УДК 004.41

ОСОБЕННОСТИ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ СРЕДНЕГО БАЛЛА АТТЕСТАТА АБИТУРИЕНТА ПРИ ОБРАБОТКЕ ИЗОБРАЖЕНИЙ АТТЕСТАТОВ

Ладогубец Т.С.,
Литвиненко П.Л., к.т.н.,
Сегол Р.И., к.н. по соц. ком.,
Финогенов А.Д., к.т.н.
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»(Україна)

В работе на примере анализа более 30 тысяч изображений приложений к аттестату, обработанных КПИ им. Игоря Сикорского в 2018 году, выделены основные проблемы, которые возникают при обработке изображений аттестатов абитуриентов во время вступительной кампании.

Различия в типах загружаемых изображений, количестве и типах представленных на них документов, размерах, качестве снимков, размещении и т.д. не позволяют напрямую использовать методы распознавания текста для выделения оценок и расчета среднего балла.

Частично проблема может быть решена введением на предварительном этапе классификатора, который определит тип представленного документа и определит дальнейшие действия по обработке изображения. Например, довольно распространенным является изображение, на котором размещены оба разворота приложения к аттестату. При наличии всего двух вариантов размещения: сверху титульный разворот страницы, а снизу внутренний или наоборот – дает возможность обрезать часть изображения, и проводить анализ внутреннего разворота. Также распространенной ошибкой является загрузка абитуриентом изображения другого документа: самого аттестата, собственной фотографии, сертификата внешнего независимого оценивания т.д. Различия между подобными изображениями и собственно внутренним разворотом приложения к аттестату столь значительны, что дают возможность отсеять данные документы на уровне классификатора.

Существенной проблемой при распознавании являются различия в разрешении изображения, т.к. наиболее популярные методы машинного обучения используют поточечный анализ.

К сложностям также приводит и огромное количество фотоснимков документа, а не их сканированных копий. В этом случае к рассматриваемым проблемам добавляются наличие фона, центрирование документа, обрезка, тени.

Т.к. анализ текста обычно выполняется на черно-белых изображениях или на изображениях с градациями серого, то контрастность документа также требует дополнительной предобработки.

В работе приведены типовые примеры изображений документов и сделаны выводы о возможности автоматического определения среднего балла на основании сканированных копий.

Ключевые слова: обработка изображений, машинное обучение, классификация документов

Постановка проблемы. Во время вступительных кампаний последних лет одной из составляющих рейтинга абитуриента является средний бал документа о полном среднем образовании (аттестат). Удельные вес среднего балла аттестата, поступающего определяется Условиями приема вы высшие учебные заведения Украины в год поступления и правилами приема каждого высшего ученого заведения в частности (в период с 2016 по 2019 годы – до 10% от максимального конкурсного балла 200). Ввиду электронной подачи документов поступающими на основе полного среднего образования на очную и заочную формы, информация о среднем бале аттестата вносится поступающим в электронную систему и подтверждается Приемными комиссиями на основании сканированной копии, которую подает абитуриент через личный электронный кабинет. С 2018 года средние учебные заведения не должны в обязательном порядке рассчитывать и вписывать средний балл в приложение к аттестату о полном среднем образовании. Расчет полностью производится поступающим, но ответственно за правильность лежит на высшем учебном заведении. В связи с ограниченным сроком подачи документов (в 2019 году – 13 дней) и трудоемкостью процедуры подсчета, актуальным является вопрос о возможности автоматического расчета среднего балла аттестата.

Анализ последних исследований и публикаций. В соответствии с «Условиями приема на обучение в высшие учебные заведения Украины в 2018 году» [1] сроки подачи заявлений абитуриентами, поступающими на основе полного общего среднего образования были ограничены датами 12.07.2018 - 26.07.2018. По результатам вступительной кампании от поступающих на 1-й курс в 7 ВУЗов было подано ≈20 000 заявлений и больше [2]:

- Киевский национальный университет имени Тараса Шевченко (37 994 заявления);
- Львовский национальный университет имени Ивана Франко (35764);
- Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского» (33324);
- Национальный университет «Львовская политехника» (28900);
- Киевский национальный торгово-экономический университет (28126);
- Национальный авиационный университет (23939);
- Харьковский национальный университет имени Василия Каразина (19982).

Определение среднего балла аттестата абитуриента как составляющей его общего рейтинга является одной из трудоемких операций при обработке заявления, что непосредственно влияет на конечное распределение абитуриента. Большое количество ошибок при определении среднего бала аттестата на уровне школ (чаще всего связанных с игнорированием балла итогов государственной аттестации (экзаменов)) привела к перекладыванию основной работы по определению среднего балла на высшие учебные заведения. Это требует необходимости определения для крупных высших учебных заведений порядка 1400-2700 средних баллов аттестатов в сутки.

Более того, при выявлении ошибки высшее учебное заведение может внести изменение в средний балл аттестата, поступающего только в том случае, если это учебное заведение первым обработало электронное заявление в Единой государственной электронной базе по вопросам образования. В случае, если первым регистрацию провело другое высшее учебное заведение, возникает проблема согласования данных, которая решается исключительно в телефонном режиме или же с помощью электронной почты, что, в свою очередь, значительно усложняет процедуру проверки данных по одному и тому же абитуриенту и выведение общей рейтинговой картины

Формулировка целей статьи. В данной статье проведен анализ изображений аттестатов на примере сканированных копий, поданных в КПИ имени Игоря Сикорского в процессе вступительной кампании 2018 года и сделаны выводы о возможности автоматического определения среднего бала аттестата.

Основная часть. Хотя приложение к аттестату является унифицированным документом, но при анализе изображений был определен набор проблем, которые затрудняют возможность его обработки, а именно:

1) Ряд абитуриентов загружают не приложение к аттестату, а отличные от него документы: собственное фото, сам аттестат или другие документы (рис. 1).

Рис.1. Примеры некорректных документов

2) Фоновое обрамление – многие из загруженных изображений получены путем не сканирования документа, а фотографирования. Кроме самого приложения аттестата часто в кадр попадает поверхность, на который расположен документ, а иногда части тела самого поступающего (рис. 2).

Рис. 2. Наличие фона

3) Расположение нескольких документов на одном изображении – часто вместе с приложением к аттестату (обратной стороны, на

которой расположена информация о ФИО, школе и оценках) на изображении присутствует и лицевая сторона и/или сам пластиковый аттестат (рис. 3).

Рис. 3. Изображения с несколькими документами

4) Расположение на картинке – встречаются как представленные в книжной, так и в альбомной ориентации. В случае, когда изображения получены с помощью фотоаппарата или мобильного телефона, документы могут быть расположены под разными углами и иметь выпуклость в месте перегиба. Кроме того есть изображения, которые расположены от «правильного» положения под углами ± 90˚ или 180˚. Отдельные приложения имели на оригинале загибы или механические повреждения (рис. 4).

Рис. 4. Размещение на листе

5) Обрезка изображения – большое количество сканированных изображений представлена на листе формата А4 в натуральную величину без обрезки. То есть само приложение к аттестату может быть расположено в верхней или нижней половине листа, в середине или ближе к одному из краев, когда ориентация (альбомная для приложения) совпадает с ориентацией листа сканирования. Кроме того, некоторые изображения приложений были слишком обрезаны, что не позволяло утверждать, что представлены все записи.

6) Наличие перекрывающих элементов, например, некоторые абитуриенты прикрепляли изображение приложения к аттестату после нотариального заверения (апостиля) и на изображении присутствует красная лента, которая скрепляет документ и перекрывает часть текста (рис. 5).

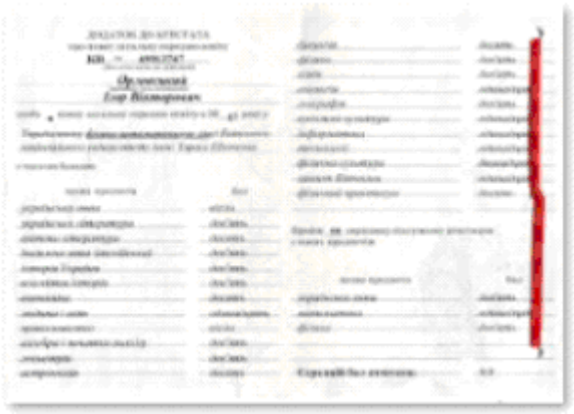

Рис. 5. Наличие перекрывающих элементов

7) Цвет – в отличие от сканированных копий, где цвет приложения к аттестату близок к циану, в зависимости от освещения съемки, встречаются изображения с цветами близкими к желтому или красному. Кроме того, значительная часть изображений предоставлена в черно-белом цвете (рис. 6).

Рис. 6. Оттенки изображений

8) Контрастность изображений – значительная часть изображений (чаще всего фотографий) представлены в «темном» виде.

9) Тени – при съемке на большом количестве фотографий имеются тени или блики фотовспышки (рис. 7).

Рис. 7. Наличие на изображениях теней

10) Качество изображений – в кабинете поступающего присутствует ограничение на размер загружаемого файла (менее 1 Мб), но отсутствуют ограничения и контроль размеров изображений. Среди анализируемых изображений размеры колеблются от 144x100 до 10200x6650 пикселей (рис. 8).

Рис. 8. Отличия в качестве изображений

11) Тип файла – хотя все приложения расположены в Единой государственной электронной базе по вопросам образования загружаются в формате JPG, но некоторые поступающие размещают файлы в формате PDF, заменив расширение файла, например, чтобы

разместить в одном файле на нескольких страницах обе стороны приложения аттестата.

Таким образом, автоматическое определение среднего балла аттестата абитуриента невозможно без предварительной обработки и классификации документов, что усложняет и замедляет процесс электронной обработки заявлений, добавляя человеческий фактор (подсчет производится вручную с изображений) и увеличивая процент ошибок. В зависимости от определенных проблем или их совокупности, последовательность действий для автоматической обработки изображения и дальнейшего его распознавания может быть крайне трудной алгоритмической задачей.

Выводы. Перед решением задачи автоматического определения среднего балла аттестата необходимо разработать классификатор, на основе анализа которого уменьшить количество «проблемных» изображений и определит порядок действий для их обработки.

Литература

1. Умови прийому на навчання до вищих навчальних закладів України у 2018 році. URL: https://zakon5.rada.gov.ua/laws/show/z1397-17/page#n17. (дата звернення: 12.02.2019).
2. Статистика ВНЗ по количеству поданных заявлений. URL: https://grade.ua/uk/news/interesnaya-statistika-vstupitelnoj-kampanii-2018/. (дата звернення : 05.04.2019).

ОСОБЛИВОСТІ АВТОМАТИЧНОГО ВИЗНАЧЕННЯ СЕРЕДНЬОГО БАЛУ АТЕСТАТУ АБІТУРІЄНТА ПРИ ОБРАБЦІ ЗОБРАЖЕНЬ АТЕСТАТІВ

Ладогубець Т.С., Литвиненко П.Л., Сегол Р.І., Фіногенов О.Д.

У роботі на прикладі аналізу більше 30 тисяч зображень додатків до атестата, поданих в КПІ ім. Ігоря Сікорського в 2018 році, виділені основні проблеми, які виникають при обробці зображень атестатів абітурієнтів під час вступної кампанії.

Відмінності в типах зображень, що завантажуються, кількості і типах представлених на них документів, розмірах, якості знімків, розміщенні і т.д. не дозволяють безпосередньо використовувати методи розпізнавання тексту для виділення оцінок і розрахунку середнього балу.

Частково проблема може бути вирішена введенням на попередньому етапі класифікатора, який визначить тип

представленого документа і визначить подальші дії по обробці зображення. Наприклад, досить поширеним є зображення, на якому розміщені обидва розвороту додатки до атестата. При наявності всього двох варіантів розміщення: зверху титульний розворот сторінки, а знизу внутрішній або навпаки - дає можливість обрізати частину зображення, і проводити аналіз внутрішнього розвороту. Також поширеною помилкою є завантаження абітурієнтом зображення іншого документа: самого атестата, власної фотографії, сертифіката зовнішнього незалежного оцінювання тощо Відмінності між подібними зображеннями і власне внутрішнім розворотом додатку до атестата настільки значні, що дають можливість відсіяти дані документи на рівні класифікатора.

Суттєвою проблемою при розпізнаванні є відмінності в розмірах зображень, тому що найбільш популярні методи машинного навчання використовують піксельний аналіз.

До складнощів також призводить і величезна кількість фотознімків документа, а не їх сканованих копій. В цьому випадку до досліджуваних проблем додаються наявність фону, центрування документа, обрізка, тіні.

Оскільки аналіз тексту зазвичай виконується на чорно-білих зображеннях або на зображеннях з градаціями сірого, то контрастність документа також вимагає додаткової попередньої обробки.

В роботі наведені типові приклади зображень документів і зроблені висновки про можливість автоматичного визначення середнього балу на підставі сканованих копій.

Ключові слова: обробка зображень, машинне навчання, класифікація документів.

SPECIAL ASPECTS IN THE AUTOMATIC DETERMINATION OF THE AVERAGE SCORE OF THE APPLICANT'S SECONDARY EDUCATION'S CERTIFICATE

Ladogubets T., Lytvynenko P., Segol R., Finogenov A.

The article includes the example analysis of over 30 thousand of annexes to the secondary education's certificate images, submitted during enrolment process to the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" in 2018, highlighted the main

problems arise when processing annexes to the secondary education's certificate images during the application campaign.

Differences in downloaded images types, documents presented number and types, size, image quality, placement, etc. do not allow direct use of text recognition methods to highlight ratings and calculate the average score.

Partially, the problem can be solved by introducing at the preliminary stage a classifier that will determine the submitted document's type and further actions on image processing. For example, it is quite common to have an image in which both spreads of the annex are placed. If there are only two options for placement: a title spread on the top of the page, and an inner one on the bottom or vice versa, it makes it possible to trim part of the image and analyze the internal spread. It is also a common mistake for an applicant to upload another document's image: the certificate itself, the applicant's own photo, an external independent assessment's certificate, etc. The differences between similar images and the actual internal spread to the secondary education's certificate are so significant that they make it possible to separate out these documents at the classifier level.

A significant problem with recognition is the differences in image resolution since the machine learning most popular methods use a point-by-point analysis.

The huge number of documents photos and not their scanned copies also leads to the difficulties. In this case, the background presence, the document's centering, cropping, and shadows are added to the problems considered.

Because text analysis is usually performed on black and white images or on images with gray-scale, then the document's contrast also requires additional pre-processing.

The paper presents typical examples of document images and made conclusions on the automatic determination possibility of the average score based on scanned copies.

Key words: image processing, machine learning, document classification.