Обзор методов распознавания человека по походке в видео

А.И. Соколова[1], А.С. Конушин[1,2]
ale4kasokolova@gmail.com|anton.konushin@graphics.cs.msu.ru
[1]Национальный Исследовательский Университет Высшая Школа Экономики, Москва, Россия;
[2]Московский Государственный Университет им. Ломоносова, Москва, Россия

Походка - важный биометрический показатель, позволяющий идентифицировать человека без личного контакта. В данной работе проводится обзор современных методов распознавания человека по походке, их анализ и сравнение, а также выявляются проблемы, препятствующие окончательному решению задачи распознавания по походке.

Ключевые слова: *походка, биометрия, силуэт, нейронные сети, идентификация.*

Review of video gait recognition methods

A. Sokolova[1], A. Konushin[1,2]
ale4kasokolova@gmail.com|anton.konushin@grapics.cs.msu.ru
[1]National Research University Higher School of Economics, Moscow, Russia;
[2]Lomonosov Moscow State University, Moscow, Russia

Human gait is an important biometrical index that allows to identify the human without any contact. In this paper, we provide a survey of state-of-the-art methods of gait recogntion, their analysis and comparison. We additionally reveal the problems that prevent the final solution of gait recognition challenge.

Keywords: *gait, biometrics, silhouette, neural networks, identification.*

1. Введение

Задача распознавания человека в видео по походке особенно актуальна в современном мире. Как показывают биометрические исследования, каждый человек обладает своей манерой движения, которую практически невозможно подделать, вследствие чего походка оказывается уникальным идентификатором, подобным сетчатке глаза или отпечаткам пальцев. С развитием систем видеонаблюдения походка становится наиболее удобной характеристикой для распознавания благодаря возможности наблюдать этот показатель без ведома человека. В связи с этим идентификация человека по походке может быть применима, например, в сфере безопасности для распознавания преступников или ограничения доступа на закрытые территории. Задача распознавания человека в видео усложняется множеством факторов, влияющих как на представление походки, так и на само движение: ракурс, разная одежда и переносимые предметы и изменение походки человека со временем.

В статье представлен обзор методов распознавания человека по походке в видео, а также их сравнение на популярных наборах данных.

На сегодняшний день существует два основных подхода к получению признаков походки и их классификации: построение признаков вручную и обучение признаков. Первый способ более традиционен и, как правило, основывается на вычислении различных свойств бинарных масок силуэта человека или на исследовании взаимного расположения суставов, относительных расстояний и скоростей и других кинетических показателей. Обучение признаков характерно для искусственных нейронных сетей, набравших популярность в последние годы благодаря выдающимся результатам в решении многих задач компьютерного зрения, таким как классификация видео и изображений, сегментация изображений, детекция объектов, визуальный трекинг и другие. Признаки, обучаемые с помощью нейронных сетей, часто обладают более высоким уровнем абстракции, необходимым для качественного распознавания. Кроме того, высокое качество идентификации достигается методами, комбинирующими два описанных подхода. На начальном этапе вручную вычисляются базовые характеристики походки, а на их основе обучается нейронная сеть, выделяющая более абстрактные признаки. Несмотря на успешность методов глубинного обучения, на данный момент наилучшего результата на некоторых наборах данных достигают неглубокие алгоритмы, поэтому оба глобальных подхода достойны внимания.

2. Базовые признаки походки

Рассмотрим сначала некоторые базовые подходы, в которых признаки походки извлекаются вручную из естественных соображений.

2.1 Бинарные силуэты человека

Наиболее распространенной характеристикой походки является изображения энергии походки (Gait Energy Image, GEI [5]). Такие изображения − усредненные по одному циклу походки бинарные маски силуэта движущегося человека. Изображения энергии походки характеризуют частоты нахождения человека в той или иной позе во время движения. Этот подход получил широкое распространение и лег в основу множества других методов распознавания походки. Кроме того, многие подходы предлагают похожую агрегацию других базовых признаков. Например, распознавание возможно по изображениям энтропии походки [1], где вместо усреднения силуэтов вычисляется энтропия каждого пикселя, или по дискретному преобразованию Фурье набора силуэтов [11].

Несмотря на простоту всех этих методов, именно изображения энергии используются и развиваются до сих пор. По таким изображениям можно вычислять дальнейшие признаки, такие как гистограммы ориентированных градиентов, или строить более сложные алгоритмы классификации, использующие специфику задачи распознавания походки.

Так, одними из наиболее успешных многоракурсных подходов являются два неглубоких метода, использующих в качестве базовых признаков изображения энергии походки. Первый из них − байесовский подход, предложенный Ли в [9]. Авторы предлагают считать

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

изображения энергии походки случайными матрицами, получающимися из собственно походки и независимого от нее шума, соответствующего различным изменяющимся условиям, причем предполагается, что оба слагаемых − нормальные случайные величины. Рассмотрение совместного распределения двух представлений походки в предположении совпадения классов или их различия сводит проблему к задаче оптимизации ковариационных матриц, решаемую с помощью EM-алгоритма. Во втором подходе [12] предлагается проводить многоракурсный дискриминантный анализ. Для признаков походки, посчитанных для каждого угла обзора, обучается отдельное вложение, так чтобы внутриклассовый разброс был минимален, а межклассовый − максимален.

Оба эти подхода интуитивно понятны и математически просты, что в дополнение к высоким результатам распознавания дает им преимущество перед многими другими методами. Общим недостатком методов, использующих GEI для многоракурсного распознавания, является необходимость вычислять изображение энергии для каждого ракурса, присутствующего в выборке. Поэтому для каждого кадра нужно знать, под каким углом он был снят, что не всегда возможно в реальных данных.

2.2 Поза человека

Помимо бинарных масок силуэта и всевозможных способов их агрегации многие исследователи предлагают обратить внимание на позу человека, а именно положение ключевых точек фигуры (основных частей тела и суставов) в каждом кадре. Так, например, в работе [22] в каждом кадре оценивается скелет человека и исследуется движение ключевых точек: период походки, скорость, траектории бедер, коленей и лодыжек, а также относительные углы между точками тела. Лу [10] в своем подходе предлагает многослойную деформируемую модель, характеризующую форму и динамику частей тела человека: их относительные размеры, позицию и ориентацию. Параметры такой модели, определяющие позу, восстанавливаются по имеющемуся (или оцененному каким-либо методом) силуэту, а по ним, в свою очередь, производится распознавание человека.

2.3 Траектории точек фигуры

Еще один неглубокий подход, показывающий высокое качество распознавания, предложен в [3], где рассматриваются траектории движения точек фигуры человека и по этим траекториям строятся дескрипторы движения Фишера, которые классифицируются методом опорных векторов.

3. Нейросетевые подходы

Несмотря на обилие структурных неглубоких подходов, сверточные нейронные сети (CNN) занимают прочную позицию во всех задачах компьютерного зрения и в том числе распознавании походки. За последние несколько лет было предложено множество нейросетевых методов идентификации по походке, отличающихся как технически (выбором архитектур сетей, функции потерь, способов обучения), так и идейно - методом обработки данных и извлечения первичных признаков, подаваемых на вход сети. В некоторых работах основным источником информации считаются сами кадры видео, и именно они подаются на вход сети. Как и для смежной задачи распознавания действий [8], классификация по отдельным кадрам дает удовлетворительный результат. Однако во многих подходах основное внимание уделяется именно движению фигуры человека, поэтому мы остановимся подробнее на нескольких более сложных моделях.

Одним из способов предобработки данных перед подачей в сеть является извлечение оптического потока − векторного поля видимого движения точек сцены. Преимущество этого подхода состоит в том, что обученная на таких данных модель не обращает внимания на цвет, яркость или контрастность кадров видео. Влияние на распознавание оказывает только движение отдельных точек фигуры, а именно это и составляет походку человека.

В нескольких работах [2, 16], появившихся практически одновременно, предлагается рассматривать блоки карт оптического потока аналогично временной составляющей классической двухпоточной модели [15] для распознавания действий. Для нескольких идущих подряд пар соседних кадров вычисляется оптический поток и строится тензор − блок из нескольких карт потока. Для большей точности из этого блока вырезается часть, во всех кадрах содержащая фигуру человека, и на таких блоках обучается нейронная сеть. На этапе тестирования сеть используется для извлечения признаков, которые потом можно классифицировать любым методом машинного обучения, например, методом опорных векторов (SVM) или методом ближайшего соседа (kNN).

Второй и наиболее популярный источник информации, на основе которого происходит обучение сетей, − бинарные маски силуэтов, о которых уже шла речь при рассмотрении неглубоких методов. В простейшем случае [24] сверточная архитектура обучается по отдельным силуэтам предсказывать человека, которому этот силуэт принадлежит. Как и предыдущих методах, сеть в дальнейшем используется для извлечения признаков, причем переход от дескрипторов отдельных кадров ко всему видео происходит путем выбора максимального отклика по циклу походки. Этот метод наиболее простой из всех глубоких подходов, т.к. при наличии масок силуэтов людей не требует практически никакой дополнительной предобработки.

Еще один метод, использующий сами силуэты, предложен в [18]. Двухэтапный алгоритм сначала определяет угол съемки видео, а потом по изначальным данным и найденному углу предсказывает человека. Для первой задачи случайным образом выбирается набор кадров видео и поступает на вход сети, предобученной для задачи классификации видео. В зависимости от предсказанного угла, данные подаются в новую сеть, свою для каждого ракурса. Чтобы учитывать не только пространственные, но и временные характеристики движения, на вход сети подаются не отдельные маски, а блоки из нескольких силуэтов, идущих подряд. Кроме того, изменчивость между кадрами учитывается благодаря архитектурам сетей в обеих подзадачах: авторы используют трехмерные сети, в которых свертки производятся не только в пространстве, но и во времени.

Трехмерные свертки применяются также в [20], где источником информации служит комбинированное трехканальное изображение: в качестве первого канала берется сам кадр, приведенный к черно-белому виду, а два других − компоненты карт оптического потока. Такая модель, благодаря своей структуре, также учитывает и пространственные характеристики видео, и временные.

Отдельно стоит выделить множество методов, совмещающих “ручное” выделение признаков, и глубинное обучение. В качестве входных данных для сетей часто используют изображения энергии походки, упоминавшиеся выше. Основанные на этой идее модели варьируются от простейших [14], где неглубокая сеть предсказывает человека по получаемым изображениям энергии, вычисленным для различных ракурсов, до более сложных, как, например, [21], где определяется степень похожести пары GEI изображений и исследуются различные методы

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

сравнения нейросетевых признаков походки. В [17, 23] также с помощью двух- и трехпоточных сиамских архитектур определяется, какие изображения походки близки и принадлежат одному человеку, а какие − разным.

Наряду с классическими сверточными сетями прямого распространения для распознавания по походке, как и для других задач анализа видео, используются рекуррентные нейросети. В них также можно подавать всевозможные данные: сами бинарные силуэты, как это делают в [19], или более сложные преобразованные данные, например, информацию о позе человека в каждом кадре [4].

Многие из обсуждаемых подходов оцениваются на одних и тех же данных при одних и тех же условиях, в этом обзоре мы приводим сравнение некоторых описанных методов и выделяем наиболее успешные решения.

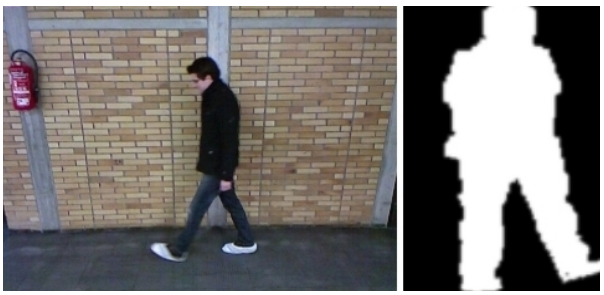

Рис. 1. Изображения кадров из базы TUM-GAID (слева) и из базы OU-ISIR (справа).

4. Базы данных для распознавания человека по походке

Наиболее широко используемыми в наше время сложными наборами данных для распознавания человека по походке являются базы OU-ISIR Large Population Dataset [7] и TUM-GAID [6]. Первый набор применим для многоракурсного распознавания, так как состоит из видеопоследовательностей для более, чем 4000 человек, снятых двумя камерами, причем ракурс съемки плавно меняется от 55° до 85°. Данные из этой коллекции распространяются в виде масок силуэтов, поэтому не все описанные методы применимы к этой базе данных.

База TUM предназначена для распознавания сбоку, все видео в ней сняты под углом 90°, она гораздо меньше предыдущей (по 10 видео для 305 человек), однако состоит из полноценных цветных видео, что делает ее применимой для большого количества подходов. Кроме того, в этой базе присутствуют данные, снятые с разницей в полгода, что дает возможность проверить устойчивость алгоритмов к изменениям походки со временем. Примеры кадров из обеих баз изображены на рис. 1.

Многие из описанных методов оценены на этих наборах данных, поэтому мы приведем сравнение подходов на них.

5. Результаты работы методов

На описанных базах данных оценка качества алгоритмов происходит следующим образом. Модель настраивается на части данных (как правило, это все видео для некоторого подмножества людей), а затем тестируется на новом наборе данных, состоящем из видео для других людей. Для баз TUM и OU-ISIR разделение на обучающую и тестовую выборки предоставлено авторами коллекций. В качестве метрики качества обычно рассматривается точность распознавания.

Метод	Точность
Sokolova [16]	97.5%
Castro, SNN+ SVM [2]	98.0%
Marín. Jiménez [13]	98.9%
Castro, дескрипторы Фишера [3]	99.2%
Zhang [24]	97.7%

Таб. 1. Результаты распознавания на базе TUM-GAID.

В таблице 1 приведено сравнение результатов распознавания на базе TUM-GAID. Наилучшего качества достигает единственный метод [3], не использующий нейронные сети, что показывает, что нейросетевые подходы пока не могут полностью вытеснить неглубокие.

Интересно также рассмотреть, насколько алгоритмы устойчивы ко времени съемки видео. Оказывается, что идентификация человека в случае, когда между несколькими съемками проходит длительный период времени, гораздо сложнее, и качество таких экспериментов оказывается невысоким. Таблица 2 отражает результаты соответствующих экспериментов. Точность каждого из представленных алгоритмов падает примерно на 30%, что говорит о том, что на сегодняшний день алгоритмы плохо переносимы во времени.

Метод	Точность
Castro, CNN + SVM [2]	59.4%
Marín. Jiménez [13]	63.6%
Castro, дескрипторы Фишера [3]	60.4%

Таб. 2. Сравнение результатов распознавания видео, снятых в разные дни, на базе TUM-GAID.

Для многоракурсных баз сравнение, как правило, производится для всевозможных пар ракурсов: данные, снятые под некоторым “тестовым” углом классифицируются алгоритмом, при настройке которого используется другой, “обучающий” угол съемки.

Метод	0°	10°	20°	30°
Zhang [24]	94.1%	71.6%	21.8%	2.9%
Shiraga [14]	94.9%	93.9%	90.5%	80.65%
Li [9]	98.3%	98.2%	97.3%	94.6%
Mansur [12]	96.8%	96.3%	94.2%	90.3%

Таб. 3. Результаты распознавания на базе OU-ISIR.

Для базы OU-ISIR есть два популярных протокола тестирования. Один из них предоставлен авторами: из коллекции выбрано 1912 человек, которые пятью способами разделяются пополам на обучающую и тестовую выборки, после чего качество моделей усредняется. Второй протокол реализует кросс-валидацию, причем модели строятся на данных для 3844 людей, для которых в базе присутствуют данные с обеих камер. Для удобства результаты сравнения агрегируют, рассматривая разности между “обучающим” и “тестовым” углами. В таблице 3 показана средняя точность методов для каждого из 4 возможных значений разности углов. Самый простой метод [24] оказывается несостоятелен, когда углы съемки сильно отличаются, однако остальные подходы показывают достаточно высокое качество распознавания. Наилучших результатов при таких экспериментах тоже достигает метод, не использующий нейронные сети.

Метод	0°	10°	20°	30°
Shiraga [14]	96.5%	95.8%	92.5%	84.9%
Wu [21]	98.9%	95.5%	92.4%	85.3%
Takemura [17]	99.3%	99.2%	98.6%	96.9%

Таб. 4. Результаты кросс-валидации на базе OU-ISIR.

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

Однако при использовании большего количества данных для обучения и тестирования нейросетевые методы оказываются очень успешны. В таблице 4 приведены результаты сравнения алгоритмов при использовании данных для почти 4 тысяч людей.

Благодаря большему размеру обучающей выборки методы, использующие такой протокол тестирования, достигают более высокой точности. Отсутствие открытых реализаций и общего протокола тестирования не дает возможности сравнить все методы и найти оптимальный, однако имеющиеся результаты показывают, что на сегодняшний день глубокие и неглубокие методы развиваются и показывают практически равное качество.

6. Заключение

Несмотря на все многообразие предлагаемых подходов, задача распознавания походки все еще не потеряла актуальности: существующие решения пока не достигли идеальной точности идентификации. Множество различных условий, влияющих на представление движения, и отсутствие достаточно больших наборов данных, способных учесть все возможные вариации походки, препятствует созданию совершенной модели.

7. Благодарности

Работа выполнена при поддержке гранта #16-29-09612 Российского фонда фундаментальных исследований.

8. Литература

[1] Bashir, K., Xiang, T., S, G. Gait recognition using gait entropy image // Proceedings of 3rd international conference on crime detection and prevention, 2009

[2] Castro, F.M., Marín.Jiménez, M.J., Guil, N., Pérez de la Blanca, N. Automatic learning of gait signatures for people identification // Advances in Computational Intelligence. 2017. pp. 257−270

[3] Castro, F.M., Marín.Jiménez, M., Medina Carnicer, R. Pyramidal Fisher Motion for multiview gait recognition // 22nd International Conference on Pattern Recognition, 2014, pp. 692−1697

[4] Dan.Liu, X.L.F.Z. Mao~Ye, Lin, L. Memory-based gait recognition // Proceedings of the British Machine Vision Conference (BMVC), 2016, pp. 82.1−82.12

[5] Han, J., Bhanu, B. Individual recognition using gait energy image // IEEE Trans Pattern Anal Mach Intell, 2006, 28, pp. 316−322

[6] Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., Rigoll, G.: The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits // J of Visual Com and Image Repres, 2014, 25(1), pp.195 − 206

[7] Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition // IEEE Trans on Information Forensics and Security, 2012, 7, Issue 5, pp.1511−1521

[8] Karpathy, A. Toderici, G. Shetty, S. Leung, T. Sukthankar, R and Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition.

[9] Li, C., Sun, S., Chen, X., Min, X. Cross-view gait recognition using joint Bayesian // Proc. SPIE 10420, Ninth International Conference on Digital Image Processing, 2017

[10] Lu, H. Plataniotis, K. N. Venetsanopoulos, A. N. A Full-Body Layered Deformable Model for Automatic Model-Based Gait Recognition // EURASIP Journal on Advances in Signal Processing, 2007, pp 261−317

[11] Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y. Gait recognition using a view transformation model in the frequency domain // Computer Vision − ECCV 2006, 2006, pp. 151−163

[12] Mansur, A., Makihara, Y., Muramatsu, D., Yagi, Y. Cross-view gait recognition using view-dependent discriminative analysis // IEEE/IAPR International Joint Conference on Biometrics, 2014

[13] Marín. Jiménez, M., Castro, F., Guil, N., de la Torre, F., Medina.Carnicer, R. 'Deep multi-task learning for gait-based biometrics'.// IEEE International Conference on Image Processing (ICIP), 2017

[14] Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y. GEINet: View-invariant gait recognition using a convolutional neural network // International Conference on Biometrics (ICB)}, 2016

[15] Simonyan, K., Zisserman, A. Two-stream convolutional networks for action recognition in videos // Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014, pp. 568−576

[16] Sokolova, A., Konushin, A. Gait recognition based on convolutional neural networks // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. vol. XLII-2/W4. 2017. pp. 207−212

[17] Takemura, N., Makihara, Y., Muramatsu, D. On input/output architectures for convolutional neural network-based cross-view gait recognition // IEEE Trans Circuits Syst Video Technol 1-1. 2017

[18] Thapar, D., Nigam, A., Aggarwal, D., Agarwal, P. VGR-net: A view invariant gait recognition network // IEEE 4th International Conference on Identity, Security, and Behavior Analysis, 2018, 1-8.

[19] Tong, S., Fu, Y., Ling, H., Zhang, E. Gait identification by joint spatial-temporal feature // Biometric Recognition, 2017, pp. 457−465

[20] Wolf, T., Babaee, M., Rigoll, G. Multi-view gait recognition using 3D convolutional neural networks // Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 4165-4169.

[21] Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T. A comprehensive study on cross-view gait based human identification with deep cnns // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39

[22] Yoo, J. and Nixon, M. S. Automated Markerless Analysis of Human Gait Motion for Recognition and Classification // ETRI Journal, 2011, 33: 259−266

[23] Zhang,C. Liu, W., Ma, H., Fu, H. Siamese neural network based gait recognition for human identification // IEEE International Conference on Acoustics, Speech and Signal Processing, 2016

[24] Zhang, X., Sun, S., Li, C., Zhao, X., Hu, Y. Deepgait: A learning deep convolutional representation for gait recognition // Biometric Recognition, 2017

Об авторах

Соколова Анна Ильинична − аспирантка департамента больших данных и информационного поиска факультета компьютерных наук НИУ ВШЭ. Email: ale4kasokolova@gmail.com.

Конушин Антон Сергеевич − к.ф.-м.н., доцент, заведующий лабораторией компьютерной графики и мультимедиа факультета ВМК МГУ имени М.В. Ломоносова, доцент департамента больших данных и информационного поиска факультета компьютерных наук НИУ ВШЭ. Email: anton.konushin@graphics.cs.msu.ru.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18