Traffic Sign Recognition on Video Sequence Using Deep Neural Networks and Matching Algorithm

Ilya Belkin   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

Sergey Tkachenko  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

Dmitriy Yudin  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction (*Heading 1*)

Глубокие нейронные сети зарекомендовали себя с лучшей стороны в самых разных задачах в различных областях знаний. В особенности это касается задач связанных с компьютерным зрением и распознаванием образов[1]. За последние годы были достигнуты впечатляющие результаты в этой области, что делает разработанные методы[2, 3, [4](http://openaccess.thecvf.com/content_cvpr_2018/html/Cai_Cascade_R-CNN_Delving_CVPR_2018_paper.html)] привлекательными с точки зрения применения в реальных приложениях. Одним из таких приложений является автономный транспорт, а точнее системы машинного зрения, настроенные на обработку и распознавание объектов дорожной инфраструктуры. Основные задачи компьютерного зрения в этой области включают в себя семантическую сегментацию[5] (например для выделения дорожного полотна), детекцию объектов (поиск дорожных знаков, светофоров), классификацию изображений (распознавание сигнала светофора), объектную сегментацию (нахождение очертаний пешеходов, автомобилей).

Как правило, большинство методов глубоко обучения для задач детекции и объектной сегментации хорошо исследованы для покадровой обработки. То есть последовательность кадров рассматривается как набор независимых кадров, а информация между соседними кадрами никак не используется.

Для использования дополнительной информации, содержащейся в последовательности, были разработаны глубокие сети особой архитектуры, принимающие на вход сразу несколько кадров, или обрабатывающие их с помощью рекуррентной нейронной сети. Такие подходы[6] не только позволили улучшить качество работы на последовательных кадрах, но и существенно увеличить скорость обработки.

Тем не менее существенным недостатком этих методов является необходимость использования размеченных последовательностей. Стандартный подход к разметке подразумевает разметку кадров с некоторым шагом, скажем 1 кадр на секунду видеопотока. При частоте 30 кадров в секунду подготовка данных для таких моделей требует выполнения в 30 раз больше работы.

Другой класс работ направлен на постобработку результатов детекции на отдельных кадров с учетом последовательной природы данных. Примером может служить метод Seq-NMS[7], являющийся адаптацией алгоритма подавления немаксимумов результатов детекции применительно к целым последовательностям детекций. Другой подход – Seq-Bbox-Matching[8], который будет подробно описан в соответствующем разделе, позволяет объединить результаты детекции, соответствующие одному объекту используя не только геометрическое расположение охватывающих прямоугольников, но и их семантическое сходство.

В данной работе исследуются подходы к детекции объектов применительно к задаче обнаружения и одновременного распознавания дорожных знаков. Получены результаты для пяти популярных архитектур. Модели были обучены на базе российских дорожных знаков RTSD[9], являющейся самой крупной базой знаков дорожного движения, находящейся в открытом доступе. Для тестирования применялся новый набор данных IceVision[10], собранный с прототипа автономного автомобиля. Получены метрики качества mAP[13], mAR а также AP, AR для различных классов. Исследован алгоритм seq-bbox-matching при тестировании на разреженной разметке, а также представлена открытая реализация этого алгоритма.

# Problem Definition

В данном разделе рассматривается постановка задачи обнаружения и распознавания дорожных знаков на видеопоследовательностях. Приводятся используемые метрики качества а также обсуждается метод оценки качества алгоритма при наличии разметки только для некоторых кадров.

## Описание задачи

Формально, нам дана упорядоченная последовательность кадров F = (f\_i)\_1^N длины N, где каждый кадр это RGB изображение фиксированного разрешения (w, h) полученное с RGB камеры, закрепленной на автомобиле. Требутеся каждому кадру сопоставить набор b\_i = (o\_i^j), где каждый o\_i^j представляет собой упорядоченную четверку чисел (x\_tl, y\_tl, w, h) задающую ограничивающий прямоугольник, где x\_tl, y\_tl координаты левого верхнего угла, а w, h – ширина и высота прямоугольника соответственно. Каждый прямоугольник суть гипотеза того, что в этой области находится некоторый объект, в нашем случае – знак дорожного движения. Кроме того, требуется определить тип знака. Из всех типов знаков, мы сосредоточили свое внимение на следующих типах: … Кроме того, мы объединили типы 5.19.1 и 5.19.2 в один тип – 5.19, а также … Идеальные изображения приведены на рис. ?. Более подробно, в качестве выхода алгоритма для каждого изображения наряду с b\_i мы считаем d\_i = (p\_i^j), где каждый p\_i^j – суть вектор длины (N + 1), каждый элемент которого представляет собой вероятность принадлежности объекта определенному классу, т.е. типу знака.

## *Оцека качества*

Стандартной метрикой оценки качества алгоритмов обнаружения объектов является mAP ([13]). При этом имеются открытые реализации этой метрики ([], []). Она позволяет оценить среднее качество детекции объектов нескольких классов. При этом она имеет несколько параметров. Параметры, с которыми были расчитаны метрики в данной работе отличаются от параметров, используемых в COCO challenge ([]). Во-первых, мы не рассматриваем при расчете метрик объекты размера меньше 20\*20 пикселей. Данная цифра была подобрана эммпирически, но она в целом отражает ситуацию на реальных изображениях: разметка, а так же детекция объектов меньшего размера вызвает серьёзные заттруднения. Повышение качества распознавания на небольших объектах связано с определёнными трудностями, в первую очередь с тем, что разрешение камер, а так же ограничения чувствительности человеческого глаза не позволяют корректно разметить такие объекты. Во-вторых, при подсчете mAP порог IOU был выбран 50% - эта цифра была выбрана в свзязи с тем, что именно такой порог фигурирует во многих исследованиях в области CV. В-третьих, при подсчёте метрик мы учитывали каждый 30 кадр. Это связано с тем, что разметка присутствовала только на каждом 30-м изображении.

Помимо mAP важной метрикой является FPS – показатель быстродействия данного алгоритма. Зачастую

выбор того или иного решения зависит от соотношения качества распознавания объектов и его производительности.

# Датасеты, использованные нами

Особое положение в задаче компьютероного зрения отведено разметке и подготовке данных. В нашем исследовании мы использовали в качестве основных два датасета с размеченными российскими дорожными знаками: RTSD( Russian Traffic Sign Dataset)[] и IceVision[].

A. RTSD

Набор данных RTSD [9] содержит самое большое количество размеченных изображений с дорожными знаками в различное время суток, разметка соответствует российским ГОСТам и правилам дорожного движения. [10]. Особенности набора данных указаны в таблице 1.

|  |  |
| --- | --- |
| Характеристика | Значение |
| разрешение | 1280x720 |
| Всего кадров | 179138 |
| Размечено кадров | 179138 |
| Число классов | 156 |
| Число боксов | 104358 |
| Кадров со знаками | 59188 |

Таблица 1. Характеристики RTSD.

Для классов 3.11, 3.12, 3.13, 3.14, 3.16, 3.24, 3.25, 3.4, 6.2 указаны дополнительные данные: числовое значение на знаке.

  
Иллюстрация 1: Пример разметки в RTSD.

База достаточно разнообразна, собрана в разные сезоны, при разных погодных условиях. Ночных изображений нет, есть вечерние. Снималось на видеорегистратор, установленный внутри автомобиля. Ездили со скоростью обычного авто, многие фото смазаны. В оригинальной базе отсутствуют знаки 8.22.1 - 8.22.3, но были размечены в лаборатории при помощи инструмента opencv/cvat[17]. Разметка осуществлена в формате MS COCO JSON[18].

B. IceVision

В рамках проводимого конкурса Ice Vision в 2019 году был опубликован набор данных, содержащий видеозаписи с размеченными дорожными знаками в зимних условиях и условиях плохой видимости[11].

Детали открытого набора данных Ice Vision представлены в таблице 2.

|  |  |
| --- | --- |
| Характеристика | Значение |
| Разрешение | 2448х2048 |
| Всего кадров | 212965 |
| Размечено кадров | 8563 |
| Число классов | 178 |
| Число боксов | 42758 |
| Кадров со знаками | 6956 |

Таблица 2. Характеристики датасета IceVision.

Все знаки также содержат дополнительную разметку в виде содержащегося на них текста. Это касается не только знаков типа ограничения скорости, но и информационных табличек. Также каждый знак имеет метку temporary (является ли знак временным, желтый фон) и occluded (является ли знак частично закрытым другим объектом).

Данные собирались исключительно в зимних условиях в Московской области. Присутствуют как ночные, так и дневные изображения. Исходные данные представляют собой последовательности кадров, записанных с частотой 30fps. При этом размечается примерно каждый 30-й кадр.

В разметке присутствуют знаки с площадью ограничивающего прямоугольника от 100 пикселей. Многие знаки при этом неразличимы. На глаз невозможно не только определить класс знака, но и само его наличие.

# Deep Learning Approach to traffic sign recognition

При решении задачи детекции объектов на изображении при помощи методов глубокого обучения встаёт вопрос о выборе оптимальной архитектуры нейросети. Выбор происходит с опором на метрики, которые тот или иной алгоритм показывает на данном датасете. Особое внимание отводится mAP, mAR и FPS. Зачастую вопрос выбора архитектуры нейронной сети сводится к тому, чтобы подобрать решение с максимальными показателями mAP и mAR и удовлетворительным FPS. Во многом именно FPS устанавливает ограничения на применения того или иного алгоритма. В нашей работе мы исследовали 5 архитектур нейростей: Mask R-CNN[3], Faster R-CNN[14], Cascade R-CNN[15], Cascade Mask R-CNN и Retina Net[16]. Все из них были протестированы на датасете IceVision и RTSD. Были проанализированы метрики для каждой архитектуры, а так же время работы на [УКАЗАТЬ ЖЕЛЕЗО].

A. Faster R-CNN

Faster R-CNN это двухступенчатный детектор, который содержит в себе две нейросети. Первая, - это RPN, которая ищет proposals, на которых могут содержаться объекты, . Вторая же представляет из себя классификатор, который определяет принадлежность объекта к определённому классу. Данный детектор отличается быстротой, а также вполне удовлетворительным качеством распознавания.

B. Mask R-CNN

Представляет из себя расширенный вариант Faster R-CNN. За счёт увеличения размера нейросети данная архитектура способна выделять не просто прямоугольные области, но и контуры объекта.

C. Cascade R-CNN

Это многоступенчатый детектор, основное отличие которого от Faster R-CNN заключается в последовательном применении сразу нескольких детекторов для повышения качества распознавания.

D. Cascade Mask R-CNN

Расширенная версия предыдущего детектора, призванная выделять дополнительно ещё и контуры объекта.

E. Retina Net

RetinaNet представляет собой единую унифицированную сеть, состоящую из backbone сети и двух специализированных subnetworks. Backbone отвечает за вычисление карты функциональных возможностей по всему входному изображению и является собственной сверточной сетью. Первая подсеть выполняет классификацию на выходе магистрали; вторая подсеть выполняет регрессию ограничивающего прямоугольника свертки.

# Matching Algorithm

Стандартной проблемой, возникающей при попытке применения подхода still image detection к задаче детекции объектов на видеопоследовательностях является периодическая потеря объекта. Как правило, объект теряется на небольшое количество кадров. Возможна и обратная ситуация, когда в последовательности детекций возникает объект, которого не было на предыдущих и нет на последующих кадрах. Возникает идея использования информации с предыдущих кадров для калибровки результатов на последующих. Один возможный подход это использование трекера объектов, для продолжения (propogation) ограничивающих прямоугольников на последующие кадры. Последние работы по детекции объектов на видео предлагают end-to-end подходы, позволяющие достигнуть значительного увеличения качества и скорости обработки. В то же время, эти подходя требуют размеченных последовательностей кадров, что является очень трудоемкой задачей. Здесь мы применяем другой подход, позволяющий сглаживать результаты still image detection на последовательных кадрах используя информацию с предыдущих кадров. Причем мы используем как геометрическую информацию, так и семантическую. Этот подход, называемый Seq-Bbox-Matching, позволяет с минимальными дополнительными расходами увеличить качество детектирования и не требует дополнительных модификаций процесса обучения/тестирования моделей.

Алгоритм Seq-Bbox-Matching заключается в объединении результатов детекции на текущем кадре с результатом на предыдущем кадре. Вводится метрика между ограничивающими прямоугольникми на друх последовательных кадрах, учитывающаяя как их геометрическое расположение посредствам IoU, так и семантическое сходство, посредствам вычисления косинусного расстояния между векторами, представляющими собой выходы сети классификации анкоров (anchor). Отметим, что в оригинальной работе использовалось скалярное произведение между векторами вероятностей. Мы обнаружли, что косинусное расстояние позволяет достигнуть лучшего результата, не требуя существенного увеличения вычислительных ресурсов. В остальном наш алгоритм повторяет оригинальную работу. Кроме того, мы предоставляем открытую реализацию метода <https://github.com/ivbelkin/icevision/blob/master/icevision/seq_bbox_matching.py>.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. [http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detectionhttp://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection](http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection)
2. <http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Hybrid_Task_Cascade_for_Instance_Segmentation_CVPR_2019_paper.html>
3. <http://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.htmlhttp://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html>
4. <http://openaccess.thecvf.com/content_cvpr_2018/html/Cai_Cascade_R-CNN_Delving_CVPR_2018_paper.html>
5. https://www.cv-foundation.org/openaccess/content\_cvpr\_2015/html/Long\_Fully\_Convolutional\_Networks\_2015\_CVPR\_paper.html
6. https://arxiv.org/abs/1710.03958
7. <https://arxiv.org/abs/1602.08465>
8. Hatem Belhassen, Heng Zhang, Fresse, Virginie, El-Bey Bourennane. Improving Video Object Detection by Seq-Bbox Matching. 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Feb 2019, Prague, Czech Republic. ⟨hal-02050829⟩
9. Лаборатория компьютерной графики и мультимедиа. URL: <http://graphics.cs.msu.ru/ru/node/1266>
10. В.И. Шахуро, А.С. Конушин. Российская база изображений автодорожных знаков, Компьютерная оптика. - 2016. - Т.40 № 2. - С. 294-300. <https://cyberleninka.ru/article/v/rossiyskaya-baza-izobrazheniy-avtodorozhnyh-znakov>
11. Artem L. Pavlov, Pavel A. Karpyshev, George V. Ovchinnikov, Ivan V. Oseledets, and Dzmitry Tsetserukou, "IceVisionSet: lossless video dataset collected on Russian winter roads with traffic sign annotations", IEEE International Conference on Robotics and Automation (ICRA), 2019.
12. <https://github.com/icevision/annotations>
13. Everingham, M., Van Gool, L., Williams, C.K.I. et al. Int J Comput Vis (2010) 88: 303. <https://doi.org/10.1007/s11263-009-0275-4>
14. <http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks>
15. <http://openaccess.thecvf.com/content_cvpr_2018/html/Cai_Cascade_R-CNN_Delving_CVPR_2018_paper.html>
16. <https://arxiv.org/abs/1708.02002>
17. <https://github.com/opencv/cvat>
18. https://arxiv.org/abs/1405.0312

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**