Алгоритм сопровождения людей на видеопоследовательностях с использованием свёрточных нейронных сетей для видеонаблюдения внутри помещений

Р.П. Богуш[1], И.Ю. Захарова[1]
[1]Полоцкий государственный университет,
211440, Республика Беларусь, г. Новополоцк, ул. Блохина, 29

Аннотация

Рассматривается алгоритм сопровождения людей в помещениях, который состоит из следующих основных этапов: обнаружение людей, формирование их признаков, установление соответствия между ними на кадрах, постобработка, индексация сопровождаемых объектов и определение их видимости на кадре. Для детектирования используется свёрточная нейронная сеть YOLO v3. Признаки людей формируются на основе гистограммы канала цветового тона пространства HSV и модифицированной СНС ResNet34. Предлагаемая структура свёрточной нейронной сети состоит из 29 свёрточных и одного полносвязного слоёв и формирует вектор из 128 значений признаков для входного изображения. Выполнено обучение данной модели свёрточной нейронной сети. Определены и представлены основные характеристики разработанного алгоритма, которые подтвердили его эффективность для видеонаблюдения внутри помещений. Эксперименты проведены по методике MOT на тестовых видеопоследовательностях, снятых в помещениях неподвижной видеокамерой. При решении задач обнаружения и сопровождения предложенный алгоритм работает в режиме реального времени с использованием технологии CUDA и видеокарты NVIDIA GTX 1060.

Ключевые слова: сопровождение людей, внутреннее видеонаблюдение, свёрточные нейронные сети.

Цитирование: **Богуш, Р.П.** Алгоритм сопровождения людей на видеопоследовательностях с использованием свёрточных нейронных сетей для видеонаблюдения внутри помещений / Р.П. Богуш, И.Ю. Захарова // Компьютерная оптика. – 2020. – Т. 44, № 1. – С. 109-116. – DOI: 10.18287/2412-6179-CO-565.

Citation: Bohush RP, Zakharava IY. Person tracking algorithm based on convolutional neural network for indoor video surveillance. Computer Optics 2020; 40(1): 109-116. DOI: 10.18287/2412-6179-CO-565.

Введение

Сопровождение объектов на видеопоследовательностях является одной из основных задач в компьютерном зрении, которая в настоящее время имеет различное количество технических применений и всё шире будет использоваться в человеческой деятельности [1, 2]: анализ окружающей обстановки в автоматизированных системах вождения транспортными средствами, оценка правильности движения в медицине и спорте, сопровождение объектов в системах технического зрения на производстве, распознавание типа активности человека в системах мониторинга и охраны и т.д.

Различают задачи сопровождения одного заданного объекта и множества объектов в видеопотоке. Второе направление отличается высокой сложностью и требует точной локализации объектов в кадре, правильной идентификации на следующем кадре и при этом высокой скорости обработки.

Сопровождение множества людей (объектов) в помещении является сложной задачей из-за: неоднородного заднего фона, фрагменты которого могут быть схожи по форме, текстуре и цвету с объектами интереса; низкого уровня освещённости; наличия теней и, соответственно, динамического изменения заднего фона; множественных перекрытий объектов; высокой схожести сопровождаемых объектов; достаточно быстрого их движения и в ряде случаев изменяющегося ускорения и резкой нелинейной трансформации траектории движения.

Анализ данных ресурса MOTChallenge [3], на котором представлены полученные при тестировании результаты на предложенных видеопоследовательностях и метрика для оценки эффективности алгоритмов сопровождения множества объектов, показывает, что наиболее эффективным является метод сопровождения через обнаружение. Этот подход использует ансамбль из детектора объектов и алгоритма для объединения результатов обнаружений на двух кадрах. Эффективное решение проблемы объединения позволяет корректно соотносить результаты обнаружения различных объектов и формировать устойчивые траектории движения для каждого из них. Понятно, что чем точнее работает детектор, тем качественнее сопровождение при таком подходе.

Для обнаружения людей могут быть использованы различные классификаторы: каскады Хаара, на основе метода опорных векторов, свёрточных нейронных сетей (СНС) и др. В настоящее время широкое развитие и применение для обнаружения объектов получили алгоритмы классификации с применением СНС, которые устойчивы к изменениям освещённости, динамическому заднему фону и позволяют осуществлять детектирование даже в случае частичных перекрытий, но требуют высоких вычислительных затрат.

Для решения проблемы объединения объектов при сопровождении могут быть использованы слабые и сильные признаки изображений. В качестве слабых признаков применяются временной сдвиг, цветовые характеристики, форма и т.д. К сильным признакам можно отнести гистограммы ориентированных градиентов (HOG), SIFT- и SURF-дескрипторы и признаки, выделенные с помощью СНС.

Таким образом, сопровождение через обнаружение характеризуется большим количеством алгоритмических решений как для обнаружения объектов в кадре, так и при последующем поиске идентичных объектов на разных кадрах. Ввиду высокой сложности данная задача не решена в полной мере даже для некоторых ограниченных практических приложений, в том числе и для сопровождения людей в помещениях.

1. Существующие подходы для решения задачи сопровождения людей

В работе [4] представлен алгоритм для обнаружения и сопровождения человека при внутреннем видеонаблюдении, который применяет для детектирования вычитание заднего фона на основе алгоритма MOG2. Для сопровождения человека используется метод оптического потока Лукаса–Канаде. Смещение обнаруженной области прогнозируется с применением фильтра Калмана. Представленный подход расширен детектированием падений на основе метода k-ближайших соседей. Однако применение данного метода рассмотрено только для одного человека в кадре. При динамическом фоне и плохой освещённости будет сложно обеспечить минимизацию ложных обнаружений детектора, а также стабильность объединения множественных объектов.

Другой метод для сопровождения основан на первоначальном детектировании головы человека, улучшенная модификация рассмотрена в [5]. После обнаружения головы для выделения фигуры человека выполняется расчёт размеров тела на основе известного статичного процентного их соотношения. Такой метод демонстрирует хорошие результаты при линейном движении пешеходов с учётом обнаружения их не на каждом кадре, а через определённый временной интервал. Таким образом, подобный подход не обеспечит качественного сопровождения людей в помещениях с нелинейной траекторией движения, множественными перекрытиями и резким изменением масштаба.

Алгоритмы, основанные на использовании СНС, обладают большей вариативностью при сложных условиях видеонаблюдения и используются при сопоставлении изображений людей, полученных с разных камер (ре-идентификация). Такой подход представляется перспективным при сопровождении, поэтому целесообразно выполнить анализ алгоритмов ре-идентификации на основе СНС в приложении к задаче сопровождения. Для решения проблемы реидентификации людей, полученных с двух разных камер, в [6] рассмотрена СНС SiameseNet. Модель использует две ветви одинаковой архитектуры, обменивающиеся признаками в процессе сравнения. Конечным результатом является бинарная классификация, которая принимает решение о схожести двух изображений на входе. В работе [7] представлена архитектура PartNet, которая предварительно выделяет наиболее информативные части человека, затем извлекает из них признаки высокого уровня с использованием СНС. Данный подход при сопровождении людей на видео, полученных с различных камер наблюдения, показывает хорошие результаты, однако требует больших вычислительных ресурсов.

В [8] для детектирования и выделения признаков использовалась СНС Faster R-CNN, которая не обучалась для задачи сопровождения людей. Однако она позволила получить всё же лучшие результаты, чем в [6]. Недостатком является то, что для сложных случаев траектории движения людей результаты не стабильны и очень сильно зависят от параметров настроек алгоритма.

Для сопровождения также применяется класс алгоритмов с использованием ключевых точек человека, которые выделяются на суставах, и дальнейшего описания их набором признаков с учётом расстояний между ними, отношений между расстояниями и др. В [9] предложена СНС PoseNet для выделения таких ключевых точек. Алгоритм основан на анализе смещения всех точек в пространстве и изменения соотношений расстояний между ними. Подобные алгоритмы характеризуются большей стабильностью при частичных перекрытиях объектов, однако требуют очень значительных вычислительных затрат.

Улучшенная модель Deep SORT [10] алгоритма SORT [11] использует СНС, которая была разработана для выделения признаков людей при их сопровождении. Для компенсации ложноотрицательных результатов детектора и предсказаний положения объекта на следующем кадре применён фильтр Калмана, который позволяет получать хорошие результаты при линейном движении объектов. Deep SORT является алгоритмом, показавшим наилучшие результаты в соревновании MOT16 [3]. При совместном применении Deep SORT для сопровождения и СНС YOLO v3 для детектирования людей скорость обработки составляет 11,5 кадров в секунду на основе NVIDIA GTX 1060 [12], что не обеспечивает работу в режиме

реального времени при стандартной частоте кадров видеоряда.

Целью статьи является разработка алгоритма сопровождения людей в помещениях с улучшенными качественными характеристиками и с возможностью обеспечения детектирования и сопровождения в режиме реального времени.

2. Обнаружение и сопровождение на основе свёрточных нейронных сетей

Предлагаемый алгоритм состоит из следующих основных этапов: обнаружение людей, формирование вектора признаков для каждого из них, установление соответствия между объектами на кадрах, постобработка, индексация людей, определение их видимости на кадре, выделение рамкой человека при его присутствии в кадре.

2.1. Обнаружение людей

Для обеспечения режима реального времени комплексной задачи обнаружения и сопровождения людей необходимо для первого этапа применять быстродействующую СНС, но с высокой точностью. Среди существующих СНС модель YOLO [13], представленная в 2016 г., направлена на уменьшение вычислительных затрат при обработке. В метрике top-5 точность YOLO составляет 88 %. Однако недостатком, ограничивающим применение для решения указанной задачи, является нестабильность работы при перекрывающихся объектах, что не позволит эффективно детектировать людей в помещениях.

Развитием данной архитектуры являются модификации YOLO v2 и YOLO 9000 [14], в которые включены: нормализация данных, позволяющая не использовать технологию отсева без опасения возникновения переобучения; повышение размерности классификатора для YOLO v2 до [448×448] для 10 эпох ImageNet; использование сети, выносящей предположение о нахождении регионов интереса по аналогии с моделью Faster R-CNN; применение метода k-средних ($k=5$) для предварительной сегментации объектов в каждой области интереса, что даёт возможность выделять пять классов объектов; многомасштабное обучение, которое позволяет сети сегментировать и классифицировать объекты при разных разрешениях; новая классификационная модель Darknet-19, которая содержит 19 свёрточных слоёв и 5 слоёв субдискретизации. YOLO 9000 имеет ту же архитектуру, что и YOLO v2, однако количество выходных гипотез ограничивается тремя. В этой модели также предусмотрено применение карт признаков разных размеров в качестве входных данных, что ведёт к лучшему обнаружению объектов, которые изменяют свой масштаб во время обработки. В метрике top-5 точность обнаружения для данной модели достигает значения 91,8 %, однако вычислительные затраты значительно больше, чем у СНС YOLO. При этом для некоторых классов объектов, например, «человек», «одежда», вероятность правильной классификации уменьшается, что накладывает ограничение на использование данной модели СНС в задаче обнаружения людей в помещениях.

Модель YOLO третьей версии [15] использует улучшенную архитектуру для выделения признаков Darknet-53, содержащую 53 слоя и 23 пропускных соединения, что позволяет при необходимости обнулить влияние слоя на результат работы детектора, т.е. даёт возможность изменять архитектуру сети так, чтобы конечное количество слоёв определялось для конкретной задачи в процессе обучения. По результатам тестирования в метрике top-5 для данной модели точность составляет 93,8 %.

Таким образом, в качестве модели СНС для детектирования объектов в предлагаемом алгоритме сопровождения через объединение используется YOLO v3, т.к. данная архитектура характеризуется хорошей точностью обнаружения и удовлетворительным временем обработки.

После применения СНС для минимизации случайных ложных обнаружений людей, которые всё же иногда присутствуют, выполняется фильтрация обнаруженных объектов по размеру.

2.2. Формирование набора признаков

Для установления соответствия между людьми на входном кадре и предыдущих кадрах необходимо сформировать набор признаков, описывающий каждый сопровождаемый объект. При сравнении людей необходимо учитывать вариативность их схожих и отличных признаков, поэтому для вычисления дескрипторов используется модифицированная СНС ResNet-34 [16], которая обладает небольшим количество слоёв, что позволит также обеспечить приемлемые вычислительные затраты. Наличие пропускных соединений в ResNet-34 позволяет изменять количество слоёв для лучшего результата обучения. Архитектура модифицированной СНС представлена на рис. 1.

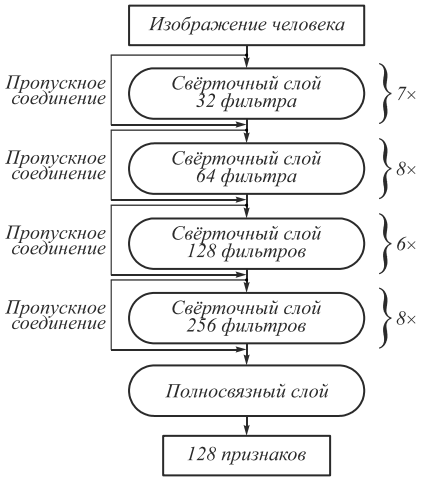

Рис. 1. Архитектура СНС для вычисления признаков сопровождаемых объектов

Необходимость изменения архитектуры ResNet-34 обусловлена отличием классификации объектов, для которой разработана данная СНС, от сопровождения. При решении данной задачи требуется выделение признаков различных людей для последующего их сравнения с учётом того, что они относятся к одному классу «человек», что не позволит сделать ResNet-34. Модификация данной СНС заключается в следующем: удаление входного свёрточного слоя с размером фильтра [7×7], т.к. применение ядер свёртки минимальных размеров [3×3] позволяет получать лучший результат при ре-идентификации [17]; количество выходов конечного полносвязного слоя уменьшено до 128, позволяющего сформировать такое же количество признаков для описания человека; сокращение числа свёрточных слоев СНС до 29 с размерами ядер для них [3×3], после каждого слоя используется пропускное соединение.

В предлагаемом наборе признаков сопровождаемого человека (Ob_{ID}), который назовём составным дескриптором, используются вычисленные для него:

- СНС-признаки при последнем правильном обнаружении (f_{det});
- гистограммные признаки при последнем правильном обнаружении (H^{det});
- координаты центра на предыдущем кадре ($l_{x,y}$);
- ширина и высота (w, h) на предыдущем кадре.

2.3. Установление соответствия между изображениями людей на соседних кадрах

Оценка схожести между сопровождаемыми объектами Ob_{tr} и обнаруженными на текущем кадре Ob_{det} выполняется на основе выражения:

$$\begin{aligned} d(Ob_{tr}, Ob_{det}) = \alpha \sqrt{\sum_{j=1}^{128} (f_{curr} - f_{det})^2} + \\ + \beta \left(\sqrt{(l_{curr} - l_{prev})^2} + \sqrt{(w_{curr} - w_{prev})^2} + \right. \\ \left. + \sqrt{(h_{curr} - h_{prev})^2} \right), \end{aligned} \quad (1)$$

где f_{curr} – СНС-признаки объекта на входном кадре; w_{curr}, w_{prev} – ширина сопровождаемого человека на текущем и предыдущем кадрах; h_{curr}, h_{prev} – высота сопровождаемого человека на текущем и предыдущем кадрах; α и β – корректирующие коэффициенты.

В результате применения выражения (1) для всех сопровождаемых людей на предыдущих кадрах и обнаруженных объектов на текущем кадре формируется матрица схожести, к которой применяется Венгерский алгоритм [18]. При этом особенностью является необходимость обработки Ob_{tr} не только с предыдущего кадра, но и с более ранних кадров, т.к. возможна кратковременная потеря оптической связи камеры с человеком. Это обусловлено тем, что в помещениях траектории движения людей достаточно часто пересекаются, а также объекты интереса перекрываются относительно камеры видеонаблюдения. Это происходит, например, при разговоре людей или при их совместном движении. Кроме этого, человек может иметь несколько точек входов и выходов в кадре, частично или полностью перекрываться другими статическими объектами.

2.4. Постобработка

Для уменьшения вероятности ложного изменения индексации после сложных случаев движения объектов с множественными перекрытиями, в составном дескрипторе Ob_{ID}, на основе которого выполняется непрерывное сопровождение, обновляются только координаты объекта, его ширина и высота. Параметр перекрытия вычисляется на основе коэффициента Жаккарда:

$$IOU = \frac{In}{w_{prev} \times h_{prev} + w_{curr} \times h_{curr} - In}, \quad (2)$$

где

$$\begin{aligned} In = & \left(\min\left(l_x^{prev} + w_{prev}, l_x^{curr} + w_{curr} \right) - \right. \\ & \left. - \max\left(l_x^{prev}, l_x^{curr} \right) \right) \times \\ & \times \left(\min\left(l_y^{prev} + h_{prev}, l_y^{curr} + h_{curr} \right) - \max\left(l_y^{prev}, l_y^{curr} \right) \right). \end{aligned} \quad (3)$$

Использование составных дескрипторов сопровождаемых людей, которые отсутствовали на предыдущих кадрах, для их поиска на текущем приводит к двум основным проблемам. Первая проблема проявляется как присвоение индекса сопровождаемого человека, который вышел из кадра, другому, вошедшему в кадр практически в том же месте. Она решается за счёт анализа ширины, высоты и координат центра объекта на текущем и предыдущем кадрах.

Второй проблемой является необходимость обнаружения момента выхода человека из кадра для прекращения его сопровождения. Для решения этой задачи в алгоритме введён параметр видимости сопровождаемого объекта. Его определение основано на комбинации сильных и слабых признаков. В качестве первых анализируются СНС-признаки. Как слабый признак используется схожесть цветовых гистограмм. Для уменьшения влияния изменения освещённости, изображения преобразовываются из цветового пространства RGB в HSV. Далее для оценки схожести используются только данные цветового тона. Величина схожести вычисляется на основе выражения [19]:

$$\begin{cases} \text{если } \left| H_i^{tr} - \overline{H^{tr}} \right| > \left| H_i^{det} - \overline{H^{det}} \right|, \\ \qquad \text{то } R = \dfrac{1}{N^2} \displaystyle\sum_{i=0}^{N-1} \dfrac{(H_i^{det} - \overline{H^{det}})}{(H_i^{tr} - \overline{H^{tr}})}, \\ \text{иначе:} \qquad R = \dfrac{1}{N^2} \displaystyle\sum_{i=0}^{N-1} \dfrac{(H_i^{tr} - \overline{H^{tr}})}{(H_i^{det} - \overline{H^{det}})}, \end{cases} \quad (4)$$

где

$$\overline{H^{tr}} = \frac{1}{N}\sum_{i=0}^{N-1} H_i^{tr}\ ;\ \overline{H^{det}} = \frac{1}{N}\sum_{i=0}^{N-1} H_i^{det}\ ;\ H_i^{tr} \text{ и } H_i^{det} -$$

элемент вектора признаков цветового тона сопровождаемого объекта и элемент вектора признаков последнего результата правильного его обнаружения соответственно; N – длина вектора.

Затем определяется видимость сопровождаемого объекта в кадре на основе правила:

$$v = \begin{cases} 0, \text{ если } d_{feat} < \varepsilon \text{ и } R > \eta; \\ 1, \text{ если } d_{feat} \geq \varepsilon \text{ или } R \leq \eta, \end{cases} \quad (5)$$

где

$$d_{feat} = \sqrt{\sum_{j=1}^{128} (f_{curr} - f_{prev})^2}\ ;$$

ε и η – пороговые значения.

Параметр видимости принимает значение «1» при наличии сопровождаемого человека на изображении либо значение «0» при потере оптической связи с объектом.

Определение постоянных коэффициентов для выражений (1) и (5) выполнено на основе проведения многопараметрического вычислительного эксперимента путём варьирования $\alpha, \beta, \varepsilon$ и η для нахождения максимального значения MOTA. При этом наибольший вес присвоен СНС-признакам, параметр α изменялся в диапазоне от 50 до 500 с дискретностью 50, величины β, ε и η варьировались от 0,1 до 1 с дискретностью 0,1. Максимальное значение MOTA зафиксировано для: $\alpha = 200$; $\beta = 0{,}3$; $\varepsilon = 0{,}3$; $\eta = 0{,}2$.

3. Обучение СНС для сопровождения

Модифицированная СНС обучена на комплексной базе данных из наборов изображений PRID [20] и iLIDS [21], которые были подготовлены для решения задачи сопровождения людей на видео, которые получены с различных видеокамер. В результате сформированный набор данных состоит из 1030 базовых объектов, т.е. различных людей. Такой подход позволил увеличить тестовый набор данных. Количество изображений для каждого человека различно и изменяется в пределах от 85 до 360. На рис. 2 представлено по 4 примера используемых изображений для людей из баз данных PRID и iLIDS соответственно.

Обучение СНС выполнено в течение 50 часов на персональном компьютере с основными характеристиками: центральный процессор Intel Core i5-8600 с тактовой частотой 3,6 ГГц, объём ОЗУ – 16 ГБ, 2 видеокарты NVIDIA GTX 1060 с основными параметрами: скорость обучения – 0,001; коэффициент инерции градиентного спуска – 0,9; количество эпох – 250 000. На тестовой выборке из изображений 300 различных людей вероятность правильной классификации – 99,7%. Среднее значение функции потерь – $7{,}46 \cdot 10^{-6}$.

4. Результаты экспериментов

Для проведения экспериментов разработанный алгоритм реализован на языке C++ с применением библиотек компьютерного зрения OpenCV 3.4 и dlib. Все процедуры обработки при обнаружении и сопровождении людей на основе СНС осуществлялись на графическом процессоре с использованием технологии параллельной обработки CUDA.

Тестирование алгоритмов выполнено на трёх видеопоследовательностях сложного типа (рис. 3).

Данные видео были получены с использованием неподвижной видеокамеры в помещениях с различным освещением, количеством людей в кадре, нелинейной траекторией их движения, полным и частичным перекрытием, кратковременным выходом людей из помещения.

а) *б)*

Рис. 2. Изображения из базы данных для обучения СНС: движущийся человек на однородном фоне в профиль из PRID (а); движущийся человек на сложном фоне с изменяющимся положением относительно видеокамеры из iLIDS (б)

На рис. 3*а* представлен кадр из первого тестового видеоряда (количество кадров 2 190) с неоднородным освещением и наличием множества теней, на котором присутствуют три человека с частичным взаимным перекрытием. Причём два из них имеют очень схожие характеристики, т.к. одежда у них практически идентичная. Кроме этого, рост, телосложение и цвет волос также схожи. Из рис. 3*б* видно, что люди расходятся, а затем пересекаются с множественным перекрытием (рис. 3*в*). Траектория движения относительно удаленности от камеры также изменяется, т.е. наиболее удалённый человек на рис. 3*а* является наименее удалённым на рис. 3*в*.

Другая тестовая видеопоследовательность с количеством кадров 1101, примеры которых приведены на рис. 3*г-е*, отличается более низким освещением и также неоднородностью освещения. На рис. 3*г* показано полное перекрытие одного человека другим на этой тестовой видеопоследовательности. Из рис. 3*д* видно, что человек, который был скрыт, перешёл в наименее

освещённую область, а затем произошло перемещение людей с изменением их масштаба на кадре и частичным перекрытием (рис. 3*е)*. Цветовые характеристики этих объектов интереса также достаточно схожи.

Рис. 3. Примеры кадров видеопоследовательностей, которые использованы при тестировании: кадры из первого видеоряда (а–в); кадры из второго видео (г–е); кадры из третьей видеопоследовательности (ж–и)

Для третьего видеоряда (число кадров 1151) показан кадр с одним частично перекрытым человеком (рис. 3*ж*). На рис. 3*з* он полностью скрыт за статичным объектом. Из рис. 3*и* видно, что размер человека с более раннего кадра (рис. 3*ж*) значительно меньше и в левом дальнем углу присутствует второй человек. Таким образом, очевидно, что для данного видеоряда необходимо сопровождение двух человек со сложной траекторией их движения и множественными перекрытиями за статическими объектами.

При оценке эффективности алгоритмов сопровождения могут быть использованы основные параметры [22]:

- IDF показывает процент правильной идентификации сопровождающих объектов;
- MOTA – учитывает количество ложноположительных результатов (FP), ложноотрицательных (FN) и IDF и характеризует точность сопровождения объектов во времени, с учётом восстановления траектории при кратковременном отсутствии объекта;
- MOTP – показывает, насколько точно был локализован объект в кадре при сопровождении без учёта обнаружения;
- F_{reg} – скорость работы алгоритма сопровождения без учёта детектирования объектов, кадров в секунду.

Для проведения экспериментов по сопровождению людей при видеонаблюдении в помещениях на трёх видеопоследовательностях (рис. 3) все объекты были размечены согласно требованиям из [3]. Подготовленные данные позволили выполнить тестирование разработанного алгоритма и наиболее эффективного по методике MOT16 Deep SORT [3]. В табл. 1 приведены характеристики, полученные как для каждой видеопоследовательности в отдельности, так и при комплексном тестировании на всех трёх видео (рис. 3*а, г, ж*).

Анализ табл. 1 показывает, что разработанный алгоритм позволяет улучшить качественные характеристики Deep SORT, за исключением параметра MOTP.

На рис. 4 показаны результаты сравнения эффективности сопровождения объектов на идентичных фрагментах кадров. Индексация людей в предложенном алгоритме начинается с 0, второму человеку присваивается индекс 1 и т.д. В Deep SORT начальным индексом является 1.

Табл. 1. Сравнение характеристик алгоритмов для различных видеопоследовательностей

Наименование параметра	Тестовый видеоряд (см.рис.3)	Deep SORT	Предложенный алгоритм
IDF	Рис.3*а*	38,3	93
	Рис.3*г*	48,7	97,2
	Рис.3*ж*	86,7	97,8
	Рис.3*а,г,ж*	52,9	**95,3**
FP	Рис.3*а*	196	111
	Рис.3*г*	10	**7**
	Рис.3*ж*	22	10
	Рис.3*а,г,ж*	228	**128**
FN	Рис.3*а*	181	199
	Рис.3*г*	59	56
	Рис.3*ж*	68	42
	Рис.3*а,г,ж*	308	**297**
MOTA	Рис.3*а*	82,5	86,4
	Рис.3*г*	94	94,5
	Рис.3*ж*	91,9	95,6
	Рис.3*а,г,ж*	87,8	**90,8**
MOTP	Рис.3*а*	79,5	78,6
	Рис.3*г*	81,4	75,6
	Рис.3*ж*	78,6	80,8
	Рис.3*а,г,ж*	**79,8**	78,4

Из рис. 4*а*, *в*, *д* видно, что предложенный алгоритм корректно выполняет сопровождение объектов без изменения их индексации. На первом кадре из примера с применением Deep SORT (рис. 4*б*) присутствует ложное обнаружение спинки стула как человека и, кроме этого, сопровождаемому человеку присвоен неправильный индекс, т.к. он должен быть равен 1.

Из рис. 4*г* видно, что при применении Deep SORT на видеопоследовательности присутствовало неоднократное некорректное изменение индексации, т.к. объект с индексом 17 должен иметь индекс 1.

На другой видеопоследовательности при изменении направления движения и кратковременной потере оптической связи с объектом предложенный алгоритм, как видно из рис. 4*д*, сопровождает объект правильно, без изменения индексации. Данный объект при входе в кадр имел индекс ноль и направлялся от видеокамеры к двери, скрылся за доской, и затем опять вошел в кадр по направлению к видеокамере. Deep SORT при работе изменяет индекс объекта до четырёх (рис. 4*е*).

Скорость работы двух алгоритмов на используемом при обучении персональном компьютере Freg = 60 кадр/с. При обнаружении и сопровождении представленный подход обеспечивает работу в режиме реального времени для пяти объектов со скоростью обработки 25 кадров в секунду.

Заключение

В статье предложен алгоритм сопровождения людей в помещении в режиме реального времени на основе выполнения следующих основных этапов: обнаружение людей, формирование составного дескриптора для каждого из них, установление соответствия между ними на соседних кадрах, постобработка, индексация людей, определение их видимости на кадре. Детектирование выполняется на основе СНС YOLO v3, а для вычисления признаков людей при установлении соответствия между ними предложена архитектура СНС путём модификации ResNet 34. Для проведения экспериментов подготовлены, в соответствии с использованием методики для множественного сопровождения объектов, три видеопоследовательности, полученные неподвижной камерой в помещении. На их основе определены основные характеристики разработанного алгоритма: IDF = 95,3; FP = 128; FN = 297; MOTA = 90,8, MOTP = 78,4; $F_{reg} = 60$, которые показывают большую его эффективность по сравнению с Deep SORT при видеонаблюдении внутри помещений. Для сокращения времени вычислений все операции СНС выполнялись на видеоадаптере NVIDIA GTX 1060 с применением технологии CUDA.

Рис. 4. Примеры сопровождения объектов на видео: на основе предложенного алгоритма (а,в,д); на основе алгоритма Deep SORT (б,г,е)

Таким образом, рассмотренный подход является перспективным для систем видеонаблюдения внутри помещений, а дальнейшие исследования планируется выполнить по адаптации алгоритма для сопровождения людей по видеопоследовательностям, полученным с разных камер внутри помещения.

Литература

1. **Forsyth, D.** Computer vision: A modern approach / D.Forsyth, J.Ponce. – 2nd ed. – Pearson Education, 2012. – 794 p.
2. **Шаталин, Р.А.** Обнаружение нехарактерного поведения в задачах видеонаблюдения / Р.А. Шаталин, В.Р. Фидельман,

П.Е. Овчинников // Компьютерная оптика. – 2017. – Т. 41, № 1. – С. 37-45. – DOI: 10.18287/2412-6179-2017-41-1-37-45.
3. MOTChallenge: The multiple object tracking benchmark [Electronical Resource]. – URL: https://motchallenge.net/ (request date 20.01.2019).
4. **Miguel, M.** Home camera-based fall detection system for the elderly / M. De Miguel, A. Brunete, M. Hernando, E. Gambao // Sensors. – 2017. – Vol. 17, Issue 12. – P. 2864-2885. – DOI: 10.3390/s17122864.
5. **Купляков, Д.А.** Алгоритм сопровождения людей в видео на основе метода Монте-Карло для Марковский цепей / Д.А. Купляков, Е.В. Шальнов, А.С. Конушин // Программирование. – 2017. – № 4. – С. 13-21. – DOI: 10.1134/S0361768817040053.
6. **Tao, R.** Siamese instance search for tracking / R. Tao, E. Gavves, A.W.M. Smeulders // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 1420-1429. – DOI: 10.1109/CVPR.2016.158.
7. **Zhao, L.** Deeply-learned part-aligned representations for person re-identification / L. Zhao, X. Li, Y. Zhuang, J. Wang // Proceedings of the IEEE International Conference on Computer Vision (ICCV). – 2017. – P. 3239-3248. – DOI: 10.1109/ICCV.2017.349.
8. **Chahyati, D.** Tracking people by detection using CNN features / D. Chahyati, M.I. Fanany, A.M. Arymurthy // Proceedings of the 4th Information Systems International Conference (ISICO 2017). – 2017. – P. 167-172.
9. **Iqbal, U.** PoseTrack: Joint multi-person pose estimation and tracking / U. Iqbal, A. Milan, J. Gall // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 4654-4663. – DOI: 10.1109/CVPR.2017.495.
10. **Wojke, N.** Simple online and realtime tracking with a deep association metric / N. Wojke, A. Bewley, D. Paulus // Proceedings of the IEEE International Conference on Image Processing (ICIP). – 2017. – P. 3645-3649. – DOI: 10.1109/ICIP.2017.8296962.
11. **Bewley, A**. Simple online and realtime tracking / A. Bewley, Z. Ge, L. Ott, F.T. Ramos, B. Upcroft // Proceedings of the IEEE International Conference on Image Processing (ICIP). – 2016. – P. 3464-3468. – DOI: 10.1109/ICIP.2016.7533003.
12. Real-time multi-person tracker using YOLO v3 and deep_sort with tensorflow [Electronical Resource]. – URL: https://github.com/Qidian213/deep_sort_yolov3 (request date 10.11.2018).
13. **Redmon, J.** You only look once: Unified, real-time object detection / J. Redmon, S.K. Divvala, R.B. Girshick, A. Farhadi // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 779-788. – DOI: 10.1109/CVPR.2016.91.
14. **Redmon, J.** YOLO9000: Better, faster, stronger / J. Redmon, A. Farhadi // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2017. – P. 6517-6525. – DOI: 10.1109/CVPR.2017.690.
15. YOLOv3: An incremental improvement [Electronical Resource]. – URL: https://arxiv.org/abs/1804.02767 (request date 10.11.2018).
16. **He, K.** Deep residual learning for image recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770-778. – DOI: 10.1109/CVPR.2016.90.
17. **Wu, L.** PersonNet: Person re-identification with deep convolutional neural networks [Electronical Resource] / L. Wu, S. Chunhua, A. Hengel // Computing Research Repository. – 2016. – URL: https://arxiv.org/pdf/1601.07255.pdf (request date 16.06.2019).
18. **Kuhn, H.W.** The hungarian method for the assignment problem / H.W. Kuhn // Naval Research Logistics Quarterly. – 1955. – Vol. 2. – P. 83-97.
19. **Bogush, R.** Minimax criterion of similarity for video information processing / R. Bogush, S. Maltsev // Proceedings of the Siberian Conference on Control and Communications. – 2007. – P. 120-127. – DOI: 10.1109/SIBCON.2007.371310.
20. Person Re-ID (PRID) dataset [Electronical Resource]. – URL: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/prid11/ (request date 10.11.2018).
21. iLIDS Video re-IDentification (iLIDS-VID) dataset [Electronical Resource]. – URL: http://www.eecs.qmul.ac.uk/~xiatian/downloads_qmul_iLIDS-VID_ReID_dataset.html (request date 10.11.2018).
22. **Keni, B.** Evaluating multiple object tracking performance: The CLEAR MOT metrics / B. Keni, R. Stiefelhagen // EURASIP Journal on Image and Video Processing. – 2008. – Vol. 1.– P. 1-10.

Сведения об авторах

Богуш Рихард Петрович, 1974 года рождения. В 1997 г. окончил радиотехнический факультет Полоцкого государственного университета, в 2002 году в Институте технической кибернетики НАН Беларуси защитил кандидатскую диссертацию, с 2006 г. доцент. Заведующий кафедрой вычислительных систем и сетей Полоцкого государственного университета. Автор более чем 140 научных публикаций, включая монографию по обработке изображений и сигналов. Область научных интересов: обработка статических и динамических изображений, информационная безопасность, интеллектуальные системы, обработка сигналов. E-mail: *bogush@mail.ru* .

Захарова Ирина Юрьевна, 1995 года рождения, магистр технических наук по специальности «Математическое моделирование, численные методы и комплексы программ», ассистент кафедры вычислительных систем и сетей. Область научных интересов: обработка изображений, сжатие информации, криптография. Автор 10 научных публикаций. E-mail: *ira9992011@yandex.ru* .

ГРНТИ: 28.23.15
Поступила в редакцию 23 апреля 2019 г. Окончательный вариант – 10 сентября 2019 г.

Person tracking algorithm based on convolutional neural network for indoor video surveillance

R. Bohush[1], I. Zakharava[1]
[1]Polotsk State University, Polotsk, Belarus

Abstract

In this paper, a person tracking algorithm for indoor video surveillance is presented. The algorithm contains the following steps: person detection, person features formation, features similarity calculation for the detected objects, postprocessing, person indexing, and person visibility determination in the current frame. Convolutional Neural Network (CNN) YOLO v3 is used for person detection. Person features are formed based on H channel in HSV color space histograms and a modified CNN ResNet. The proposed architecture includes 29 convolutional and one fully connected layer. As the output, it forms a 128-feature vector for every input image. CNN model was trained to perform feature extraction. Experiments were conducted using MOT methodology on stable camera videos in indoor environment. Main characteristics of the presented algorithm are calculated and discussed, confirming its effectiveness in comparison with the current approaches for person tracking in an indoor environment. Our algorithm performs real time processing for object detection and tracking using CUDA technology and a graphics card NVIDIA GTX 1060.

Keywords: person tracking, indoor video surveillance, convolutional neural networks.

Citation: Bohush RP, Zakharava IY. Person tracking algorithm based on convolutional neural network for indoor video surveillance. Computer Optics 2020; 40(1): 109-116. DOI: 10.18287/2412-6179-CO-565.

References

[1] Forsyth D, Ponce J. Computer vision: A modern approach. 2nd Ed. Pearson Education; 2012.

[2] Shatalin RA, Fidelman VR, Ovchinnikov PE. Abnormal behavior detection method for video surveillance applications. Computer Optics 2017; 41(1): 37-45. DOI: 10.18287/2412-6179-2017-41-1-37-45.

[3] MOTChallenge: The multiple object tracking benchmark Source: ⟨https://motchallenge.net⟩.

[4] Miguel MD, Brunete A, Hernando M, Gambao E. Home camera-based fall detection system for the elderly. Sensors 2017; 17(12): 2864-2885. DOI: 10.3390/s17122864

[5] Kuplyakov D, Shalnov E, Konushin A. Markov chain Monte Carlo based video tracking algorithm. Programming and Computer Software 2017; 43(4): 224-229. DOI: 10.1134/S0361768817040053.

[6] Tao R, Gavves E., Smeulders AW. Siamese instance search for tracking. IEEE Conf Comp Vis Pattern Recogn 2016: 1420-1429. DOI: 10.1109/CVPR.2016.158.

[7] Zhao L, Li X, Zhuang Y, Wang J. Deeply-learned part-aligned representations for person re-identification. IEEE Int Conf Comp Vis 2017: 3239-3248. DOI: 10.1109/ICCV.2017.349.

[8] Chahyati D, Fanany MI, Arymurthy A. Tracking people by detection using CNN features. Proc 4th Inform Sys Int Conf 2017: 167-172.

[9] Iqbal U, Milan A, Gall J. PoseTrack: Joint multi-person pose estimation and tracking. IEEE Conf Comp Vis Pattern Recogn 2017: 4654-4663. DOI: 10.1109/CVPR.2017.495.

[10] Wojke N, Bewley A, Paulus D. Simple online and real time tracking with a deep association metric. IEEE Int Conf Image Process 2017: 3645-3649. DOI: 10.1109/ICIP.2017.8296962.

[11] Bewley A, Ge Z, Ott L, Ramos FT, Upcroft B. Simple online and real time tracking. IEEE Int Conf Image Process 2016: 3464-3468. DOI: 10.1109/ICIP.2016.7533003.

[12] Real-time Multi-person tracker using YOLO v3 and deep_sort with tensorflow. Source: ⟨https://github.com/Qidian213/deep_sort_yolov3⟩.

[13] Redmon J, Divvala SK, Girshick RB, Farhadi A. You only look once: Unified, real-time object detection. IEEE Conf Comp Vis Pattern Recogn 2016: 779-788. DOI: 10.1109/CVPR.2016.91.

[14] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. IEEE Conf Comp Vis Pattern Recogn 2017: 6517-6525. DOI: 10.1109/CVPR.2017.690.

[15] YOLOv3: An incremental improvement. Source: ⟨https://arxiv.org/abs/1804.02767⟩.

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conf Comp Vis Pattern Recogn 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[17] Wu L, Chunhua S, Hengel A. PersonNet: Person re-identification with deep convolutional neural networks. Source: ⟨https://arxiv.org/pdf/1601.07255.pdf⟩.

[18] Kuhn HW. The hungarian method for the assignment problem. Naval Research Logistics Quarterly 1955; 2: 83-97.

[19] Bogush R, Maltsev S. Minimax criterion of similarity for video information processing. Proc Siberian Conf Control Commun 2007: 120-127. DOI: 10.1109/SIBCON.2007.371310.

[20] Person Re-ID (PRID) Dataset. Source: ⟨https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/prid11/⟩.

[21] iLIDS Video re-IDentification (iLIDS-VID) dataset. Source: ⟨http://www.eecs.qmul.ac.uk/~xiatian/downloads_qmul_iLIDS-VID_ReID_dataset.html⟩.

[22] Keni B, Stiefelhagen R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP J Image Video Process 2008; 1: 1-10.

Authors' information

Rykhard Bohush, (b. 1974) graduated from Polotsk State University in 1997. In 2002 he got his PhD in the field of Information Processing at the Institute of Engineering Cybernetics, the National Academy of Sciences of Belarus. Head

of Computer Systems and Networks department of Polotsk State University. He is a member of the National Qualifications Framework of Higher Education of Belarus in IT and Electronics Science. His scientific interests include image and video processing, object representation and recognition, intelligent systems, digital steganography. Author of approximately 140 works, including one book on image processing. E-mail: *bogushr@mail.ru* .

Iryna Zakharava, (b. 1995) graduated from Polotsk State University in 2017. In 2019 has got MSc in the field of Mathematical Modeling, Numerical Methods and Complexes of Programs. Assistant at Computer Systems and Networks department. Scientific interests include image processing, information compression, cryptography. Author of 10 scientific publications. E-mail: *ira9992011@yandex.ru* .

Received April 23, 2019. The final version – September 10, 2019.