержены окклюзиям. Окклюзиями называется частичное

Таблица 2. Результаты обучения модели с учетом различных пропорций

Группа	MPJPE	P-MPJPE
Руки (локоть–кисть)	37.73	28.32
Корпус (голова–таз)	37.95	28.10
Корпус (шея–таз)	38.22	28.58
Ноги (колено–лодыжка)	38.62	29.61
Бедра	40.60	31.12
Плечи	41.44	32.21
Без регуляризатора	42.87	32.30

или полное перекрытие одной или нескольких ключевых точек. Появление такого эффекта обусловлено возможностью съемки с разных ракурсов. Примерами таких групп могут быть бедра и плечи.

Заключение. Временные ряды, характеризующие движение различных ключевых точек тела человека, проанализированы на наличие линейных корреляций. В результате были выявлены наиболее и наименее скоррелированные группы ключевых точек. Кроме того, исследовали процесс обучения фиксированной модели при помощи различных пропорций в регуляризационном слагаемом. Результаты показали, что при использовании регуляризатора, основанного на пропорциях человека, линейная корреляция может быть критерием выбора, однако не является единственным фактором, влияющим на эффективность регуляризационного слагаемого.

СПИСОК ЛИТЕРАТУРЫ

1. *Vapnik V., Vashist A.* A New Learning Paradigm: Learning Using Privileged Information // Neural Networks. 2009. V. 22. P. 544–557.
2. *Lehrmann A., Gehler P., Nowozin S.* A Non-parametric Bayesian Network Prior of Human Pose // Proc. IEEE Intern. Conf. On Computer Vision. Sydney, 2013. P. 1281–1288.
3. *Ionescu C., Papava D., Olaru V., Sminchisescu C.* Human3. 6m: Large Scale Datasets and Predictive Methods for 3d Human Sensing in Natural Environments // IEEE Trans. On Pattern Analysis And Machine Intelligence. 2013. V. 36. P. 1325–1339.
4. *Ignatov A., Strijov, V.* Human Activity Recognition Using Quasiperiodic Time Series Collected from a Single Tri-axial Accelerometer // Multimedia Tools And Applications. 2016. V. 75. P. 7257–7270.
5. *Katrutsa A., Strijov V.* Stress Test Procedure for Feature Selection Algorithms // Chemometrics And Intelligent Laboratory Systems. 2015. V. 142. P. 172–183.
6. *Cliff O., Lizier J., Tsuchiya N., Fulcher B.* Unifying Pairwise Interactions in Complex Dynamics // ArXiv 2022. *ArXiv Preprint ArXiv:2201.11941.*
7. *Trumble M., Gilbert A., Malleson C., Hilton A., Collomosse J.* Total Capture: 3d Human Pose Estimation Fusing Video and Inertial Sensors // Proc. Of 28th British Machine Vision Conf. London, 2017. P. 1–13.
8. *Márquez-Neila P., Salzmann M., Fua P.* Imposing Hard Constraints on Deep Networks: Promises and Limitations // *ArXiv Preprint ArXiv:1706.02025* (2017).
9. *De Luca G., Lampoltshammer T., Scholz, J.* How Many Equations of Motion Describe a Moving Human? // *ArXiv Preprint ArXiv:2207.14331* (2022).
10. *Zheng C., Zhu S., Mendieta M., Yang T., Chen C., Ding, Z.* 3d Human Pose Estimation with Spatial and Temporal Transformers // Proc. IEEE/CVF Intern. Conf. On Computer Vision. Montreal, 2021. P. 11656–11665.