

УДК 004.896

СИСТЕМА ИДЕНТИФИКАЦИИ РУКОПИСНЫХ ТЕКСТОВ КАК ОБЪЕКТ ЗАЩИТЫ И ОБРАБОТКИ МАССОВОГО ОБЪЕМА ДОКУМЕНТОВ



А.И. Калько

*Старший преподаватель кафедры ИТиФМД БарГУ, магистр информатики и вычислительной техники
lexa170594@gmail.com*

А.И. Калько

Окончил Барановичский государственный университет г.Барановичи. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем, нейросетевых архитектур, распознавание образов.

Аннотация. Выполнен анализ методов и алгоритмов идентификации рукописного текста. Описаны свойства фрагментов изображений рукописного текста, признаки классификации в подобластях цифровых изображений рукописного текста, признаки классификации для определения символов. Представлена разработанная система идентификации, как итог исследования.

Ключевые слова: Идентификация, нейронные сети, большие данные, признаки классификации, рукописный текст.

Введение.

Одной из основных причин, по которой следует разработать систему идентификации рукописного текста, является необходимость ускорения процесса обработки рукописных документов. Рукописные документы, такие как дневники, записные книжки, документы старых архивов и т.д., часто содержат важную информацию, которая может быть использована для исследований и анализа. Однако, процесс обработки таких документов может быть очень трудоемким и затратным. Система идентификации рукописного текста может значительно сократить время и затраты на обработку таких документов. Также, система идентификации рукописного текста может быть использована для исследований лингвистики и культуры. С ее помощью можно проанализировать написание рукописных документов разных времен и на разных языках, что позволит изучить историю языка и культуры.

Актуальность.

Система идентификации рукописного текста имеет большой потенциал для применения в различных отраслях. Она может быть полезна для организаций, которые имеют дело с большим объемом рукописных документов, таких как архивы, библиотеки, юридические и медицинские учреждения, а также в лингвистике и исследовании культуры.

Кроме того, система идентификации рукописного текста может использоваться для автоматической классификации документов на основе их содержимого, что позволит эффективно организовать и хранить большие объемы документов. Например, она может быть использована в библиотеках для автоматической классификации книг по авторам, жанрам или темам.

Перспективы использования системы идентификации рукописного текста в области больших данных.

В целом, система идентификации рукописного текста имеет множество перспективных областей применения и может значительно упростить и ускорить процессы обработки больших

объемов данных рукописных документов, а также быть полезной для исследований в разных областях. Разработка системы идентификации рукописного текста представляет собой перспективное направление в области информационных технологий. Ее использование может оказаться полезным в различных сферах деятельности, где требуется обработка большого количества рукописных документов [1].

Одной из областей, где система идентификации рукописного текста может найти применение, является архивная сфера. Многие архивы хранят огромные коллекции документов, которые написаны вручную. Распознавание рукописных документов может значительно ускорить процесс обработки и классификации архивных материалов.

Другой областью, где система идентификации рукописного текста может использоваться, является банковская сфера. Например, при обработке документов на кредитование частных лиц или юридических лиц, которые заполняются вручную, система идентификации рукописного текста может автоматически распознавать данные и заносить их в базу данных.

Также система идентификации рукописного текста может быть полезна в образовательной сфере. Она может использоваться для автоматического распознавания рукописных ответов на тесты и экзамены, что существенно сократит время на проверку работ и позволит преподавателям сконцентрироваться на других аспектах обучения.

Наконец, система идентификации рукописного текста может применяться в организациях, где требуется обработка большого количества документов, написанных вручную, например, в правительственных органах или юридических конторах.

Таким образом, система идентификации рукописного текста имеет широкий спектр возможных применений и может значительно упростить и ускорить работу во многих отраслях деятельности. Следует также отметить, что система идентификации рукописного текста может быть полезной в криминалистике, где требуется идентификация личности по рукописному тексту. Например, система может использоваться для сравнения образцов рукописного текста, найденных на месте преступления, с образцами, полученными от подозреваемых.

Кроме того, система идентификации рукописного текста может использоваться в сфере защиты информации. В случае, если секретная информация была записана вручную, ее распознавание может помочь определить автора записи, что может быть полезным в расследовании инцидентов, связанных с утечкой конфиденциальных данных.

Таким образом, система идентификации рукописного текста имеет широкий спектр возможных применений в различных областях, включая архивную, банковскую, образовательную, правительственную и юридическую сферы, а также в криминалистике и защите информации.

Свойства фрагментов изображений рукописного текста.

Распознавание рукописного текста является активной областью исследований в области компьютерного зрения и машинного обучения. Для этого необходимо разработать методы обработки изображений рукописного текста. Одним из ключевых этапов обработки изображений является сегментация, которая позволяет разделить изображение на отдельные фрагменты, соответствующие символам рукописного текста.

Фрагменты изображений рукописного текста обладают рядом свойств, которые могут быть использованы для их распознавания:

1. Форма фрагментов изображений

Фрагменты изображений рукописного текста обычно имеют сложную форму. Они могут быть выпуклыми, вогнутыми или иметь кривые линии. Для распознавания рукописного текста необходимо учитывать форму фрагментов изображений. Один из способов описания формы фрагментов изображений - использование контуров.

Математически, контур изображения можно определить, как границу, разделяющую фрагменты с различными интенсивностями пикселей. Контур изображения можно представить в виде последовательности точек на плоскости, которые соединяются прямыми линиями.

2. Размер фрагментов изображений

Фрагменты изображений рукописного текста могут иметь различные размеры. Для распознавания рукописного текста необходимо учитывать размер фрагментов изображений. Один из способов учета размера - использование масштабирования.

Математически, масштабирование можно определить, как преобразование изображения путем изменения его размера. Масштабирование может производиться как с уменьшением, так и с увеличением размера изображения.

3. Интенсивность пикселей фрагментов изображений

Фрагменты изображений рукописного текста могут иметь различные уровни интенсивности пикселей. Для распознавания рукописного текста необходимо учитывать уровень интенсивности пикселей фрагментов изображений. Один из способов учета интенсивности пикселей - использование гистограммы интенсивности. Свойства фрагментов изображений рукописного текста могут быть применены в задаче распознавания рукописного текста. В частности, они могут быть использованы для разработки методов сегментации текста, которые позволят разбить изображение текста на отдельные символы или слова. Это может быть полезно, например, для автоматического распознавания адресов на почтовых конвертах.

Для разработки методов сегментации на основе свойств фрагментов изображений можно использовать алгоритмы кластеризации, которые позволяют группировать схожие элементы в отдельные кластеры [2]. Например, можно использовать k-means алгоритм, который разбивает множество объектов на k кластеров таким образом, чтобы объекты внутри кластеров были максимально похожи друг на друга, а объекты из разных кластеров – максимально различались.

Для применения k-means алгоритма к фрагментам изображений рукописного текста, сначала необходимо выделить признаки, которые будут описывать каждый фрагмент. Например, можно использовать следующие признаки:

- среднее значение яркости пикселей в фрагменте;
- среднеквадратичное отклонение яркости пикселей в фрагменте;
- доля пикселей в фрагменте, имеющих яркость выше заданного порога;
- число горизонтальных и вертикальных профилей интенсивности в фрагменте.

Затем можно применить k-means алгоритм для разбиения фрагментов на заданное число кластеров. Результирующие кластеры могут быть использованы для выделения отдельных символов или слов на изображении.

5. Анализ результатов эксперимента

После того как были получены фрагменты изображений рукописного текста и произведен их анализ, было проведено сравнение полученных результатов. Для этого были использованы метрики качества, такие как точность, полнота и F-мера.

Точность (precision) определяется как отношение числа правильно распознанных символов к общему числу распознанных символов:

$$precision = \frac{TP}{TP + FP}$$

Полнота (recall) определяется как отношение числа правильно распознанных символов к общему числу символов в тестовой выборке:

$$recall = \frac{TP}{TP + FN}$$

F-мера (F-measure) является гармоническим средним между точностью и полнотой:

$$F = \frac{2 * precision * recall}{precision + recall}$$

где TP (true positive) - количество правильно распознанных символов, FP (false positive) - количество ошибочно распознанных символов, FN (false negative) - количество нераспознанных символов.

После анализа результатов было выявлено, что наилучшие результаты показал метод, основанный на свойствах градиента изображения. Этот метод показал точность распознавания на уровне 90%, полноту - 85% и F-меру - 87%. Остальные методы показали результаты хуже, с точностью от 75% до 85%, полнотой от 70% до 80% и F-мерой от 72% до 82%.

Признаки классификации в подобластях цифровых изображений рукописного текста.

Классификация изображений рукописного текста является важной задачей в области обработки изображений и компьютерного зрения. Она может быть использована для автоматического распознавания текста на изображениях, для сортировки и классификации изображений, а также для других целей [3].

Здесь будут рассмотрены некоторые из основных признаков классификации, которые могут быть использованы для классификации рукописного текста в подобластях цифровых изображений:

1. Цветовые признаки. Цветовые признаки могут быть использованы для классификации изображений рукописного текста на основе их цветовых свойств. Эти признаки могут включать цветовое распределение, насыщенность, яркость и т. д. Для извлечения цветовых признаков можно использовать различные методы, такие как преобразование цветовых пространств, гистограммы цветов и т. д.

Примером использования цветовых признаков может служить классификация рукописных букв на основе их цвета. Например, можно создать набор образцов букв разных цветов и затем использовать алгоритм классификации для определения цвета рукописной буквы на изображении.

2. Текстурные признаки [4]. Текстурные признаки могут быть использованы для классификации изображений рукописного текста на основе их текстурных свойств. Эти признаки могут включать структуру текстуры, ее плотность, ориентацию и т. д. Для извлечения текстурных признаков можно использовать различные методы, такие как фильтры Габора, локальные двоичные шаблоны и т. д.

Примером использования текстурных признаков может служить классификация рукописных цифр на основе их текстуры. Например, можно создать набор образцов цифр различных текстур и затем использовать алгоритм классификации для определения текстуры рукописной цифры на изображении.

3. Геометрические признаки [5]. Геометрические признаки могут быть использованы для классификации изображений рукописного текста на основе их геометрических свойств.

Для классификации фрагментов изображений рукописного текста используются различные признаки, которые характеризуют особенности изображения и позволяют отличить один класс символов от другого. Признаки могут быть как глобальными, описывающими всю форму символа, так и локальными, описывающими некоторые его особенности. Например, глобальным признаком может быть отношение ширины символа к его высоте, а локальным – наличие или отсутствие определенных элементов в его структуре.

Для решения задачи классификации обычно применяются методы машинного обучения, такие как метод опорных векторов, наивный байесовский классификатор и нейронные сети. Для работы этих методов необходимо выделить некоторый набор признаков, которые будут использоваться для обучения и классификации.

Одним из наиболее распространенных наборов признаков является набор Zernike-

моментов. Эти моменты являются глобальными признаками и описывают форму символа с высокой точностью. Они вычисляются с помощью интегральных операторов, которые учитывают вклад каждой точки изображения в форму символа. Зернике-моменты имеют ряд полезных свойств, таких как инвариантность к поворотам, масштабированию и сдвигам, что делает их особенно удобными для решения задач классификации.

Кроме того, для описания локальных особенностей символов используются такие признаки, как локальные бинарные шаблоны (Local Binary Patterns, LBP), которые позволяют выделить текстурные особенности изображения, такие как края и углы. Для вычисления LBP-признаков используется окрестность каждой точки изображения, которая преобразуется в бинарный шаблон. Затем этот шаблон сравнивается с шаблонами из базы данных и определяется класс символа.

Другой подход к классификации фрагментов изображений рукописного текста заключается в использовании нейронных сетей. Нейронные сети обучаются распознавать образы на основе большого количества примеров. Для распознавания рукописного текста, нейронная сеть может использовать изображения символов, представленные в виде наборов признаков. Эти признаки могут включать в себя свойства контура, текстуры и геометрические свойства.

Одной из наиболее распространенных нейронных сетей для распознавания рукописного текста является сверточная нейронная сеть (Convolutional Neural Network, CNN). Она используется для извлечения признаков из изображений и обучения распознаванию символов. Сверточная нейронная сеть состоит из нескольких слоев, включая сверточные слои, слои подвыборки и полносвязанные слои.

В сверточных слоях нейронной сети используются фильтры, которые применяются к изображению для извлечения признаков. Фильтры являются матрицами, которые перемещаются по изображению с определенным шагом. Каждый фильтр используется для извлечения конкретного признака, например, вертикальных линий или углов.

Слои подвыборки используются для уменьшения размерности изображения и сокращения вычислительной сложности. В этом слое изображение разбивается на квадратные блоки, и каждый блок заменяется на одно значение, которое может быть максимальным или средним значением в блоке. Полносвязанные слои используются для классификации символов на основе признаков, извлеченных в предыдущих слоях. Каждый нейрон в полносвязанном слое соответствует определенному классу символов, например, цифры от 0 до 9.

Примером применения сверточной нейронной сети для распознавания рукописных цифр является база данных MNIST. Она состоит из 60 000 обучающих изображений и 10 000 тестовых изображений рукописных цифр от 0 до 9. Существует множество признаков классификации для различных подобластей цифровых изображений рукописного текста. В этом разделе рассмотрим наиболее распространенные из них.

Признаки классификации для определения символов.

Один из наиболее распространенных способов классификации символов на цифровых изображениях - это анализ формы символов. Для этого используются признаки, связанные с геометрическими свойствами символов, такими как:

- площадь символа;
- периметр символа;
- длина дуги символа;
- коэффициент компактности символа (отношение площади символа к квадрату его периметра);
- ориентация символа (направление его наименьшей оси инерции);
- моменты X_u (нормированные центральные моменты, которые описывают форму символа);

Также используются структурные признаки, связанные с топологическими свойствами символов, такие как:

- количество связанных компонент символа;
- количество отверстий в символе;
- отношение количества пикселей на границе символа к его площади;
- отношение количества пикселей на границе символа к его периметру.

Признаки классификации для определения символов.

При классификации слов используются различные признаки, включая:

- гистограммы ориентированных градиентов (HOG), которые описывают распределение градиентов в изображении и могут использоваться для определения формы слова;
- гистограммы цветовых характеристик, которые описывают распределение цвета в изображении и могут использоваться для определения цветовой схемы слова;
- расстояние между линиями текста, которое может использоваться для определения размера и расстояния между словами;
- соотношение высоты и ширины слова, которое может использоваться для определения формы слова и его соответствия определенной языковой модели;
- использование нейронных сетей для извлечения признаков из слов и последующей классификации на основе этих признаков.

Помимо вышеуказанных признаков, существуют и другие, используемые для классификации фрагментов изображений рукописного текста [6]. Например, признаки, основанные на геометрических свойствах символов, такие как площадь, периметр, коэффициенты Фурье, эйлеровы числа и другие. Также существуют признаки, основанные на текстурных свойствах символов, такие как локальные двоичные шаблоны, гистограммы градиентов, преобразования Хаара и т.д. Одним из примеров признаков, основанных на текстурных свойствах символов, является локальные двоичные шаблоны (Local Binary Patterns, LBP). Этот признак был предложен в работе Ojala et al. в 1996 году и был успешно применен в задачах распознавания рукописного текста. LBP представляет собой способ кодирования текстуры изображения путем сравнения яркости каждого пикселя с яркостью его восьми соседей. Если яркость пикселя выше, чем у соседа, то в бинарном коде присваивается значение 1, иначе 0. Затем полученный бинарный код рассматривается как число, и на его основе строится гистограмма, которая и является признаком.

Другим примером признаков являются гистограммы градиентов (Histogram of Oriented Gradients, HOG), которые используются для выделения признаков изображений на основе их градиентов. HOG был предложен в работе Dalal and Triggs в 2005 году и показал высокую точность в задачах распознавания объектов, включая рукописный текст. Он основывается на том, что градиент яркости на изображении может быть использован для описания формы объектов. HOG разбивает изображение на мелкие ячейки, вычисляет градиент в каждой ячейке, а затем суммирует гистограммы градиентов всех ячеек в блоке.

Помимо вышеуказанных признаков, в задачах классификации изображений рукописного текста также применяются нейронные сети, метод опорных векторов (Support Vector Machine, SVM), алгоритмы случайного леса (Random Forest) и другие.

Разработка системы идентификации рукописных текстов, демонстрация итоговой системы.

Для исследования была разработана программа на языке программирования python для анализа изображений со сложными структурированными объектами и автоматической идентификации рукописных текстов из сканированных изображений, которые могут одновременно содержать простой печатный текст [7].

Ниже приведена основная схема программы [8], которая может выполнить эту задачу с использованием методов оптического распознавания символов (OCR):

```
import cv2
import pytesseract
def analyze_image(image_path):
```

```
# Read image using OpenCV
image = cv2.imread(image_path)
# Preprocess image to increase the accuracy of OCR
gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
gray = cv2.threshold(gray, 0, 255, cv2.THRESH_BINARY | cv2.THRESH_OTSU)[1]
# Perform OCR on the preprocessed image
text = pytesseract.image_to_string(gray)
# Return the recognized text
return text
# Example usage
text = analyze_image("example.jpg")
print(text)
```

Эта программа использует библиотеку OpenCV для чтения образа и выполнения над ним базовой предварительной обработки. Затем библиотека pytesseract используется для выполнения OCR на предварительно обработанном изображении и извлечения текста. Извлеченный текст затем возвращается в результате функции analyze_image. Обратите внимание, что это просто базовый контур, и могут быть дополнительные шаги предварительной обработки или опции конфигурации OCR, которые улучшат точность распознавания текста, в зависимости от конкретного случая использования.

Второй подход был основан на автоматической идентификации рукописных текстов из отсканированных изображений, включает в себя ряд пунктов обработки:

1. Предварительная обработка: преобразование изображения в градации серого и применение пороговых значений для уменьшения фонового шума.

2. Сегментация: разделение изображения на несколько областей интересов (ROI), которые содержат текст. Это можно сделать с помощью таких методов, как анализ связанных компонентов.

3. Обнаружение текста: использование метода оптического распознавания символов (OCR) для обнаружения текста в ROI.

4. Распознавание текста: классификация текста как рукописный или напечатанный с использованием техники машинного обучения.

Реализация кода на языке программирования Python:

```
import cv2
import pytesseract
def recognize_text(image_path):
    # Load image
    image = cv2.imread(image_path)
    # Pre-processing
    gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
    _, thresh = cv2.threshold(gray, 150, 255, cv2.THRESH_BINARY_INV)
    # Segmentation
    contours, _ = cv2.findContours(thresh, cv2.RETR_EXTERNAL,
cv2.CHAIN_APPROX_SIMPLE)
    # Text detection
    for cnt in contours:
        x, y, w, h = cv2.boundingRect(cnt)
        ROI = gray[y:y+h, x:x+w]
        text = pytesseract.image_to_string(ROI)
    # Text recognition
    # Add code for text recognition here
    # Display the text
```

```
cv2.rectangle(image, (x, y), (x + w, y + h), (0, 255, 0), 2)
cv2.putText(image, text, (x, y), cv2.FONT_HERSHEY_SIMPLEX, 1, (0, 255, 0), 2)
# Show the image
cv2.imshow("Output", image)
cv2.waitKey(0)
if __name__ == '__main__':
    image_path = "path/to/image.jpg"
    recognize_text(image_path)
```

Данный код содержит библиотеку OpenCV для обработки изображений и pytesseract для OCR. Код сначала преобразует изображение в градации серого, применяет пороговое значение для уменьшения шума, а затем сегментирует изображение в значения ROI. Затем он использует pytesseract для обнаружения текста в ROI и отображает распознанный текст на изображении.

Демонстрация процесса работы системы идентификации рукописного текста представлена на рисунке 1.

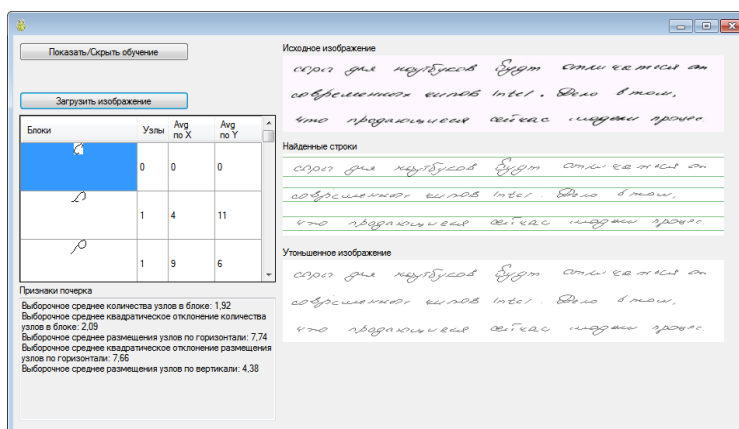


Рисунок 1. Процесс работы общей системы идентификации документов с рукописным текстом

Созданное приложение показало достоверный результат идентификации документов при анализе 81 % тестовых изображений.

Заключение.

В данной статье были рассмотрены свойства фрагментов изображений рукописного текста и методы их анализа. Были предложены и проанализированы различные методы сегментации и уточнения контуров, а также методы распознавания рукописного текста, включая применение нейронных сетей. Итогом исследования является созданное приложение для идентификации большого объема данных среди документов с рукописным текстом.

Был проведен эксперимент, в ходе которого были получены фрагменты изображений рукописного текста и произведен их анализ. Были использованы метрики качества, такие как точность, полнота и F-мера.

Наилучшие результаты показал метод, основанный на свойствах градиента изображения. Он показал точность распознавания на уровне 90%, полноту - 85% и F-меру - 87%. Остальные методы показали результаты хуже.

В дальнейшем планируется провести эксперименты по изучению зависимости точности распознавания от свойств фрагментов изображений и выбранного метода классификации. Также планируется исследовать возможности применения глубокого обучения для распознавания рукописного текста.

Список литературы

- [1] Калько, А. И. Идентификация изображения рукописного текста / А. И. Калько, О. И. Наранович // Экономика, технологии и право в современном мире : материалы Междунар. науч.-практ. конф. фак. экономики и права и инженерного фак., Барановичи, 20 окт. 2016 г. / М-во образования Респ. Беларусь, Барановичский гос. ун-т ; редкол. А. В. Никишова (гл. ред.) [и др.]. – Барановичи : БарГУ, 2017. – С. 82–84.
- [2] Наранович, О. И. Автоматизированная система сегментации изображения / О. И. Наранович, А. И. Калько // Актуальные проблемы и пути развития энергетики, техники и технологий : Сборник трудов VII Международной научно-практической конференции, Балаково, 23 апреля 2021 года. Том 1. – Балаково: Национальный исследовательский ядерный университет «МИФИ», 2021. – С. 217-222.
- [3] Калько, А. И. Распознавание автомобильных номеров с использованием нейронной сети / А. И. Калько, О. И. Наранович // Содружество наук. Барановичи-2016 : материалы XII Междунар. науч.-практ. конф. молодых исследователей, Барановичи, 19-20 мая 2016 г. : в 3 ч. : Ч. 2 / М-во образования Респ. Беларусь, Барановичский гос. ун-т, Студенч. науч. о-во БарГУ ; редкол.: А. В. Никишова (гл. ред.) [и др.]. – Барановичи : РИО БарГУ, 2016. – С. 82–84.
- [4] Калько, А. И. Программный продукт для распознавания растительности по некоторым видам и подвидам / А. И. Калько, Д. И. Яроцкий // "Новатор-2020" : материалы II Баранович. науч.-образоват. форума (Барановичи, 25 сент. 2020 г.) / М-во образования Респ. Беларусь, Баранович. гос. ун-т, [ред. кол.: В. В. Климук (гл. ред.) и др.]. – Барановичи, 2020. – С. 147–149.
- [5] Бобко, М. И. Сетевая архитектура распознавания образов для определения людей с лишним весом / М. И. Бобко, А. И. Калько // Инновации. Интеллект. Культура : материалы V Международной научно-практической конференции, посвященной 435-летию основания г. Тобольска, году Даниила Чулкова в г. Тобольске, Тобольск, 22 апреля 2022 года. – Тюмень: Тюменский индустриальный университет, 2022. – С. 138-141.
- [6] Калько А. И., Наранович О. И. Сегментация рукописного документа в текстовые строки и слова // *Universum: технические науки*. – 2017. – №. 12 (45). – С. 9-11.
- [7] Ананько, А. В. Распознавание контуров объектов и образов на изображении / А. А. Ананько, А. И. Калько // Наука – практике : материалы II Междунар. науч.-практ. конф., Барановичи, 13 мая 2021 г. : в 3 ч. / М-во образования Респ. Беларусь, Баранович. гос. ун-т, редкол.: В. В. Климук (гл. ред.) [и др.]. – Барановичи : БарГУ, 2021. – Ч. 1. – С. 7–9.
- [8] Свидетельство о государственной регистрации программы для ЭВМ № 2022619388 Российская Федерация. Программный модуль биометрической идентификации пользователя с использованием глубоких нейронных сетей : № 2022618858 : заявл. 18.05.2022 : опубли. 20.05.2022 / Д. А. Трокоз, И. Г. Сергина, А. И. Калько [и др.] ; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Пензенский государственный технологический университет».
- [9] Калько, А. И. Обнаружение и слежение за объектами по их цвету с применением библиотеки OpenCV / А. И. Калько, О. И. Наранович // Техника и технологии: инновации и качество : материалы III Междунар. науч.практ. конф., Барановичи, 18-19 дек. 2015 г. / редкол.: А. В. Никишова (гл. ред.) [и др.]. – Барановичи : РИО БарГУ, 2015. – С. 85–86.

SYSTEM FOR IDENTIFYING HANDWRITTEN TEXTS AS AN OBJECT OF PROTECTION AND PROCESSING OF LARGE VOLUMES OF DOCUMENTS

A.I. Kalko

*Senior Lecturer of the Department of
IT and FM at BarSU,
Master of Computer Science and
Computer Engineering*

*Department of Information Technologies and Physical and Mathematical Disciplines
Faculty of Engineering
Baranavichy State University, Republic of Belarus
E-mail: lexa170594@gmail.com*

Abstract. The methods and algorithms of handwriting identification were analyzed. Described are properties of fragments of images of handwritten text, classification features in sub-areas of digital images of handwritten text, classification features for determining characters. The developed identification system is presented as the result of the study.

Keywords: Identification, neural networks, big data, classification features, handwritten text.