Разработка подсистемы автоматизированного анализа читаемости иллюстрированного материала в студенческих работах

Ключевые слова: читаемость изображений, автоматизация оценивания, студенческие работы.

Аннотация

В работе рассмотрена проблема автоматизации оценивания студенческих работ с акцентом на анализ читаемости изображений. Исследованы методы для оценки читаемости изображений в студенческих работах, включая Tesseract OCR, EasyOCR, Laplacian (Blur Detection), BRISQUE и Entropy-based Quality. Проведен сравнительный анализ этих методов по выбранным критериям. По результатам сравнения определены наиболее подходящие методы для интеграции в подсистему автоматизированного анализа читаемости иллюстрированного материала, обеспечивающие комплексный подход к оценке читаемости изображений.

Введение

На сегодняшний день цифровизация играет ключевую роль в повышении эффективности образовательного процесса, однако уровень внедрения цифровых технологий в вузах остается неравномерным [1]. В связи с этим для развития цифровой инфраструктуры возникает необходимость в системах автоматизации учебного процесса, особенно в условиях нарастающей конкуренции как среди отечественных, так и международных образовательных учреждений. Одной из основных подзадач таких систем является автоматизация проверки студенческих работ. Внедрение подсистемы интеллектуального анализа способствует повышению объективности оценки и формированию внутренней виртуальной образовательной среды, что соответствует общим тенденциям цифровизации высшего образования [2].

Целью данной работы является изучение и анализ инструментов для определения читаемости иллюстрированного материала для внедрения в подсистему интеллектуального анализа студенческих работ. Объектом исследования являются инструменты интеллектуального анализа студенческих работ, а предметом — читаемость иллюстрированного материала в студенческих работах.

В ходе проведения исследования были решены задачи, перечисленные ниже.

1. Обзор существующих методов для анализа изображений.
2. Сравнение найденных решений по выбранным критериям.
3. Выбор наиболее подходящего решения по результатам сравнения.

Обзор предметной области

Принцип отбора аналогов

В качестве аналогов рассматривались инструменты для автоматизации анализа качества изображений и распознавания текста. Аналоги отобраны с учетом их популярности, эффективности, совместимости с Python и применимости для оценки читаемости изображений в студенческих работах. Под эффективностью инструмента понимается его способность достигать требуемых результатов с минимальными ошибками и низкими временными затратами при обработке изображений. Инструменты для распознавания текста на изображениях были выбраны в качестве применимых для оценки читаемости, так как плотность текста на изображении и другая информация, полученная в ходе анализа распознанного текста, может быть полезной для определения критериев читаемости изображения. Поиск аналогов осуществлялся с использованием электронной библиотеки КиберЛенинка, ресурсов Google Scholar и GitHub. Для подбора аналогов использовались следующие запросы: “Optical Character Recognition (OCR)”, “Image Quality Assessment (IQA)”, “Laplacian”, “Оценка качества изображений на основе энтропии”.

EasyOCR

EasyOCR — библиотека для оптического распознавания текста (OCR), основанная на методах глубокого обучения, таких как сверточные нейронные сети (CNN) и долговременная краткосрочная память (LSTM). EasyOCR поддерживает более 80 языков и может быть легко интегрирован в различные приложения для анализа текстовой информации [3].

Tesseract OCR

Tesseract OCR — система оптического распознавания текста (OCR) с открытым исходным кодом, поддерживаемая Google [4]. Система использует метод адаптивного распознавания символов, он применяет комбинацию предварительной обработки изображений, классификации символов, а также контекстного анализа текста с учетом языковой модели. Tesseract OCR эффективно работает с традиционными шрифтами и поддерживает более чем 100 языков, а наличие оболочки для Python значительно упрощает интеграцию в приложения [5].

Laplacian (Blur Detection)

Метод использует вычисление дисперсии лапласиана изображения для оценки его резкости [6]. Дисперсия лапласиана позволяет выявить степень размытия изображения, однако результат может быть искажен из-за шума. Этот метод используется для предварительной оценки резкости изображений и реализован в библиотеке OpenCV [7].

BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator)

BRISQUE — мера, используемая для оценки качества изображений без эталонного изображения. Она основана на статистике естественных сцен (Natural Scene Statistics, NSS) и анализирует структурные и текстурные особенности изображения. Естественность сцены в рамках метода BRISQUE подразумевает, что статистические свойства изображения соответствуют характеристикам реальных, ненарушенных

природных или искусственных сцен, таких как сбалансированное распределение яркости, текстур и контраста, без чрезмерных искажений или шумов. Это важно, так как BRISQUE оценивает качество изображения, основываясь на отклонении его статистических характеристик от этих естественных эталонов [8].

Entropy-based Quality

Метод оценки качества на основе энтропии изображения используется для определения степени информативности изображения. Энтропия Шеннона измеряет количество информации, которое можно извлечь из изображения, анализируя распределение значений пикселей, что позволяет оценить его сложность и детализацию. Чем выше качество информации, передаваемой в изображении, тем больше его энтропия [9].

Критерии сравнения аналогов

Направленность анализа

Этот критерий оценивает, на какой аспект качества изображений или распознавания текста ориентирован анализ каждого аналога. Анализ может быть направлен на такие характеристики, как резкость изображения, распознавание текста, а также оценка структуры, текстуры, шума и других факторов. Необходимо, чтобы выбранный инструмент фокусировался на тех аспектах, которые наиболее важны для оценки читаемости изображений в контексте студенческих работ.

Ограничения

Этот критерий описывает возможные ограничения [3, 6, 8, 9, 10] каждого метода для получения максимально точных результатов анализа качества изображений или распознавания текста. Он учитывает, какие ограничения существуют при применении метода к различным типам изображений, включая наличие шумов, особенности шрифтов, размер текста, контраст и другие факторы. Описание ограничений помогает выявить, в каких ситуациях метод может работать неэффективно.

Время обработки

Измеряется время обработки, которое требуется инструменту для выполнения анализа качества изображения или распознавания текста, с учетом характеристик системы. Чем быстрее обрабатываются изображения, тем более эффективно инструмент будет работать при массовом анализе студенческих работ. Для сравнения аналогов время обработки измерялось на системе с процессором AMD Ryzen 5 3550H с тактовой частотой 2.10 GHz, оперативной памятью 16.0 ГБ и 64-разрядной операционной системой. В качестве тестовых изображений использовались изображения из дипломных работ по техническим специальностям прошлых лет. Изображения включали графики, диаграммы и схемы, типичные для технических исследований. Средняя ширина изображений составила 14.28 см, а высота — 8.46 см, что приблизительно соответствует четверти страницы формата А4. Среднее количество пикселей для изображений — 1113879, а средний вес — 314.13 КБ.

Таблица сравнения аналогов

Таблица 1 – Сравнение аналогов по критериям.

Метод	Направленность анализа	Ограничения	Время обработки (11 изображений)
EasyOCR	Распознавание текста.	Эффективность может снижаться при работе с рукописным текстом.	16.97 секунд
Tesseract OCR	Распознавание текста.	Эффективность может снижаться при низком качестве изображения или при наличии сложных шрифтов. Требует предварительной обработки.	11.06 секунд
Laplacian (Blur Detection)	Анализ резкости изображения и выявление степени размытия.	Для повышения эффективности требуется предварительная обработка, а именно уменьшение шума и выполнение сглаживания.	0.45 секунд
BRISQUE	Оценка визуальных искажений (размытие, шум, артефакты сжатия).	Может быть менее точным на изображениях, не соответствующих критериям естественности сцены.	8.71 секунд
Entropy-based Quality	Оценка информативности изображения на основе энтропии Шеннона.	Неэффективен для изображений с высоким уровнем шума или сглаживания, так как эти факторы влияют на коэффициент энтропии.	1.99 секунд

Выводы по итогам сравнения

На основе проведенного анализа и полученных в таблице 1 результатов сформулированы основные выводы.

- EasyOCR и Tesseract OCR оба предназначены для распознавания текста, однако их особенности отличаются. EasyOCR эффективно справляется с текстами различных шрифтов и текстур. В свою очередь, Tesseract OCR больше подходит для четких изображений и эффективно работает с традиционными шрифтами.

- Laplacian (Blur Detection) подходит для первичной оценки четкости изображения, но его эффективность снижается при наличии шума. Этот метод полезен для определения степени размытия, однако для повышения эффективности его работы требуется предварительная обработка изображений.

- BRISQUE — мера, которая дает комплексную оценку качества изображения, учитывая различные искажения, такие как размытие, шум и артефакты сжатия. Однако её эффективность может снижаться, если изображение существенно отличается от “естественных” сцен.

- Entropy-based Quality — метод, оценивающий количество информации, может быть полезен для выявления сложных и детализированных изображений.

Однако на изображениях с высоким уровнем шума или сглаживания он теряет свою эффективность.

Для определения читаемости изображений в студенческих работах требуется комплексный анализ, который учитывает сразу несколько факторов: четкость и детализацию, наличие искажений и шумов, а также количество и содержание распознанного текста. Приведенные аналоги, каждый из которых фокусируется на одном аспекте, не могут обеспечить полноценную оценку читаемости изображений. Несмотря на то что BRISQUE — мера, дающая комплексную оценку качества изображения, она не подходит для данной задачи, так как изображения в дипломах и студенческих работах часто не соответствуют критериям “естественных сцен”, на которых BRISQUE демонстрирует свою максимальную эффективность. Однако, объединив часть инструментов в единую систему, можно обеспечить эффективный подход для комплексной оценки читаемости изображений в студенческих работах.

Выбор метода решения

Решение должно представлять собой подсистему для автоматизированного анализа читаемости иллюстрированного материала в студенческих работах. Данная подсистема в дальнейшем должна быть интегрирована в уже существующую систему проверки студенческих работ, поэтому аналоги отбирались с учетом совместимости с Python. Подсистема должна учитывать как качество изображения (четкость, шумы, искажения), так и свойства распознанного текста (плотность, содержание), обеспечивая объективную оценку. В студенческих работах изображения часто содержат текст, выполненный в стандартных шрифтах, поэтому инструмент для их распознавания должен эффективно работать с такими текстами.

Для оценки качества работы и практической применимости решения необходимо провести несколько экспериментов:

- Сравнение автоматической оценки читаемости изображения с результатами визуального восприятия человеком.
- Оценка влияния интеграции подсистемы анализа читаемости иллюстрированного материала на общую производительность системы проверки студенческих работ.

Описание метода решения

Подсистема автоматизированного анализа читаемости иллюстрированного материала в студенческих работах будет основана на интеграции нескольких существующих инструментов для анализа качества изображений и распознавания текста, что обеспечит комплексный подход. Для распознавания текста выбран инструмент Tesseract OCR, так как он демонстрирует более высокую эффективность по сравнению с EasyOCR при работе с традиционными шрифтами. Для оценки четкости изображений будет применяться метод дисперсии лапласиана. Дополнительно для оценки качества изображения будет использован метод на основе энтропии, который измеряет детализацию и информативность изображения.

Анализ читаемости изображений будет осуществляться следующим образом:

1. Извлечение изображений: Для извлечения изображений используется библиотека python-docx. Из документа извлекаются все изображения, после чего для каждого изображения определяются его размеры в сантиметрах относительно листа формата А4, что позволяет точно учитывать их масштаб на странице. Эта информация будет полезна для дальнейшего анализа.

2. Распознавание текста: С помощью инструмента Tesseract OCR производится распознавание текста на изображениях. Для распознавания используются два языка — русский и английский.

3. Оценка качества изображения: Для каждого изображения рассчитываются значения дисперсии лапласиана и энтропии.

4. Анализ распознанного текста: Выполняется проверка на наличие символов, входящих в заданное множество, а также расчет плотности текста на изображении или его сегменте. Заданное множество — выборка редких символов, наличие которых свидетельствует о некорректном распознавании текста, возникающем из-за низкого качества изображения или его нечитаемости.

5. Оценка читаемости: На основании полученных данных — дисперсии лапласиана, энтропии, плотности текста, количества символов из множества — выполняется итоговая оценка читаемости. Сравнив эти характеристики с заранее установленными пороговыми значениями, система выносит решение о том, является ли изображение читаемым или нет.

Для оценки качества работы и практической применимости решения будет проведено тестирование подсистемы с использованием дипломных работ по техническим специальностям прошлых лет в качестве тестовых данных. Сравнение автоматической оценки читаемости изображений с результатами визуального восприятия человека позволит оценить объективность разработанного решения и при необходимости откорректировать пороговые значения для повышения эффективности оценки читаемости.

Производительность подсистемы будет проверена с помощью нагрузочных тестов, имитирующих реальные условия эксплуатации. В ходе тестирования будут анализироваться следующие показатели:

- Среднее время обработки одного изображения, включая извлечение изображения, оценку его качества, распознавание текста и вынесение итоговой оценки читаемости.
- Среднее время обработки всех изображений в одной дипломной работе, что позволит оценить влияние подсистемы на общее время проверки.

Результаты тестирования производительности позволят определить, насколько эффективно подсистема выполняет свои функции, оценить её влияние на работу всей системы проверки студенческих работ, а также выявить возможные проблемные места, требующие дальнейшей оптимизации.

Заключение

В рамках исследования был проведен обзор существующих методов для анализа изображений. В ходе работы были отобраны пять аналогов и определены три критерия для их сравнения: направленность анализа, ограничения и время обработки. В ходе сравнения аналогов было выявлено, что существующие методы имеют слишком узкую направленность для комплексного анализа читаемости изображений. Кроме того, ограничения некоторых методов не позволят эффективно использовать их для автоматизации проверки студенческих работ.

По результатам сравнения было принято решение — объединить Laplacian (Blur Detection), Entropy-based Quality и Tesseract OCR в подсистеме автоматизированного анализа читаемости иллюстрированного материала в студенческих работах для обеспечения комплексного анализа читаемости изображений.

Направление дальнейших исследований включает реализацию описанной подсистемы, внедрение подсистемы в систему интеллектуального анализа студенческих работ, поиск подходящих пороговых значений для числовых характеристик, а также формирование множества редких символов.

Список использованных источников

1. Аксенова Н. И., Черняков М. К., Усачева О. В. Рейтинговая оценка состояния цифровизации вузов //Образование и наука. – 2024. – Т. 26. – №. 7. – С. 88-115.
2. Родионова Е. В. Цифровизация высшего образования Российской Федерации: тренды и перспективы //Научные труды Вольного экономического общества России. – 2023. – Т. 243. – №. 5. – С. 64-84.
3. JaidedAI. EasyOCR. – [Электронный ресурс]. – URL: https://github.com/JaidedAI/EasyOCR (дата обращения: 02.12.2024).
4. Google. Tesseract OCR. – [Электронный ресурс]. – URL: https://github.com/tesseract-ocr/tesseract (дата обращения: 02.12.2024).
5. Бобров К. А., Шульман В. Д., Власов К. П. Анализ технологий распознавания текста из изображения //Международный журнал гуманитарных и естественных наук. – 2022. – №. 3-2. – С. 124-128.
6. Замула А. И., Горячкин Б. С. МНОГОКРИТЕРИАЛЬНАЯ ОЦЕНКА КАЧЕСТВА ФОТОГРАФИЙ //E-Scio. – 2020. – №. 7 (46). – С. 332-346.
7. OpenCV Documentation. Laplacian Blur Detection. – [Электронный ресурс]. – URL: https://docs.opencv.org (дата обращения: 02.12.2024).
8. Mittal A., Moorthy A. K., Bovik A. C. No-reference image quality assessment in the spatial domain //IEEE Transactions on image processing. – 2012. – Т. 21. – №. 12. – С. 4695-4708.
9. Александров А. А., Мизюков Г. С., Бутакова М. А. ОЦЕНКА КАЧЕСТВА СЛИЯНИЯ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ ЭНТРОПИИ ШЕННОНА И КОЭФФИЦИЕНТА ПОЛЕЗНОЙ ИНФОРМАЦИИ ХАРТЛИ //Известия ЮФУ. Технические науки. – 2024. – №. 5.
10. Tesseract OCR Documentation. Improve Quality. – [Электронный ресурс]. – URL: https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html (дата обращения: 12.12.2024).