УДК 519.688
DOI: 10.15827/0236-235X.117.106-111
Дата подачи статьи: 01.08.16
2017. Т. 30. № 1. С. 106–111

ИСПОЛЬЗОВАНИЕ ГЕОМЕТРИИ СЦЕНЫ ДЛЯ УВЕЛИЧЕНИЯ ТОЧНОСТИ ДЕТЕКТОРОВ

Е.В. Шальнов, аспирант, eshalnov@graphics.cs.msu.ru
(Московский государственный университет им. М.В. Ломоносова, Ленинские горы, 1-52, г. Москва, 119991, Россия);
А.С. Конушин, к.ф.-м.н., доцент, ktosh@graphics.cs.msu.ru
(Московский государственный университет им. М.В. Ломоносова, Ленинские горы, 1-52, г. Москва, 119991, Россия;
Национальный исследовательский университет «Высшая школа экономики», ул. Мясницкая, 20, г. Москва, 101000, Россия)

Ключевым элементом любых систем интеллектуальной видеоаналитики является алгоритм выделения, или детектирования объектов в видео. Недостаточно высокие скорость и точность существующих алгоритмов детектирования являются существенными сдерживающими факторами распространения технологий видеоаналитики.

В данной работе предлагается новый алгоритм повышения скорости и точности работы детекторов, основанных на подходе скользящего окна, за счет учета геометрических свойств сцены. В зависимости от расположения камеры относительно сцены для каждой области изображения можно определить, какого размера может быть изображение искомого объекта в данной области. Окна других размеров не могут соответствовать искомым объектам в сцене, поэтому их можно пропускать и за счет этого увеличивать скорость работы детектора. Предлагаемый алгоритм позволяет определить допустимые размеры объекта для каждой области изображения. Сутью алгоритма является нейронная сеть, которая для таких заданных параметров, как калибровка камеры, размер и положение объекта на снимке, определяет, правдоподобна ли данная сцена. Нейронная сеть обучается на множестве синтетических сцен, что позволяет ей работать для произвольных камер. С помощью нейронной сети для каждой видеопоследовательности строится карта допустимых размеров объектов. Детектор затем применяется только к допустимым фрагментам, которые составляют часть от всего множества фрагментов.

Экспериментальная оценка предложенного алгоритма на реальных данных показала, что он позволяет повысить скорость работы детектора на 70 % при одновременном увеличении точности его работы.

Ключевые слова: *нейронные сети, компьютерное зрение, обнаружение объектов, видеоаналитика, компьютерная графика, распознавание образов.*

Обнаружение и локализация объектов интереса в видео – классическая задача компьютерного зрения. Несмотря на долгую историю развития методов обнаружения [1–3], ключевой подход к решению, называемый методом скользящего окна и заключающийся в применении классификатора к разным частям (окнам) изображения, остается неизменным. Классификатор определяет, содержит ли рассматриваемая область изображения объект интереса или нет. Обычно рассматривают окна, отличающиеся не только положением, но и размерами.

Количество таких окон может быть очень велико. Например, при поиске голов людей на изображении fullHD разрешения (1920×1080) (рис. 1) алгоритм [2] классифицирует 1 310 859 окон. Их обработка требует значительных вычислительных ресурсов. Эта проблема существенно ограничивает широкое применение алгоритмов видеоаналитики на практике.

В то же время можно заметить, что в большинстве анализируемых окон размеры человека в сцене неправдоподобны. Например, маловероятно в сцене, представленной на рисунке 1, увидеть голову человека, занимающую половину изображения. Такие априорные знания могут существенно сократить количество обрабатываемых окон и увеличить скорость обработки видеоданных. Для этого необходимо разработать алгоритм классификации окон изображений на ложные и правдоподобные в зависимости от положения камеры. К сожалению, в настоящий момент для произвольной сцены такого алгоритма фильтрации обнаружений не существует.

В данной работе представлен алгоритм определения ложных срабатываний детектора головы человека, использующий информацию о наблюдаемой сцене. Авторы ограничиваются рассмотрением только простых сцен:

– камера снимает плоскую сцену, то есть наблюдаемая поверхность земли является горизонтальной плоскостью;

– люди могут находиться только на поверхности земли;

– дисторсией камеры можно пренебречь.

Эти предположения существенно ограничивают набор рассматриваемых сцен. Так, сцена, содержащая лестницу, не подходит под описанные ограничения. В то же время рассматриваемый класс сцен является наиболее распространенным в видеонаблюдении.

Основной вклад данной работы:

– предложен метод построения классификатора обнаружений на *невозможные* с геометрической точки зрения и *правдоподобные*;

– показан способ построения синтетической выборки для обучения классификатора голов человека;

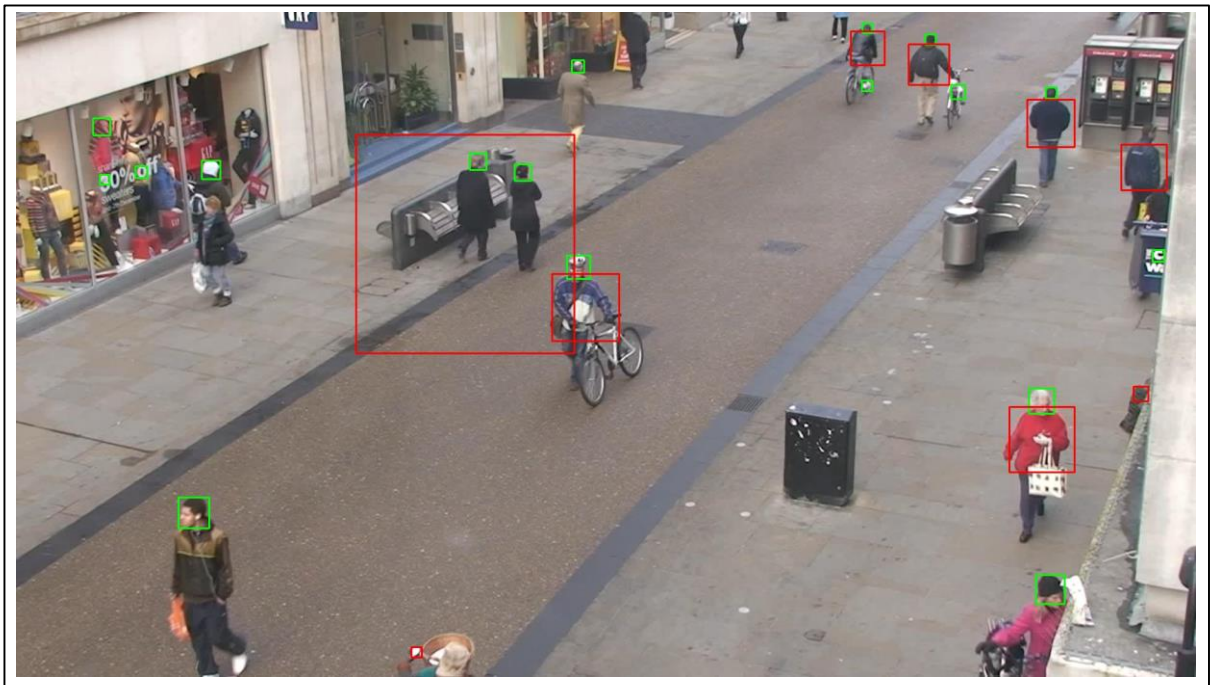


Рис. 1. Результаты работы детектора голов людей на изображении (красным цветом обозначены обнаружения, классифицированные предложенным методом как ложные)

Fig. 1. An image of human head detector results. False detections are shown in red

– показано, что построенный алгоритм классификации обнаружений позволяет повысить точность обнаружения голов людей;

– предложен способ интеграции классификатора обнаружений с используемым детектором голов, который позволил повысить скорость и точность обработки данных.

Смежные работы

Задача построения детектора объектов на изображении всегда интересовала исследователей в области компьютерного зрения. Обычно на разрабатываемые алгоритмы накладывались требования по времени работы и количеству ложных срабатываний. Эти требования зачастую противоречили друг другу. Действительно, часто повышение точности классификатора окон приводит к повышению его вычислительной сложности. Для практического применения в видеонаблюдении скорость обработки данных является ключевым параметром, поэтому многие разработчики стремились найти способы понижения вычислительной сложности детекторов при сохранении их качества. Можно выделить два основных направления работы в этой области: построение быстрого классификатора и уменьшение количества рассматриваемых окон.

Первые работы по ускорению детектирования посвящены ускорению применяемого классификатора. Авторы [1] предложили использовать каскад простых классификаторов для детектирования лиц на изображениях. Первые этапы каскада отбрасывают большое количество простых окон, не содержащих лиц. Предложенная идея оказалась настолько эффективной, что каскадные детекторы стали применяться даже в цифровых фотоаппаратах. Одним из важных недостатков такого подхода является отсутствие возможности изменять соотношение точность/полнота для уже построенного классификатора. Авторы работы [4] преодолели это ограничение, изменив структуру каскада. Они разделили построение простых классификаторов на каждом этапе каскада и выбор границы для разделения положительных и отрицательных примеров. В [5] ускорение классификатора достигнуто за счет вычисления признаков лишь на разреженной пирамиде изображений. На промежуточных слоях предложено восстанавливать признаки с помощью интерполяции. В данной работе предложен алгоритм понижения вычислительной сложности детектора, который не зависит от типа используемого классификатора окон, поэтому его можно использовать совместно с быстрыми классификаторами.

Другое направление по ускорению обнаружения на изображениях – уменьшение количества рассматриваемых окон. Авторы работы [6] используют корреляцию откликов классификатора в соседних окнах для выделения регионов изображения, где могут находиться объекты. Для этого разреженное множество окон классифицируют на первых этапах обработки. В связи с существенным развитием нейросетевых алгоритмов классификации изображений [7–10] сверточные нейронные сети стали применять и для задачи обнаружения

объектов на изображении. Обычно нейросетевые классификаторы требуют больших вычислительных ресурсов, поэтому в работе [11] было предложено классифицировать лишь небольшое подмножество выбранных окон. В работах [3, 12] авторы развили предыдущую идею и предложили разбить классификатор на этапы классификации и уточнения положения объекта. Это позволило увеличить размер окон и уменьшить их количество.

Предложенный в данной работе алгоритм может быть интегрирован с любым из методов уменьшения количества обрабатываемых окон, поскольку дает априорную оценку областей, где могут находиться объекты интереса.

Предложенный метод

В данной работе создан алгоритм классификации обнаружений детектора голов людей на корректные и невозможные с геометрической точки зрения. Построенный алгоритм опирается на информацию о положении и параметрах камеры. Важно отметить, что ключевым требованием к разработанному классификатору является инвариантность к классификатору окон, используемому детектором. Это позволяет построить алгоритм фильтрации обнаружений для любого детектора. Формально входом алгоритма являются положение камеры и признаки классифицируемого обнаружения. Под обнаружением понимается прямоугольник, ограничивающий изображение объекта, найденного детектором.

Предложенный метод предусматривает три этапа построения классификатора результатов детектора:

1) построение синтетической выборки изображений;

2) построение признаков обнаружений, инвариантных для синтетических и реальных данных;

3) обучение классификатора.

Рассмотрим подробнее каждый их этих этапов для построения классификатора обнаружений детектора голов людей.

Построение обучающей выборки

Для обучения классификатора обнаружений необходима обучающая выборка. Авторы не могли использовать размеченные данные видеонаблюдения из-за сложности их сбора, поэтому с помощью компьютерной графики построили синтетическую выборку, моделирующую сценарий видеонаблюдения. Выборка состоит из 30 179 сцен, все они определяются положением камеры. Каждая сцена состоит из плоскости земли, стоящих на ней людей и камеры. Использовалась упрощенная модель камеры, которая определяется углами наклона и поворота камеры, фокусным расстоянием и высотой камеры над плоскостью земли. Для построения каждой сцены параметры камеры выбирались случайным образом из равномерного распределения (см. таблицу).

Параметры камеры

Camera parameters

Обозначение	**Параметр**	**Диапазон значений**
h	Высота	[0; 20]
P	Наклон	$\left[\frac{\pi}{2}; \frac{11\pi}{12}\right]$
R	Поворот	$\left[-\frac{\pi}{12}; \frac{\pi}{12}\right]$
F	Фокусное расстояние	[0; 5000]

Для моделирования людей в сцене использовались результаты работы [13]. Каждая построенная сцена содержит не менее 200 человек, размещенных в случайных положениях плоскости земли.

Извлечение признаков обнаружений

Обучение алгоритмов машинного обучения только на синтетической выборке может привести к эффекту переобучения, поэтому важным этапом обработки является построение признаков, инвариантных к типу обрабатываемых данных. В качестве таких признаков выбраны параметры ограничивающего прямоугольника, обнаруженного детектором. Для задачи фильтрации обнаружений головы была использована оптимизированная версия детектора голов людей [2] из-за высокой скорости его работы.

Модель [13] позволила исключить ложные обнаружения из обучающей выборки за счет информации о положении ключевых точек на человеке.

Построение классификатора

Формально задача классификации обнаружений ложных срабатываний детектора по построенной выборке относится к классу задач поиска аномалий. Отличительной особенностью таких задач является отсутствие примеров отрицательного класса (ошибки детектора). Сведем данную задачу к задаче классификации, предложив простую модель распределения ошибочных результатов работы детектора, состоящую из смеси двух распределений.

Первое моделирует случайные ошибки детектора. Для этого используется равномерное распределение обнаружений (их положений и размеров) по изображению. В качестве второго компонента смеси взято распределение обнаружений в построенной выборке. Это позволяет алгоритму отличить обнаружения, характерные для сцен. Эксперименты показали, что наилучшие результаты классификации получаются при смешивании этих распределений в отношении 1:9.

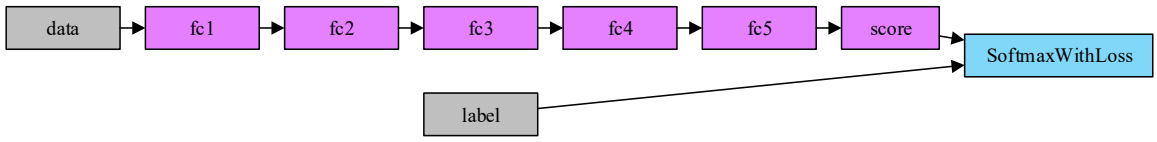


Рис. 2. Визуализация архитектуры используемой сети

Fig. 2. Neural network architecture

Для построения классификатора обнаружений детектора использована нейронная сеть, представленная на рисунке 2. Она состоит из полносвязных слоев, после которых используется нелинейная функция ReLu.

Тестирование и оценка

Предложенная сеть обучена на синтетической выборке, состоящей из 24 143 сцен, с помощью библиотеки caffe [14]. Скорость обучения понижалась каждые 1 500 итераций с параметром γ, равным 0,95.

Проведены оценка качества предложенного алгоритма на синтетических и реальных данных, а также интеграция его с используемым детектором.

Тестирование на синтетических данных. Тестирование проведено на синтетической выборке. Тестовая выборка состояла из 6 036 различных сцен, не использованных при обучении. На рисунке 3а показано качество классификации обнаружений на синтетической тестовой выборке.

Тестирование на реальных данных. Для проведения тестирования необходимо знать положение камеры для каждого кадра выборки. Это затрудняет использование стандартных баз размеченных изображений.

Использована выборка TownCentre [15], так как она содержит большую часть необходимой информации. К сожалению, параметры положения камеры, представленные в выборке, оказались неправдоподобными. Это может быть связано с использованием единицы измерения в мировой системе координат, отличной от метра. В данной сцене высота камеры над землей приблизительно равна 8 метрам.

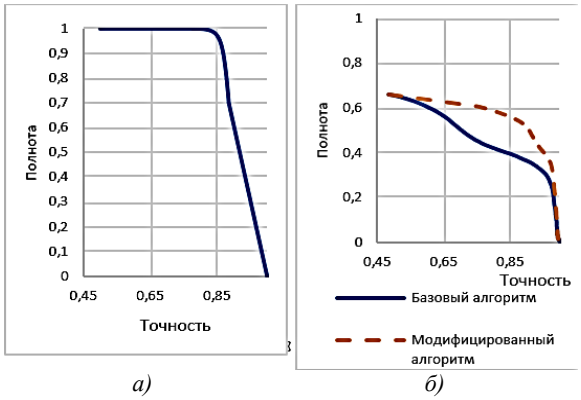
Авторы применили фильтрацию обнаружений детектора с порогом 0,25. Обнаружение считалось правильным, если его пересечение с регионом головы человека в разметке занимало не менее 25 % их объединения. Сравнение качества исходного и полученного детекторов (рис. 3б) выявило, что предложенный алгоритм фильтрации позволяет увеличить точность без существенного уменьшения полноты обнаружения.

Интеграция с детектором

Использовался построенный фильтр обнаружений детектора для ускорения его работы. Действительно, можно оценить, какую область изображения необходимо обрабатывать на каждом кадре.

Построены маски областей, которые необходимо обрабатывать на каждом уровне пирамиды изображений для камеры, соответствующей выборке TownCentre (рис. 4, первая строка). Результаты показывают, что в сцене необходимо обработать только небольшое подмножество окон. Например, для сцены TownCentre достаточно обработать лишь 21,44 % всех окон.

Базовый детектор расширен возможностью обрабатывать только те строки изображения на каждом уровне пирамиды, где правдоподобно обнаружение головы человека. Пример обрабатываемых областей представлен на рисунке 4 (вторая строка). Это соответствует обработке 24,03 % всех окон.



а) *б)*

Рис. 3. Качество фильтрации обнаружений на синтетической тестовой выборке (а) и качество работы детектора до и после использования фильтра обнаружений (б)

Fig. 3. Filtration quality on the synthetic test set (a) and detector performance quality with and without filtration (b)

Выделение областей интереса является эффективным способом повышения производительности детектора. Обработка области интереса на сцене TownCentre позволила повысить скорость обработки данных с 20,03 до 34,36 кадра в секунду.

Полученный результат наиболее важен для систем видеонаблюдения, где параметры камеры резко изменяются и могут быть оценены один раз.

Заключение

В данной работе представлен эффективный способ фильтрации обнаружений детектора, осно-

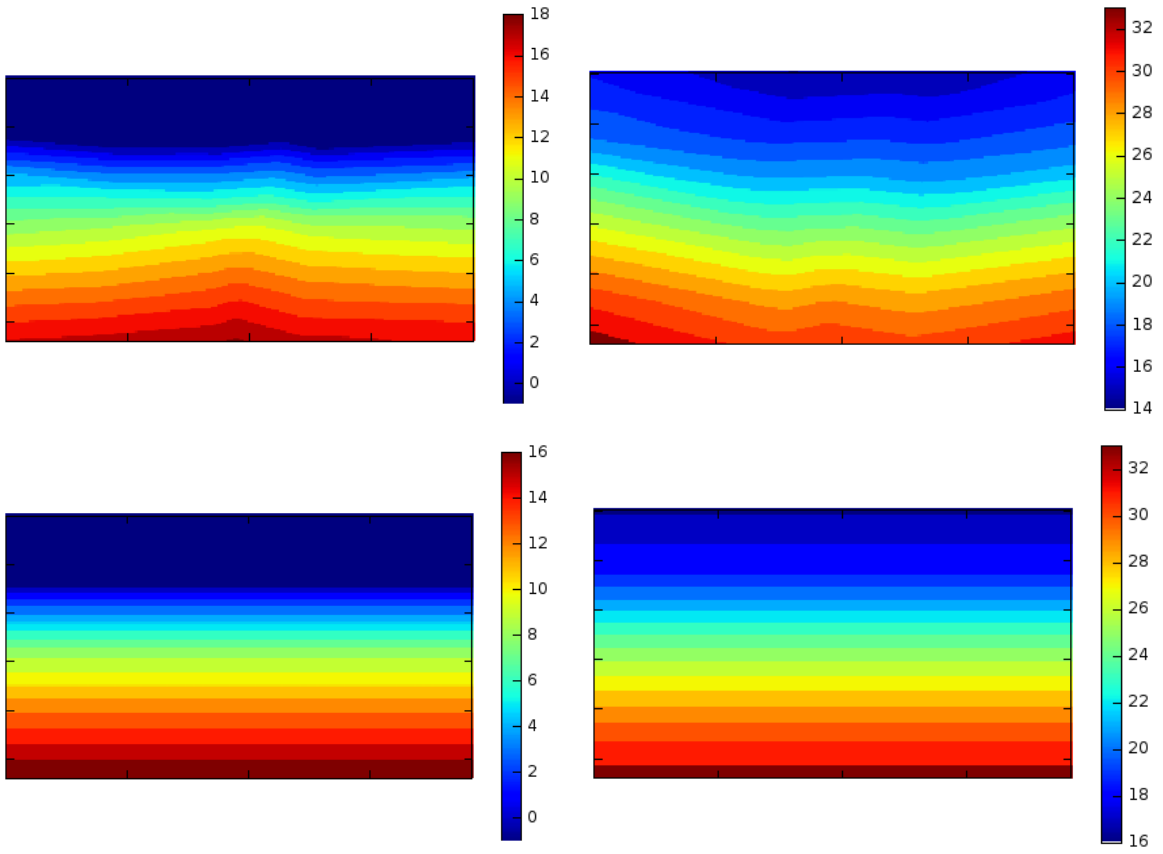


Рис. 4. Уровни пирамиды, соответствующие положительным обнаружениям для сцены TownCentre: а) минимальный, б) максимальный (в первой строке указана предсказанная область интереса, во второй – область интереса, используемая в детекторе)

Fig. 4. Pyramid layers associated to true positive detections on each image location for TownCentre scene: a) minimum, b) maximum (the first row presents the predicted range of interest, the second row presents the one incorporated with the detector)

ванный на использовании информации о положении камеры.

Предложенный метод позволил повысить точность и скорость работы детектора.

Работа выполнена при поддержке РФФИ, грант № 15-31-20596, и Сколковского института науки и технологии, договор № 081-R приложение А2.

Литература

1. Viola P., Jones M. IEEE. Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition, 2001, CVPR 2001, Proc. IEEE Comp. Society Conf. 2001, vol. 1, p. 511.

2. Prisacariu V., Reid I. FastHOG-a real-time GPU implementation of HOG. Department of Eng. Sc., 2009, vol. 2310, no. 9, pp. 325–332.

3. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 2015, pp. 91–99.

4. Bourdev L., Brandt J. IEEE. Robust object detection via soft cascade. Proc. 2005 IEEE Computer Society Conf. CVPR'05, 2005, vol. 2, pp. 236–243.

5. Dollár P., Belongie S., Perona P. The fastest pedestrian detector in the west. BMVC, 2010, vol. 2, p. 7.

6. Dollár P., Appel R., Kienzle W. Crosstalk cascades for frame-rate pedestrian detection. Springer, 2012, pp. 645–659.

7. Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

8. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

9. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.

10. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. Rethinking the inception architecture for computer vision. arXiv preprint arXiv:1512.00567, 2015.

11. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. IEEE Conf. on Comp. Vision and Pattern Recognition, 2014, pp. 580–587.

12. Girshick R. Fast r-cnn. Proc. IEEE Intern. Conf. on Comp. Vision, 2015, pp. 1440–1448.

13. Pishchulin L., Wuhrer S., Helten T., Theobalt C., Schiele B. Building statistical shape spaces for 3d human modeling. arXiv preprint arXiv:1503.05860, 2015.

14. Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S., Darrell T. ACM. Caffe: Convolutional architecture for fast feature embedding. Proc. 22nd ACM Intern. Conf. on Multimedia, 2014, pp. 675–678.

15. Benfold B., Reid I., IEEE. Stable multi-target tracking in real-time surveillance video. Proc. IEEE Conf. CVPR, 2011, pp. 3457–3464.

Software & Systems
DOI: 10.15827/0236-235X.117.106-111

Received 01.08.16
2017, vol. 30, no. 1, pp. 106–111

SCENE GEOMETRY FOR DETECTOR PRECISION IMPROVEMENT

E.V. Shalnov[1], *Postgraduate Student, eshalnov@graphics.cs.msu.ru*
A.S. Konushin[1, 2], *Ph.D (Phisics and Mathematics), Associate Professor, ktosh@graphics.cs.msu.ru*

[1] *Lomonosov Moscow State University, Leninskie Gory, Moscow, 119991, Russian Federation*
[2] *National Research University, Higher School of Economics,*
Myasnitskaya St., 20, Moscow, 101000, Russian Federation

Abstract. Object detection algorithms are the key component of any intelligent video content analysis systems. High computation requirements and low precision of existing methods restrain widespread acceptance of intelligent video content analysis.

The paper introduces a novel algorithm that accelerates existing sliding window object detectors and increases their precision. This approach is based on the geometric properties of an observed scene. If the camera position in the scene is known, we can determine feasible sizes of detected objects in each location of an input image. Windows of other sizes cannot correspond to objects in a scene and thus could be skipped. It significantly decreases computation time. The proposed algorithm estimates feasible sizes of object for each location of an input image. We apply Neural Network (NN) to solve this task. A NN takes camera calibration parameters and window parameters as the input and determines if this configuration feasible or not. We train the NN on the synthetic dataset. It allows us to take into account a huge range of camera calibration parameters. We apply the NN to construct a map of feasible object sizes for the input scene.

Thus the detector processes the feasible subset of windows. The performed evaluation reveals that the proposed algorithm accelerates processing by 70 % and increases precision of a detector.

Keywords: neural networks, computer vision, object detection, video content analysis, computer graphics, pattern recognition.

Acknowledgements. *The research has been supported by RFBR, grant no. 15-31-20596 and Skolkovo Institute of Science and Technology, the contract no. 081-R, Annex A2.*

References

1. Viola P., Jones M. Rapid object detection using a boosted cascade of simple features. *Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2001)*. 2001, vol. 1, p. 511.

2. Prisacariu V., Reid I. *FastHOG-a real-time GPU implementation of HOG*. Department of Eng. Sc., 2009, vol. 2310, no. 9, pp. 325–332.

3. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*. 2015, pp. 91–99.

4. Bourdev L., Brandt J. Robust object detection via soft cascade. *Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)*. 2005, vol. 2, pp. 236–243.

5. Dollár P., Belongie S., Perona P. The Fastest Pedestrian Detector in the West. *BMVC*. 2010, vol. 2, p. 7.

6. Dollár P., Appel R., Kienzle W. *Crosstalk cascades for frame-rate pedestrian detection*. Springer Publ., 2012, pp. 645–659.

7. Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105.

8. Simonyan K., Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv preprint arXiv:1409.1556, 2014.

9. He K., Zhang X., Ren S., Sun J. *Deep Residual Learning for Image Recognition*. arXiv preprint arXiv:1512.03385, 2015.

10. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. *Rethinking the Inception Architecture for Computer Vision*. arXiv preprint arXiv:1512.00567, 2015.

11. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. 2014, pp. 580–587.

12. Girshick R. Fast r-cnn. *Proc. IEEE Int. Conf. on Computer Vision*. 2015, pp. 1440–1448.

13. Pishchulin L., Wuhrer S., Helten T., Theobalt C., Schiele B. *Building Statistical Shape Spaces for 3d Human Modeling*. arXiv preprint arXiv:1503.05860, 2015.

14. Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S., Darrell T. Caffe: Convolutional architecture for fast feature embedding. *Proc. 22nd ACM Int. Conf. on Multimedia*. 2014, pp. 675–678.

15. Benfold B., Reid I. Stable multi-target tracking in real-time surveillance video. *Computer Vision and Pattern Recognition (CVPR 2011)*, IEEE Conf., 2011, pp. 3457–3464.