Трекинг объектов в видеопотоке на основе сверточных нейронных сетей и фрактального анализа

Е.Ю. Минаев[1,2], В.В. Кутикова[1,2], А.В. Никоноров[1,2]

[1]Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001
[2]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. Проблема трекинга множества объектов в видеопотоке является одной из наиболее востребованных в сфере технического зрения. На ее основе решается множество прикладных задач - видеонаблюдения, беспилотного транспорта, взаимодействия человек-компьютер и других. В данной работе исследуется возможность применения методов фрактального сжатия для оценки близости детекций объектов при решении задачи трекинга. При этом сами детекции объектов требуемых классов получаются на каждом кадре при помощи сверточной нейронной сети YOLOv2. В результате показано, что предложенное сочетание сверточных нейронных сетей и методов фрактального анализа позволяет успешно решать задачу трекинга объектов. Для экспериментального исследования было использовано видео из открытой тестовой базы данных по множественному трекингу.

1. Введение

Под задачей трекинга понимается задача автоматизированного распознавания и дальнейшего отслеживания объектов на протяжении последовательности кадров видеопотока. Как правило, различают трекинг одного и нескольких объектов. Трекинг нескольких объектов в видеопотоке наиболее общая и практически полезная постановка задачи трекинга. Данная проблема актуальна во многих сферах технического зрения – в задачах видеонаблюдения, отслеживания трафика, при создании систем беспилотного транспорта, интерфейсов человек-компьютер и многих других. В последнее время опубликован ряд тестовых наборов данных, позволяющих тестировать различные алгоритмы технического зрения, в различных областях, таких как детектирование объектов, обнаружение пешеходов, 3D реконструкция, вычисление оптического потока. В 2015 году была опубликована обширная и постоянно пополняющаяся база данных для тестирования алгоритмов трекинга множества объектов [1].

Задача трекинга объекта в видеопотоке заключается в сопоставлении детекций объекта на последовательности кадров видеопотока треку объекта. Задача трекинга множества объектов (MOT) заключается в трекинге нескольких различных объектов [2, 3, 4]. В последнее время появилось множество надежных алгоритмов для трекинга одного объекта, однако при наличии нескольких объектов, необходимо отслеживать соответствие объектов текущего фрейма объектам предыдущих фреймов. Отслеживание нескольких объектов является сложной задачей, особенно в случае их частичного перекрытия (окклюзии) или значительного сходства нескольких объектов между собой.

На базе тестовых данных [1] сравнивается достаточно большое число алгоритмов решения задачи MOT. Многие из них, такие как алгоритм SORT [5] решают задачу трекинга в два этапа, сначала решается задача детектирования объектов требуемых классов в кадре, а потом выполняется их сопоставление с детекциями, полученными на предыдущих кадрах. Каждая детекция описывается ограничивающей ее прямоугольной областью интереса, как показано на рисунке 1. В случае совпадения детекций одного и того же класса объектов на последовательных кадрах, эти детекции относятся к одному треку. Основной признак принадлежности детекций одному треку – близкое расположение ограничивающих их прямоугольников на последовательных кадрах [5]. Однако у такого критерия близости есть недостатки, в частности, он неустойчиво работает в случае скопления объектов. В ряде работ предлагается в качестве критерия близости детекций использовать меры сходства, рассчитанные непосредственно по участку изображения, внутри детекций. Подобная мера близости, в частности, была предложена в [6] в качестве дополнения к алгоритму из работы [5].

В настоящей работе исследуется возможность использования методов фрактального анализа в качестве меры близости детекций на различных кадрах для формирования трека объекта. Применяется подход фрактального распознавания, успешно использованный авторами для локализации объектов на изображениях в работе [7].

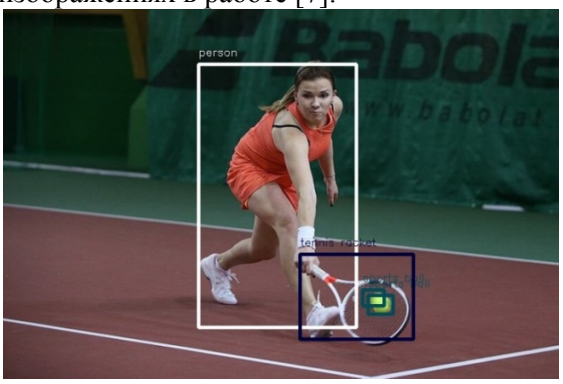

Рисунок 1. Детектирование объектов с помощью сверточной сети YOLOv2.

В настоящей работе применяется следующая общая схема трекинга множества объектов в видеопотоке. Сначала на поступившем кадре выполняется детектирование объектов требуемых классов при помощи сверточной нейронной сети YOLOv2 [8]. Решение задачи трекинга выполняется на основе методов фрактального распознавания. Далее в работе описан применяемый к детектированию подход на основании сверточной нейронной сети YOLOv2, метод фрактального распознавания, используемый для трекинга, описываются результаты проведенного экспериментального исследования.

2. Детектирование объектов на основе сверточной нейронной сети

Детектирование объектов на изображении проводится с помощью сверточной нейронной сети YOLOv2 (You Only Look Once) [8], реализованной на основе библиотеки Tensorflow. YOLOv2 – улучшенная модель YOLO [9]. Скорость обработки изображений позволяет применять YOLOv2 для обработки видео с высокой частотой кадров. Так, на GTX Titan X скорость обработки варьируется от 40 FPS на изображениях с разрешением 544 × 544 до 90 FPS на изображениях более низкого разрешения 288 × 288. Кроме высокой скорости YOLOv2 на определённых наборах данных превосходит по точности распознавания такие детекторы, как SSD [10] и Faster R-CNN [11]. На рисунке 2 приведено сравнение YOLOv2 с SSD и Faster R-CNN по точности распознавания (mean average precision) и количеству кадров, обрабатываемых за секунду (FPS), на наборе VOC 2007.

Обучение сети проводилось на наборах данных PASCAL, VOC и COCO. Применение нового мульти-масштабного подхода к обучению [8] позволило получить одну модель для обработки изображений различных размеров. Результат работы сети представлен на рисунке 3.

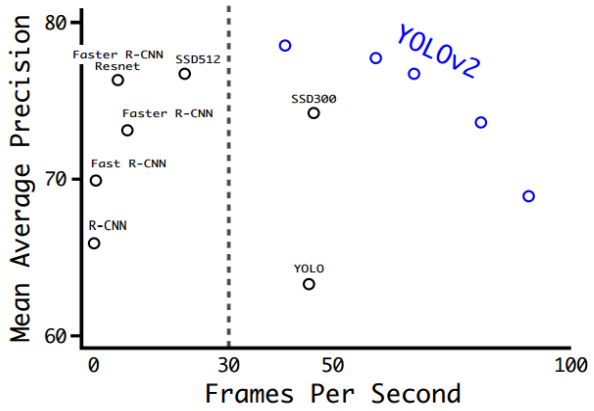

Рисунок 2. Сравнение результатов работы детекторов на наборе VOC 2007.

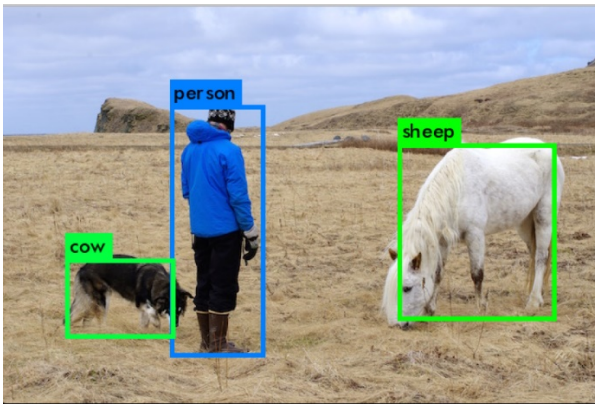

Рисунок 3. Пример работы yolov2.

На рисунке 4 показан принцип работы нейронной сети. На входное изображение накладывается регулярная сетка, разделяющая его на S×S областей. Для каждой области нейронная сеть определяет 5 ограничивающих рамок объекта, уровень достоверности (confidence score) обнаружения рамки, которая отражает степень уверенности модели в том, что поле содержит объект:

$$confidence = \Pr(\text{object}) * IOU(b, \text{object}), \quad (1)$$

где $IOU = \frac{S_\cap}{S_U}$, $S_\cap$ – площадь пересечения эталонной ограничивающей области и поля b, S_U – площадь их объединения. Если объект в области b не существует, то *confidence* будет равен нулю. В результате работы сети выбирается область с наибольшим уровнем достоверности ответа сети (1).

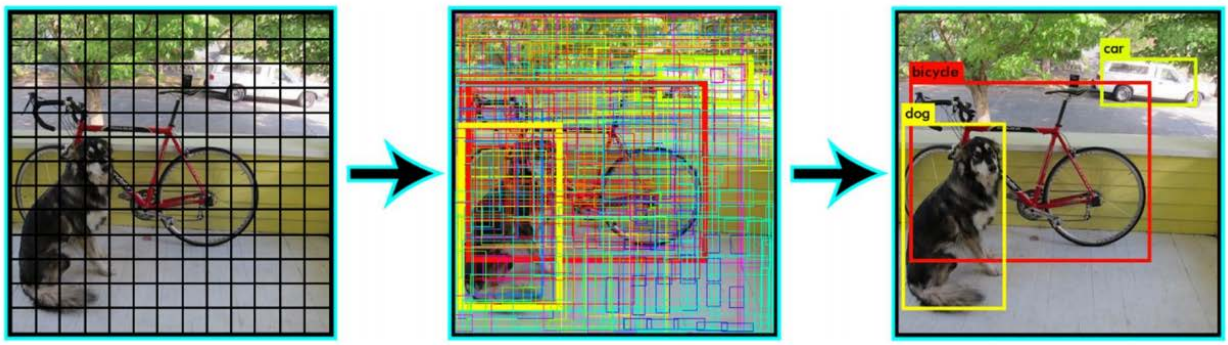

Рисунок 4. Принцип работы YOLOv2.

Кроме значения *confidence* для каждой ограничивающей области YOLOv2 также оценивает координаты центра объекта (t_x, t_y), ширину и высоту объекта (t_w, t_h). Если ячейка сетки, в которую попал центр объекта, смещена на c_x по оси X и на c_y по оси Y, то координаты левого верхнего угла, ширина и высота ограничивающей области вычисляется следующим образом (рисунок 5):

$$\begin{aligned} b_x &= \sigma(t_x) + c_x, \\ b_y &= \sigma(t_y) + c_y, \\ b_w &= p_w e^{t_w}, \\ b_h &= p_h e^{t_h}, \end{aligned} \quad (2)$$

где $\sigma(t_x)$ и $\sigma(t_y)$ – расстояния от центра объекта до верхней и левой границ ячейки сетки, p_w, p_h – априорная информация о ширине и высоте рамки, которая получена на основе обучающей выборки [8].

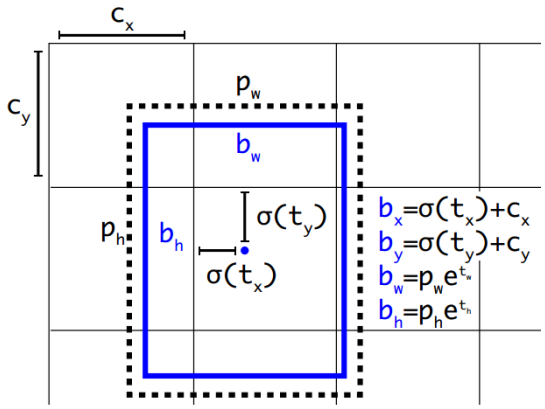

Рисунок 5. Вычисление координат ограничивающей рамки.

3. Трекинг объектов на основе методов фрактального анализа

Для отслеживания и различения областей интереса в видеопотоке в данной работе используется метод фрактального сжатия на основе систем итерированных функций. Основная идея анализа изображений с помощью систем итерированных функций (СИФ) заключается в следующем. Используется схема, которая применялась авторами в предыдущих работах по распознаванию артефактов на цветных изображениях [7]. Исходное изображение разбивается на квадратные непересекающиеся области, называемые ранговыми, и на более крупные квадратные области, называемые доменными. Выделяются две основные стадии анализа изображения: фрактальное сжатие и распознавание. Алгоритм сжатия на основе СИФ для каждой ранговой области производит поиск лучшего преобразования из доменной в ранговую область. В результате, несколько наборов аффинных преобразований кодируют исходное изображение. Аффинные преобразования описываются следующей формулой:

$$f_i\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix} \quad (3)$$

где вектор (x_1,x_2) – координаты пикселя в доменной области, a_i, b_i, c_i, d_i, e_i, f_i – коэффициенты преобразования, и результат преобразования вектор $f_i(x_1,x_2)$ – координаты пикселя в ранговой области. Возможные наборы коэффициентов a_i, b_i, c_i, d_i при этом фиксированы и предопределены заранее, и соответственно выбраны таким образом, что преобразование является сжимающим. Для преобразования компонента яркости используется следующее выражение:

$$r_j = u_i \cdot d_j + v_i, \quad (4)$$

где r_j – яркость пикселя в ранговой области, d_j – яркость пикселя в доменной области, u_i – параметр контраста, v_i – сдвиг яркости.

Результат стадии сжатия заключается в нахождении наилучшего преобразования из доменной области в ранговую. В результате, каждое исходное изображение представляется фрактальным изображением, полученный итерированием аффинных преобразований. Полученные фрактальные изображения используются непосредственно для различения областей интереса на видеоизображении. В качестве критерия соответствия друг другу областей интереса видеоизображения на разных временных отсчетах используется следующий параметр:

$$D = \frac{d(\bigcup_i f_i(p_j), q_i)}{I_h \cdot I_w}, \quad (5)$$

где q_i – фрактальное изображение первой области и $\bigcup_i f_i$ набор соответствующих ему преобразований, p_j – фрактальное изображение второй области. Соответственно, если D=0, то области интереса полностью эквиваленты.

4. Результаты экспериментов

В экспериментах использовались видеоизображения из базы данных [1] (рисунок 6), на которых предварительно было проведено детектирование объектов при помощи предобученной сети YOLOv2.

Рисунок 6. Пример найденных областей интереса на видео из базы MOT challenge [1].

Для обнаруженных областей интереса детекций были сформированы фрактальные изображения. Использовалось 8 фиксированных наборов аффинных преобразований и два варианта разбиения на ранговые области 16x16 и 32x32. Пространство доменных областей было организовано с взаимопересечениями по 15x15 и 31x31 областей. Было рассчитано значение параметра D между различными областями на разных видеокадрах, результаты представлены в таблице 1.

Получено, что изменение точности разбиения исходных изображений на ранговые и доменные области несущественно повлияло на значения параметра близости D, и для однозначного определения соответствующих областей достаточно задать порог параметра $D < 0.055$.

Таблица 1. Значения параметра D для разбиения ранговых областей 16x16(32x32).

t=12	Область 1 t=0	Область 2 t=0	Область 3 t=0	Область 4 t=0	Область 5 t=0
Обл. 1	0.041(0.043)	0.135(0.141)	0.242(0.240)	0.210(0.215)	0.207(0.209)
Обл. 2	-	0.049(0.047)	0.193(0.199)	0.166(0.159)	0.183(0.187)
Обл. 3	-	-	0.043(0.043)	0.301(0.312)	0.145(0.133)
Обл. 4	-	-	-	0.043(0.044)	0.224(0.211)
Обл. 5	-	-	-	-	0.047(0.045)

5. Благодарности

Работа выполнена при поддержке гранта Президента Российской Федерации МД-2531.2017.9 и грантов РФФИ (проекты 16-47-630721 р_а, № 17-29-03112-офи-м, № 18-07-01390-А, № 16-07-00729-А, № 18-37-00457-мол_а).

6. Заключение

В работе рассмотрена задача трекинга множества объектов в видеопотоке. В работе предложено использовать фрактальный анализ для оценки схожести детекций объектов для

посторения их треков по последовательности кадров. Сами детекции объектов различных классов обнаруживаются на каждом кадре при помощи сверточной сети. Проведенное исследование показало эффективность предложенной комбинации нейронной сети и методов фрактального анализа. Экспериментальное исследование было проведено по одному из представленных в базе [1] тестовых видео.

Настоящая работа отражает лишь начало исследований в области трекинга объектов на основе комбинации нейросетевых и фрактальных методов, подтверждая возможность такой комбинации. Более глубокое теоретическое и экспериментальное исследование являются направлениями дальнейших исследований. Так, необходимо провести детальный анализ качества трекинга на основе метрик, предложенных в [1], а также проанализировать производительности предлагаемого решения. Представляет интерес исследование возможности интеграции нейросетевых и фрактальных алгоритмов в единую структуру.

7. Благодарности

Работа выполнена при поддержке Федерального агентства научных организаций (соглашение № 007-ГЗ/Ч3363/26).

8. Литература

[1] Milan, A. Mot16: A benchmark for multi-object tracking / A. Milan, L. Leal-Taixe, I. Reid, S. Roth, K. Schindler, 2016. – P. 1-12. arXiv:1603.00831.

[2] Dicle, C. The way they move: Tracking multiple targets with similar appearance / C. Dicle, M. Sznaier, O. Camps // International Conference on Computer Vision, 2013. – P. 2304-2311.

[3] Rezatofighi, S.H. Joint Probabilistic Data Association Revisited / S.H. Rezatofighi, A. Milan, Z. Zhang, A. Dick, Q. Shi, I. Reid // International Conference on Computer Vision, 2015. – P. 3047-3055.

[4] Kim, C. Multiple Hypothesis Tracking Revisited / C. Kim, F. Li, A. Ciptadi, J.M. Rehg // International Conference on Computer Vision, 2015. – P. 4696-4704. DOI: 10.1109/ ICCV.2015.533.

[5] Bewley, A. Simple online and realtime tracking / A. Bewley, G. Zongyuan, F. Ramos, B. Upcroft // ICIP. – 2016. – P. 3464-3468.

[6] Wojke, N. Simple online and realtime tracking with a deep association metric / N. Wojke, A. Bewley, D. Paulus // arXiv preprint, 2017. – arXiv:1703.07402.

[7] Minaev, E.Y. Fractal Recognition of Compact Artifacts on Color Images / E.Y. Minaev, A.V. Nikonorov // Pattern Recognition and Image Analysis. – 2013. – Vol. 23, № 4. – P. 455-458.

[8] Redmon, J. YOLO9000: Better, Faster, Stronger / J. Redmon, A. Farhadi // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 6517-6525.

[9] Redmon, J. You only look once: Unified, real-time object detection / J. Redmon, S. Divvala, R. Girshick, A. Farhadi // arXiv preprint, 2015. – arXiv:1506.02640.

[10] Liu, W. SSD: single shot multibox detector / W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.E. Reed // CoRR, 2015. – N. 1512.02325.

[11] Ren, S. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks / S. Ren, K. He, R. Girshick, J. Sun // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 39, № 6. – P. 1137-1149.

Multiple object tracking based on convolutional neural network and fractal analysis

E.Y. Minaev[1,2], V.V. Kutikova[1,2], A.V. Nikonorov [1,2]

[1]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001
[2]Samara National Research University, Moskovskoye shosse, 34, Samara, Russia, 443086

Abstract. This paper explores a multiple object tracking focusing on possibility of using fractal analysis as closeness measure. We used YOLOv2 convolutional neural network for object detection and then apply fractal-based measure to determine when the different detections belong to the same track. Experimental evaluation on the sample video from the MOT challenge benchmark confirms efficiency of the proposed combination of convolutional neural network and fractal analysis.

Keywords: Convolutional neural network, object detection, object tracking.