Использование HOG дескриптора для сверточной нейронной сети в задаче детектирования пешеходов*

Томинин*[1] *В. Д., Томинин*[1] *Я. Д., Демидова*[1] *Ю. О., Дудоров*[1] *Н. А., Ерлыгин*[1] *Л. О., Гнеушев*[1,2] *А. Н.

tominin.vd@phystech.edu, tominin.yad@phystech.edu, demidova.ua@phystech.edu, zoom_ccss@mail.ru, erlygin.la@phystech.edu, gneushev@ccas.ru

[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9

[2]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Рассматривается подход обобщения алгоритма HOG (Histograms of Oriented Gradients) для детектирования пешеходов на изображении путем замены линейного бинарного классификатора SVM на сверточную нейронную сеть небольшой глубины. Для возможности применения входных сверточных слоев нейросети в пространстве интегральных признаков в статье предлагается преобразовать интегральный вектор HOG дескриптора в трехмерный тензор, имеющий две пространственные и одну спектральную размерности, в соответствии с его внутренней блочной структурой. Учет локальной структуры HOG дескриптора позволяет выделять дополнительные локальные признаки первыми сверточными слоями нейросети с целью оптимизации пространства HOG признаков для классифицирующего слоя. В статье исследуются простые сверточные архитектуры нейросети, проведены вычислительные эксперименты и представлены результаты сравнения работы базового алгоритма HOG-SVM и предлагаемых обобщений HOG-DNN и HOG-CNN на базах изображений пешеходов INRIA и Cityscapes.

Ключевые слова: *HOG, CNN, гистограмм ориентированных градиентов, сверточные нейросети, дектироапние пешехода.*

1 Введение

Автоматическое детектирование и распознавание объектов на изображениях является одной из основных задач компьютерного зрения. Задача локализации человека на видеоизображениях широко востребована в таких областях, как мониторинг и анализ дорожных ситуаций, обнаружение дорожно-транспортных происшествий, контроль за соблюдением правил дорожного движения, системы безопасности и следящие системы, беспилотные автомобили, робототехника, системы помощи водителю.

Основные сложности обнаружения человека на изображении связаны с несколькими причинами: изменение структуры изображения вследствие движения человека, неравномерная освещенность изображения, большая вариабельность изображений человека из-за разных ракурсов (поз, размеров, углов поворота), частичные перекрытия фигуры человека другими объектами.

Как правило, задача детектирования объекта разделяется на две подзадачи: выделение характерных свойств изображения объекта и бинарная классификация. Характерные свойства изображения объекта - это набор признаков, приближенно описывающий интересующий объект. Выбранный набор признаков является важнейшим фактором, влияющим на качество классификации и ее устойчивость. Чем лучше признаки обладают разделительной способностью (возможно, сложнее устроены), тем проще устроено признаковое

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Матвеев И. А. Задачу поставил и консультировал: Гнеушев А. Н.

пространство, и классификатор может иметь простой вид. Наоборот, чем менее уникальны признаки (но проще по структуре), тем сложнее устроено признаковое пространство и требуется более сложный классификатор для его успешного разделения. Наиболее эффективным в настоящее время является подход, объединяющий оба этап анализа изображения: одновременное выделение множества признаков и их классификация многослойными сверточными нейронными сетями (CNN), в которых процедура выделения признаков осуществляется в начальных слоях, структура признаков формируется автоматически в процессе его обучения и определяется моделью и архитектурой сети. Чем больше признаков необходимо использовать для характеристики целевых объектов, тем больше параметров требуется для задания модели сети, тем она вычислительно сложнее и требует больше вычислительных ресурсов и объема обучающей выборки изображений. Это связано с тем, что в CNN признаковая модель объекта формируется на основе только той информации, которая содержится в обучающей базе и общих регуляризационных ограничениях. Привлекая априорную, более детальную экспертную информацию о характере конкретного класса объектов на изображении для построения модели пространства признаков, можно уменьшить количество параметров нейросети, ее вычислительную сложность и потребность в большом объеме обучающей выборки.

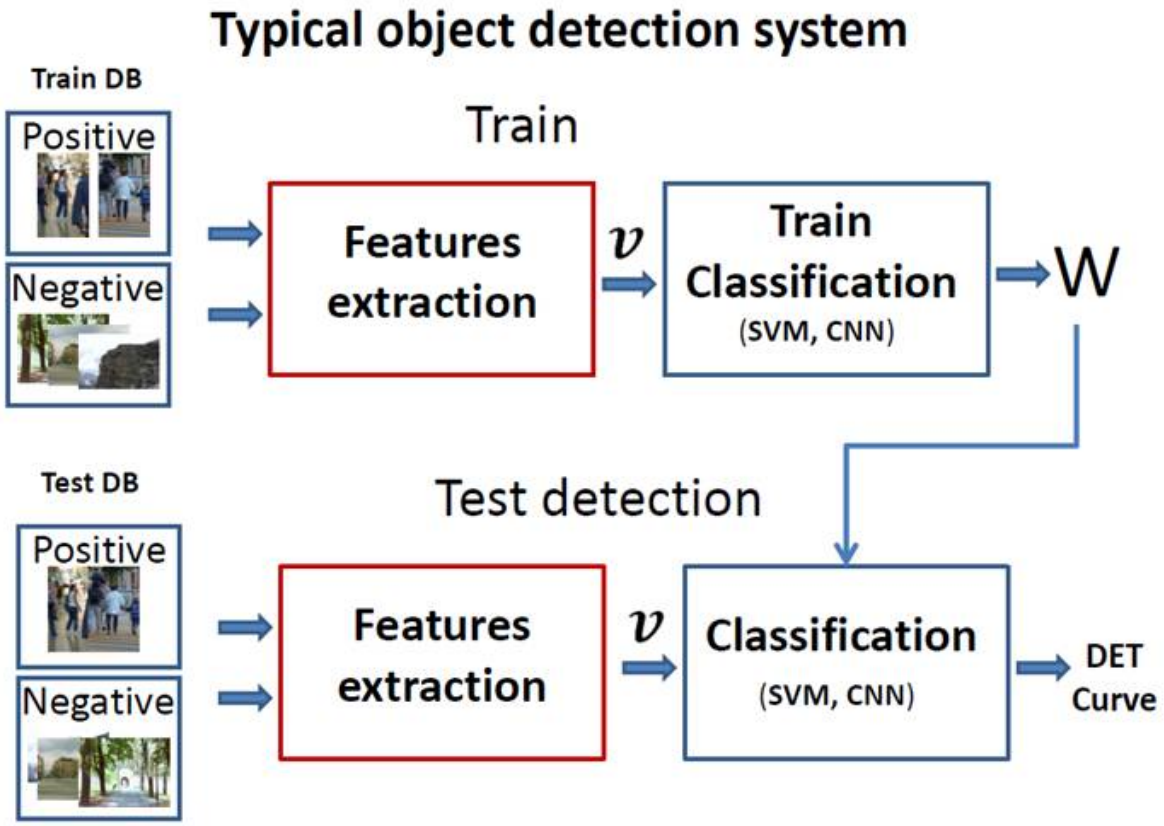

Рис. 1 Типичная система детектирования объектов.

Множество признаков, которые используются для распознавания объекта, определяются его характерной структурой на изображении. Изображение человека может быть представлено совокупностью контуров, силуэтов частей тела, которые являются контурными признаками и представляются на изображении как максимальные перепады зна-

чений яркости [1]. Эффективный алгоритм детектирования пешеходов, основанный на градиентах яркости изображения и учитывающий совокупность контурных признаков предложен в методе HOG [3] и подходах, которые его развивают [2, 4]. HOG-дескриптор обладает рядом преимуществ, он вычислительно эффективен, показывает лучшие в своем классе результаты, однако для детектирования используется линейный классификатор SVM (Support Vector Machine), который предполагает линейную разделимость пространства признаков. Использование нелинейного классификатора может уменьшить ошибки детектирования особенно в сложных случаях с частичным перекрытием объектов, большой вариабельностью ракурса, в линейно неразделимом пространстве признаков. Таким образом, использование несложной сверточной нейронной сети в качестве нелинейного классификатора для пространства HOG дескрипторов с одной стороны позволяет задействовать все современные преимущества CNN, с другой стороны минимизировать сложность и вычислительную сложность детектора пешеходов в целом.

Обычно, для задачи классификации в интегральных признаковых пространствах используются полносвязные нейронные сети (DNN), так как в интегральных дескрипторах-векторах отсутствует локальная связность их компонент. Однако HOG дескриптор имеет внутреннюю блочную структуру по построению, и каждый определенный участок интегрального вектора HOG дескриптора является локальным дескриптором блока изображения, локальной о области. В данной работе предлагается использовать эту информацию и преобразовать интегральный вектор HOG дескриптора в трехмерный тензор, имеющий две пространственные и одну спектральную размерности, и сохраняющий внутренней блочную структуру дескриптора. Данный подход позволяет использовать сверточные входные слои нейросети и, следовательно, использовать глубокое обучение на HOG пространстве. Учет локальной структуры HOG дескриптора позволяет выделять дополнительные локальные признаки первыми слоями нейросети с целью оптимизации пространства HOG признаков для финального классифицирующего слоя.

2 Существующие подходы

Исследователи используют несколько подходов к решению задачи.

3 Постановка задачи

Рассмотрим структуру интегрального вектора признаков $\boldsymbol{g}(f) = (g_1, \ldots, g_n)^{\mathrm{T}}$ HOG дескриптора [3] из признакового пространства G для входного изображения f, где n — размерность пространства признаков. В соответствии со схемой HOG, изображение $f(x, y)$, где x, y — координаты точки изображения размером $X \times Y$, разбивается на смежные области — квадратные ячейки размером $K \times K$ пикселей, $I = X/K$ ячеек по горизонтали и $J = Y/K$ ячеек по вертикали. В каждой ячейке (i, j), где i, j — индексы ячейки по вертикали и горизонтали соответственно, выделяются локальные признаки, которые определяются дескриптором ячейки — вектором $\boldsymbol{u}_{i,j} = (u_1, \ldots, u_l)^{\mathrm{T}}$, где l — размерность вектора. Четыре смежные ячейки объединяются в пересекающиеся блоки. Таким образом, каждый блок содержит $2K \times 2K$ пикселей и имеет общие ячейки с соседними блоками. Дескриптор блока определяется объединением дескрипторов собственных ячеек, вектором $\boldsymbol{v}_{i,j} = \boldsymbol{u}_{i,j} \cup \boldsymbol{u}_{i+1,j} \cup \boldsymbol{u}_{i,j+1} \cup \boldsymbol{u}_{i+1,j+1}$, где i,j — индексы блока. Общее количество блоков — $(I-1)(J-1)$. Под операцией объединения $\cup$ двух векторов $\boldsymbol{u}_1$ и $\boldsymbol{u}_2$ с размерностью l_1 и l_2 соответственно будем понимать результирующий вектор из пространства размерности $l_1 + l_2$, первые l_1 компонент которого являются компонентами вектора первого аргумента, последние l_2 — компонентами вектора второго аргумента.

Дескриптор блока нормируется с помощью одной из двух норм: L_2 нормы

$$\tilde{\boldsymbol{v}} = N_{L_2}(\boldsymbol{v}) = \frac{\boldsymbol{v}}{\sqrt{\|\boldsymbol{v}\|_2^2 + \varepsilon^2}}$$

либо L_2–hys нормы [3]

$$\tilde{\boldsymbol{v}} = N_{L_2-\mathrm{hys}}(\boldsymbol{v}) = N_{L_2}\left(\min(N_{L_2}(\boldsymbol{v}), h)\right),$$

где операция min применяется покомпонентно к вектору первого аргумента; h — пороговое значение, которое используется для ограничения значений компонент вектора в операции min. В данной работе, как и в работе [3], в качестве порогового значения используется $h = 0{,}2$.

Интегральный дескриптор определяется объединением всех блочных дескрипторов, т. е. вектором

$$\boldsymbol{g} = \bigcup_{j}^{J-1} \bigcup_{i}^{I-1} \tilde{\boldsymbol{v}}_{i,j}\,. \tag{1}$$

Интегральные дескрипторы $\boldsymbol{g}$ множества изображений формируют признаковое пространство G.

В работе ставится задача преобразования признакового векторного пространства G в трехмерное тензорное пространство с размерностью $2(I-1) \times 2(J-1) \times l$, восстанавливающее пространственную смежность дескрипторов ячеек $\boldsymbol{u}$ из блоков $\boldsymbol{v}$ вектора $\boldsymbol{g}$. Используя тензорное представление вектора $\boldsymbol{g}$ реализовать сверточную нейронную сети одной из известных архитектур для решения задачи разбиения этого пространства на два непересекающихся класса: первый характеризует пешеходов; второй — фон, не содержащий пешехода.

Множество весов и параметров нейросетевого классификатора находятся из процедуры обучения по специально подготовленной обучающей выборке из базы изображений INRIA, CityScapes [6] содержащие два подмножества изображений: с положительными примерами, содержащими пешеходов, и отрицательными примерами, содержащими фон.

Критерием качества классификации на специально подготовленной тестовой выборке из базы изображений INRIA [6] и CityScapes будем считать отношение $\mathrm{MR} = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN})$ — доля неверно отвергнутых классификатором изображений (Miss Rate), к $\mathrm{FPPW} = \mathrm{FP}/(\mathrm{TN} + \mathrm{FP})$ — доля неверно принятых изображений (False Positive Per Window), где FN — количество неверно отвергнутых классификатором положительных примеров; TP — количество верно классифицированных положительных примеров; TN — количество верно классифицированных отрицательных примеров; FP — количество неверно классифицированных отрицательных примеров.

4 Тензорное представление HOG дескриптора для CNN

Полученный интегральный дескриптор $\mathbf{g}$(Рис. 2) является объединением всех блочных дейскрипторов. Признаковое пространство G имеет размерность $2(I-1)2(J-1)l$. Для востановления пространственной смежности ячеек нужно перейти в новое трехмерное тензорное пространство с размерностью $2(I-1) \times 2(J-1) \times l$.

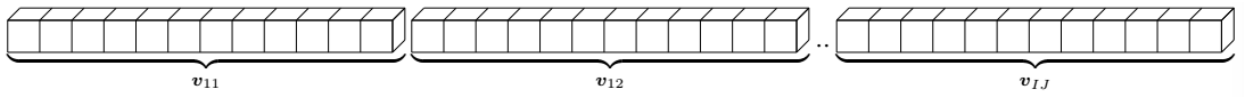

Рис. 2 Изначально вектор g имеет вид

Далее проводится операция с каждым вектором размерности 4ℓ. Рассмотрим произвольный вектор блока $\boldsymbol{v}_{i,j} = \boldsymbol{u}_{i,j} \cup \boldsymbol{u}_{i+1,j} \cup \boldsymbol{u}_{i,j+1} \cup \boldsymbol{u}_{i+1,j+1}$, где i,j — индексы блока; $\boldsymbol{u}_{ij}$-вектор ячейки (i,j). Состаим тензор $\mathbf{Z}_{ij}$ размерности $2 \times 2 \times \ell$, причем $Z_{ij}[0][0] = \boldsymbol{u}_{i,j}$, $Z_{ij}[0][1] = \boldsymbol{u}_{i,j+1}$, $Z_{ij}[1][0] = \boldsymbol{u}_{i+1,j}$, $Z_{ij}[1][1] = \boldsymbol{u}_{i+1,j+1}$.(Рис. 3)

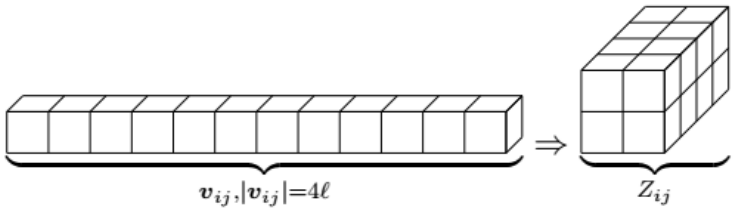

Рис. 3 Изменение размерности вектора длины 4ℓ

Для каждого $j = 1 \div \mathbf{J} - 1$ составим тензор $\mathbf{W}_j = \mathbf{Z}_{1j} \cdot \mathbf{Z}_{2j} \cdot ... \cdot \mathbf{Z}_{(I-1)j}$(Рис. 4), здесь под операцией $\mathbf{Z}_{ij} \cdot \mathbf{Z}_{we}$ подразумевается покомпонентное объединение векторов $\mathbf{Z}_{ij}[:][k][q] \cup \mathbf{Z}_{we}[:][k][q]$, где $k = 1 \div 2; q = 1 \div \ell$

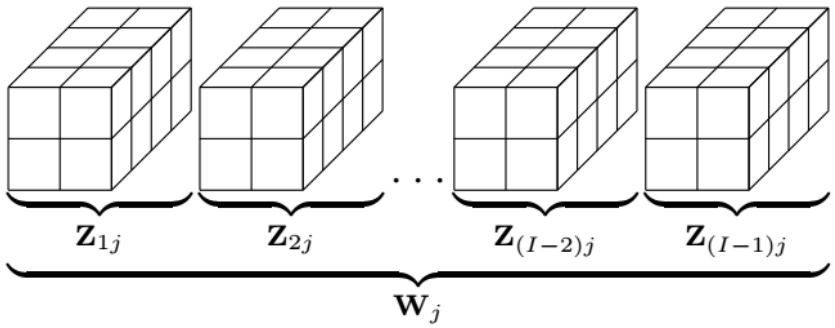

Рис. 4 Получение тензора $\mathbf{W}$ размерности $2\mathbf{I} \times 2 \times \ell$

Далее получим тензор $\mathbf{R}$, который будет соответстовать трехмерному тензорному пространству с размерностью $2(I-1) \times 2(J-1) \times l$ по формуле $\mathbf{R} = \mathbf{W}_1 \circ \mathbf{W}_2 \circ ... \circ \mathbf{W}_{J-1}$ (Рис. 5), где под операцией $\mathbf{W}_j \circ \mathbf{W}_i$ подразумевается покомпонентное объединение векторов $\mathbf{W}_j[k][:][q] \cup \mathbf{W}_i[k][:][q]$, $k = 1 \div 2(\mathbf{I} - 1); q = 1 \div \ell$.

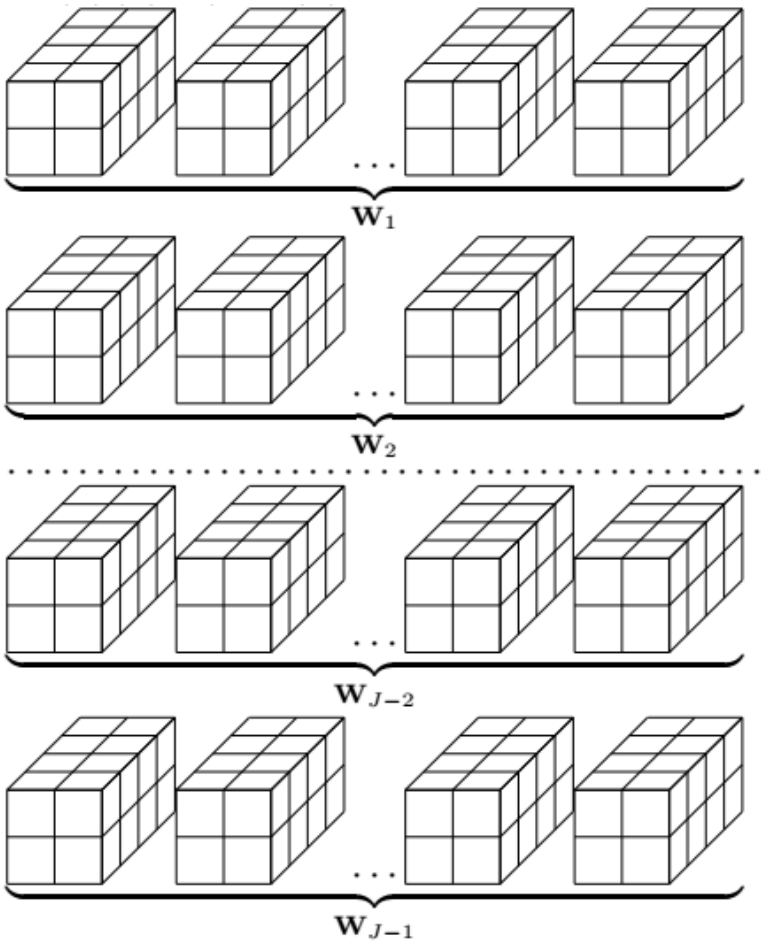

Рис. 5 Трехмерное тензорное пространство $\mathbf{R}$с размерностью $2(I-1) \times 2(J-1) \times l$

Полученный тензор $\mathbf{R}$, имеющий размерность $2(I-1) \times 2(J-1) \times l$, будет подаваться в нейронную сеть.

5 Результаты вычислительных экспериментов

Целью вычислительных экспериментов является проверка работы предлагаемого в статье метода на реальных данных, а также сравнение его качества детектирования и времени работы с аналогичными показателями HOG дескриптора. В качестве обучающей выборки была использована база изображений INRIA. В базу входит 2478 изображений пешеходов размером 96×160 пикселей. Из каждого положительного изображения выбирается окно размером 64 пикселей, центр которого совпадает с центром изображения. Таким образом, 2478 изображений составляют положительную часть обучающей выборки. Также база содержит 1218 изображений фона, из которых в отрицательную чать обучающей выборки выделяется 12180 окон размером 64 пикселей, центр окна определяется случайно. Для обучения классификатора была использована библиотека OpenCV, гласификатор обучался с помощью линейного метода SVM.

6 Заключение

В работе представлен подход обобщения алгоритма HOG для детектирования пешеходов на изображении путем замены линейного бинарного классификатора SVM на свер-

точную нейронную сеть небольшой глубины. Представлены результаты вычислительного эксперимента с использованием сверточной нейронной сети небольшой глубины и их сравнение с HOG-дескриптором по производительности. Учет локальной структуры HOG дескриптора позволяет выделять дополнительные локальные признаки первыми сверточными слоями нейросети с целью оптимизации пространства HOG признаков для классифицирующего слоя. Основным недостатком предлагаемого подхода является довольно низкая производительность, вызванная оптимизацией весов нейронной сети на каждом шаге обучения.

Neural convolutional network based on HOG features for pedestrian detection

Tominin*[1] *V. D.*, *Tominin*[1] *Y. D.*, *Demidova*[1] *U. O.*, *Dudorov*[1] *N. A.*, *Erlygin*[1] *L. A.*, *Gneushev*[1,2] *A. N.

tominin.vd@phystech.edu, tominin.yad@phystech.edu, demidova.ua@phystech.edu, zoom_ccss@mail.ru, erlygin.la@phystech.edu, gneushev@ccas.ru

[1]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow, Russia

[2]Federal Research Center "Computer Science and Control" of RAS, 44/2 Vavilova Str., Moscow, Russia

HOG ¡...¿.

Keywords: *HOG, CNN, Histograms of Oriented Gradients, convolutional network, pedestrian detection.*

Литература

[1] Gneushev, A. N., and A. B. Murynin. 2003. Adaptive gradient method for extracting contour features of objects in images of real-world scenes. J. Comput. Sys. Sci. Int. 42(6):973–980.

[2] Samsonov N. A., Gneushev, A. N. 2017. Textural descriptor in the Hough accumulator space of the gradient field for detecting pedestrians. Machine Learning and Data Analysis 3(3):203–215.

[3] Dalal, N., and B. Triggs. 2005. Histograms of oriented gradients for human detection. IEEE CVPR. San Diego, CA.

[4] Felzenszwalb, P. F., B. R. Girshick, D. McAllester, and D. Ramanan. 2010. Object detection with discriminatively trained part based models. IEEE T. Patt. Anal. 32(9):1627–1645.

[5] Sun, D., and J. Watanada, 2015. Detecting pedestrians and vehicles in traffic scene based on boosted HOG features and SVM. IEEE 9th Symposium (International) on Intelligent Signal Processing.

[6] INRIA Person Dataset. Avalaible at: http://pascal.inrialpes.fr/data/human/ (accessed June 4, 2017).

[7] Open Source Computer Vision Library. Avalaible at: http://opencv.org/releases.html (accessed May 16, 2017).

[8] Vorontsov, K. V. 2007. Lectures on support vector machine. http://www.ccas.ru/voron/download/SVM.pdf (accessed June 25, 2017).