Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«**СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**»

Гуманитарный институт
Кафедра информационных технологий
в креативных и культурных индустриях

УТВЕРЖДАЮ
И. о. заведующего кафедрой

_____ М. А. Лаптева

«_____» _____ 2023 г.

БАКАЛАВРСКАЯ РАБОТА

Компьютерное зрение для разделения изображений и текста в журнале «Курьер ЮНЕСКО»

Направление подготовки: 09.03.03 Прикладная информатика

Наименование программы: 09.03.03.30 Прикладная информатика

Руководитель канд. культурологии, доц., ст. науч. сотр. И. А. Кижнер

Выпускник А. С. Дяченко

Нормоконтролер И. Р. Нигматуллин

Красноярск 2023

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Долгое время компьютерный анализ текста являлся наиболее актуальным и важным направлением исследований в цифровых гуманитарных науках. Ученые добились значительных результатов в компьютерном анализе текста. Появились такие технологии как OCR (оптическое распознавание текста), тематическое моделирование, поиск именованных сущностей и другие. Тем не менее ученые, концентрируясь на тексте, упускают огромное значение визуальных форм репрезентации. За последние 10 лет быстрое развитие технологий компьютерного зрения позволило цифровым гуманитариям заниматься изучением изображений. Стали использоваться нейронные сети для выявления тенденций в больших коллекциях изображений [1], что позволило говорить о дальнем чтении для визуальных форм.

Однако, изучение текстов и изображений в исторических документах, таких как газета, журнал, книга, не может быть изучено отдельно, т.е. как только визуальное или только текстовое представление. Французский семиотик Ролан Барт уже в 1961 году заметил, что фотография не существует в изоляции, она всегда связана по меньшей мере с одной другой структурой, а именно - с текстом [2]. Поэтому следующий этап цифровых гуманитарных исследований – это совместное изучение текстов и изображений.

Для исследований, где будет изучаться текст и изображения совместно, требуется подготовка: извлечение текстовой и визуальной части объекта исследования. В данной работе извлекается визуальная и текстовая часть журнала «Курьер ЮНЕСКО».

«Курьер ЮНЕСКО» - это журнал, выпускаемый с 1948 г. по наши дни, созданный Организацией Объединенных Наций для образования, науки и культуры (ЮНЕСКО), основной задачей которого является информирование общественности о важных событиях в области образования, науки, культуры и коммуникаций. Исследование материалов журнала, особенно сочетание изучения как визуальной части журнала, так и текстовой, дает возможность получить целостную картину, как и предполагает классическая гуманитарная

традиция. Однако возникает вопрос, как получить текст и изображения из журналов так, чтобы была возможность изучать их совместно, а также быть уверенными в том, что в данных нет пропусков? Получение текста не является само по себе сложной задачей, но можем ли мы получить изображения быстрым способом и сразу перейти к анализу?

В этой работе мы утверждаем, что извлечение изображений из сложных исторических документов, таких как журнал «Курьер ЮНЕСКО», неуниверсальная и трудная задача, требующая создания собственного метода, настроенного конкретно под объект исследования.

Выпускная квалификационная работа посвящена разделению текста и изображений на примере журнала «Курьер ЮНЕСКО».

Достоверность результатов, полученных в ходе научного исследования, обусловлена большим набором данных, точностью создания выборки, сочетанием разных способов извлечения изображений, сравнением результатов с результатами работы других алгоритмов.

Проблема исследовательской работы заключалась в том, что не существует универсального способа извлечения изображений из исторических документов.

Объект исследования – цифровые копии журналов «Курьер ЮНЕСКО», выпущенные в 1960-х и 1990-х годах.

Предметом исследования являются способы извлечения изображений и текстов из журналов «Курьер ЮНЕСКО», выпущенных в 1960-х и 1990-х годах.

Цель настоящей работы – извлечь изображения и текст из журналов «Курьер ЮНЕСКО» для создания наборов данных, с помощью которых можно проводить мультимодальный анализ.

Задачи исследования:

— проанализировать стилистические особенности журналов «Курьер ЮНЕСКО», выпущенных в 1960-х и в 1990-х гг.;

— создать алгоритм по извлечению изображений из журналов «Курьер ЮНЕСКО» с точностью (accuracy) не менее 90%;

— создать алгоритм по извлечению 100% текста из журналов «Курьер ЮНЕСКО»;

— подготовить текст и изображения к совместному изучению: в названии извлеченных изображений отображать страницу, с которой они были взяты;

— сравнить результаты созданного алгоритма по извлечению изображений с другими методами.

Научная новизна настоящего исследования заключается в создании нового подхода к извлечению изображений из исторических документов. Новый подход состоит в создании алгоритма с использованием технологий компьютерного зрения и нейронный сетей.

Выпускная квалификационная работа состоит из введения, четырёх разделов, заключения, списка использованной литературы и приложения.

Первый раздел «Источник данных: изображения и тексты в журнале "Курьер ЮНЕСКО"» описывает историю создания и развития ЮНЕСКО, журнала «Курьер ЮНЕСКО», а также повествует о многоплановости изображений и мультимодальном анализе.

Второй раздел «Анализ и извлечение изображений компьютерными способами» описывает опыт исследователей в цифровых гуманитарных науках в извлечении изображений из исторических документов.

Третий раздел «Методология процесса извлечения изображений и текста из журналов «Курьер ЮНЕСКО» и последующая их группировка» описывает процесс создания алгоритма для извлечения изображений и текстовых блоков с помощью языка программирования Python, а также процесс их последующей группировки для мультимодального анализа.

Четвертый раздел «Результаты исследования» включает в себя сравнение созданного алгоритма с другими существующими способами извлечения изображений. Раздел повествует о том, как компьютерное зрение справляется со сложными случаями дизайнерских решений в журналах «Курьер ЮНЕСКО».

Кроме того, в разделе приводится описание того, как полученные результаты соотносятся с предыдущими исследованиями в данной области.

Заключение подводит итоги исследования. Список литературы даёт библиографическое описание цитируемых в работе источников. В приложении представлены таблицы для расчета F-score для созданного алгоритма.

1. Источник данных: изображения и тексты в журнале «Курьер ЮНЕСКО»

В то время как задача разделения изображений и текстов для типовых случаев может считаться решенной, экспериментальные дизайнерские решения журналов второй половины 20 века все еще представляют сложности. Выбор журнала определялся необычными решениями для распределения цвета и дизайна страницы, соединения текстов и изображений в одном пространстве и последующими задачами совместного анализа изображений и текстов для определения расхождений, диалогических ситуаций или, наоборот, точного совпадения содержания изображений и текстов. Последнюю задачу особенно интересно выполнить, используя статьи и изображения из журнала, задачей которого было представление разнообразного набора ценностей и обнаружение соответствия этих ценностей общечеловеческим представлениям о перспективах развития объединенных наций, важной задаче послевоенного времени.

1.1 История создания журнала «Курьер ЮНЕСКО»

Организация Объединённых Наций по вопросам образования, науки и культуры (ЮНЕСКО), созданная под эгидой ООН, начала свое существование 16 ноября 1945 года. Идеалы организации сформировались после осознания ужасов войны, поэтому неудивительно, что самой главной их целью было поддержание мира. Создатели ЮНЕСКО считали, что культурная и научная неосведомленность, особенно касающаяся расового разнообразия людей, — это ключевая причина конфликтов и войн человечества. Поэтому ЮНЕСКО стремилось к равенству и миру путем «свободного обмена идеями и знаниями (...) в целях взаимопонимания и более истинного и совершенного знания о жизни друг друга» [3]. Таким образом, создатели ЮНЕСКО верили, что широкое понимание культуры и культурного разнообразия может изменить мышление людей и предотвратить войны. Впоследствии сформировались идеалы ЮНЕСКО, которые организация продвигает и по сей день: просвещение, демократия, равенство.

Как ЮНЕСКО продвигала свои ценности? Сначала (1945-1970гг) организация была информационным центром, где мировые лидеры, деятели искусства и науки планировали просвещать общество с помощью трансляции общепризнанных классических произведений искусства. Тогда члены организации придерживались мнения, что это путь к совершенному и образованному обществу. Такой путь размышления понятен и в наше время: пусть люди увидят все самое лучшее, созданное человечеством, и будут жить в мире и спокойствии, вдохновленные прекрасным. Для достижения своей цели члены организации собирались вместе в главном штабе ЮНЕСКО и формировали глобальный культурный канон, который и будет транслироваться в общество. Впоследствии члены организации на форумах по международному культурному обмену в странах-участниках ЮНЕСКО просвещали население по сформированной ими культурной программе. ЮНЕСКО хотела передать все «лучшее» из мировой культуры людям, но «лучшее» было одинаковым для всех народов мира. Хотя смысл такого подхода понятен, ранний взгляд организации на культуру очевидно показывает, что большую роль в формировании глобального культурного канона играли элиты, выбиравшие произведения искусства для трансляции в общество [4].

Однако ситуация, когда группа людей формирует единственно верный культурный канон для всего человечества, стала подвергаться критике в 1960-1980 годы. Это время, когда количество стран-участников ЮНЕСКО стало увеличиваться (с 59 в 1950 г. до 153 в 1980 г.), и все большее распространение получает мнение о том, что нужно транслировать многообразие человеческой культуры. В результате организация перешла от идеи создания глобального культурного канона к идее о разнообразии культур, т.е. от «универсальных культурных объектов» перешла к «уникальным культурным практикам» [5].

В 1990-е и на протяжении 2000-х годов культурное разнообразие и межкультурный диалог становятся главными приоритетами ЮНЕСКО [6]. Позиция ЮНЕСКО в области культуры изменилась. Озабоченность общим культурным наследием человечества (произведениями искусства, памятниками,

книгами) сменилось стремлением видеть культуру повсюду [7]. Теперь в организации считали, что культура должна позволять людям развиваться и жить в гармонии, а признание и поощрение культурного разнообразия приведет человечество к равенству и миру [8]. В отличии от своей изначальной идеи о создании глобального культурного канона с привлечением элит мира, новая идеология ЮНЕСКО рассматривает культуру как целостный образ жизни, который интересен сам по себе.

В то же время, несмотря на транслирование идеи о разнообразии культур, можно заметить, что не все культурные идеологии рассматриваются ЮНЕСКО как «правильные». «Правильные» культуры должны обладать ценностями, которые продвигает организация: терпимость, диалог, открытость, демократия. От закрытых и консервативных культур ЮНЕСКО предпочитает дистанцироваться. Но, если следовать такой логике, замечает Бьярк Нельсон в статье, которая предлагает анализ политики ЮНЕСКО, закрытые культурные сообщества виновны в том, что не соответствуют другим «правильным культурам» [7]. Большинство специалистов ЮНЕСКО знают, что они "следуют иллюзии" [9], сужая мир до «правильных» культур. Как заметил Бьярк Нельсон в интервью с сотрудниками ЮНЕСКО, они не могут ничего с этим поделать, ведь люди, занимающие руководящие должности, имеют право отклонять предложения специалистов, которые не соответствуют стандартам ЮНЕСКО [7]. Стремясь воплотить безусловно благородные и амбициозные цели по сохранению мира и гармонии в обществе, ЮНЕСКО закрывает глаза на «неправильные культуры», что на самом деле сужает культурный плюрализм, к которому организация стремится и который продвигает.

Практически с самого основания в ЮНЕСКО появляются различные проекты, распространяющие идеалы и цели организации. В том числе появляются периодические журналы, такие как Unesco Bulletin for Libraries (1947-1978), Museum (1948- 1992), Journal of World History (1953-1972), The Unesco Courier (1948-) и другие. Журнал «Курьер ЮНЕСКО» – самый популярный и важный журнал ЮНЕСКО. Этот журнал всегда был первым по

количеству и разнообразию читателей, а сейчас это единственное периодическое издание, которое все еще выпускает ЮНЕСКО. Для ЮНЕСКО «Курьер» всегда был главной платформой для продвижения идеалов организации и трибуной для международных дебатов [10]. Поэтому, изучая «Курьер ЮНЕСКО», мы изучаем идеалы всей организации.

История журнала «Курьер ЮНЕСКО» начинается с работы молодого журналиста Сэнди Коффлера, принятого в организацию ЮНЕСКО в качестве нового специалиста. До работы в ЮНЕСКО, в военные годы, Коффлер был в рядах армии США и работал в подразделении по ведению психологической войны (Psychological Warfare Branch). Для этого он проходил подготовку в правительственном Бюро по вопросам военной информации (Office of War Information) и изучал современные методы пропаганды для распространения пацифистских идей. Его направили в Италию, где Коффлер создаст свои первые «курьеры». Курьеры – ряд газет в Италии, которые оповещали население о продвижении союзников и об установлении мира. Названия этих газет варьировались в зависимости от названия освобожденного города или региона: «Римский курьер», «Вениционский курьер», «Курьер Венето», «Курьер Эмилии» и.т.д.

После окончания войны Коффлер отправится в свою любимую Францию, где в юные годы он проходил обучение, но которую ему пришлось покинуть в военное время из-за еврейского происхождения. В это время в интеллектуальных кругах Франции много говорят о новой международной организации ЮНЕСКО. Молодой журналист Сэнди Коффлер был впечатлен и вдохновлен ценностями организации, и не мог не обратить внимание на ЮНЕСКО. Коффлера с охотой приняли в штаб ЮНЕСКО 26 октября 1947 года.

Сэнди Коффлер был вдохновлен своей новой работой и уже 19 ноября представил Гарольду Каплану, первому директору Бюро информации, проект газеты с подробным описанием редакционной политики, периодичности, основных рубрик, с указанием числа колонок и даже шрифта. Вот что Коффлер

писал об этом в своем дневнике: «Деятельность ЮНЕСКО очень разнообразна, ее программа охватывает множество крайне важных тем в области образования, науки и культуры, и я уверен, что собрать достаточное количество интересных и актуальных статей не составит труда». Так зарождалось главное международное издание ЮНЕСКО.

В газете «Курьер ЮНЕСКО» Сэнди Коффлер предлагает печатать обзоры иностранной прессы, интервью с деятелями науки и культуры, выдающимися сотрудниками ЮНЕСКО, а также статьи экспертов со всего мира, посвященные анализу важных общемировых проблем и вопросов. Сэнди Коффлер говорил: «Курьер ЮНЕСКО» — это «окно в мир», через которое читатели могут «открыть новые горизонты» [11]. Для перевода на другие языки, Коффлер планировал приглашать редакторов из других стран, чтобы у всех читателей, независимо от их места проживания, был одинаково качественный материал в газете. Но с 1948 по 1953 годы Коффлер и ЮНЕСКО только искали свой формат, и в это время выпускались только описания мероприятий ЮНЕСКО (см. таблицу 1).

Сэнди Коффлер был уверен в своих идеях и успехе журнала «Курьер ЮНЕСКО». Он понимал, что в журнале «Курьер ЮНЕСКО» должно быть что-то большее, чем описания мероприятий организации. Благодаря уверенности Коффлера первоначальный газетный формат 1948 года был переработан в 1954 году: в журнале появились цветные иллюстрации и фото, расширился список тем. Сравнение тем за 1948-1953 гг. с темами за 1954 года журналов «Курьер ЮНЕСКО» представлено в таблице 1.

Таблица 1 – Сравнение десяти заголовков журналов «Курьер ЮНЕСКО» за 1948-1953гг и десяти заголовков за 1954 г.

Заголовки журналов «Курьер ЮНЕСКО» за 1948-1953гг	**Заголовки журналов «Курьер ЮНЕСКО» за 1954г**
Конференция ЮНЕСКО провела специальную сессию	Обещание атомной энергии
3000 делегатов соберутся в этом месяце на региональной конференции ЮНЕСКО	Редкие шедевры мирового искусства
Бейрут готовится принять конференцию ЮНЕСКО	Заключенные — это люди; право на гуманное обращение
Какие культурные обязанности у государства: тема форума ЮНЕСКО на сентябрь	Последние рубежи цивилизации

Международный театральный институт создан в Праге; Париж назван первой штаб-квартирой	S.O.S. из прошлого; сохранение нашего наследия в камне
Специальный выпуск: кино	Американский н(егр)

Продолжение таблицы 1

Культура живет за счет обмена, а не за счет принудительных займов: урок Абоссоло Симиона, африканского ремесленника и художника	Молодежь в новой Японии; насколько изменились молодые люди?
40 000 детей-беженцев с Ближнего Востока заявили о своем праве на образование благодаря школам ЮНЕСКО в пустыне	Свобода информации: жертва 20-го века?
Организация Объединенных Наций в действии; международный эксперимент по оказанию технической помощи	Гаити: 150 лет независимости
Южная Италия ведет борьбу с неграмотностью	Языки: мост или барьер?

Сэнди Коффлер не ошибся, и новый журнал обрел популярность. Тираж вырос с 40 тысяч экземпляров в 1949 году до 500 тысяч в 1980-х годах. Спустя 40 лет «Курьер ЮНЕСКО» уже был доступен на 35 языках, включая русский (с 1957 г.). Сегодня «Курьер ЮНЕСКО» издается на шести официальных языках Организации (английском, французском, испанском, арабском, русском и китайском), а также на португальском, эсперанто и сардинском. В 2006 году была запущена онлайн-версия журнала, а также был создан архив всех выпусков на английском языке. Несмотря на то, что журнал сильно изменился за годы издания как по содержанию, так и по форме, его миссия осталась та же: продвигать идеалы ЮНЕСКО, поддерживать платформу для диалога между культурами и обеспечивать форум для международных дискуссий [10].

На момент исследования в цифровой библиотеке ЮНЕСКО находится более чем 700 выпусков журналов «Курьер ЮНЕСКО» [12]. На сайте есть интерфейс навигации по журналам (рисунки 1,2).

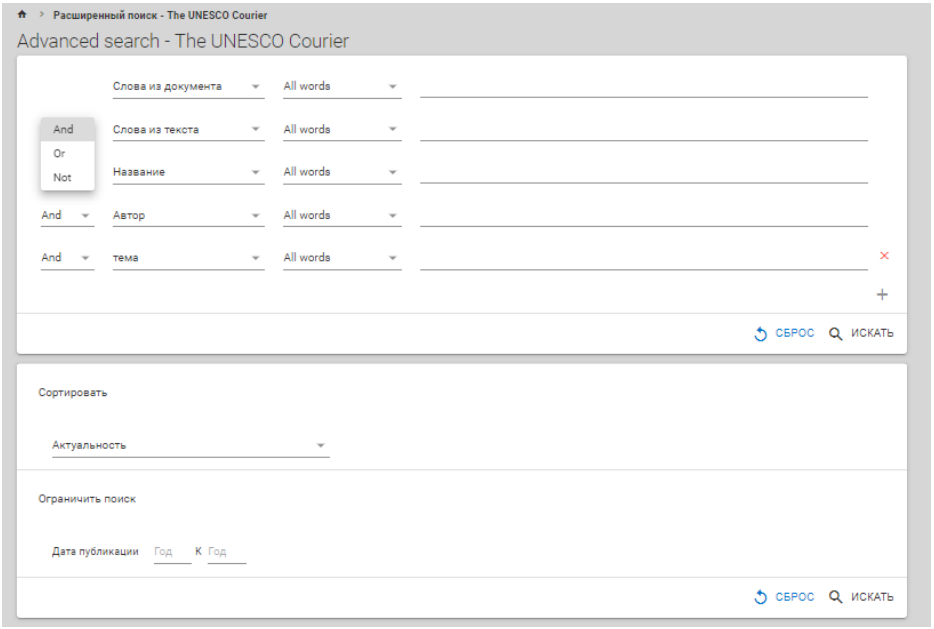


Рисунок 1 – Расширенный поиск по журналам «Курьер ЮНЕСКО»

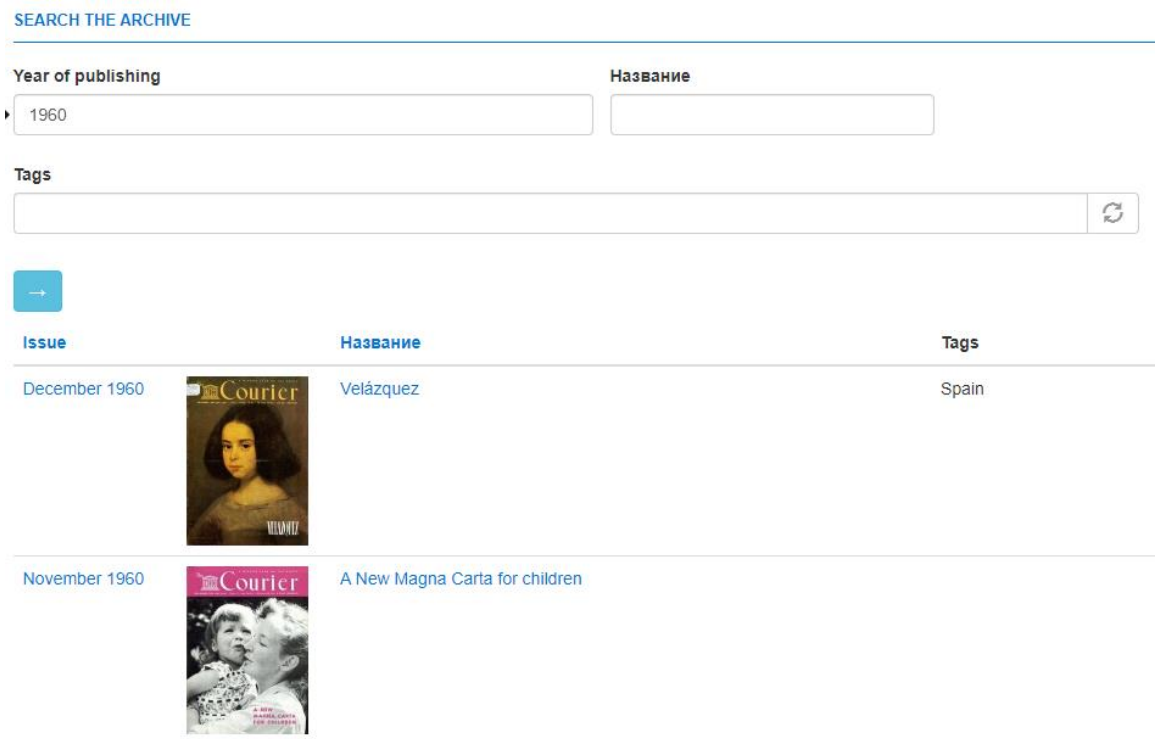


Рисунок 2 – Интерфейс простого поиска по журналам «Курьер ЮНЕСКО»

1.2 Многоплановость изображений в «Курьере ЮНЕСКО»

Уже в 1977 году философ Роланд Барт вводит понятие «риторика изображения (image = образ)» [13]. Отличительной чертой этой риторики является то, что она включает в себя по крайней мере два уровня языка: собственный (обозначаемый) и образный (коннотируемый), тем самым доказывая, что изображение содержит не только буквальный смысл, а включает

в себя еще и дополнительные сложные образы. В своей работе Роланд Барт анализирует, как изображения передают смысл через свою структуру, композицию, контекст и отношения между элементами. Анализировать образы из изображений, по мнению Роланда Барта, означает использовать науку о знаках – семиотику. Так, в работе [13] Роланд Барт анализирует рекламу товаров от фирмы «Panzani» для французской аудитории. В фотографии Роланд находит три смысловых слоя, два из которых анализируются семиотически (рисунок 3).

Образ	Три сообщения	Значение
	Лингвистическое сообщение	"Panzani" звучит по-итальянски
	Закодированное сообщение в фото	1) Сетчатая сумка для продуктов ассоциируется с возвращением с рынка, где продаются натуральные свежие продукты. Поход на рынок ассоциируется с неиндустриальным обществом, где между покупателем и производителем нет посредника. 2) Цвета на фото: желтый, зеленый, красный, ассоциируются с Италией. 3) Из всех продуктов можно приготовить блюдо. Среди набора продуктов есть консервы, но рядом лежат свежие овощи, что показывает нам натуральность этих консервов. 4) Эстетика в расположении продуктов - напоминает натюрморт (культурный фон)
	Незакодированное сообщение (изображение рассматривается буквально)	Сообщение без кода (все элементы, которые мы видим на изображении, и есть послание)

Рисунок 3 – Анализ рекламного фото продуктов «Panzani» для французских покупателей

С помощью закодированных сообщений в рекламной фотографии, покупатель получает представление о высоком качестве представленных продуктов и их происхождении. Так Роланд Барт продемонстрировал важность изучения смыслов изображений.

Далее, в 1981 году известный философ-постмодернист Жан Бодрийяр в книге "Simulacra and Simulation" [14], также рассуждает о изображениях. Он показывает, как искажается реальность, и как такие фото используются для манипуляции общественным мнением. Таким образом, Роланд Барт и Жан Бодрийяр признают важность изображений в культуре и исследуют, как они используются для передачи смысла и воздействия на зрителя.

В статье [15] изучается, как повторение и распространение изображений в фотокнигах и туристических путеводителях по Израилю создают образ и национальный брендинг страны. В работе прослеживается, как с помощью повторяющихся изображений формируется нужное представление об Израиле. Например, образ юной пионерки в шляпе тембель (израильский национальный символ) присутствует в журналах в начале становления государственности Израиля, позднее девушка изображается уже без шляпы, а в выпусках фотокниг, ближе к нашему времени (2018 г), снова ее «надевает». Таким образом, изучение того, как на промежутке времени меняются или повторяются изображения в важных документах, может рассказать исследователям о новых аспектах истории или общества, а также подтвердить или опровергнуть какие-либо гипотезы.

1.3 Мультимодальный анализ: исследование взаимодействия изображений и текстов

До этого мы говорили об анализе изображений отдельно от текста. Однако, существуют и исследования об анализе текста и изображений совместно. Так в 1996 г. Дэвид Флеминг в работе «Могут ли картинки быть аргументами?» [16] рассуждает о самостоятельности изображений как таковых, автор обсуждает, можем ли мы считать изображения аргументами без текстового сопровождения. Флеминг приходит к выводу, что изображения являются поддержкой для текста и они не могут быть полностью независимыми. Также Флеминг говорит о том, что изображение должно быть переведено в лингвистическое сообщение, и в этом случае «либо визуальное не имеет значения (поскольку теперь оно дублируется языком), либо вербальное является такой редукцией визуального, что представляет собой совершенно новую мысль» [16]. Так Флеминг предлагает сравнивать текстовое описание изображения как такового и текста, который сопровождает это изображение, что может привести нас к выводу о значении изображений в изучаемом тексте, их взаимодействии.

Рассуждает о взаимодействия текстов и иллюстраций из литературы 19 века Джулия Томас в книге «Nineteenth-century illustration and the digital: studies in word and image» (2017 г) [17]. В работе рассматривается влияние перехода книги в цифровую среду на ее изучение практик чтения и на литературоведческе исследования. Ключевым вопросом является диалог между словом и иллюстрацией. Автор замечает, что иллюстрация в книгах 19 века — это не просто интерпретация текста в изображении, а «тонкое соотношение взаимодополняемости и конфликта, сходств и различий». Эта книга прекрасный пример того, как изучение текстов и изображений в книге без предвзятости (например, что иллюстрация — это всегда «бледное отражение текста») может дать исследователю интересные результаты.

Все перечисленные ранее работы доказывают важность изучения изображений как отдельно, так и в совокупности с текстами. Однако, в данном разделе мы имели дело с гуманитарными работами, где используется близкое чтение изображений. Далее в разделе 2 речь пойдет о том, как используя классическую гуманитарную традицию цифровые гуманитарии стали изучать визуальный контент с помощью компьютерных технологий, т.е. перешли к дальнему чтению изображений.

2. Анализ и извлечение изображений компьютерными способами

За последние 10 лет быстрое развитие компьютерного зрения позволило цифровым гуманитариям изучать миллионы изображений как из оцифрованных документов, так и из цифровых коллекций, созданных в цифровую эпоху. Исследования в цифровых гуманитарных науках перестали концентрироваться только на тексте, как это было до 2010 года [18], а стали обращать внимание на визуальную часть изучаемого материала, опираясь на многочисленные классические гуманитарные труды о важности изучения изображений (см. раздел 1.2). Например, в исследовании [1] используются нейронные сети для анализа исторических изображений из голландских газет, благодаря чему ученым удалось создать классификацию газетных изображений, тем самым показав, что CNN можно использовать для изучения и анализа больших коллекций визуальных данных без необходимости просматривать такие архивы вручную. В работе [19] предлагается вариант дальнего чтения больших визуальных корпусов и утверждается, что такие подходы необходимы, ведь нужно уравновесить сильную текстовую ориентацию в цифровых гуманитарных науках, которая исключает множество неязыковых явлений, традиционно представляющих интерес для гуманитарных наук. Таким образом, текстовые исследования все еще превалируют, но тенденция изучать изображения растет с каждым годом.

Однако, исследования, где описывается процесс предварительного извлечения изображений из стилистически сложных по дизайну документов (газеты, журналы и.т.д.), проводились гораздо реже. Это, как нам кажется, связано с тем, что нет универсального способа извлечения визуальной части из исторических документов, и гораздо удобнее взять уже готовые изображения. Но, если исследователь будет работать с журналами, газетами или с другими сложными документами, сначала нужно получить обрезанные по своим границам изображения, что может представлять сложность, ведь в таких документах текст и изображения связаны (например, текст накладывается на изображения или изображение используется как подложка для текста). Каждое

изображение является смысловой единицей для исследователя, а анализируя всю страницу целиком, без извлечения, мы рискуем получить обобщенный и неверный результат (например, когда на одной странице находится сразу несколько изображений). Поэтому перед анализом изображений из сложных исторических документов исследователи предпочитают сначала их извлечь. Так, в работе [20] для извлечения изображений из американских газет привлекались добровольцы, которые вручную выделяли весь визуальный материал на 3437 страницах газет (на них содержалось 32 424 изображения), далее эти данные использовались для обучения нейронной сети, которая, в свою очередь, правильно вырезала изображения из газет.

Однако, как говорилось в разделе «О чем могут рассказать изображения?», изображения и текст важно изучать вместе. Как цифровые гуманитарии, используя технические знания, будут изучать текст и изображения совместно? В работе [21] авторы поднимают тему важности мультимодальных исследований в сфере цифровых гуманитарных наук. Мультимодальные исследования в контексте цифровых гуманитарных наук – это изучение изображений и текстов из одного исследуемого объекта (например, из книги, или из серии книг) вместе, в совокупности, как и описано в классических гуманитарных трудах (см. раздел 1.2). Исследователи описывают как с помощью нейронной сети для продвинутого описания изображений CLIP [22] прийти к анализу набора изображений, не аннотируя вручную большое количество данных. CLIP создала исследовательская лаборатория OpenAI для изучения визуальных материалов. В качестве обучающей выборки используется 400 миллионов комбинаций изображения и текста [23], что позволяет CLIP изучать немаркированные данные и прогнозировать подписи к ним, т.е. давать текстовое описание изображений. Также, CLIP’у можно дать «подсказки» по набору данных, и система сама найдет в нем изображения по тексту-подсказке. Например, по запросу «изображение семьи» система сама найдет по немаркированным данным нужные изображения (рисунок 1).

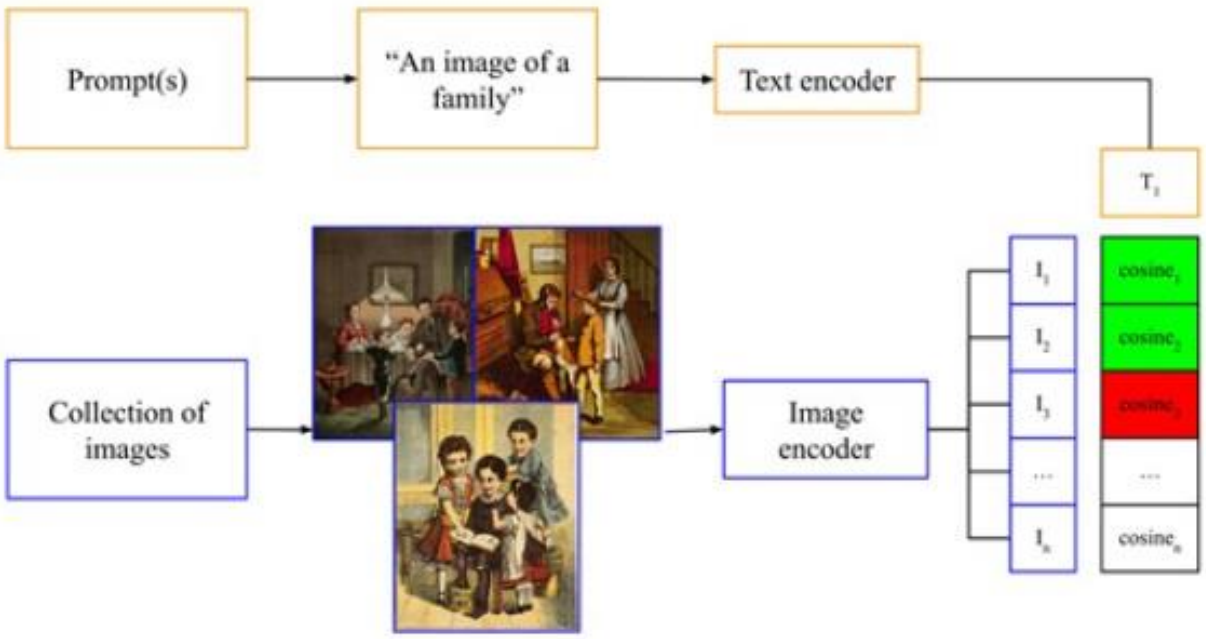


Рисунок 4 – Схема работы нейронной сети для продвинутого описания изображений CLIP, представленная в работе [21].

Это исследование предлагает решение большой проблемы анализа исторических изображений: перевод их в текстовый (вербальный) формат с помощью CLIP. Однако, авторами не был описан способ сравнения текстового и визуального содержания данных (например, сравнение текста из книги с вербальным представлением картинок из этой же книги), а также не описан способ извлечения изображений из документа для дальнейшей их загрузки в CLIP.

Таким образом, в работах [1, 15, 21] а также в других исследованиях, где изучаются изображения, либо не описан процесс извлечения картинок, либо берется готовый набор данных. В работах [1], [21] подразумевается извлечение изображений, но описание данного процесса, а также процент успешного извлечения не указан. Напротив, в работе [20] авторы приводят подробное описание извлечения изображений, и мы видим, насколько это было трудозатратно (привлечение добровольцев и ручная пометка 3437 страниц). Поэтому в нашем исследовании поднимается тема извлечения изображений для дальнейшего их анализа из небольшого по объему журнала, данных которого не хватит для обучения нейронной сети, как было в работе [20].

Поскольку в данной работе подготавливаются материалы для изучения визуальной и текстовой части журнала «Курьер ЮНЕСКО» совместно, мы также постараемся найти способ автоматической группировки текста и картинок из журнала, т.е. сделать систему, которая знает какой текст и изображения появляются на одной и той же странице, чтобы в будущем прийти к мультимодальности, описанной в работе [21].

3. Методология процесса извлечения изображений и текста из журналов «Курьер ЮНЕСКО» и последующая их группировка

3.1 Создание датасетов и выбор библиотек для анализа

Исследование проводилось на журналах «Курьер ЮНЕСКО», вышедших в 60-е и 90-е годы. Эти два десятилетия были выбраны для того, чтобы проверить созданный программный код извлечения картинок на разных по дизайну журналах. В нашу выборку вошли: 109 журналов за 60-е и 110 журналов за 90-е годы.

Журналы 60-х годов отличаются практически полным черно-белым форматом выпуска и своими уникальными стилистическими особенностями оформления (рисунок 5). В журналах 90-х есть как черно-белый формат выпуска, так и полностью цветной, а также совершенно новое оформление (рисунок 6). Стилистические особенности этих двух десятилетий представлены в таблице 2.

Рисунок 5 – Пример оформления страниц в журналах 60-х годов

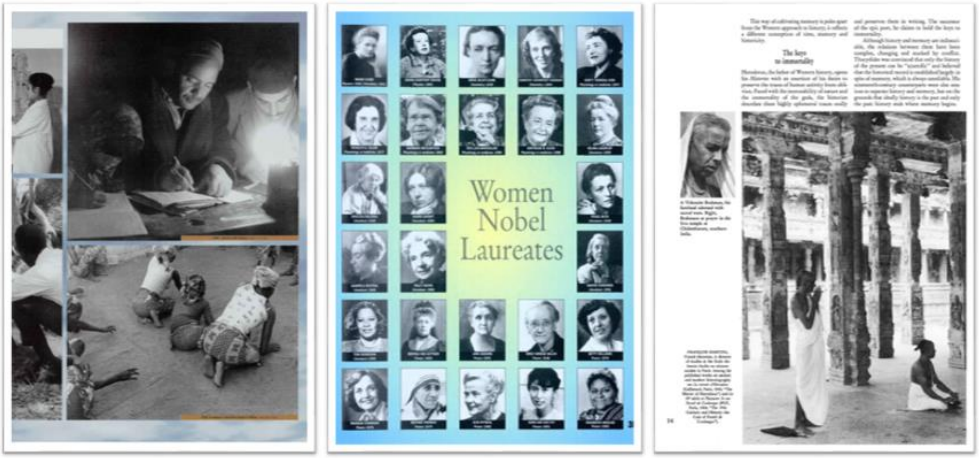


Рисунок 6 – Пример оформления страниц в журналах 90-х годов

Таблица 2 – Сравнение стилистических особенностей журналов 60-х и 90-х годов

Стилистические особенности журналов 60-х годов	**Стилистические особенности журналов 90-х**
Черно-белый формат выпуска	Цветной формат выпуска, но иногда есть черно-белые страницы или фото
Использование черных и серых подложек под текст/изображения	Использование однотонных цветных или градиентных цветных подложек под текст/изображения
Преобладание черно-белых фотографий	Преобладание цветных фотографий

Для извлечения картинок из журналов «Курьер ЮНЕСКО» был использован язык программирования Python версии 3.10, а также библиотеки PyPDF2, PyMuPDF, Fast.ai, OpenCV и PIL.

Библиотеки PyPDF2 и PyMuPDF предназначены для работы с файлами в PDF-формате. Они включают в себя множество функций, такие как разделение, объединение, обрезка и другие преобразования страниц в формате PDF. Эти библиотеки нам нужны, т.к. все журналы «Курьер ЮНЕСКО» представлены в PDF-формате. Этих библиотек было недостаточно, т.к. для документов, которые изначально не были созданы на компьютере, а являются отсканированными историческими документами, простая функция извлечения картинок из PDF-файла, встроенная в эти две библиотеки, работать не будет.

OpenCV – это библиотека алгоритмов компьютерного зрения, обработки изображений и численных алгоритмов общего назначения с открытым кодом [24]. Эта библиотека требуется для того, чтобы искать границы изображений на страницах журналов.

Fast.ai — это библиотека глубокого обучения, которая предоставляет как высокоуровневые компоненты, которые могут быстро и легко обеспечивать современные результаты в стандартных областях глубокого обучения, так и низкоуровневые компоненты, которые можно смешивать и сопоставлять, строить новые подходы [25]. Использование данной библиотеки в исследовании обусловлено тем, что в журнале «Курьер ЮНЕСКО» присутствуют сложные для компьютерного зрения дизайнерские решения (см. раздел 4.3), для которых только OpenCV будет недостаточно.

PIL (Python Imaging Library) — библиотека языка программирования Python, предназначенная для различных преобразований с растровой графикой. Используется в исследовании для работы с уже вырезанными картинками.

Для извлечения картинок из журналов «Курьер ЮНЕСКО» был создан алгоритм, который можно описать в несколько этапов. На рисунке 7 представлена схема работы всего алгоритма.

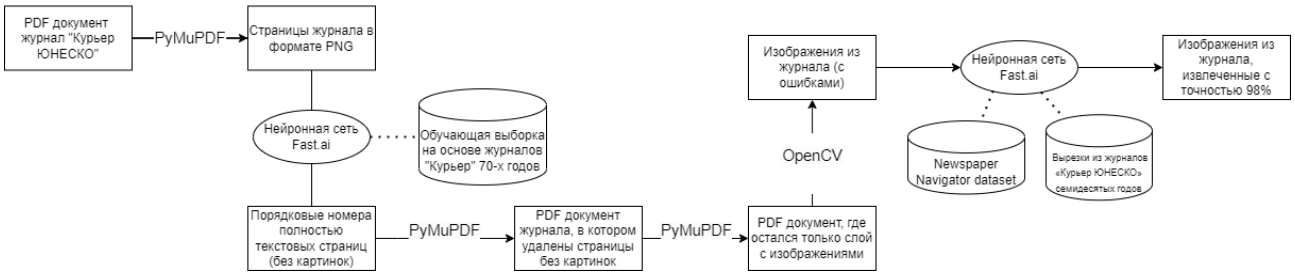


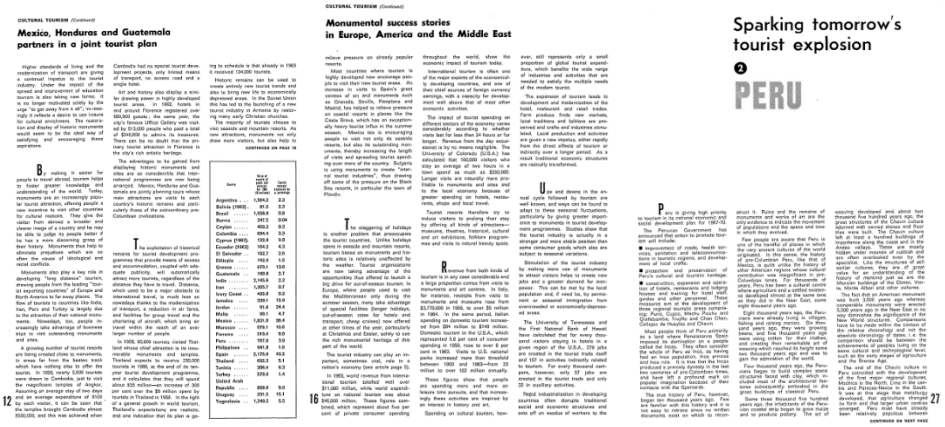
Рисунок 7 – Схема работы созданного алгоритма по извлечению картинок из журналов «Курьер ЮНЕСКО»

3.2 Этапы создания алгоритма

На первом этапе в программу загружается журнал в формате PDF. С помощью библиотек PyPDF2 и PyMuPDF изначальный файл преобразуется в набор картинок в формате PNG. Перевод страниц журнала в картинки обусловлен тем, что на втором этапе программы используется нейронная сеть, обученная на картинках в формате PNG. Результат первого этапа показан на рисунке 8.

Рисунок 8 – Преобразование pdf-файла журнала в картинки формата PNG

На втором этапе требуется удалить полностью текстовые страницы без картинок. Это поможет в дальнейшем избежать ошибок при извлечении изображений. Пример полностью текстовых страниц представлен на рисунке 9, а ошибки, которых удастся избежать, представлены на рисунке 10. Нейронная сеть для извлечения полностью текстовых страниц создана с помощью библиотеки глубокого обучения Fast.ai. Способ разделения картинок на полностью текстовые и картинки с иллюстрациями описан в уроке Computer Vision for the Humanities [26].


CULTURAL TOURISM (Continued)

Mexico, Honduras and Guatemala partners in a joint tourist plan

Monumental success stories in Europe, America and the Middle East

Sparking tomorrow's tourist explosion

PERU

Рисунок 9 – Пример полностью текстовых страниц из журнала «Курьер ЮНЕСКО»

Рисунок 10 – Примеры ошибок, которых удастся избежать благодаря использованию нейронной сети.

На рисунке 10 мы видим пример ошибки, которая могли бы оказаться в итоговом результате извлечения картинок. Благодаря использованию Fast.ai практически все (98%) текстовые страницы удалятся (включая элементы, обозначенные на рисунке 10), что позволит не загружать программу на дальнейших этапах работы большим количеством неверных результатов обрезки, а также исключит попадание таких ошибок в итоговый результат извлечения. Для обучения нейронной сети Fast.ai была создана обучающая выборка, в которую входили 1000 страниц из журналов семидесятых годов (рисунок 11). Семидесятые годы были взяты для того, чтобы исключить пересечения обучающей выборки с исследовательской выборкой шестидесятых

и девяностых годов. Таким образом при проверке нейронной сети мы были уверены, что результат ее работы не является простым «запоминанием» такой же страницы из обучающей выборки, а она может определить по совершенно «новой» для нее странице правильный результат. В обучающей выборке мы ручным способом пометили полностью текстовые страницы как «-», а страницы, где помимо текста встречается изображение, «+».

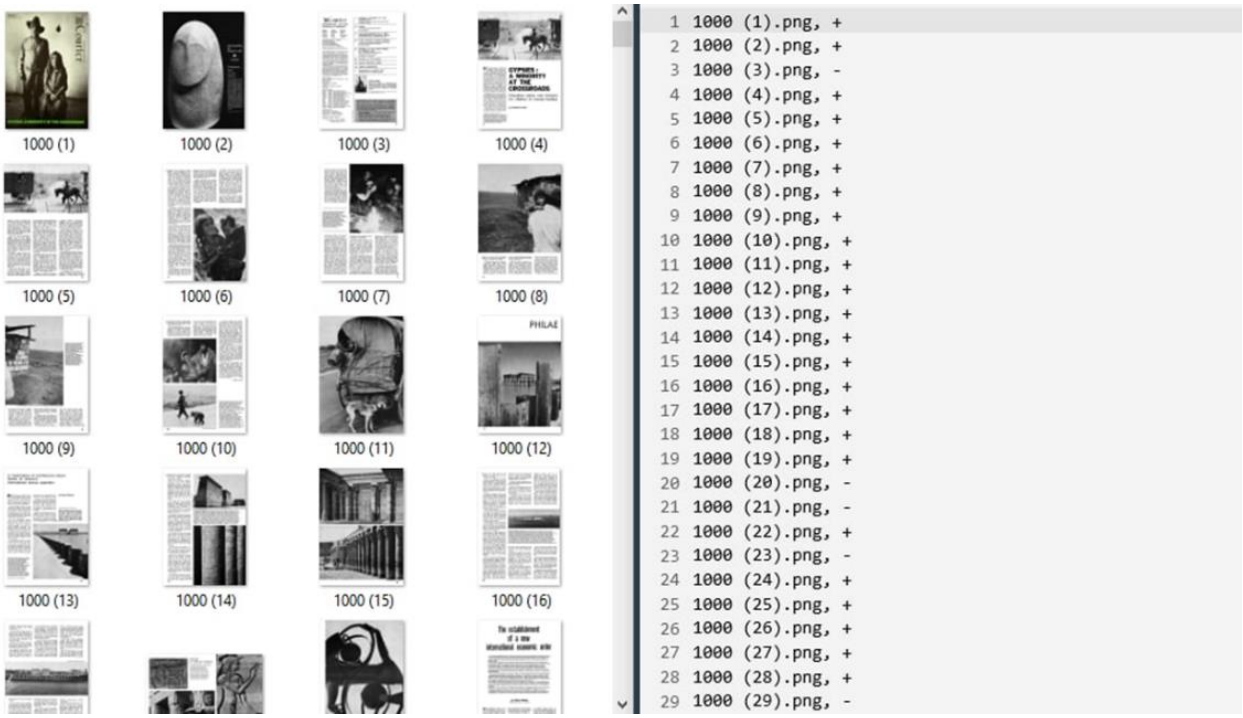


Рисунок 11 – Фрагмент обучающей выборки

Таким образом, по завершению второго этапа алгоритма, мы узнали номера страниц, где есть только текст, и удалили их в изначальном PDF-файле журнала. Для удаления страниц в PDF-файле использовалась библиотека PyMuPDF.

На третьем этапе была проведена работа с журналами, где удалены все текстовые страницы. У библиотеки PyMuPDF есть функция отделения картинок из PDF-документов. Но в результате мы получим не готовые картинки, обрезанные по их фактическим границам, а страницы журнала, где удален текстовый слой. Текстовый слой удаляется не везде (у приблизительно 72% страниц одного журнала). К тому же мы получаем дополнительно ошибочный технический (скрытый) скан-слой (рисунок 13). Результат работы данной библиотеки представлен на рисунке 12.

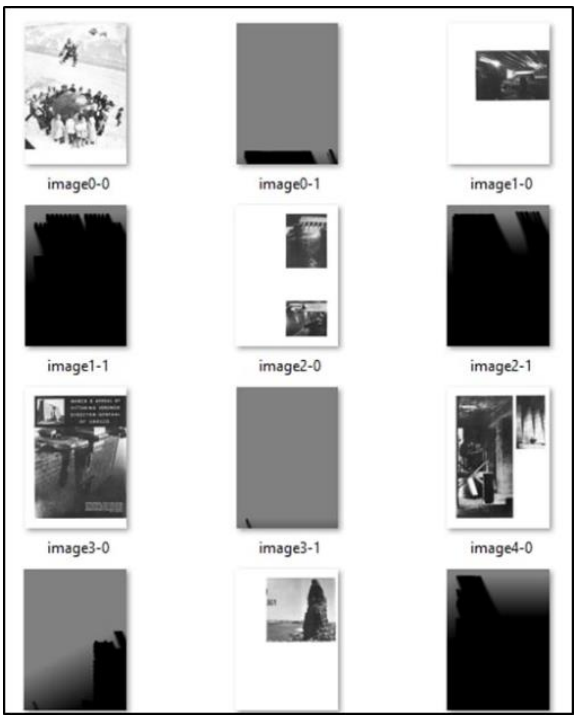


Рисунок 12 – Результат отделения слоя с изображениями

Как мы видим, программа успешно отделила слой с картинками (у 72% страниц одного журнала). Помимо слоя с картинками, появляется ошибочно выделенный слой (рисунок 13).

Рисунок 13 – Ошибочно выделенный слой

Все ошибочные страницы помечены в названии файла «-1», их мы удалили программным способом с помощью стандартных функций Python.

Таким образом, к концу третьего этапа мы получаем журнал, где у 72% страниц текстовый слой удален. Это поможет избежать ошибок на следующем этапе: поиск границ и извлечение изображений. Ошибки, которых удастся избежать, аналогичны тем, что представлены на рисунке 10.

На четвертом этапе работы программы происходит извлечение картинок с помощью библиотеки компьютерного зрения для анализа и обработки изображений OpenCV. PDF-журнал мы переводим в набор PNG-картинок с помощью PyMuPDF, т.к. библиотека OpenCV работает с изображениями. Функция поиска контуров данной библиотеки пытается найти потенциальные границы изображений, обрезать их и сохранить. В результате получено огромное множество картинок (100 и больше) большинство из которых обрезаны неправильно и являются очень маленькими изображениями. Функция сохраняет все варианты обрезки изображения, в том числе и правильные. Для того, чтобы удалить маленькие неправильные варианты обрезки, используется модуль image библиотеки PIL. Так мы удалим изображения длина или ширина которых меньше 150 пикселей. Размер 150 пикселей был выбран из-за того, что по результат проверки 10% журналов из 60-х и 90-х годов, всего 1% картинок, длина и ширина которых меньше 150 пикселей, являются важными для извлечения. Остальные представляют собой логотипы и рекламу, которые не представляют для нас исследовательского интереса. Таким образом, установив условие по извлечению картинок «удалять изображения длина или ширина которых меньше 150 пикселей», мы избавимся от неправильных маленьких обрезков и не интересующих нас картинок. Результат четвертого этапа представлен на рисунке 14.

Рисунок 14 – Результат работы после применения OpenCV и PIL

Мы видим, что в папке все еще есть фрагменты текста, которые выделены с помощью серой подложки (рисунок 15). Картинки с текстом на серых,

черных, цветных подложках составляют 11% в журналах 60-х годов и 8% в журналах 90-х годов. Такие картинки являются текстовой частью журнала и не представляют для нас исследовательского интереса.

Рисунок 15 – Фрагменты текста на подложке

Для решения возникшей проблемы в программе есть пятый этап, где мы еще раз используем нейронную сеть, созданную с помощью Fast.ai. Нам требуется обучить нейронную сеть находить полностью текстовые фрагменты на цветной подложке (рисунок 16). Помимо этого, в журнале часто встречаются картинки вместе с текстом (рисунок 17), их нейронная сеть должна оставить.

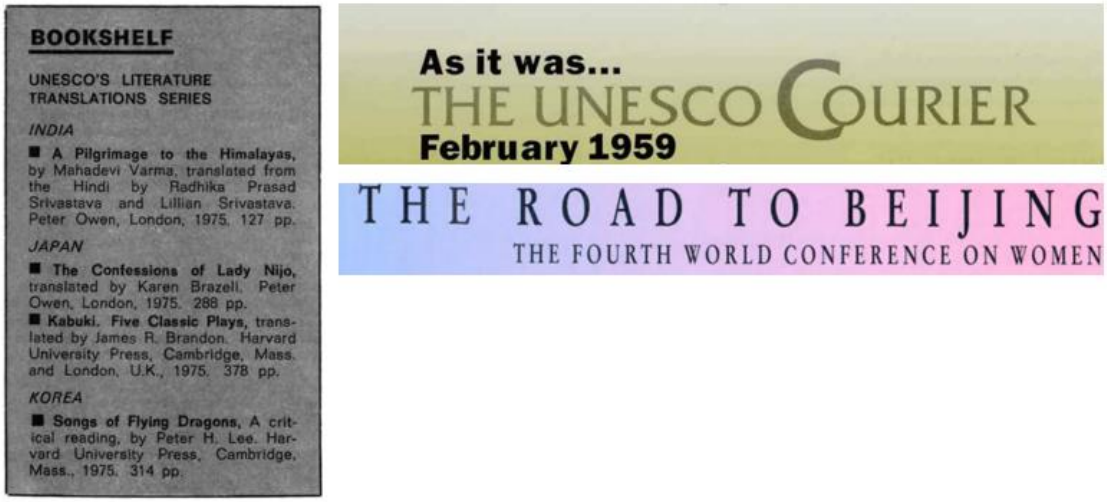


Рисунок 16 – Примеры текста на цветной подложке, которые определяются программой как картинки

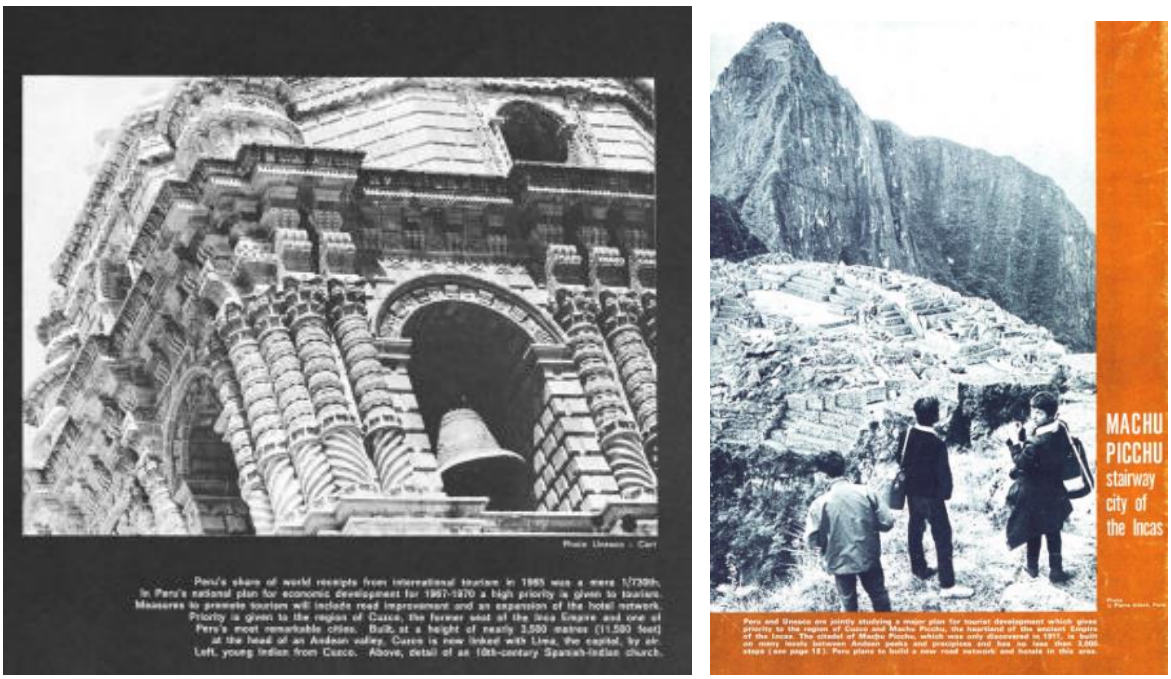


Рисунок 17 – Пример изображений вместе с текстом на одной большой цветной подложке

В таблице 3 представлено процентное соотношение сложных случаев в журнале.

Таблица 3 – Процент сложных случаев в журналах 60-х и 90-х годов

	Текст на цветной подложке	Картинки на цветной подложке
Журналы «Курьер ЮНЕСКО» за 60-е годы	11%	19,1%
Журналы «Курьер ЮНЕСКО» за 90-е годы	8%	11,5%

Важно заметить, что большинство (88%) картинок на цветной подложке с текстом программа вырезает именно по границе подложки, а не самой картинки. Таким образом, мы получаем не изображение, обрезанное по своим границам и пригодное для дальнейшего анализа, а текст вместе с изображением, либо несколько изображений на одной подложке, что помешает правильному анализу.

Исходя из того, что у нас есть как текст на подложке (рисунок 16), так и картинка вместе с текстом на подложке (рисунок 17), мы не можем удалить не интересующие нас тексты на цветной подложке с помощью фильтрации по обнаружению текста. Удалив картинки, где обнаружится много текста, мы потеряем также и картинки на цветной подложке. Для нас приоритетнее сохранить данные картинки в формате с подложкой и текстом, чем совсем

потерять их в итоговом результате. Поэтому удалить нужно только текст на цветной подложке.

Для решения данной задачи была создана обучающая выборка для нейронной сети, которая представляет собой комбинацию между двумя видами данных: вырезки журналов «Курьер ЮНЕСКО» семидесятых и двухтысячных годов и вырезки из проекта Newspaper Navigator dataset [27]. В выборке Newspaper Navigator dataset находятся подходящие данные для обучения нашей нейронной сети: 603 рекламные вырезки из газет Америки 1905 года с пометками о том, является реклама полностью текстовой, или там присутствует иллюстрация (рисунок 18).

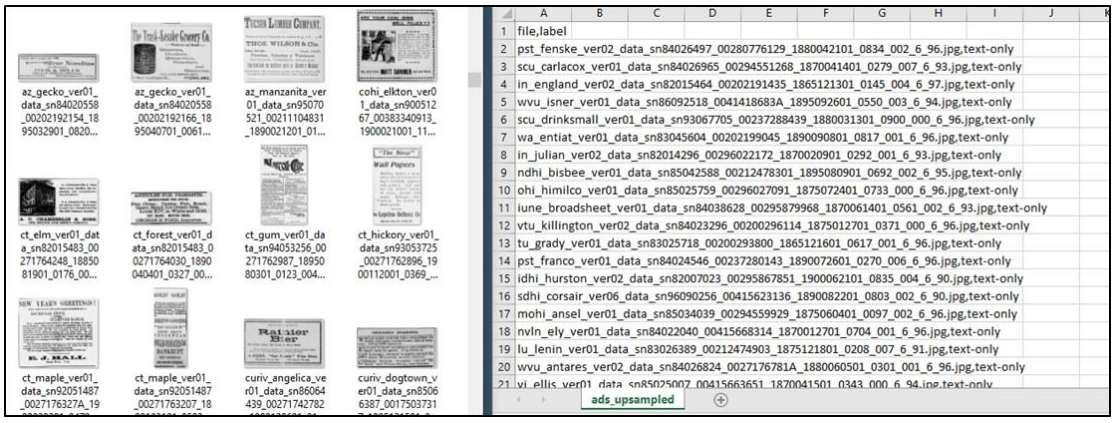


Рисунок 18 – Фрагмент данных для обучающей выборки из Newspaper Navigator dataset

При использовании в качестве обучающей выборки только данных из Newspaper Navigator dataset, нейронная сеть часто ошибалась (24% случаев правильного результата), когда этот инструмент применялся к вырезкам из журнала «Курьер ЮНЕСКО». Было решено добавить в обучающую выборку 714 вырезок из журналов «Курьер ЮНЕСКО» семидесятых и двухтысячных годов (рисунок 19). Вырезки из журналов 2000-х годов были добавлены в обучающую выборку из-за того, что в журналах 70-х годов не было примеров оформления текста на градиентных цветных подложках, что было популярно в журналах 90-х годов.

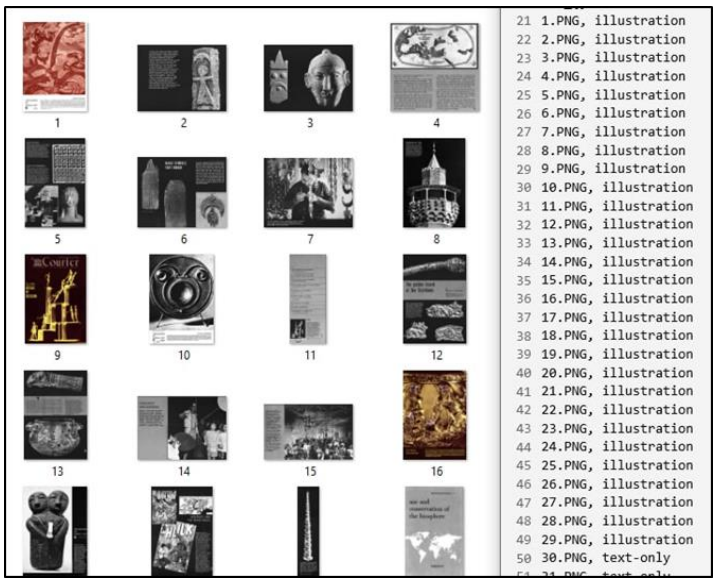


Рисунок 19 – Фрагмент данных для обучения из журналов «Курьер ЮНЕСКО» семидесятых и двухтысячных годов

Эта часть обучающей выборки была помечена нами вручную: если на вырезке только текст, мы помечали данные как «text-only», а если на вырезке присутствует изображение, то «illustration».

Результат работы по удалению текстовых фрагментов представлен на рисунке 20.

Рисунок 20 – Результат работы пятого этапа программы: удаление текстовых фрагментов

В результате выполнения всех пяти этапов алгоритма, мы получили извлеченные картинки из журналов 60-х годов с Accuracy 0,96, F-score 0,69, а для 90-х годов с Accuracy 0,97, F-score 0,7. В журналах 1960-х было найдено 4285 картинки, а в журналах 1990-х 5841 картинка.

Таблица 4 – Результаты алгоритма по извлечению картинок в журналах за 60-е и 90-е годы

	Было найдено картинок	Accuracy	F-score
Журналы «Курьер ЮНЕСКО» за 60-е годы	4285	0,96	0,69
Журналы «Курьер ЮНЕСКО» за 90-е годы	5841	0,97	0,7

Для того, чтобы получить весь текстовый корпус журналов в файле формата txt, мы использовали Python, который обращался к Tesseract. Tesseract - это свободная компьютерная программа для распознавания текстов. Функция распознавания нам не требовалась, поскольку все журналы «Курьер ЮНЕСКО» представлены на сайте «ЮНЕСКО» уже с распознанным текстом. Нам нужно было объединить весь текст за десятилетие в один текстовый корпус. Результаты данной работы представлены в таблице 5.

Таблица 5 – Текстовый корпус журналов за 60-е и 90-е годы

	Текстовый корпус (слов)
Журналы «Курьер ЮНЕСКО» за 60-е годы	1 951 161
Журналы «Курьер ЮНЕСКО» за 90-е годы	2 623 578

Поскольку данная работа является подготовкой к мультимодальному анализу журнала «Курьер ЮНЕСКО», т.е. изучению текстов и картинок совместно (см. раздел 2), нам нужно было найти способ группировки текста и изображений с одной страницы.

В нашем алгоритме процесс поиска изображений на каждой странице осуществляется с помощью цикла. Программа не перейдет на следующую страницу до тех пор, пока не найдет все изображения на текущей странице. Поэтому, мы решили присвоить каждому найденному изображению номер страницы, который равен порядковому номеру итерации цикла. Такой подход позволяет нам учитывать то, что на одной странице может быть сразу несколько изображений. Так, если на второй странице журнала будет три картинки, они будут помечены в названии цифрой «1» (отсчет начинается с 0, с обложки журнала). Таким образом, в названии всех изображений пишется число обозначающее номер страницы, с которой это изображение было взято

(рисунок 21). Данная часть кода нашей программы в перспективе позволяет попарно анализировать текст и фото с одной и той же страницы, а далее и сравнивать.

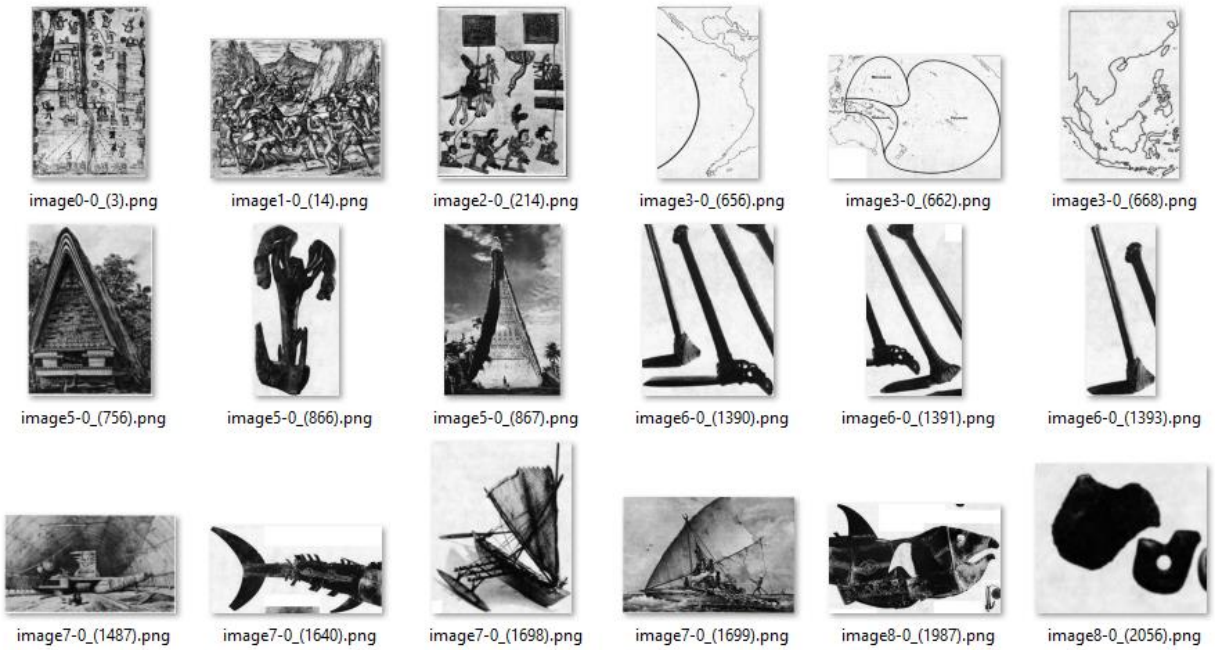


Рисунок 21 – Извлечённые изображения из журнала, где в названии указано, на каких страницах они были обнаружены

Код программы: https://github.com/Alina-meow/courier.

4. Результаты исследования

4.1 Сравнение созданного алгоритма с другими методами

Для оценки нашего алгоритма мы сравнивали его с другими инструментами для выделения картинок в тексте. Все методы, включая наш, проверялись на двух случайных выборках: 10% журналов (11 выпусков) из 60-х и 10% журналов (11 выпусков) из 90-х. Таким образом мы смогли составить таблицу эффективности всех проверенных методов.

Какие инструменты были проверены? Во-первых, мы взяли простой и быстрый инструмент для работы с PDF-файлами «PDF24 Tools» [28]. Среди функций этой программы есть извлечение картинок из файла формата PDF. Нам хотелось узнать, как «универсальная» программа справится с историческими документами и сложными стилистическими особенностями дизайна журнала.

Во-вторых, мы использовали Transkribus. Transkribus — это комплексная платформа для оцифровки, распознавания текста с помощью искусственного интеллекта, его расшифровки и разметки исторических документов [29]. Важно отметить, что с помощью Transkribus можно получить разметку страницы, но не готовые вырезанные картинки. Чтобы получить картинки, нужен дополнительный программный код, который будет использовать данные о разметке из Transkribus и вырезать картинки. Таким образом, требуется как работа с Transkribus, так и с программным кодом, с помощью одной только платформы решить нашу задачу не получится.

В Transkribus мы использовали два метода создания разметки страницы: Printed Block Detection и P2PaLa. Первый работает без обучения, для второго нужна обучающая выборка. С помощью первого метода можно автоматически искать «печатные блоки» (картинки, абзацы) на странице и размечать ее. Это звучит удобно, даже если для реализации данного способа требуется написать программный код в дальнейшем.

Второй метод также делает разметку страницы, но для его обучения мы использовали 100 страниц из журналов «Курьер ЮНЕСКО», размеченных

ручным способом в Транкрибусе. Количество страниц для обучения было выбрано исходя из инструкции на сайте Транкрибуса релевантной для этого метода [30]. Этот способ требует ручной разметки и составления обучающей выборки, но при хорошем результате эти усилия были бы оправданы.

Все инструменты для выделения картинок в тексте, которые были попробованы, представлены в таблице 6.

Таблица 6 – Инструменты для выделения картинок в тексте

PDF24 Toolbox	Многофункциональная и простая в использовании программа для работы с PDF, с помощью которой можно извлекать картинки из PDF-файла
Transcribus Printed Block Detection	Автоматическая функция Транскрибуса по поиску текстовых блоков и разметки страницы
Transcribus P2PaLa	Метод по разметке страницы, где можно использовалось предварительное обучение на наших данных

Для сравнения нашего метода с другими инструментами по выделению картинок в тексте используются метрики Accuracy, Precision, Recall и F-score. Результаты сравнения представлены в таблице 7 и 8.

Таблица 7 – Сравнение способов: 60-е годы

	Accuracy	Precision	Recall	F-score
Наш метод	0.96	0,69	0,7	0,69
PDF24 Toolbox	0.06	0,01	0,05	0,02
Transcribus Printed Block Detection	0.34	0,07	0,54	0,12
Transcribus P2PaLa (модель обученная на моих данных)	0.08	0,01	0,05	0,02

Таблица 8 – Сравнение способов: 90-е годы

	Accuracy	Precision	Recall	F-score
Наш метод	0.97	0,7	0,69	0,7
PDF24 Toolbox	0.08	0,02	0,06	0,03
Transcribus Printed Block Detection	0.41	0,1	0,61	0,17
Transcribus P2PaLa (модель обученная на моих данных)	0.08	0,02	0,06	0,03

Accuracy показывает долю правильных ответов алгоритма. Это интуитивно понятная метрика, где мы извлекаем все правильно определенные картинки с помощью алгоритма и делим на общее количество картинок в

журнале. Так, программа PDF24 Tools верно сохранит 6% картинок из журналов 60-х годов, Transcribus Printed Block Detection 34%, а наш метод 96%. Эта метрика показала нам, что все методы, кроме нашего, специально настроенного для анализа журнала «Курьер ЮНЕСКО», вырезают картинки не очень качественно. Но, Accuracy не отразит то, что есть разные классы проблем при извлечении картинок, и все представленные инструменты справляются с ними по-своему. Так, 34% Accuracy у Transcribus Printed Block Detection не покажет то, что данный метод будет определять текст на цветной подложке как изображение, внутри картинки выделит лишние картинки и пропустит иллюстрации. Для нашего метода, метрика не будет учитывать логотипы и рекламные изображения, которые появятся среди конечного результата извлечения картинок. Таким образом, нам нужна метрика, которая учтет все классы ошибок при извлечении картинок.

F-score объединит в себе информацию о точности (Precision) и полноте (Recall) нашего алгоритма. Точность (Precision) системы рассчитывается как раз в пределах определенного класса (картинка с четкими границами, логотип/реклама, текст на цветной подложке и.т.д.) – так мы учтем долю картинок, которую метод неправильно включил в результат, а также потери картинок. Полнота (Recall) системы – это доля найденных правильных картинок относительно всех найденных картинок. Таким образом, посчитав F-score мы оценим эффективность использованных методов точнее.

Для расчета F-score использовалась матрица неточностей (Confusion Matrix). Эта матрица представляет собой особый макет таблицы, который позволяет визуализировать производительность алгоритма. Матрицы неточностей для 1960-х и 1990-х годов представлены в приложении А.

Таблица 7 для 60-х годов показывает, что лучшая эффективность работы у нашего метода (69%). У PDF24 Tools и Transcribus P2PaLa мы видим одинаковый результат эффективности 2%. Эти методы определили каждую страницу журнала целиком как картинку, т.е. не идентифицировали никакой структуру внутри страницы. Откуда возник результат 2% эффективности у этих

методов? Дело в том, что в журналах «Курьер ЮНЕСКО» иногда картинка занимает всю страницу, и когда программа вырежет все страницы, мы получим, что некоторые из них действительно окажутся правильно вырезанными картинками.

У Transcribus Printed Block Detection результат 16% из-за множества ошибок при определении границ изображений, дробления одной картинки на несколько, игнорирования многих картинок. Все вертикальные линии на картинках Transcribus определяет как «separator», т.е. как разделители в таблицах и включает это в структуру страницы. Примеры ошибок Transcribus Printed Block Detection представлены на рисунке 37.

Рисунок 34 – Примеры ошибок при определении картинок с помощью Transcribus Printed Block Detection

Таблица 3 показывает, что для 90-х получились аналогичные результаты. Лучшее качество у нашего метода 70%, далее идет Transcribus Printed Block Detection (17%) и PDF24 Toolbox с Transcribus P2PaLa (3%). Однако, по рисунку 3 видно, что Transcribus хорошо определяет абзацы, количество строчек, т.е. все полностью текстовые блоки (на рисунке 3 выделены синем). Таким образом, инструменты платформы Transcribus отлично справляются с разметкой и распознаванием текста, что также видно во множестве исследований на текстах с использованием Transcribus, но с изображениями возникают проблемы. Возможно, инструментам платформы Transcribus требуется обучение на большем объеме данных, чтобы распознавать

изображения (мы использовали для обучения метода Transcribus P2PaLa 100 вручную размеченных страниц журналов «Курьер ЮНЕСКО», что соответствует рекомендациям, представленным на сайте Transcribus [30]). Однако в данном случае мы хотели проверить сопоставимые по трудовым затратам методы.

Что означают эти результаты? Разрыв между методами, которые не были специально настроены (или были недостаточно настроены) для журнала «Курьер ЮНЕСКО», и нашим методом достаточно большой. Это говорит о неуниверсальности задачи извлечения картинок из исторических документов (т.е. из документов, которые не были подвергались компьютерной обработке в момент создания). Задача извлечения также осложняется стилистическими особенностями (см. раздел 4.3) оформления журнала «Курьер ЮНЕСКО», которые мешают отделить текст от картинки компьютерным способом. Таким образом, если мы хотим получить хороший результат извлечения картинок из исторического документа, где применяются необычные дизайнерские решения, нам нужен собственный метод, хорошо подготовленный и настроенный для конкретного объекта исследования.

4.2 Неуниверсальность извлечения изображений в рамках нашего метода

На рисунке 22 представлен пример того, как разработанный алгоритм справляется со страницей журнала «Курьер ЮНЕСКО».

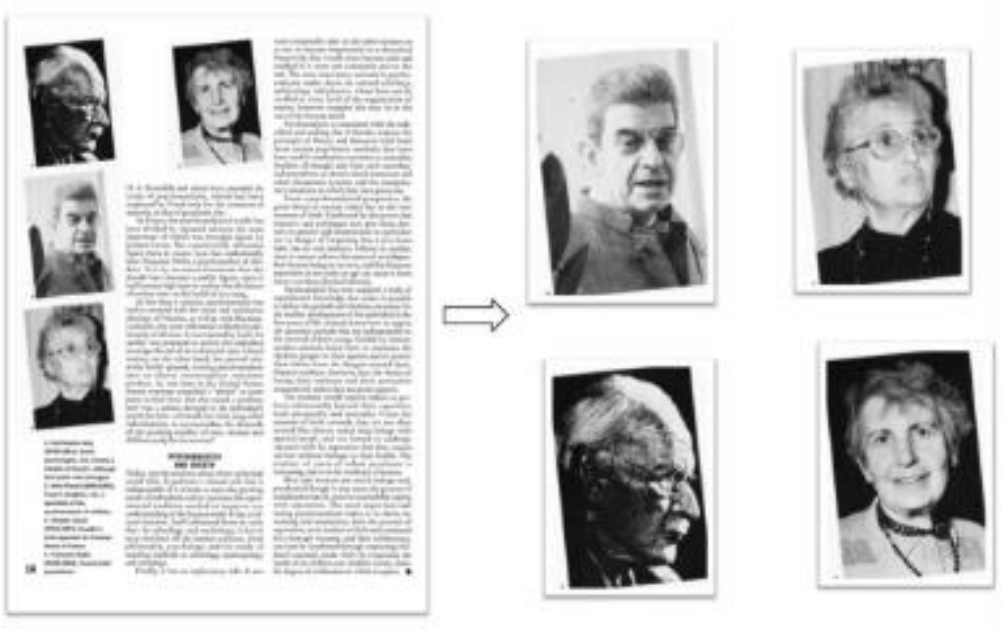


Рисунок 22 – Результат работы созданного алгоритма

Наш метод вырезает (12%) картинок из цветных подложек правильно. В остальных случаях он принимает цветную подложку за часть картинки и вырезает всё одним кадром (рисунок 23). Результат работы нашего алгоритма

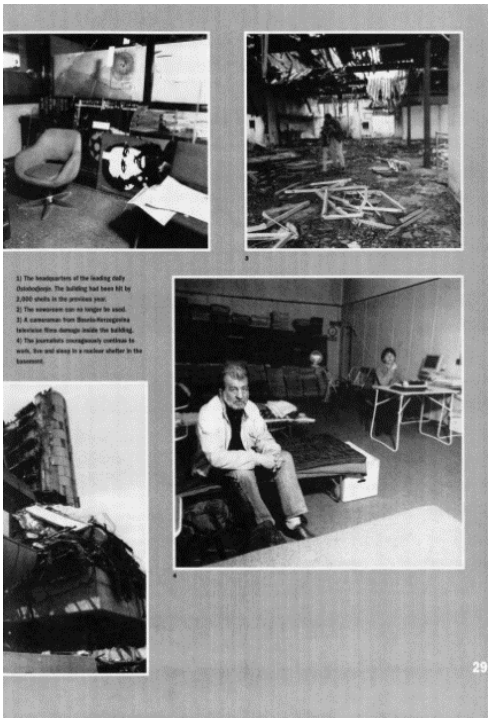


Рисунок 23 – Пример картинок, расположенных на одной цветной подложке, которые наш алгоритм не смог разделить

Эта особенность оформления журналов осложняется еще и тем, что при сканировании и преобразовании страницы в PDF-файл в дальнейшем, однородный цвет подложки (серый, черный, цветной) на самом деле не будет являться одним цветом или небольшим промежутком оттенков. В однородной

на первый взгляд подложке цветов будет не меньше, чем в самих картинках, поэтому эксперименты с попытками настраивать алгоритм по цветовым промежуткам подложек были неуспешны (эксперимент ниже).

Стоит отметить, что, подбирая разные значения в обработке страницы журнала (контраст, блюр, инверсия, различные цветовые маски) в нашем алгоритме (в части, где работает библиотека OpenCV), мы можем вырезать картинки и с таких сложных цветных подложек. Например, чтобы вырезать картинки с достаточно темных подложек, нужно инвертировать цвета страницы журнала с такими картинками. Но главная сложность заключается в том, как научить алгоритм определять: какие настройки нужно сейчас применить к странице, чтобы правильно вырезать из нее картинки. Ведь вариантов сложного стилистического оформления в журнале много (см. примеры в разделе 4.3), а также на одной странице может быть сразу несколько классов сложных случаев, к которым нужен разный подход.

Для того, чтобы проверить, можно ли уменьшить количество ошибок в нашем алгоритме, мы возьмем большой класс сложного стилистического оформления «картинки на серой подложке», где на первый взгляд все изображения очень похожи, и попробуем подобрать к ним универсальные параметры для извлечения.

На рисунке 24 представлен случайных набор четырех серых картинок. Настройки для правильной обрезки подбирались к первой картинке, а потом применялись ко всем остальным.

Рисунок 24 – Набор случайных четырех картинок на серых подложках

Эти изображения расположены на очень похожем сером фоне. Мы применим настройки контрастности и размытия, которые получили при эксперименте с первой картинкой ко всем остальным (рисунок 25).

Рисунок 25 – Применение настроек ко всем картинкам

В результате мы получим, что вторая картинка при вырезке раздробилась на 13 маленьких кусочков, из второй успешно получились две правильные картинки, а третья раздробилась на 3 неправильных варианта обрезки (рисунок 26).

Рисунок 26 – Результат применения настроек

Что мы получим на более объемной выборке? Мы взяли 100 неправильно вырезанных картинок на серых положках похожих оттенков. Далее также опытным путем сделали подходящие настройки для 5 случайных из этих 100 картинок, а потом применили их ко всем. Таким образом в результате мы

получили 32 правильно вырезанные картинки, что составляет 32% правильных решений.

Полученные результаты говорят нам о том, что даже в рамках одного узконаправленного метода сложно достичь универсальности в применении алгоритма к с сложной стилистической конструкции исторического документа.

4.3 Классификация стилистических особенностей оформления в журналах и как с ними справляется компьютерное зрение

В журнале «Курьер ЮНЕСКО» есть некоторые стилистические особенности оформления, которые мешают компьютерному зрению эффективно вырезать картинки. Все эти особенности требуют индивидуального решения. Далее будут представлены выделенные особенности.

1. Страницы с серой рамкой

Этот визуальный эффект не позволяет вырезать картинки, находящиеся внутри рамки по отдельности, поскольку алгоритм считает крайнюю серую рамку границей всего изображения (рисунок 27).

Рисунок 27 – Пример страниц с серой рамкой (весь текст удален)

Такая особенность встречается только в некоторых журналах 60-х годов (3% от всех картинок в журналах 60-х). Алгоритм не справляется с такими страницами.

2. Картинки на серых подложках

Это большая группа картинок из журналов 60-х годов (10%), где на одной серой подложке представлены сразу несколько картинок, либо картинка и текст вместе (рисунок 28).

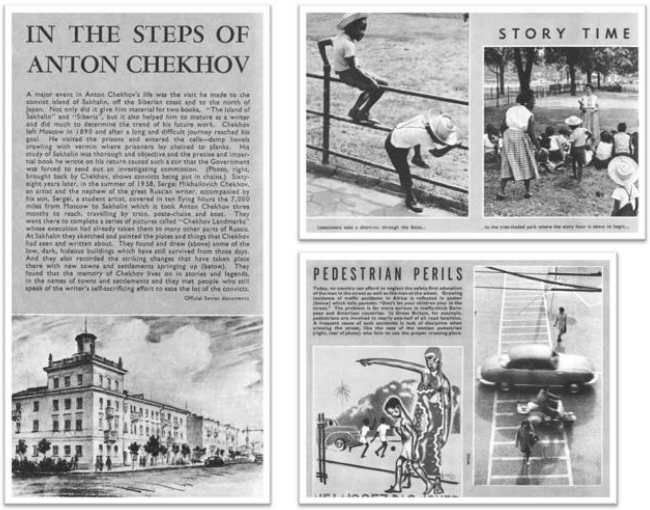


Рисунок 28 – Пример картинок на серых подложках

Алгоритм видит границу изображения по краю серой подложки, тем самым вырезая весь коллаж как одну картинку. Стоит отметить, что алгоритм иногда справляется со светло-серыми подложками (3% картинок из всех серых картинок вырезается правильно), пример такой страницы представлен на рисунке 29.

Рисунок 29 – Пример картинки с серой подложкой, с которой справился алгоритм

Мы предполагаем, что это происходит из-за того, что на этапе размытия картинки (см. раздел 4.2) светлые серые подложки сливаются с белым фоном, и алгоритму удается в дальнейшем вырезать картинки правильно. Но, как видно

из раздела о неуниверсальности в рамках нашего метода, другие картинки на более темных серых подложках (даже визуально похожие) требуют разного подхода к настройке для извлечения. Таким образом, сложный класс с картинками на серых подложках, на самом деле внутри имеет еще неизвестное нам число подклассов, которые не определяются визуально и требуют к себе индивидуальной настройки.

3. Картинки на черных подложках в 60-е годы

Класс аналогичный картинкам на серых подложках, только подложки черного цвета (6% от всех картинок за 60-е года). Примеры таких коллажей из картинок, либо картинок и текста на черных подложках, представлены на рисунке 30.

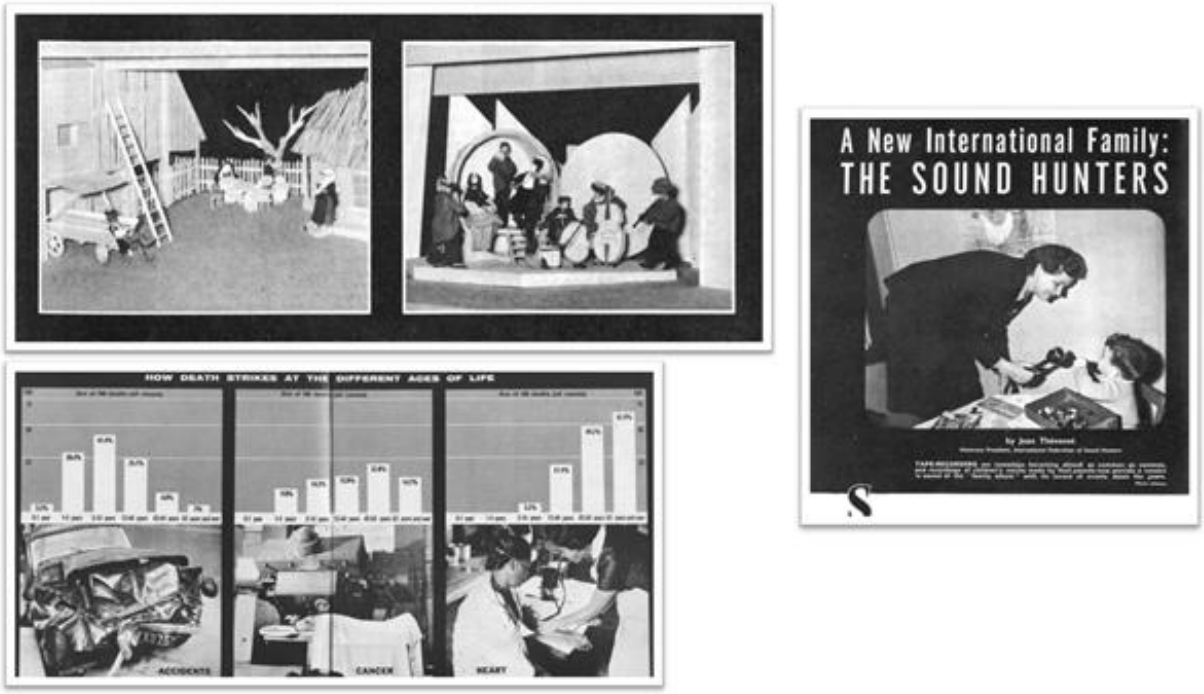


Рисунок 30 – Картинки на черных подложках

Мы знаем, как с ними работать: нужно инвертировать цвета у всего вырезанного коллажа, найти координаты границ внутренних картинок и на оригинальном (неинвертированном) коллаже их извлечь. Вся сложность заключается в том, что нужно работать с уже вырезанными картинками, поскольку инвертировать нужно только их, а не всю страницу, где могут быть и другие случаи. Следовательно, мы должны работать с уже вырезанными картинками в папке. Но нам не удалось отделить картинки с черной подложкой

от всех остальных по превосходству темных оттенков на изображении, ведь черный контур может занимать лишь малую часть картинки или обычная картинка (без черной подложки) сама по себе может быть темной. Таким образом, наш алгоритм не справляется с извлечением внутренних картинок с черных подложек.

4. Картинки на цветных (градиентных) подложках

Такое оформление картинок стало встречаться в журналах «Курьер ЮНЕСКО» в 90-х годах (11% от всех картинок в журналах 90-х). Оно похоже на класс с картинками на черных и серых подложках, только теперь подложки цветные, и чаще всего это градиент (рисунок 31).

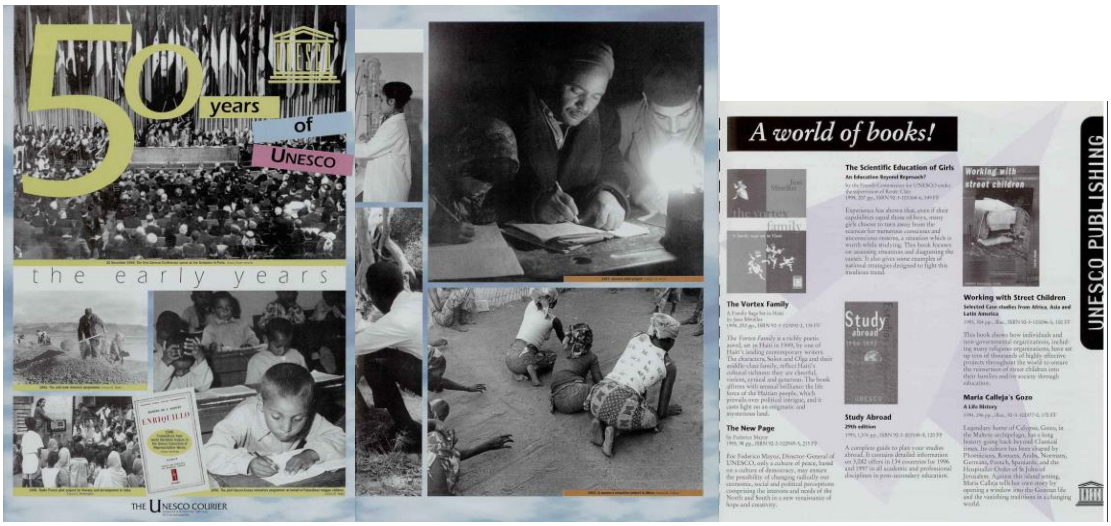


Рисунок 31 - Картинки на цветных подложках (градиентах)

Градиент еще больше усложняет работу алгоритма, ведь в подложке теперь большое количество оттенков, что по своей сути не отличает его от самих изображений, и алгоритм вырезает весь коллаж как целую картинку. Иногда градиент стоит под одной картинкой и без текста, такие случаи мы не считаем ошибкой алгоритма (рисунок 32).

Рисунок 32 – Картинка на градиенте без текста

Таким образом, наш алгоритм не справляется с картинками на градиентном фоне.

5. Текст на цветных (серых, черных, градиентных) подложках

Помимо картинок на подложках, в журналах существуют тексты на подложках. Стилизация текста на различных подложках — это также очень заметная особенность журналов «Курьер ЮНЕССКО» (10% от всех картинок в 60-е годы, 11% в 90-е годы). Текст располагается на черных, серых, цветных, градиентных подложках (рисунок 33).

Рисунок 33 – Примеры текста на разных подложках

Изначально это было проблемой нашего алгоритма, ведь по набору пикселей текст на подложке ничем не отличается от картинок. Однако благодаря способу, который описан в разделе «Методы», мы научились отделять текстовые картинки от всех остальных, тем самым увеличив точность нашего метода. Таким образом, наш алгоритм справляется с 99% таких случаев.

4.4 Сравнение результатов исследования с предшествующими работами

Все чаще в цифровых гуманитарных исследованиях ученые обращаются к изучению изображений. Однако, как показывают наши эксперименты по извлечению изображений из журнала «Курьер ЮНЕСКО» с помощью «универсальных инструментов» (PDF24 Tools, Transcribus), получить изображения из исторических документов — это сложная задача. Так, точность (Accuracy) извлечения на журналах 1960-х годов программы PDF24 Toolbox составляет 6%, Transcribus Printed Block Detection - 34%, Transcribus P2PaLa – 8%, у нашего специально созданного метода – 96%. К аналогичному выводу можно прийти, рассмотрев работу Бена Ли [20], где нейронная сеть для извлечения изображений из исторических газет Америки обучалась на данных ручной разметки (добровольцы вручную выделяли весь визуальный материал на 3437 страницах газет, это 32424 изображения). Рассматривая относительно небольшой объект для изучения, данных которого не хватит для обучения нейронной сети, мы видим, что создание собственного алгоритма является наилучшим и, пожалуй, единственным возможным решением.

В работе [1] авторы продемонстрировали прекрасный пример работы с изображениями путем создания классификатора изображений голландских исторических газет, которые были их объектом. Однако мы хотим обратить внимание на то, что авторы упускают процент успешного извлечения изображений. Процент успешного извлечения позволяет понять, возможно ли проводить исследования, делать выводы и формулировать гипотезы на основе этих данных. Так, например, можно проверить, есть ли в данных «ямы», когда изображения плохо извлекаются для определенного типа страниц, периода

выпуска газет, или из-за другой особенности. По этой причине, в нашей работе мы занимаемся подготовкой изображений из журналов «Курьер ЮНЕСКО» к анализу, четко понимая, какой процент правильно извлеченных изображений в итоге получим, с какими сложными случаями алгоритм не справляется, из чего можно сделать вывод о том, где будет больше всего пропусков в данных, что безусловно важно для дальнейшей работы с этими данными.

Мы считаем, что задача извлечения картинок все еще является сложным и не универсальным этапом подготовки к анализу изображений, который в определенной степени препятствует изучению визуального материала отсканированных книг, журналов, газет и других исторических документов. Также этот этап необходим для проведения мультимодальных исследований, описанных в работе [21]. В мультимодальных исследованиях анализируется объект (книга, журнал, газета и.т.д.) целиком, без разделения на анализ только текста или только картинок. Так, в рамках мультимодального исследования возможно провести сравнение результатов компьютерного анализа изображений и текста с одной страницы журнала. Для этого в алгоритм встроен счетчик изображений, а в названии помечается, на какой странице они были обнаружены.

Исследование изображений, сопоставление их с текстом - важный новый этап цифровых гуманитарных исследований. Однако мы должны понимать, что исследования, с возможностью формулирования гипотез или выводов, должны быть проведены на надежном материале: изображения должны быть максимально точно извлечены, их количество должны быть близко к тому, сколько их в исходном объекте, мы должны учитывать особенности объекта исследования, влияющие на процесс извлечения изображений. Так мы сможем прийти к дальнему чтению изображений и их сопоставлению текстам.

ЗАКЛЮЧЕНИЕ

В этой работе был создан алгоритм для извлечения изображений и текстовых блоков из журналов «Курьер ЮНЕСКО». Кроме того, программа группирует изображения и тексты для мультимодального анализа. Алгоритм представляет собой комбинацию из инструментов библиотеки компьютерного зрения OpenCV, нейронной сети, созданной с помощью Fast.Ai и других вспомогательных библиотек: PyPDF2, PyMuPDF и PIL. Такой подход позволил достичь максимальной точности извлечения изображений (97% Accuracy), что и представляло основную сложность. Созданный алгоритм учитывает особенности журналов «Курьер ЮНЕСКО», например, то, что текст располагается на цветных подложках, которые можно было бы принять за изображения. Разработанный подход дает возможность успешно отделить сложные случаи оформления текстовых блоков и не включать их в результат извлечения изображений. При этом работа демонстрирует, что инструменты платформы Transcribus, программа PDF24 Tools, т.е. более универсальные способы, не справляются с извлечением изображений из исторических документов («Курьер ЮНЕСКО»), и дают не более 41% точности (Accuracy).

Не менее важно и то, что работа показывает неуниверсальность задачи извлечения изображений не только с помощью «ненастроенных» методов, но и в рамках созданного алгоритма. Эксперимент, где были взяты изображения на похожих серых подложках, показал, что настроить алгоритм для одной из таких картинок, не значит настроить для всех таких случаев. Это показывает, что задача извлечения изображений все еще является сложным и важным этапом подготовки к мультимодальному анализу.

Анализ работ в области изучения изображений показал, что этапу извлечения изображений отводится мало внимания. Однако процент извлеченных изображений является ключевой информацией для того, чтобы определить возможность проведения исследований, формулирования выводов и гипотез, на основе предоставленных данных. Благодаря этому станет возможно осуществить анализ изображений и их сопоставление с текстовыми данными.

Код программы: https://github.com/Alina-meow/courier

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Wevers M., Smits T. The visual digital turn: Using neural networks to study historical images //Digital Scholarship in the Humanities. – 2020. – Т. 35. – №. 1. – С. 194-207.

2. Barthes R. Le message photographique //Communications. – 1961. – Т. 1. – №. 1. – С. 127-138.

3. «Устав ООН» [Электронный ресурс] – Режим доступа: https://www.un.org/en/about-us/un-charter/full-text – Дата доступа: 2023.

4. UNESCO and the Issue of Cultural Diversity – Review and Strategy 1946–2000. A Study Based on Official Documents. [Электронный ресурс] – Режим доступа:https://unesdoc.unesco.org/ark:/48223/pf0000125248?posInSet=1&queryId=243c8e04-7cec-4e80-84ba-43e5c752ed98 – Дата доступа: 2023

5. Stoczkowski W. UNESCO's doctrine of human diversity: a secular soteriology? //Anthropology Today. – 2009. – Т. 25. – №. 3. – С. 7-11.

6. UNESCO. Draft Programme and Budget, 2002-2003: General Conference, Thirty-first Session, Paris 2001. – UNESCO, 2001.

7. Nielsen B. UNESCO and the ‘right’ kind of culture: Bureaucratic production and articulation //Critique of Anthropology. – 2011. – Т. 31. – №. 4. – С. 273-292.

8. Our Creative Diversity: Report of the World Commission on Culture and Development. [Электронный ресурс] – Режим доступа: https://unesdoc.unesco.org/ark:/48223/pf0000105586?posInSet=4&queryId=ca5ea1b5-b3c3-4b73-8fb7-fa11149c0095 – Дата доступа: 2023

9. Zizek S. The sublime object of ideology Verso. – 1989.

10. «Курьер ЮНЕСКО» [Электронный ресурс] – Режим доступа: https://ru.unesco.org/courier/about – Дата доступа: 2023

11. «UNESDOC» [Электронный ресурс] – Режим доступа: https://unesdoc.unesco.org/ark:/48223/pf0000261279_rus – Дата доступа: 2023.

12. «Архив журналов "Курьер ЮНЕСКО"» [Электронный ресурс] – Режим доступа: https://ru.unesco.org/courier/archives – Дата доступа: 2023.

13. Barthes R. Rhetoric of the Image //Visual culture: The reader. – 1999. – С. 33-40.

14. Baudrillard J. Simulacra and simulation. – University of Michigan press, 1994.Baudrillard J. Simulacra and simulation. – University of Michigan press, 1994

15. Yanoshevsky G., Michaeli M. On recurring images and nation branding: the case of Israel's albums and tourist guidebooks //Image & Narrative. – 2021. – Т. 22. – №. 2.

16. Fleming D. Can pictures be arguments? //Argumentation and advocacy. – 1996. – Т. 33. – №. 1. – С. 11.

17. Thomas J. Nineteenth-Century Illustration and the Digital //Studies in Word and Image. Cham. – 2017.

18. Champion E. M. Digital Humanities is text heavy, visualization light, and simulation poor //Digital Scholarship in the Humanities. – 2017. – Т. 32. – №. suppl_1. – С. i25-i32.

19. Münster S., Terras M. The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures //Digital Scholarship in the Humanities. – 2020. – Т. 35. – №. 2. – С. 366-389.

20. Lee B. C. G. et al. The newspaper navigator dataset: extracting and analyzing visual content from 16 million historic newspaper pages in chronicling America //arXiv preprint arXiv:2005.01583. – 2020.

21. Smits T., Wevers M. A multimodal turn in Digital Humanities: using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections //Digital scholarship in the humanities: a journal of the Alliance of Digital Humanities Organizations.- Oxford, 2015, currens. – 2023. – С. 1-14.

22. «CLIP» [Электронный ресурс] – Режим доступа: https://openai.com/research/clip – Дата доступа: 2023

23. Radford A., Kim J. W., Hallacy C., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, PMLR, virtual, pp. 8748–63

24. «Библиотека OpenCV» [Электронный ресурс] – Режим доступа: https://opencv.org/ – Дата доступа: 2023.

25. «Библиотека Fast.Ai» [Электронный ресурс] – Режим доступа: https://www.fast.ai/ – Дата доступа: 2023.

26. Strien D. et al. Computer Vision for the Humanities: An Introduction to Deep Learning for Image Classification (Part 1) //Programming Historian. – 2022. – T. 11.van Strien D. et al. Computer Vision for the Humanities: An Introduction to Deep Learning for Image Classification (Part 1) //Programming Historian. – 2022. – T. 11.

27. «Newspaper Navigator» [Электронный ресурс] – Режим доступа: https://github.com/LibraryOfCongress/newspaper-navigator – Дата доступа: 2023

28. PDF24 Tools [Электронный ресурс] – Режим доступа: https://tools.pdf24.org/ru – Дата доступа: 2023

29. «Transkribus» [Электронный ресурс] – Режим доступа: https://transkribus.ai/ – Дата доступа: 2023.

30. «Transkribus p2pala» [Электронный ресурс] – Режим доступа: https://readcoop.eu/transkribus/docu/p2pala/ – Дата доступа: 2023.

ПРИЛОЖЕНИЕ А

Таблица А.1 – Матрица неточностей для журналов 1960-х

	Картинки с четкой рамкой	Логотипы/реклама	Текст на цветной подложке	Картинка на серой подложке
Картинки с четкой рамкой	63	1	0	19
Логотипы/реклама	0	4	0	0
Текст на цветной подложке	0	0	11	0
Картинка на серой подложке	0	0	0	0

P1	0,76	0,69
P2	1,00	
P3	1,00	
P4	0,00	
R1	1	0,7
R2	0,8	
R3	1	
R4	0	
F-score	0,69	

Таблица А.2 - Матрица неточностей для журналов 1990-х

	Картинки с четкой рамкой	Логотипы/реклама	Текст на цветной подложке	Картинка на серой подложке
Картинки с четкой рамкой	81	2	0	11
Логотипы/реклама	0	6	0	0
Текст на цветной подложке	0	0	8	0
Картинка на серой подложке	0	0	0	0

P1	0,86	0,72
P2	1,00	
P3	1,00	
P4	0,00	
R1	1	0,6875
R2	0,75	
R3	1	
R4	0	
F-score	0,70	

Министерство науки и высшего образования РФ
Федеральное государственное автономное
образовательное учреждение высшего образования
«**СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**»

Гуманитарный институт
Кафедра информационных технологий
в креативных и культурных индустриях

УТВЕРЖДАЮ
И. о. заведующего кафедрой
_____ М. А. Лаптева
«27» июня 2023 г.

БАКАЛАВРСКАЯ РАБОТА

Компьютерное зрение для разделения изображений и текста в журнале «Курьер ЮНЕСКО»

Направление подготовки: 09.03.03 Прикладная информатика

Наименование программы: 09.03.03.30 Прикладная информатика

Руководитель канд. культурологии, доц., ст. науч. сотр. И. А. Кижнер

Выпускник А. С. Дяченко

Нормоконтролер И. Р. Нигматуллин

Красноярск 2023