Нижегородский государственный технический университет им. Р. Е. Алексеева

МОДЕЛЬ И АЛГОРИТМЫ ОБНАРУЖЕНИЯ ТЕКСТОВЫХ БЛОКОВ В ИЗОБРАЖЕНИИ ПЕЧАТНЫХ ДОКУМЕНТОВ

СТУДЕНТ: ЗАЙЦЕВ АНТОН АЛЕКСАНДРОВИЧ

НАУЧНЫЙ РУКОВОДИТЕЛЬ: К.Т.Н., ДОЦЕНТ ГАЙ ВАСИЛИЙ ЕВГЕН

Нижний Новгород, 2020 год

Цель и задачи

Целью исследования работы является разработка и исследование методов обнаружения текстовых блоков в изображении печатных документов.

Задачи:

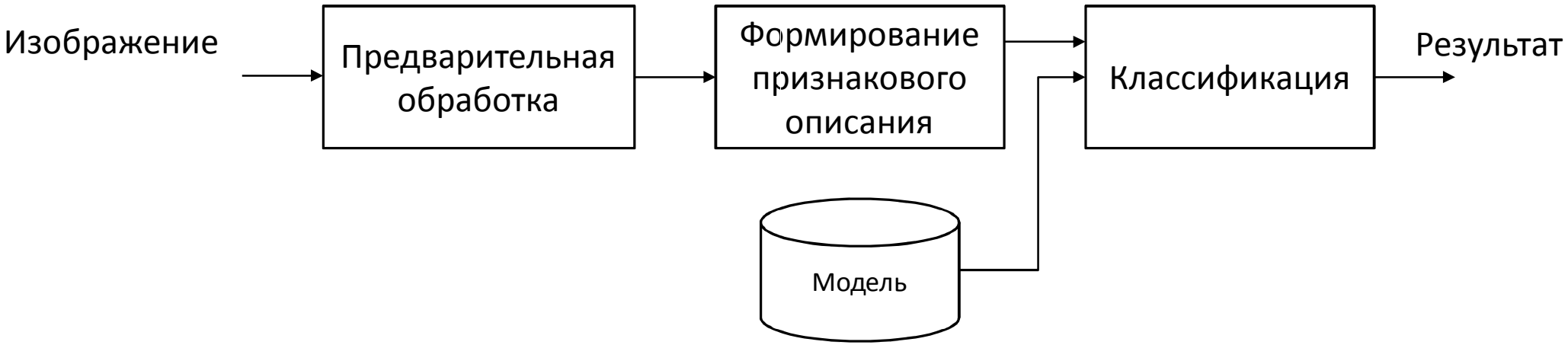
- Исследовать существующие методы обнаружения текстовых блоков в изображении;
- разработка модели и алгоритмов извлечения и хранения признакового описания (формирования входных данных) для алгоритмов классификации;
- разработка и тестирование программной реализации предложенного метода.

Существующие методы

- методы с использованием признаковых классификаторов;
- статистические методы;
- методы с использованием структурных составляющих.

Этапы решения задачи

Общая структурная схема



Предварительная обработка

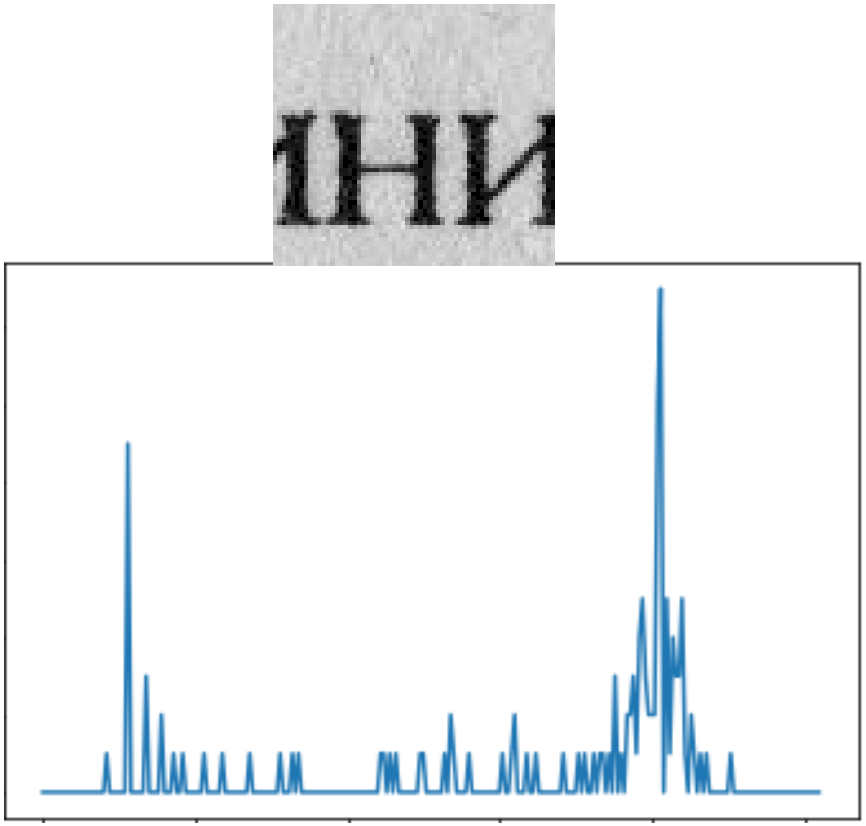
- Приведение к градациям серого

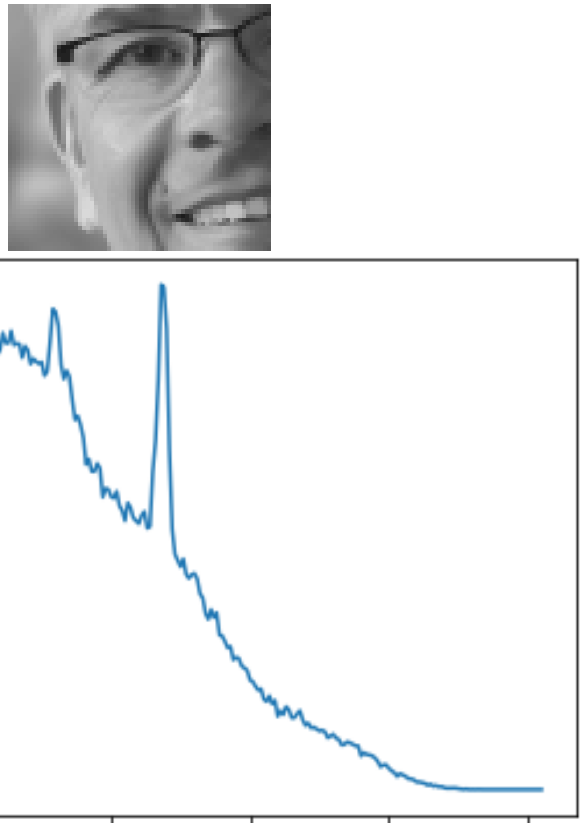
Формула для перевода изображения в градации серого:

$Y = 0.299 * R + 0.587 * G + 0.144 * B,$

где R, G, B – красный, зеленый и синий каналы исходного изображения соответственно

Метод Гистограмм





Формирование признакового описания

- Изображения разбиваются на блоки размером 64x64
- Строится гистограмма яркости блока
- Деление гистограммы на 8 частей
- Суммирование значений в каждой части
- Формирование вектора признаков (8 значений)

Принятие решения

- Метод опорных векторов
- К-ближайших соседей
- Алгоритм случайного леса
- Метод решающих деревьев

Вычислительный эксперимент

- Размер базы изображений : 100 фотографий каждого класса
- Итоговая выборка : 11000 изображений размера 64x64

$$x[(k+1)T] = e^{FT}x[kT] + \int_0 e^{F(T-\tau)}d\tau Du[kT] =$$

$$= e^{FT}x[kT] + \sum_{p=0}^{S-1}\int_{pT_0}^{(p+1)T_0} e^{F(T-\tau)}d\tau Du[kT+pT_0] =$$

$$= e^{FT}x[kT] + \sum_{p=0}^{S-1}\int_{pT_0}^{(p+1)T_0} e^{FT}e^{-F\tau}d\tau Du[kT+pT_0] =$$

$$= e^{FT}x[kT] + \sum_{p=0}^{S-1} e^{FT}\int_{pT_0}^{(p+1)T_0} e^{FT}d\tau Du[kT+pT_0] =$$

$$= e^{FT}x[kT] + \sum_{p=0}^{S-1} e^{FT}(-e^{-F\tau}\Big|_{pT_0}^{(p+1)T_0})F^{-1}Du[kT+$$

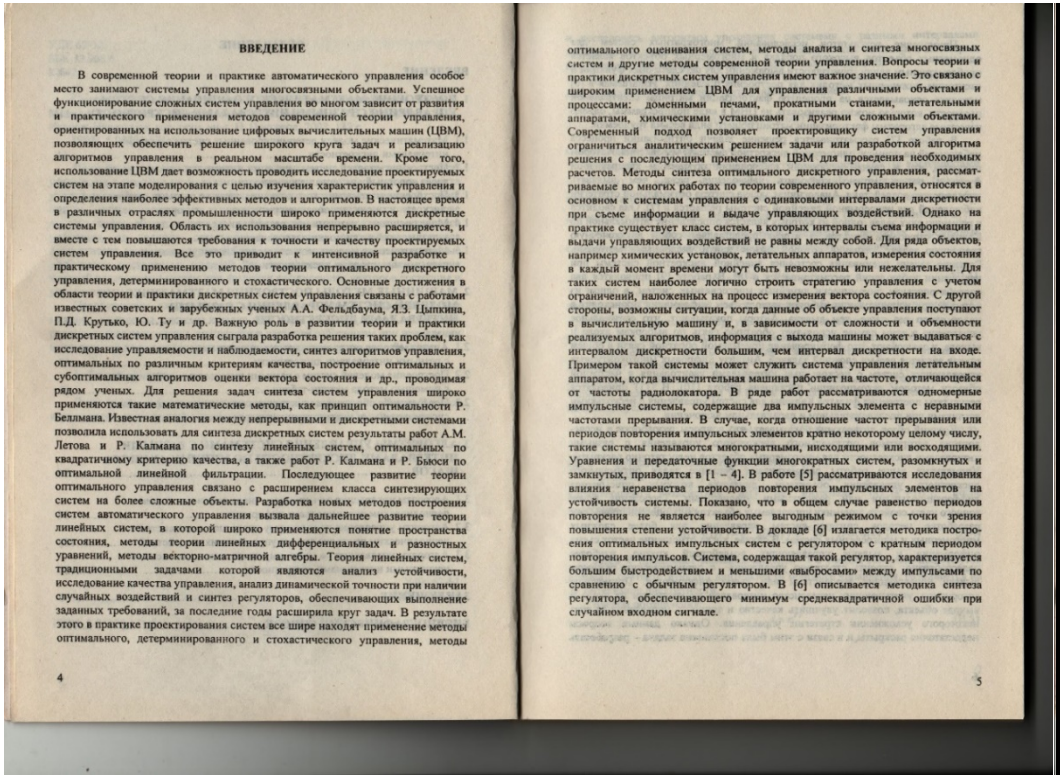
$$= e^{FT}x[kT] + \sum_{p=0}^{S-1} e^{FT}(-e^{-F(p+1)T_0} + e^{-FpT_0})F^{-1}Du$$

Результаты вычислительного эксперимента

Алгоритм	Accuracy	Precision	Recall	F1
sklearn.svm.SVC	0.9754	-	-	-
sklearn.neighbors.KNeighborsClassifier	0,9821	0.98	0,98	**0,98**
sklearn.tree.DecisionTreeClassifier	0.9862	0.99	0.99	**0.99**
sklearn.ensemble.RandomForestClassifier	**0.9904**	**0,99**	**0,99**	**0,99**

Тестирование проводилось методом гистограмм на изображениях с простым графическим фоном

Результаты вычислительного эксперимента



ВВЕДЕНИЕ

В современной теории и практике автоматического управления особое место занимают системы управления многосвязными объектами. Успешное функционирование сложных систем управления во многом зависит от развития и практического применения методов современной теории управления, ориентированных на использование цифровых вычислительных машин (ЦВМ), позволяющих обеспечить решение широкого круга задач и реализацию алгоритмов управления в реальном масштабе времени. Кроме того, использование ЦВМ дает возможность проводить исследование проектируемых систем на этапе моделирования с целью изучения характеристик управления и определения наиболее эффективных методов и алгоритмов. В настоящее время в различных отраслях промышленности широко применяются дискретные системы управления. Область их использования непрерывно расширяется, и вместе с тем повышаются требования к точности и качеству проектируемых систем управления. Все это приводит к интенсивной разработке и практическому применению методов теории оптимального дискретного управления, детерминированного и стохастического. Основные достижения в области теории и практики дискретных систем управления связаны с работами известных советских и зарубежных ученых А.А. Фельдбаума, Я.З. Цыпкина, П.Д. Крутько, Ю. Ту и др. Важную роль в развитии теории и практики дискретных систем управления сыграла разработка решения таких проблем, как исследование управляемости и наблюдаемости, синтез алгоритмов управления, оптимальных по различным критериям качества, построение оптимальных и субоптимальных алгоритмов оценки вектора состояния и др., проводимая рядом ученых. Для решения задач синтеза систем управления широко применяются такие математические методы, как принцип оптимальности Р. Беллмана. Известная аналогия между непрерывными и дискретными системами позволила использовать для синтеза дискретных систем результаты работ А.М. Летова и Р. Калмана по синтезу линейных систем, оптимальных по квадратичному критерию качества, а также работ Р. Калмана и Р. Бьюси по оптимальной линейной фильтрации. Последующее развитие теории оптимального управления связано с расширением класса синтезирующих систем на более сложные объекты. Разработка новых методов построения систем автоматического управления вызвала дальнейшее развитие теории линейных систем, в которой широко применяются понятие пространства состояния, методы теории линейных дифференциальных и разностных уравнений, методы векторно-матричной алгебры. Теория линейных систем, традиционными задачами которой являются анализ устойчивости, исследование качества управления, анализ динамической точности при наличии случайных воздействий и синтез регуляторов, обеспечивающих выполнение заданных требований, за последние годы расширила круг задач. В результате этого в практике проектирования систем все шире находят применение методы оптимального, детерминированного и стохастического управления, методы

4

оптимального оценивания систем, методы анализа и синтеза многосвязных систем и другие методы современной теории управления. Вопросы теории и практики дискретных систем управления имеют важное значение. Это связано с широким применением ЦВМ для управления различными объектами и процессами: доменными печами, прокатными станами, летательными аппаратами, химическими установками и другими сложными объектами. Современный подход позволяет проектировщику систем управления ограничиться аналитическим решением задачи или разработкой алгоритма решения с последующим применением ЦВМ для проведения необходимых расчетов. Методы синтеза оптимального дискретного управления, рассматриваемые во многих работах по теории современного управления, относятся в основном к системам управления с одинаковыми интервалами дискретности при съеме информации и выдаче управляющих воздействий. Однако на практике существует класс систем, в которых интервалы съема информации и выдачи управляющих воздействий не равны между собой. Для ряда объектов, например химических установок, летательных аппаратов, измерения состояния в каждый момент времени могут быть невозможны или нежелательны. Для таких систем наиболее логично строить стратегию управления с учетом ограничений, наложенных на процесс измерения вектора состояния. С другой стороны, возможны ситуации, когда данные об объекте управления поступают в вычислительную машину и, в зависимости от сложности и объемности реализуемых алгоритмов, информация с выхода машины может выдаваться с интервалом дискретности большим, чем интервал дискретности на входе. Примером такой системы может служить система управления летательным аппаратом, когда вычислительная машина работает на частоте, отличающейся от частоты радиолокатора. В ряде работ рассматриваются одномерные импульсные системы, содержащие два импульсных элемента с неравными частотами прерывания. В случае, когда отношение частот прерывания или периодов повторения импульсных элементов кратно некоторому целому числу, такие системы называются многократными, нисходящими или восходящими. Уравнения и передаточные функции многократных систем, разомкнутых и замкнутых, приводятся в [1 – 4]. В работе [5] рассматриваются исследования влияния неравенства периодов повторения импульсных элементов на устойчивость системы. Показано, что в общем случае равенство периодов повторения не является наиболее выгодным режимом с точки зрения повышения степени устойчивости. В докладе [6] излагается методика построения оптимальных импульсных систем с регулятором с кратным периодом повторения импульсов. Система, содержащая такой регулятор, характеризуется большим быстродействием и меньшими «выбросами» между импульсами по сравнению с обычным регулятором. В [6] описывается методика синтеза регулятора, обеспечивающего минимум среднеквадратичной ошибки при случайном входном сигнале.

5

ВВЕДЕНИЕ

В современной теории и практике автоматического управления особое место занимают системы управления многосвязными объектами. Успешное функционирование сложных систем управления во многом зависит от развития и практического применения методов современной теории управления, ориентированных на использование цифровых вычислительных машин (ЦВМ), позволяющих обеспечить решение широкого круга задач и реализацию алгоритмов управления в реальном масштабе времени. Кроме того, использование ЦВМ дает возможность проводить исследование проектируемых систем на этапе моделирования с целью изучения характеристик управления и определения наиболее эффективных методов и алгоритмов. В настоящее время в различных отраслях промышленности широко применяются дискретные системы управления. Область их использования непрерывно расширяется, и вместе с тем повышаются требования к точности и качеству проектируемых систем управления. Все это приводит к интенсивной разработке и практическому применению методов теории оптимального дискретного управления, детерминированного и стохастического. Основные достижения в области теории и практики дискретных систем управления связаны с работами известных советских и зарубежных ученых А.А. Фельдбаума, Я.З. Цыпкина, П.Д. Крутько, Ю. Ту и др. Важную роль в развитии теории и практики дискретных систем управления сыграла разработка решения таких проблем, как исследование управляемости и наблюдаемости, синтез алгоритмов управления, оптимальных по различным критериям качества, построение оптимальных и субоптимальных алгоритмов оценки вектора состояния и др., проводимая рядом ученых. Для решения задач синтеза систем управления широко применяются такие математические методы, как принцип оптимальности Р. Беллмана. Известная аналогия между непрерывными и дискретными системами позволила использовать для синтеза дискретных систем результаты работ А.М. Летова и Р. Калмана по синтезу линейных систем, оптимальных по квадратичному критерию качества, а также работ Р. Калмана и Р. Бьюси по оптимальной линейной фильтрации. Последующее развитие теории оптимального управления связано с расширением класса синтезирующих систем на более сложные объекты. Разработка новых методов построения систем автоматического управления вызвала дальнейшее развитие теории линейных систем, в которой широко применяются понятие пространства состояния, методы теории линейных дифференциальных и разностных уравнений, методы векторно-матричной алгебры. Теория линейных систем, традиционными задачами которой являются анализ устойчивости, исследование качества управления, анализ динамической точности при наличии случайных воздействий и синтез регуляторов, обеспечивающих выполнение заданных требований, за последние годы расширила круг задач. В результате этого в практике проектирования систем все шире находят применение методы оптимального, детерминированного и стохастического управления, методы

4

оптимального оценивания систем, методы анализа и синтеза многосвязных систем и другие методы современной теории управления. Вопросы теории и практики дискретных систем управления имеют важное значение. Это связано с широким применением ЦВМ для управления различными объектами и процессами: доменными печами, прокатными станами, летательными аппаратами, химическими установками и другими сложными объектами. Современный подход позволяет проектировщику систем управления ограничиться аналитическим решением задачи или разработкой алгоритма решения с последующим применением ЦВМ для проведения необходимых расчетов. Методы синтеза оптимального дискретного управления, рассматриваемые во многих работах по теории современного управления, относятся в основном к системам управления с одинаковыми интервалами дискретности при съеме информации и выдаче управляющих воздействий. Однако на практике существует класс систем, в которых интервалы съема информации и выдачи управляющих воздействий не равны между собой. Для ряда объектов, например химических установок, летательных аппаратов, измерения состояния в каждый момент времени могут быть невозможны или нежелательны. Для таких систем наиболее логично строить стратегию управления с учетом ограничений, наложенных на процесс измерения вектора состояния. С другой стороны, возможны ситуации, когда данные об объекте управления поступают в вычислительную машину и, в зависимости от сложности и объемности реализуемых алгоритмов, информация с выхода машины может выдаваться с интервалом дискретности большим, чем интервал дискретности на входе. Примером такой системы может служить система управления летательным аппаратом, когда вычислительная машина работает на частоте, отличающейся от частоты радиолокатора. В ряде работ рассматриваются одномерные импульсные системы, содержащие два импульсных элемента с неравными частотами прерывания. В случае, когда отношение частот прерывания или периодов повторения импульсных элементов кратно некоторому целому числу, такие системы называются многократными, нисходящими или восходящими. Уравнения и передаточные функции многократных систем, разомкнутых и замкнутых, приводятся в [1 – 4]. В работе [5] рассматриваются исследования влияния неравенства периодов повторения импульсных элементов на устойчивость системы. Показано, что в общем случае равенство периодов повторения не является наиболее выгодным режимом с точки зрения повышения степени устойчивости. В докладе [6] излагается методика построения оптимальных импульсных систем с регулятором с кратным периодом повторения импульсов. Система, содержащая такой регулятор, характеризуется большим быстродействием и меньшими «выбросами» между импульсами по сравнению с обычным регулятором. В [6] описывается методика синтеза регулятора, обеспечивающего минимум среднеквадратичной ошибки при случайном входном сигнале.

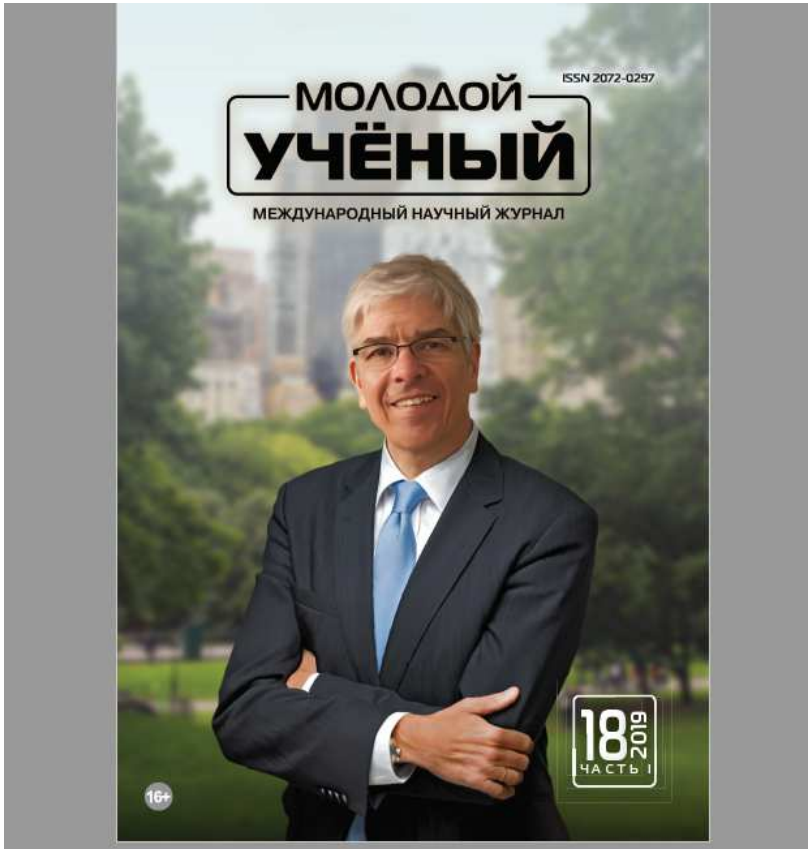
5

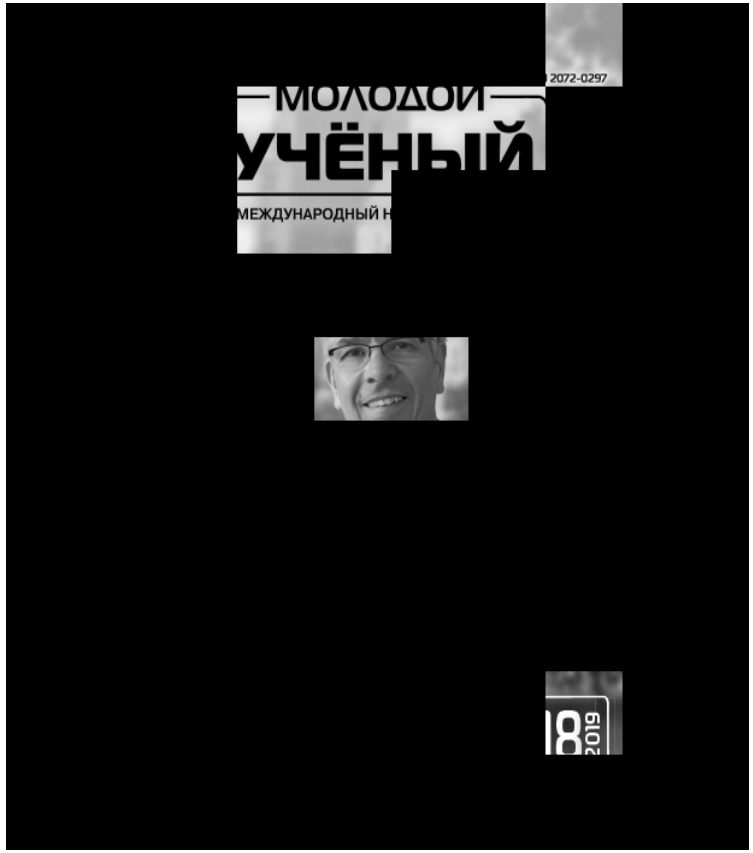
Результаты вычислительного эксперимента

Алгоритм	Accuracy	Precision	Recall	F1
sklearn.svm.SVC	0.6868	-	-	-
sklearn.neighbors.KNeighborsClassifier	0,6918	0.69	0,78	0,69
sklearn.tree.DecisionTreeClassifier	0.7006	0.70	0.81	0.70
sklearn.ensemble.RandomForestClassifier	0.7026	0.70	0.81	0.70

Тестирование проводилось методом гистограмм на изображениях с сложным графическим фоном

Результаты вычислительного эксперимента



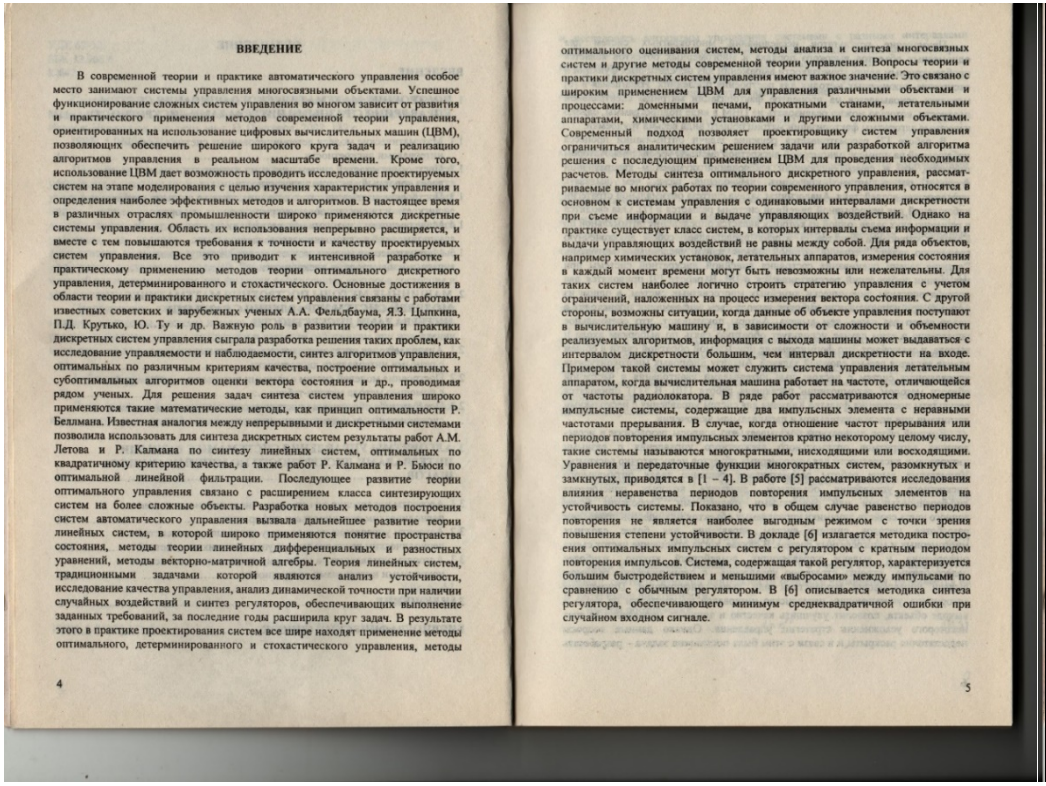


Результаты вычислительного эксперимента

Алгоритм	Accuracy	Precision	Recall	F1
sklearn.svm.SVC	0.9779	-	-	-
sklearn.neighbors.KNeighborsClassifier	0,9891	0.99	0,99	0,98
sklearn.tree.DecisionTreeClassifier	0.9883	0.99	0.99	0.99
sklearn.ensemble.RandomForestClassifier	0.9933	0,99	0,99	0,99

Тестирование проводилось методом Харалика на изображениях с простым графическим фоном

Результаты вычислительного эксперимента



ВВЕДЕНИЕ

В современной теории и практике автоматического управления особое место занимают системы управления многосвязными объектами. Успешное функционирование сложных систем управления во многом зависит от развития и практического применения методов современной теории управления, ориентированных на использование цифровых вычислительных машин (ЦВМ), позволяющих обеспечить решение широкого круга задач и реализацию алгоритмов управления в реальном масштабе времени. Кроме того, использование ЦВМ дает возможность проводить исследование проектируемых систем на этапе моделирования с целью изучения характеристик управления и определения наиболее эффективных методов и алгоритмов. В настоящее время в различных отраслях промышленности широко применяются дискретные системы управления. Область их использования непрерывно расширяется, и вместе с тем повышаются требования к точности и качеству проектируемых систем управления. Все это приводит к интенсивной разработке и практическому применению методов теории оптимального дискретного управления, детерминированного и стохастического. Основные достижения в области теории и практики дискретных систем управления связаны с работами известных советских и зарубежных ученых А.А. Фельдбаума, Я.З. Цыпкина, П.Д. Крутько, Ю. Ту и др. Важную роль в развитии теории и практики дискретных систем управления сыграла разработка решения таких проблем, как исследование управляемости и наблюдаемости, синтез алгоритмов управления, оптимальных по различным критериям качества, построение оптимальных и субоптимальных алгоритмов оценки вектора состояния и др., проводимая рядом ученых. Для решения задач синтеза систем управления широко применяются такие математические методы, как принцип оптимальности Р. Беллмана. Известная аналогия между непрерывными и дискретными системами позволила использовать для синтеза дискретных систем результаты работ А.М. Летова и Р. Калмана по синтезу линейных систем, оптимальных по квадратичному критерию качества, а также работ Р. Калмана и Р. Бьюси по оптимальной линейной фильтрации. Последующее развитие теории оптимального управления связано с расширением класса синтезирующих систем на более сложные объекты. Разработка новых методов построения систем автоматического управления вызвала дальнейшее развитие теории линейных систем, в которой широко применяются понятие пространства состояния, методы теории линейных дифференциальных и разностных уравнений, методы векторно-матричной алгебры. Теория линейных систем, традиционными задачами которой являются анализ устойчивости, исследование качества управления, анализ динамической точности при наличии случайных воздействий и синтез регуляторов, обеспечивающих выполнение заданных требований, за последние годы расширила круг задач. В результате этого в практике проектирования систем все шире находят применение методы оптимального, детерминированного и стохастического управления, методы

4

оптимального оценивания систем, методы анализа и синтеза многосвязных систем и другие методы современной теории управления. Вопросы теории и практики дискретных систем управления имеют важное значение. Это связано с широким применением ЦВМ для управления различными объектами и процессами: доменными печами, прокатными станами, летательными аппаратами, химическими установками и другими сложными объектами. Современный подход позволяет проектировщику систем управления ограничиться аналитическим решением задачи или разработкой алгоритма решения с последующим применением ЦВМ для проведения необходимых расчетов. Методы синтеза оптимального дискретного управления, рассматриваемые во многих работах по теории современного управления, относятся в основном к системам управления с одинаковыми интервалами дискретности при съеме информации и выдаче управляющих воздействий. Однако на практике существует класс систем, в которых интервалы съема информации и выдачи управляющих воздействий не равны между собой. Для ряда объектов, например химических установок, летательных аппаратов, измерения состояния в каждый момент времени могут быть невозможны или нежелательны. Для таких систем наиболее логично строить стратегию управления с учетом ограничений, наложенных на процесс измерения вектора состояния. С другой стороны, возможны ситуации, когда данные об объекте управления поступают в вычислительную машину и, в зависимости от сложности и объемности реализуемых алгоритмов, информация с выхода машины может выдаваться с интервалом дискретности большим, чем интервал дискретности на входе. Примером такой системы может служить система управления летательным аппаратом, когда вычислительная машина работает на частоте, отличающейся от частоты радиолокатора. В ряде работ рассматриваются одномерные импульсные системы, содержащие два импульсных элемента с неравными частотами прерывания. В случае, когда отношение частот прерывания или периодов повторения импульсных элементов кратно некоторому целому числу, такие системы называются многократными, нисходящими или восходящими. Уравнения и передаточные функции многократных систем, разомкнутых и замкнутых, приводятся в [1 – 4]. В работе [5] рассматриваются исследования влияния неравенства периодов повторения импульсных элементов на устойчивость системы. Показано, что в общем случае равенство периодов повторения не является наиболее выгодным режимом с точки зрения повышения степени устойчивости. В докладе [6] излагается методика построения оптимальных импульсных систем с регулятором с кратным периодом повторения импульсов. Система, содержащая такой регулятор, характеризуется большим быстродействием и меньшими «выбросами» между импульсами по сравнению с обычным регулятором. В [6] описывается методика синтеза регулятора, обеспечивающего минимум среднеквадратичной ошибки при случайном входном сигнале.

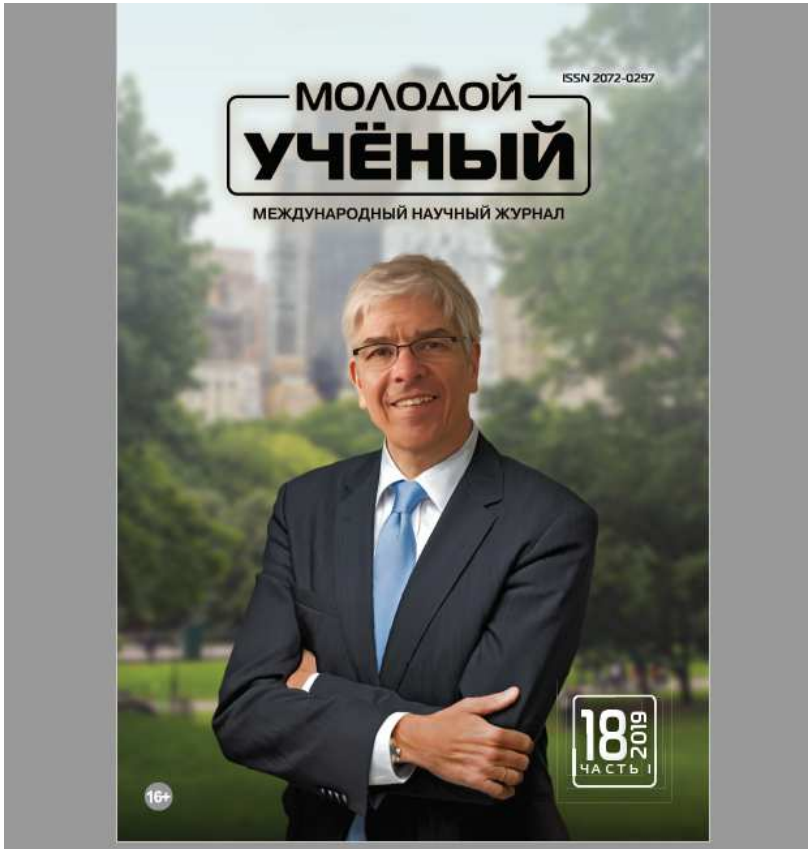
5

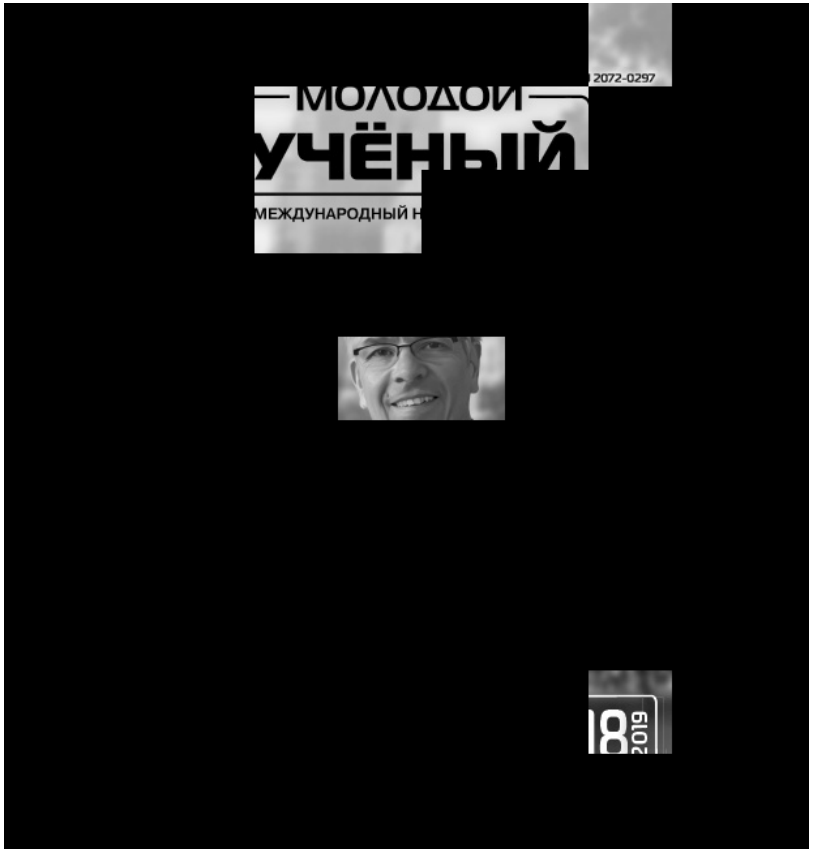
Результаты вычислительного эксперимента

Алгоритм	Accuracy	Precision	Recall	F1
sklearn.svm.SVC	0.6868	-	-	-
sklearn.neighbors.KNeighborsClassifier	0,6918	0.69	0,78	**0,69**
sklearn.tree.DecisionTreeClassifier	0.7005	0.70	0.81	**0.70**
sklearn.ensemble.RandomForestClassifier	**0.7026**	**0.70**	**0.81**	**0.70**

Тестирование проводилось методом гистограмм на изображениях с сложным графическим фоном

Результаты вычислительного эксперимента





Заключение

- Выполнен обзор существующих алгоритмов и методов обнаружения текстовых блоков
- Разработан новый метод решения задачи обнаружения текстовых блоков в изображении методом гистограмм
- Проведен вычислительный эксперимент, подтверждающий работоспособность и корректность данного подхода

Публикации

А.А. Зайцев, В.Е. Гай - Модель и алгоритмы обнаружения текстовых блоков на изображении печатных документов // Материалы XXVI международной научно-технической конференции «Информационные системы и технологии - 2020», ИСТ-2020, Россия, Н. Новгород, 2020г.

Спасибо за внимание!