Санкт-Петербургский государственный университет

Кафедра системного программирования

Группа 21.Б07-мм

Распознавание текста на видео в ADAS для учета контекста

Артеменко Софья Алексеевна

Отчёт по учебной практике
в форме «Теоретическое исследование»

Научный руководитель:
старший преподаватель кафедры СП, к. т. н., Ю. В. Литвинов

Консультант:
инженер-программист АО «Кама», М. С. Осечкина

Санкт-Петербург
2024

Оглавление

Введение

ADAS (Advanced Driver-Assistance Systems) — системы помощи водителю, которые помогают безопасно управлять транспортным средством. Их главная задача состоит в ассистировании в сложных ситуациях, возникающих во время движения. Архитектура ADAS подразумевает наличие датчиков (лидары, радары, камера), процессора восприятия и основного контроллера. Процессор восприятия предназначен для распознавание полос движения, профиля дороги, пешеходов, знаков и текста на них, автомобилей. Основной контроллер обрабатывает информацию и принимает решения относительно действий системы на основании результата работы процессора восприятия. Существуют символьные и текстовые дорожные знаки. Распознавание последних используется как для навигации, так и для дополнения карт. Над текстовыми типами знаков работает очень мало исследователей из-за трудностей, одной из которых является отсутствие общедоступных наборов данных. Более того, распознавание текстовых дорожных знаков затрудняют сложный фон, шум, условия освещения, различные шрифты, геометрические искажения знаков, а также погодные условия.

Некоторые методы распознавания текста достигли значительных успехов, однако качество их работы существенно падает при работе со сложными погодными условиями, потому становятся совершенно непригодными в таких обстоятельствах. Так как основной задачей ADAS является поддержка водителей в сложных условиях, необходимо, чтобы система могла быть столь же эффективна, например, в дождь и туман, как в ясную погоду.

Существующих найденных единичных решений данной проблемы недостаточно, потому что они направлены на решение более узких задач. Общее у них – разделение на две подзадачи, а именно: улучшение качества и только затем работа с текстом.

Предположение состоит в том, что решение проблемы состоит в комбинировании готовых решений между собой или с собственными и попытке улучшить результаты.

1. Постановка задачи

Целью является реализация инструмента для распознавания текста на видео в плохих погодных условиях для систем ADAS.

Для её выполнения были поставлены следующие задачи:

1. проанализировать существующие решения;
2. проанализировать возможности трекинга для улучшения качества распознавания текста;
3. изучить и реализовать фильтрацию изображений плохого качества для распознавания текста в плохих погодных условиях;
4. изучить методы ускорения распознавания.

2. Обзор

Готовых подходов к решению поставленной задачи было найдено два: FTDNet [7] и DIP-5 [15]. Общая идея состоит в улучшении изображения и только потом распознавании текста.

2.1. Существующие полные решения

2.1.1. FTDNet

FTDNet [7] — решение, предназначенное для обнаружения текста в условиях тумана. В его основе лежит уже существующий алгоритм DBNet [12] для нахождения текста и оригинальная модель улучшения качества изображения. Данная подсеть состоит из модулей CB (Common Block) и FVE (Feature Visibility Enhancement). Первый отвечает за ввод информации и передачу карт признаков после обработки исходного изображения на следующий этап. Это происходит с помощью остаточной сверточной нейронной сети ResNet-50 [3]. Модуль FVE, в свою очередь, состоит из двух частей. Полученные на предыдущем этапе блоки масштабируются и объединяются, используя набор весов, вычисляемых по определенной формуле. После этого дополнительно используются три сверточных слоя, чтобы уменьшить количество каналов на карте признаков с 256 до 3. Также далее используется пропускное соединение со сверткой 1×1, чтобы избежать чрезмерной потери информации. В конце обработки применяется билинейная интерполяция для увеличения выборки карт признаков до требуемого масштаба, затем используется сверточный слой и готовый объект подставляется в итоговую формулу.

В условиях тумана FTDNet имеет значительное преимущество в чувствительности (чувствительность — отношение верно отобранных объектов ко всем подходящим)[1] по сравнению с другими алгоритмами обнаружения текста. В наборе данных с естественным туманом показатель чувствительности более чем на 2% выше, чем инструмента на

[1] https://en.wikipedia.org/wiki/Precision_and_recall (Дата доступа: 27/12/2023)

втором месте, а в наборе данных с синтетическим туманом он выше более чем на 1,2%. FTDNet также достигает оптимальных результатов по F-мере (F-мера — среднее гармоническое чувствительности и точности)[2]. По сравнению с DBNet, то есть без добавления подсети улучшения видимости, обеспечивается улучшение для естественных данных 3,55% в чувствительности, 2,99% в точности и 3,35% в F-мере, и для синтетических — 3,86%, 4,63% и 4,21% соответственно. Код закрыт.

2.1.2. DIP-5

DIP-5 [15] предназначен для распознания текста в условиях тумана, также в учет берется плохое качество исходных изображений. Последнее действительно является проблемой, так как уже существующие методы нацелены на изображения с высоким разрешением и направлены на восстановление на уровне пикселей. Однако в обрезанных текстовых изображениях отсутствует изначальная информация, что затрудняет удаление дымки с изображения. Более того, распознавание текста — это высокоуровневая задача, которую трудно улучшить путем восстановления на уровне пикселей. Следовательно, готовые методы удаления дымки неприменимы для распознавания текста, поэтому авторы тоже предлагают разбить задачу на две. Эта нейронная сеть состоит из нескольких наборов блоков цифровой обработки изображений (DIP), которые содержат дифференцируемые фильтры, параметры которых оцениваются с помощью параллельного последовательного остаточного блока (PSRB). Вся нейронная сеть может быть обучена сквозным образом с помощью модуля улучшения текста и изображений и модуля распознавания. Код закрыт.

Из того, что оба решения предназначены для устранения дымки, следует закономерный вопрос: будет ли тот же инструмент работать в условиях дождя? Понятно, что физические свойства этих явлений различны, на некоторых ресурсах [1] указано, что видимость во время

[2] https://en.wikipedia.org/wiki/F-score (Дата доступа: 27/12/2023)

дождя ниже, из чего делается вывод, что решение для тумана подходит и для дождя. Однако, в статье, сопровождающей код раздельных алгоритмов для дождя и тумана (NVDeRainNet и NVDeHazeNet [11]) приводится математическое доказательство различий подходов.

2.2. Улучшение изображения

2.2.1. NVDeRainNet

Первый этап NVDeRainNet [11] состоит из начального слоя свертки 5×5 с 16 ядрами, который просто выводит максимальное значение в каждом месте изображения. Созданная карта выходного фильтра из четырех каналов создает узкое место, которое анализирует нелинейное отображение исходного изображения RGB. Затем этот результат передается в банк многомасштабных слоев свертки, состоящий из 16 ядер, каждое из которых содержит фильтры 3×3, 5×5 и 7×7. Максимальное объединение с двумя различными размерами окон 11×11 и 21×21 выполняется для размещения больших дождевых полос, а также нескольких масштабов. Результат объединения слоев и выходные данные предыдущего слоя свертки объединяются перед передачей на другой слой свертки размером 6x6 и состоящий из 128 ядер. Получившийся слой, наконец, передается в слой нелинейного отображения 1×1, который сопоставляет его со слоем из 3 каналов. Этот слой ограничен значениями от -1 до $+1$ и добавляется к исходному изображению для генерации выходных данных первого этапа. Следующий этап повторяет предыдущие операции (пропускается только повтор выбора максимальных значений), чтобы получить окончательное очищенное изображение.

2.2.2. NVDeHazeNet

Сначала в NVDeHazeNet [11] размер ядра первого сверточного слоя устанавливается равным 1×1, чтобы превратить его в ядро нелинейного отображения. Это сопоставляет значение RGB каждого пикселя с

16 различными взвешенными комбинациями. Следующие два слоя аналогичны тому, что используется в модели для избавления от дождя с двумя слоями максимального объединения. Затем добавляются еще два сверточных слоя для получения трехканальной карты передачи. В отличие от DehazeNet, изучается не только карта передачи, но также и значение атмосферной освещенности. Это делается для каждого пикселя с помощью другого сверточного слоя. Таким образом устраняется необходимость вручную вычислять значение глобальной освещенности атмосферы. Наконец, выходное изображение без дымки вычисляется с помощью формулы $J(x) = (I(x) - a(1 - t(x)))/t(x)$.

Возможности использовать эти инструменты нет, так как код закрыт, однако в сравнение им в сопровождающей статье [11] приводятся решения [13], [5] с уступающими результатами по таким показателям, как пиковое отношение сигнала к шуму, среднеквадратичная ошибка, индекс структурного сходства и некоторым другим, но открытым кодом.

2.2.3. DehazeNet

Предлагаемая модель DehazeNet [5] для избавления изображения от тумана состоит из каскадных сверточных и уплотняющих слоев, с соответствующими функциями нелинейной активации, используемыми после некоторых из этих слоев. Слои и функции активации[3] DehazeNet предназначены для реализации четырех последовательных операций для оценки пропускания среды, а именно: выделение признаков (свертка и взятие максимального значения пикселей), многомасштабное отображение, локальный экстремум (снова взятие максимального значения и сжатие) и нелинейная регрессия.

[3]Функции активации — нелинейные функции, которые применяются поэлементно, способствуют извлечению более информативных признаков даже при малом количестве ядер.

2.2.4. Deep Detail Network

Deep Detail Network [13], метод для устранения дождя, тоже основан на сверточной нейронной сети. Непосредственная тренировка сети на первоначальных изображениях представляется недостаточной, так как в значительной мере страдают финальные оттенки цветов, а также пропадает градиент. Первое объясняется тем, что диапазон отображения охватывает все возможные значения пикселей, что затрудняет обучение функции регрессии. Далее следует заметить, что у разности чистого и изначального изображения разнообразие оттенков становится меньше. Таким образом, найденный остаток испольуется в качестве выходных данных слоев параметров. Поскольку дождь имеет тенденцию проявляться на изображениях в виде белых полос, большинство значений разности обладают склонностью быть отрицательными. После ведется сравнение с результатами ResNet [4], недостатком которых является недостаточно точное обнаружение дождя, то есть страдают просто мелкие детали изображения. Лучшие результаты показывает структура ResNet вкупе с вышеописанными негативными остатками. В отличие от исходного подхода ResNet, используется уровень детализации в качестве входных данных для слоев параметров. Так, дождливое изображение моделируется как $X = X_{detail} + X_{base}$. Базовый слой можно получить с помощью низкочастотной фильтрации X, который подавляет флуктуационные шумы и делает изображение более плавным. После вычитания базового слоя из изображения фон удаляется, и в слое детализации остаются только полосы дождя и структуры объектов. В итоге, комбинирование X_{base} и разности чистого и изначального изображений в качестве параметров ResNet приводит к наилучшему результату.

Целью является работа с видео. Так, вышеописанные инструменты также подходят при покадровой обработке, однако при задействовании видео целиком можно использовать большее количество информации.

2.3. Улучшение видео

2.3.1. Rain or Snow Detection in Image Sequences Through Use of a Histogram of Orientation of Streaks

В Rain or Snow Detection in Image Sequences Through Use of a Histogram of Orientation of Streaks [2] для отделения переднего плана от фона в видео как последовательности изображений, используется классическая модель гауссовой смеси. Модель переднего плана служит для обнаружения дождя и снега (гистограмма[4] направления полос осадков), поскольку это погодные явления, вид которых меняется в течение времени. В статье предложены правила отбора, основанные на фотометрии и размере, чтобы выделить потенциальные дождевые полосы. Затем рассчитывается гистограмма направлений полос дождя или снега, оцененная методом геометрических моментов, которая, по предположению, соответствует модели гауссовой однородной смеси. Распределение Гаусса представляет ориентацию дождя или снега, тогда как равномерное распределение представляет ориентацию шума. Для разделения этих двух распределений используется алгоритм максимизации ожидания. После проверки согласия распределение Гаусса сглаживается во времени, и его амплитуда позволяет определить наличие дождя или снега. При обнаружении присутствия дождя или снега гистограмма направления полос осадков позволяет обнаружить пиксели дождя или снега на изображениях переднего плана и оценить интенсивность выпадения дождя или снега.

2.3.2. Detection and Removal of Rain from Videos

В статье Detection and Removal of Rain from Videos [8] сначала для обнаружения предположительных пикселей, подвергшихся воздействию дождя, в каждом кадре видео используются ограничения, полученные с использованием фотометрической модели. В теоретической части статьи было показано, что падение капли вызывает положительную флуктуацию интенсивности длительности кадра. Следовательно, чтобы най-

[4]Гистограмма в данном контексе представляет распределение яркостей изображения.

ти возможные дождевые пиксели во втором кадре, нам нужно рассматривать только интенсивности пикселей в текущем, прошлом и следующем кадре. Если фон остается неподвижным в этих трех кадрах, то интенсивности в прошлом и следующем должны быть равны, а изменение интенсивности из-за капли дождя в n-м кадре должно быть меньше значения, представляющего минимальное изменение интенсивности из-за падения, обнаруживаемого в присутствии шума. В итоге, выбранные пиксели включают почти все пиксели, на которые повлиял дождь. Для исключения ошибочных пикселей для каждой отдельной полосы дождя в кадре проверяется, линейно ли связаны изменения интенсивности вдоль полосы с интенсивностью фона прошлого кадра. Оценивается наклон линейной аппроксимации, и затем полосы, которые не удовлетворяют ограничению линейности или наклоны которых выходят за пределы допустимого диапазона, отклоняются. На следующем этапе еще больше уменьшается количество ложных срабатываний с помощью динамической модели. Ранее в теоретии было показано, что в бинарном поле, созданном дождем, существует сильная временная корреляция между соседними пикселями в направлении дождя. Используя оцененное двоичное поле, вычисляется временная корреляция пикселя с каждым из его соседей в локальной окрестности по набору кадров. Яркие области указывают на сильную корреляцию, то есть далее остальные претенденты отсеиваются.

2.4. Работа с текстом

2.4.1. DBNet

DBNet [12] — инструмент, который является частью выше описанного FTDNet, направлен на обнаружение текста. Особенность решения состоит во вставке операции бинаризации в сеть сегментации для совместной оптимизации. Таким образом, пороговое значение в каждом месте изображения может быть адаптивно предсказано, что позволяет полностью отличать пиксели текста от переднего плана и фона. Однако стандартная функция бинаризации не является дифференцируемой,

вместо этого представлена приближенная функция для бинаризации, называемая дифференцируемой бинаризацией, которая полностью дифференцируема при обучении ее вместе с сетью сегментации.

Как было упомянуто ранее, другие алгоритмы обнаружения текста уступают FTDNet по ряду параметров. В частности в сопровождающей статье в сравнении были упомянуты инструменты PSE [14] и PAN [6]. Особенность первого состоит в ориентации на обнаружение текстов, вписанных в произвольные формы. Далее предлагается алгоритм расширения масштаба, с помощью которого можно идентифицировать соседние экземпляры текста, то есть каждому экземпляру сопоставляется несколько прогнозируемых областей сегментации, которые для простоты обозначаются как «ядра». Каждое ядро имеет форму, аналогичную исходному текстовому экземпляру, но разные масштабы. Второй же состоит из двух частей — модуля улучшения пирамиды признаков (FPM) и модуля объединения признаков (FFM). FPM предназначен для введения многоуровневой информации для улучшения сегментации. FFM же может объединить характеристики, присущие слоям различной глубины, в конечный объект для сегментации. Постобработка реализована с помощью Pixel Aggregation, которая может точно агрегировать текстовые пиксели по прогнозируемым векторам сходства.

3. Эксперимент

3.1. Условия эксперимента

Готовых датасетов с текстом в плохих условиях найти не удалось. В изученных статьях используются как единичные фрагменты фильмов, самостоятельно собранные данные, так и модифицированные изображения из существующих датасетов. Так, в ранее упомянутом инструменте для обнаружения текста в условиях тумана FTDNet используются ICDAR 2015(IC15) [9], 2019 MLT [10] и некоторые другие датасеты с последующим добавлением тумана, а также вручную отобранные подходящие изображения, найденные на различных ресурсах.

3.2. Исследовательские вопросы

- Насколько точно итоговый инструмент справляется с распознаванием текста в плохих условиях?
- Работает ли данный алгоритм быстрее, нежели аналоги?

3.3. Метрики

- Улучшение изображения: MSE (средний квадрат ошибки определения какой-либо величины, является квадратом среднеквадратического отклонения[5]), NRMSE (среднеквадратическое отклонение), PSNR (пиковое отношение сигнала к шуму, наиболее часто используется для измерения уровня искажений при сжатии изображений), SSIM (индекс структурного сходства).
- Обнаружение текста: чувствительность, точность, F-мера.
- Распознавание теста: точность.
- Скорость работы.

[5] https://ru.wikipedia.org/wiki/Среднеквадратическое_отклонение

Заключение

За первый семестр удалось:

- сделать подробный обзор существующих решений;
- рассмотреть инструменты для улучшения качества изображения;
- изучить разницу между подходами для улучшения видео и изображений.

Первым было найдено решение FTDNet. В нем используется описанный DBNet, сверточная нейронная сеть применена для избавления от тумана. Сравнение результатов ведется с инструментами для обнаружения текста в обычных условиях, поэтому превосходство FTDNet по всем метрикам весьма предсказуемо, однако провести более содержательное сравнение невозможно ввиду отсутствия аналогов. Важной деталью является проведение отдельных экспериментов для синтетических и естественных данных, для китайского и английского языков.

Кроме тумана обнаружение и распознавание текста может быть осложнено дождем. Были найдены отдельные алгоритмы для решения этих задач с достаточно подробным объяснением, почему это необходимо. Использовать эти инструменты не представляется возможным, так как код закрыт. Каждому из алгоритмов в сравнение был приведен аналог с открытым кодом. Больший интерес вызывает Deep Detail Network, так как здесь проводится предварительная работа с изображением перед использованием сверточной нейронной сети. В сопровождающей статье к этому инструменту было обращено внимание на то, что работа с видео — более простая задача, нежели с кадрами по отдельности. Действительно, предложенные примеры подходов существенно отличаются от ранее изученнных, но является ли их применение более эффективным, учитывая, что они достаточно стары?

Результаты DIP-5, второго найденного решения, также сравниваются с результатами аналогов, которые и не заявлены, как подходящие для сложных погодных условий. Улучшение по точности распознава-

ния составляет от 2,38% до 9,68% в зависимости от условий проведения эксперимента.

Скорость работы инструмента не упоминалась ни в одной из сопровождающих статей.

Остальные задачи, таким образом, переходят на следующий семестр, а именно:

1. анализ области интереса;
2. анализ возможности трекинга;
3. реализация фильтрации изображений плохого качества для распознавания текста в плохих погодных условиях;
4. изучение методов ускорения распознавания.

Список литературы

[1] Auto Vision News. FAIR WEATHER FRIEND: HOW DO LIDAR SYSTEMS COPE IN RAIN & FOG? — 2020. — URL: https://www.autovision-news.com/adas/lidar-systems-rain-fog/ (дата обращения: 2023-12-13).

[2] Bossu Jérémie, Hautière Nicolas, Tarel Jean-Philippe. Rain or Snow Detection in Image Sequences Through Use of a Histogram of Orientation of Streaks // International Journal of Computer Vision. — 2011. — Vol. 93, no. 3. — P. 348–367. — URL: https://doi.org/10.1007/s11263-011-0421-7.

[3] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep Residual Learning for Image Recognition. — 2015. — 1512.03385.

[4] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep Residual Learning for Image Recognition. — 2015. — 1512.03385.

[5] DehazeNet: An End-to-End System for Single Image Haze Removal / Bolun Cai, Xiangmin Xu, Kui Jia et al. // IEEE Transactions on Image Processing. — 2016. — Vol. 25, no. 11. — P. 5187–5198. — URL: http://dx.doi.org/10.1109/TIP.2016.2598681.

[6] Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network / Wenhai Wang, Enze Xie, Xiaoge Song et al. // CoRR. — 2019. — Vol. abs/1908.05900. — arXiv : 1908.05900.

[7] FTDNet: Joint Semantic Learning for Scene Text Detection in Adverse Weather Conditions / Jiakun Tian, Gang Zhou, Yangxin Liu et al. // Document Analysis and Recognition - ICDAR 2023 / Ed. by Gernot A. Fink, Rajiv Jain, Koichi Kise, Richard Zanibbi. — Cham : Springer Nature Switzerland, 2023. — P. 137–154.

[8] Garg K., Nayar S.K. Detection and removal of rain from videos // Proceedings of the 2004 IEEE Computer Society Conference on Computer

Vision and Pattern Recognition, 2004. CVPR 2004. — IEEE. — URL: http://dx.doi.org/10.1109/CVPR.2004.1315077.

[9] ICDAR 2015 competition on Robust Reading / Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou et al. // 2015 13th International Conference on Document Analysis and Recognition (ICDAR). — IEEE, 2015. — . — URL: http://dx.doi.org/10.1109/ICDAR.2015.7333942.

[10] Nayef Nibal, Patel Yash, Busta Michal et al. ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition – RRC-MLT-2019. — 2019. — 1907.00945.

[11] Mukhtarjee Jashojit, Praveen Kyatham, Madumbu Venugopala. Visual Quality Enhancement Of Images Under Adverse Weather Conditions // 2018 21st International Conference on Intelligent Transportation Systems (ITSC). — 2018. — P. 3059–3066. — URL: https://api.semanticscholar.org/CorpusID:54460186.

[12] Real-time Scene Text Detection with Differentiable Binarization / Minghui Liao, Zhaoyi Wan, Cong Yao et al. // CoRR. — 2019. — Vol. abs/1911.08947. — arXiv : 1911.08947.

[13] Removing Rain from Single Images via a Deep Detail Network / Xueyang Fu, Jiabin Huang, Delu Zeng et al. // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2017. — P. 1715–1723. — URL: https://api.semanticscholar.org/CorpusID:17115407.

[14] Wang Wenhai, Xie Enze, Li Xiang et al. Shape Robust Text Detection with Progressive Scale Expansion Network. — 2019. — 1903.12473.

[15] Text Enhancement: Scene Text Recognition in Hazy Weather / En Deng, Gang Zhou, Jiakun Tian et al. // Document Analysis and Recognition - ICDAR 2023 / Ed. by Gernot A. Fink, Rajiv Jain, Koichi Kise, Richard Zanibbi. — Cham : Springer Nature Switzerland, 2023. — P. 122–136.