В. В. Арлазаров, Методы комбинирования множественных результатов распознавания текста, *Искусственный интеллект и принятие решений*, 2022, выпуск 3, 106–116

DOI: 10.14357/20718594220309

Методы комбинирования множественных результатов распознавания текста

В. В. Арлазаров[I,II]

[I] Федеральный исследовательский центр «Информатика и управление» Российской академии наук», Москва, Россия
[II] ООО «Смарт Энджинс Сервис», Москва, Россия

Аннотация. Задача комбинирования результатов распознавания текста на множестве изображений является важным компонентом систем распознавания документов в видеопоследовательности. В настоящее время все еще нет общего подхода к решению этой задачи, дающего высокую точность распознавания текста. В работе проведен сравнительный анализ известных подходов к межкадровому комбинированию результатов распознавания полей идентификационных документов. Показано, что различные подходы имеют преимущество на разных частях пакетов данных, при этом потенциальный идеальный результат распознавания может значительно превосходить результаты, полученные проанализированными методами.

Ключевые слова: распознавание текста, анализ документов, распознавание видеопоследовательности, методы комбинирования, OCR, обработка изображений.

DOI 10.14357/20718594220309

Введение

В последние годы бурное развитие методов компьютерного зрения и машинного обучения позволяет строить методы решения все более сложных задач автоматизированного анализа и распознавания документов. Среди таких задач можно выделить задачу разбора сложно структурированных документов [1, 2], распознавание произвольно деформированного текста на естественных изображениях [3, 4], автоматическую проверку подлинности документов и поиск компрометирующих признаков [5, 6] и многие другие. Однако в области современного распознавания документов существуют нетривиальные задачи, относящиеся не столько непосредственного к компьютерному зрению, сколько к вопросам построения автоматизированных систем и к процессу автоматического принятия решений, которым в литературе не уделяется достаточно внимания. Одной из таких задач является комбинирование множественных результатов распознавания одного и того же текстового объекта.

Применительно к системам распознавания текста, задача комбинирования множественных результатов возникает в нескольких случаях. Иногда в системах массового ввода документов возникает возможность (или даже необходимость) повторного сканирования документов — как правило, в случае, если для архивного хранения документа качество загруженного в первый раз изображения недостаточно, либо какой-либо фрагмент изображения или все изображение не обладало достаточным качеством. Хотя самый очевидный способ обработки таких изображений является выбор одного варианта, который будет подаваться на распознавание и дальнейшую обработку. Некоторого увеличения точности распознавания можно до-

✉ *Арлазаров Владимир Викторович. E-mail: vva777@gmail.com*

стичь, распознав все варианты сканирования одного и того же документа и улучшить качество распознавания, используя комбинирование этих вариантов. К примеру, в задаче оцифровки сканированных страниц книг такой подход применялся в работе [7]. Похожая возможность возникает и при обработке идентификационных документах в KYC (Know Your Customer, Знай Своего Клиента) системах, где для удаленной идентификации человека (к примеру, для банковского обслуживания, страхования, и т.п.) пользователю необходимо загрузить фотографию или сканированное изображение своего паспорта или иного документа, удостоверяющего личность. Требования регуляторов, накладываемые на поставщиков таких сервисов, определяют параметры качества изображения, которые могут использоваться для обработки (в особенности в тех случаях, когда регулятор также требует от сервиса хранить изображение документа), и загруженное пользователем фотография или скан может этим требованиям не соответствовать. В таких случаях пользователю сообщается, что идентификационный документ необходимо загрузить повторно. И снова возникает возможность увеличить точность распознавания реквизитов документа, либо увеличить полноту их извлечения за счет использования множества изображений.

В наиболее явном виде задача комбинирования множественных результатов распознавания текста возникает при анализе видеопоследовательностей. Так, к примеру, при построении автоматизированных систем помощи водителю и систем полностью автоматического управления автомобилем [8, 9] нужно распознавать и анализировать текстовые элементы, присутствующие на дорожной сцене (к таким элементам могут относиться указатели, дорожные знаки, вывески, адреса улиц, регистрационные номера других автомобилей, текстовые элементы дорожной разметки и т.п.). Распознавание таких текстовых объектов должно проводиться в реальном времени (поскольку результат должен быть обработан системой принятия решений как можно раньше). Наличие множественных кадров, на которых присутствует текстовый объект, позволяет уточнять результат распознавания и повышать общее качество функционирования системы [10]. Другим примером является распознавание на мобильных устройствах — к примеру, распознавание документов, удостоверяющих личность, при помощи смартфонов [11]. При распознавании на мобильном устройстве имеется возможность использовать не одиночные фотографии, а захватываемую в реальном времени видеопоследовательность, для того чтобы увеличить ожидаемую точность распознавания и повысить надежность извлечения. При съемке идентификационных документов зачастую возникает ситуация, когда ни на одном одиночном кадре целевой реквизит не присутствует целиком (к примеру, из-за бликов, возникающих на отражающей поверхности документа [12]).

1. Задача и методы комбинирования результатов распознавания текста

Одни из наиболее ранних упоминаний задач, напоминающих задачу комбинирования результатов распознавания текста из нескольких источников, возникают в связи с задачей распознавания устной речи. В методах нередко использовались алгоритмы совмещения результатов распознавания фраз несколькими системами с целью увеличения общей точности распознавания [13, 14]. Постановку задачи комбинирования результатов распознавания нескольких изображений одного и того же объекта действительно можно выразить в том же виде, как выражают постановку задач комбинации классификаторов (или задачу «ансамблирования»). В таких постановках, как правило, рассуждают следующим образом [15]: выход каждого классификатора является шумной оценкой принадлежности распознаваемого объекта к некоторому классу, и шум от различных классификаторов можно считать независимым. Совмещая результаты от разных классификаторов можно попытаться минимизировать шум, получим тем самым наиболее точную оценку. В случае распознавания одной системой различных изображений, можно аналогично считать, что шум оценки принадлежности, полученной на каждом кадре, зависит от шума изображения и различается от изображения к изображению. Тогда можно попытаться минимизировать этот шум, совмещая результаты распознавания различных изображений. Следует отметить, что хотя задачу комбинирования результатов распознавания часто формулируют аналогично задаче построения ансамблей классификаторов, некоторые методы решения последней задачи, такие как методы групповой экспертной классифика-

ции многопризнаковых объектов [16, 17], согласно доступной литературе, к этой задаче не применялись.

Подходы к комбинированию межкадровой информации, описанные в литературе, можно условно разделить на две группы:

1. Методы, основанные на комбинировании нескольких *изображений*, ставящие в качестве цели получение единого представления объекта с более высоким «качеством», что позволило бы использовать классификатор одиночного изображения и достигнуть более высокой ожидаемой точности;

2. Методы, основанные на комбинировании *результатов* классификации нескольких одиночных изображений.

К методам первой группы можно отнести методы выбора наиболее информативного кадра [18, 19], методы “супер-разрешения”, получающие изображение с более высоким эффективным разрешением из нескольких кадров [20-22], методы слежения за конкретными объектами на кадрах видеопоследовательности и комбинирования локальных областей конкретных объектов [23], методы компенсации конкретных локальных искажений, таких как смазывание, путем замены локальных областей на соответствующие им области из других кадров, и методы, основанные на глубоком машинном обучении, принимающие на вход сразу множество изображений [24]. Здесь стоит отметить, что в системах мобильного распознавания, производящих съемку и вычисления на мобильном устройстве, в качестве входных данных можно использовать не только непосредственно кадры, полученные с камеры мобильного устройства, но и измерения с других сенсоров, таких как акселерометр и гироскоп. Однако даже для современных мобильных устройств ошибки измерения сенсоров могут быть настолько значительными, что использование таких данных для реконструкции изображений высокого качество может быть затруднено или невозможно, в особенности, если конкретное устройство неизвестно заранее [25]. Методы первой группы, предполагающие комбинирование нескольких изображений, могут обладать высокой трудоемкостью, чувствительностью к геометрическим искажениям и быть трудно расширяемыми на случай неизвестной заранее длины последовательностей изображений.

Методы второй группы используют правила комбинирования распределений, аналогичные правилам, используемым при ансамблировании классификаторов (правило суммы, произведения, максимума, медианы и т.п. [15, 26-28]). В отличие от методов первой группы, на методы второй группы могут сильно влиять свойства оценок принадлежности, порождаемые классификаторами одиночных изображений. Соответственно, выбор конкретного метода комбинирования может сильно зависеть от структуры и свойств классификатора.

В работах [29, 30] рассмотрена задача межкадрового комбинирования результатов распознавания текстовых полей паспорта РФ и приведены результаты экспериментального сравнения методов на основе подсчета кластеров идентичных результатов и на основе выравнивания входных строк при помощи модифицированного редакционного расстояния. В работе [31] построен более общий метод комбинирования результатов распознавания текстовых строк на основе предварительного выравнивания с учетом оценок принадлежности символов к классам. В работе [32] рассмотрены различные стратегии взвешивания одиночных результатов на уровне целиковых текстовых полей и на уровне отдельных символов. В ряде работ анализируется, как распределяются оценки принадлежности в видеопотоке, и как построить эффективную предиктивную модель этих оценок. Так, в работах [33, 34] была построена и исследована вероятностная модель результатов распознавания образов символов в видеопоследовательности и было предложено использовать распределение Коннора-Мосиманна для моделирования последовательности оценок принадлежности символов в видеопотоке.

Стоит, однако, отметить, что хотя и существуют работы, рассматривающие вопрос комбинирования множественных результатов распознавания по существу, в литературе все еще не предложено общего принципа, на основе которого можно было бы строить такие алгоритмы, и применять их в различных доменах с предсказуемым качеством. Возможно, именно это приводит к тому, что зачастую задачу межкадрового комбинирования авторы работ обходят стороной, упоминая лишь о том, что какой-либо метод комбинирования может быть применим, или рассматривают только какой-то один метод без какого-либо обоснования. Так, в

рамках упомянутой выше задачи распознавания текста на дорожной сцене, авторы массивного пакета данных RoadText1K [35] заявляют, что опубликованный ими пакет данных предназначен для исследований в области распознавания текстовых объектов дорожной сцены, дают подробное описание методов поиска и непосредственно распознавания текста, однако упоминают лишь подход голосования как способ комбинирования результатов из множества кадров.

2. Подходы к распознаванию текста документов

Выбор конкретного метода комбинирования результатов распознавания текста документов в видеопоследовательности может быть обусловлен требованиями, накладываемыми в целом на систему распознавания. Рассмотрим в качестве примера задачу распознавания текстовой строки как последовательности символов. Сравним три подхода к комбинированию покадровых результатов распознавания.

Подход № 1. Естественно предполагать, что на изображениях, на которых искажения вида «смаз» и «размытие» менее выражены, ожидаемая точность классификации будет выше [36]. Пусть задана функция $f : \mathbf{I} \to [0,1]$, реализующая оценку сфокусированности изображения, где 0 означает наименее сфокусированное изображение, 1 наиболее сфокусированное, с некоторым набором промежуточных градаций (к примеру, при помощи алгоритма, анализирующего гистограммы модуля градиента изображения в нескольких направлениях [37]). Выберем из множества изображений объекта $\{I_1(x), I_2(x), ..., I_n(x)\}$ единственное изображение $I_f(x)$, обладающее максимальной оценкой фокусировки: $I_f(x) = \arg\max_{k=1}^{n} f(I_k(x))$. В качестве результата распознавания видеопоследовательности примем результат распознавания выбранного изображения $I_f(x)$:

$$R_1^{(n)}(I_1(x), I_2(x), ..., I_n(x)) = r\left(\arg\max_{k=1}^{n} f(I_k(x))\right), \quad (1)$$

где r — классификатор одиночного изображения.

Подход № 2. Значения оценок принадлежности к классам каждого отдельного символа также можно использовать как апостериорный критерий качества, а именно: чем выше максимальное значение оценки принадлежности к классу, тем выше ожидаемая точность классификации. Пусть, при распознавании текстовой строки x, составленной из нескольких символов, на изображении $I(x)$ результатом является последовательность результатов классификации одиночных символов:

$$r(I(x)) = \begin{Bmatrix} \{(c_{11}, q_{11}), (c_{12}, q_{12}), ..., (c_{1M}, q_{1M})\}, \\ \{(c_{21}, q_{21}), (c_{22}, q_{22}), ..., (c_{2M}, q_{2M})\}, \\ ... \\ \{(c_{K1}, q_{K1}), (c_{K2}, q_{K2}), ..., (c_{KM}, q_{KM})\} \end{Bmatrix} \quad (2)$$

где K — число результатов классификации одиночных символов в ответе, M — число классов, $c_{ij} \in C$ — класс (символ из некоторого алфавита), $q_{ij} \in [0,1]$ — оценка принадлежности i-го символа к классу c_{ij}. Положим в качестве оценки качества результата распознавания строки x минимальное значение максимальной оценки принадлежности символа:

$$q(r(I(x))) = \min_{i=1}^{K} \max_{j=1}^{M} q_{ij}. \quad (3)$$

Выберем из множества результатов распознавания строки на одиночных изображениях $\{I_1(x), I_2(x), ..., I_n(x)\}$ единственный результат, обладающий максимальной оценкой качества, и в качестве результата распознавания видеопоследовательности примем этот результат:

$$R_2^{(n)}(I_1(x), I_2(x), ..., I_n(x)) = \operatorname{argmax}_{k=1}^{n} q(r(I_k(x))). \quad (4)$$

Подход № 3. Используем прямое комбинирование одиночных результатов распознавания текстовой строки методом ROVER [31] с правилом усреднения:

$$R_3^{(n)}(I_1(x), I_2(x), ..., I_n(x)) = \\ = \mathrm{ROVER}(r(I_1(x)), ..., r(I_n(x))). \quad (5)$$

В первую очередь эмпирически проанализируем характеристики достигаемой точности распознавания строк в видеопоследовательностях, с использованием трех описанных подходов. Сравним теперь характеристики точности распознавания строк в видеопоследовательно-

стях, достигаемые в указанных подходах. Для этого воспользуемся двумя открытыми пакетами данных MIDV-500 [38] и MIDV-2019 [39], разработанных для проведения исследований в области мобильного распознавания документов в видеопотоке. Пакет данных MIDV-500 состоит из 500 видеопоследовательностей идентификационных документов, снятых при помощи мобильных устройств, с незначительными геометрическими искажениями. Пакет MIDV-2019 содержит 200 видеопоследовательностей, снятых в условиях низкого освещения (подмножество MIDV-2019-L) и с сильными проективными искажениями (подмножество MIDV-2019-D). Проанализируем четыре группы текстовых полей – номер документа, даты, имя держателя документа, написанное латинским алфавитом, и машиночитаемые зоны. При сравнении распознанных значений игнорировался регистр, а также игнорировались различия между цифрой «0» и латинской буквой «O». В качестве метрики качества использовалось нормализованное расстояние Левенштейна [40] до истинного значения распознаваемой строки. Для распознавания текстовых строк применялся двухпроходный алгоритм [41].

Поскольку два из трех описанных подходов сводятся к выбору одного лучшего результата, для контроля также рассмотрим точность распознавания строки, которая достигалась бы при идеальном выборе (т.е. при выборе строки, заведомо наиближайшей к истинному значению).

Поскольку два из трех описанных подходов сводятся к выбору одного лучшего результата, для сравнения была также рассчитана возможная точность распознавания строки, достигаемая при идеальном выборе, т.е. при выборе строки, заведомо наиближайшей к истинному значению.

На Рис. 1, 2 и 3 представлены графики зависимости достигнутой точности распознавания строки с применением описанных выше подходов к комбинированию, и с контрольным методом идеального выбора, на пакетах данных MIDV-500, MIDV-2019-L и MIDV-2019-D, соответственно. Средние достигнутые значения нормализованного расстояния Левенштейна до истинного ответа после комбинирования на полях этих пакетов данных приведены в Табл. 1, 2 и 3.

На Рис. 1, 2 и 3 представлены графики зависимости точности распознавания строки достигнутой тремя методами комбинирования и методом идеального выбора на текстовых пакетах данных MIDV-500, MIDV-2019-L и MIDV-2019-D, соответственно. Средние значения нормализованного расстояния Левенштейна до истинного ответа для пакетов данных приведены в Табл. 1, 2 и 3.

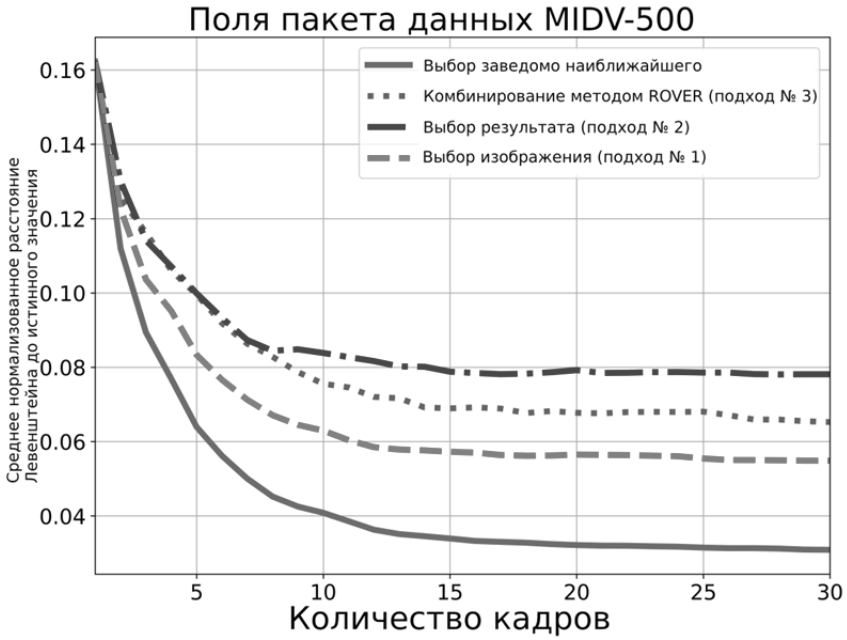

Рис. 1. Достигнутая точность распознавания строки при разных методах комбинирования, пакет данных MIDV-500

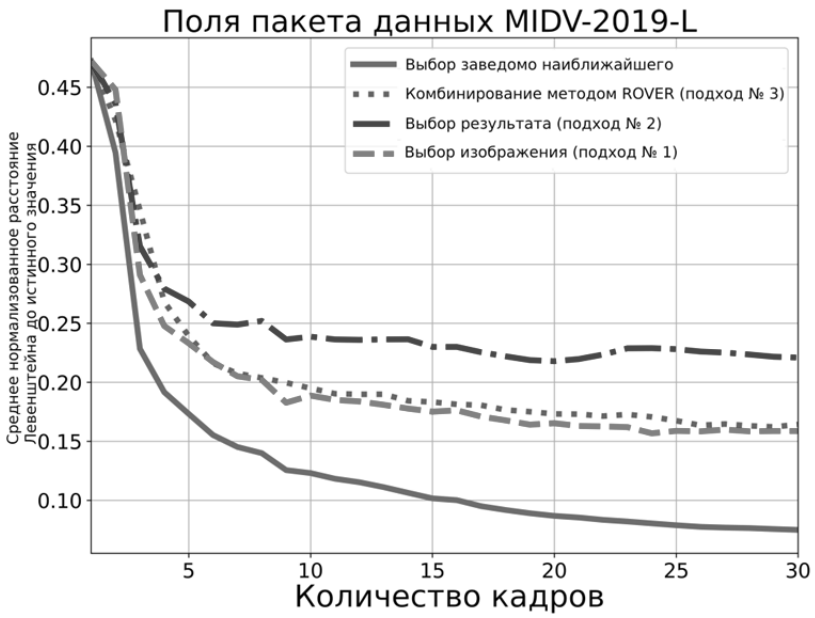

Рис. 2. Достигнутая точность распознавания строки при разных методах комбинирования, пакет данных MIDV-2019-L

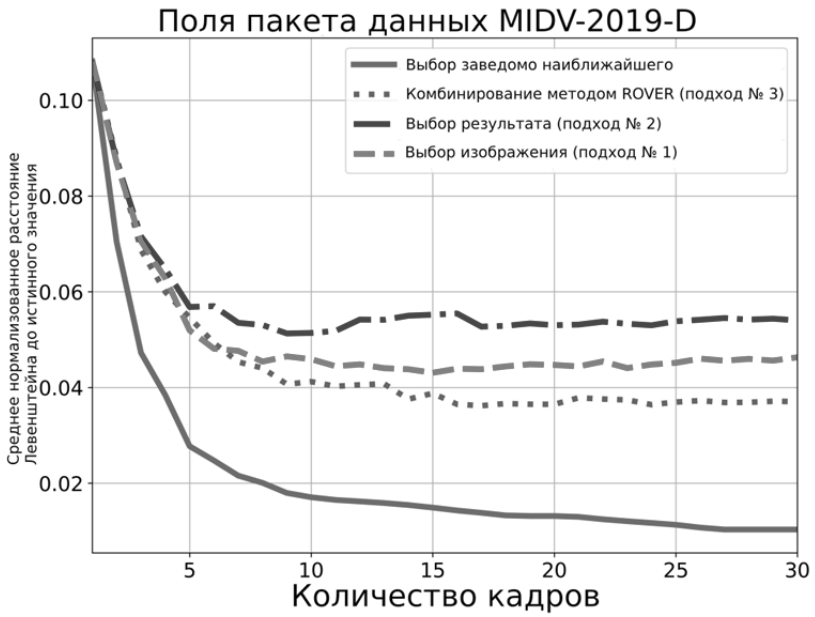

Рис. 3. Достигнутая точность распознавания строки при разных методах комбинирования, пакет данных MIDV-2019-D

Табл. 1. Среднее значение нормализованного расстояния Левенштейна до истинного ответа при разных методах комбинирования, пакет данных MIDV-500

Метод комбинирования	Количество комбинированных кадров					
	5	10	15	20	25	30
Подход № 1 (выбор изображения)	**0,0833**	**0,0629**	**0,0573**	**0,0565**	**0,0554**	**0,0549**
Подход № 2 (выбор результата)	0,0999	0,0838	0,0788	0,0792	0,0786	0,0781
Подход № 3 (ROVER)	0,0995	0,0756	0,0689	0,0677	0,0680	0,0652
Выбор заведомо наиближайшего	0,0639	0,0408	0,0339	0,0321	0,0315	0,0309

Табл. 2. Среднее значение нормализованного расстояния Левенштейна до истинного ответа при разных методах комбинирования, пакет данных MIDV-2019-L

Метод комбинирования	Количество комбинированных кадров					
	5	10	15	20	25	30
Подход № 1 (выбор изображения)	**0,2331**	**0,1888**	**0,1751**	**0,1653**	**0,1588**	**0,1587**
Подход № 2 (выбор результата)	0,2685	0,2386	0,2300	0,2179	0,2281	0,2209
Подход № 3 (ROVER)	0,2392	0,1950	0,1833	0,1732	0,1675	0,1643
Выбор заведомо наиближайшего	0,1733	0,1231	0,1017	0,0868	0,0789	0,0750

Табл. 3. Среднее значение нормализованного расстояния Левенштейна до истинного ответа при разных методах комбинирования, пакет данных MIDV-2019-D

Метод комбинирования	Количество комбинированных кадров					
	5	10	15	20	25	30
Подход № 1 (выбор изображения)	**0,0519**	0,0459	0,0431	0,0447	0,0452	0,0463
Подход № 2 (выбор результата)	0,0568	0,0514	0,0552	0,0530	0,0538	0,0540
Подход № 3 (ROVER)	0,0546	**0,0412**	**0,0388**	**0,0365**	**0,0370**	**0,0371**
Выбор заведомо наиближайшего	0,0277	0,0171	0,0149	0,0131	0,0113	0,0103

Как можно увидеть из Рис. 1 и Табл. 1, на пакете данных MIDV-500 наиболее высокое качество распознавания достигается при комбинировании методом выбора единственного изображения с максимальной оценкой сфокусированности (подход № 1). Тот же эффект наблюдается на подмножестве видеопоследовательностей в условиях низкого освещения из пакета данных MIDV-2019 (Рис. 2, Табл. 2), хотя различия между этим подходом и подходом полного комбинирования методом ROVER (подход № 3) уже становятся несущественными. На подмножестве видеопоследовательностей с сильными проективными искажениями из пакета данных MIDV-2019 (Рис. 3, Табл. 3) картина меняется: наиболее высокая точность распознавания достигается при полном комбинировании результатов распознавания методом ROVER. Тем самым, специфические свойства потока входных данных (или их распределение), может влиять на выбор оптимальной стратегии комбинирования результатов распознавания в видеопоследовательности.

Стоит также заметить, что во всех трех случаях наилучшее качество давала стратегия выбора одного результата (заведомо наиближайшего), не реализуемая на практике. Это говорит о том, что все еще может существовать критерий выбора единственного наилучшего результата, который бы консистентно показывал наилучшие результаты на всех трех пакетах данных.

Заключение

Задача комбинирования множественных результатов распознавания текста важна как для построения серверных автоматизированных систем обработки документов, включающих функцию уточнения результатов распознавания за счет использования множества изображений, так и для увеличения точности распознавания

текста в видеопоследовательности. К настоящему моменту уже накоплено достаточно открытых пакетов экспериментальных данных и описаны несколько различных подходов к решению задачи комбинирования, в том числе достаточных для внедрения в конкретные системы. Вместе с тем остается много нерешенных вопросов — какова общая модель решения такого рода задач, наиболее подходящая к различным сценариям, каким образом выбирать конкретную стратегию комбинирования для той или иной системы.

Как было показано в работе, комбинирование результатов распознавания позволяет значительно снизить среднюю ошибку, что говорит о необходимости более детально изучить методы и подходы к комбинированию результатов, а также уточнить связь между этой задачей и задачей ансамблирования классификаторов, для которой существуют эффективные методы, еще не опробованные в системах распознавания видеопоследовательностей.

Литература

1. S. C. Kosaraju, M. Masum, N. Z. Tsaku, P. Patel, T. Bayramoglu, G. Modgil, M. Kang. DoT-Net: Document layout classification using texture-based CNN" // International Conference on Document Analysis and Recognition (ICDAR), 2019, P. 1029-1034. DOI: 10.1109/ICDAR.2019.00168.
2. D. He, D. Cohen, B. Price, D. Kifer, C. L. Giles. Multi-scale multi-task FCN for semantic page segmentation and table detection" // International Conference on Document Analysis and Recognition (ICDAR), 2017, P. 254-261. DOI: 10.1109/ICDAR.2017.50.
3. F. Jia, C. Shi, Y. Wang, C. Wang, B. Xiao. "Grayscale-projection based optimal character segmentation for camera-captured faint text recognition" // International Conference on Document Analysis and Recognition, 2017, P. 1301-1306. DOI: 10.1109/ICDAR.2017.214.
4. J. Baek et al., "What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis" // IEEE/CVF International Conference on Computer Vision (ICCV), 2019, P. 4714-4722. DOI: 10.1109/ICCV.2019.00481.
5. H. Li, S. Wang, A. C. Kot. "Image recapture detection with convolutional and recurrent neural networks" // Electronic Imaging, 2017, P. 87-91. DOI: 10.2352/ISSN.2470-1173.2017.7.MWSF-329.
6. N. Yusoff, L. Alamro. "Implementation of feature extraction algorithms for image tampering detection" // International Journal of Advanced Computer Research, 2019, 9(43), P. 197-211. DOI: 10.19101/IJACR.PID37.
7. D. Wemhoener, I. Z. Yalniz, R. Manmatha, "Creating an Improved Version Using Noisy OCR from Multiple Editions" // International Conference on Document Analysis and Recognition (ICDAR), 2013, P. 160-164. DOI: 10.1109/ICDAR.2013.39.
8. R. Wang, S. M. Pizer, J.-M. Frahm, "Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth" // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, P. 5555–5564.
9. J. Jeong, Y. H. Yoon, J. H. Park, "Reliable road scene interpretation based on itom with the integrated fusion of vehicle and lane tracker in dense traffic situation" // Sensors 20, 2020, Article No. 2457. DOI:10.3390/s20092457.
10. K. B. Bulatov, N. V. Fedotova and V. V. Arlazarov, "An approach to road scene text recognition with per-frame accumulation and dynamic stopping decision," // International Conference on Machine Vision(ICMV 2020), 2021, V. 11605, P. 116051S1-116051S10. DOI: 10.1117/12.2586912.
11. K. B. Bulatov, P. V. Bezmaternykh, D. P. Nikolaev and V. V. Arlazarov, "Towards a unified framework for identity documents analysis and recognition" // Computer Optics, 2022, V. 46, N. 3, P. 436-454, DOI: 10.18287/2412-6179-CO-1024.
12. Д. В. Полевой, К. Б. Булатов, Н. С. Скорюкина, Т. С. Чернов, В. В. Арлазаров, А. В. Шешкус. Ключевые аспекты распознавания документов с использованием малоразмерных цифровых камер // Вестник РФФИ. 2016. № 4. С. 97-108. DOI: 10.22204/2410-4639-2016-092-04-97-108.
13. T. Kohonen, "Median strings" // Pattern Recognition Letters, V. 3, N. 5, 1985, P. 309-313. DOI: 10.1016/0167-8655(85)90061-3.
14. J. G. Fiscus. "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)." // IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, 1997, P. 347-354.
15. J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, "On combining classifiers" // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, V. 20, N. 3, P. 226-239, DOI: 10.1109/34.667881.
16. А. Б. Петровский. Методы групповой классификации многопризнаковых объектов (часть 1) // Искусственный интеллект и принятие решений. 2009. №. 3. С. 3-14.
17. А. Б. Петровский. Методы групповой классификации многопризнаковых объектов (часть 2) // Искусственный интеллект и принятие решений. 2009. №. 4. С. 3-14.
18. D. V. Polevoy, M. A. Aliev, D. P. Nikolaev. "Choosing the best image of the document owner's photograph in the video stream on the mobile device" // International Conference on Machine Vision (ICMV 2020), 2021, V. 11605, P. 1-9. DOI: 10.1117/12.2586939.
19. C. Zhanzhan, L. Jing, N. Yi, P. Shiliang, W. Fei, Z. Shuigeng. "You only recognize once: Towards fast video text spotting" // 27th ACM International Conference, 2019, P. 855-863. DOI: 10.1145/3343031.3351093.
20. В. Л. Арлазаров, О. А. Славин, В. В. Фарсобина В. В. "Алгоритмы поиска оптимального положения образов при их суммировании" // Искусственный интеллект и принятие решений. 2015. № 2. С. 25-34.

21. M. Haris, G. Shakhnarovich, N. Ukita. “Recurrent backprojection network for video super-resolution” // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019, P. 3897-3906. DOI: 10.1109/CVPR.2019.00402.
22. K. Mehregan, A. Ahmadyfard, H. Khosravi. “Super-resolution of license-plates using frames of low-resolution video” // 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), 2019, P. 1-6. DOI: 10.1109/ICSPIS48872.2019.9066104.
23. C. Merino-Gracia, M. Mirmehdi. “Real-time text tracking in natural scenes” // IET Computer Vision, 2014, 8(6), P. 670-681. DOI: 10.1049/iet-cvi.2013.0217.
24. S. Zhang, P. Li, Y. Meng, L. Li, Q. Zhou, X. Fu. “A video deblurring algorithm based on motion vector and an encoder-decoder network” // IEEE Access, 2019, V. 7, P. 86778-86788. DOI: 10.1109/ACCESS.2019.2923759.
25. V. V. Myasnikov, E. A. Dmitriev. “The accuracy dependency investigation of simultaneous localization and mapping on the errors from mobile device sensors” // Computer Optics, 2019, V. 43, N. 3, P. 492-503. DOI: 10.18287/2412-6179-2019-43-3-492-503.
26. К. Б. Булатов. “Выбор оптимальной стратегии комбинирования покадровых результатов распознавания символа в видеопотоке” // Информационные технологии и вычислительные системы, 2017, Т. 3, С. 45-55.
27. R. Polikar. “Ensemble based systems in decision making” // IEEE Circuits and Systems Magazine, 2006, V. 6, N. 3, P. 21-45. DOI: 10.1109/MCAS.2006.1688199.
28. Z. H. Zhou “Ensemble methods: Foundations and algorithms”. New York: Chapman and Hall/CRC, 2012, ISBN: 978-1-4398-3003-1.
29. К. Б. Булатов, В. Ю. Кирсанов, В. В. Арлазаров, Д. П. Николаев, Д. В. Полевой. “Методы интеграции результатов распознавания текстовых полей документов в видеопотоке мобильного устройства” // Вестник РФФИ, 2016, № 4, С. 109-115. DOI: 10.22204/2410-4639-2016-092-04-109-115.
30. Т. И. Булдакова, О. А. Славин, Д. Н. Путинцев. “Алгоритмы интеграции результатов распознавания в видеопоследовательностях полей документов, удостоверяющих личность” // Международный журнал прикладных и фундаментальных исследований, 2017, № 7, часть 2, С. 172-175.
31. K. B. Bulatov, “A Method to Reduce Errors of String Recognition Based on Combination of Several Recognition Results with Per-Character Alternatives” // Bulletin of the South Ural State University, Series: Mathematical Modelling, Programming and Computer Software, V. 12, N 3, P. 74-88, 2019, DOI: 10.14529/mmp190307.
32. O. Petrova, K. Bulatov, V. L. Arlazarov, “Methods of weighted combination for text field recognition in a video stream” // International Conference on Machine Vision (ICMV 2019), 2020, V. 11433, 11433 2L, P. 1-7, DOI: 10.1117/12.2559378.
33. E. Andreeva, V. V. Arlazarov, O. Slavin, I. Janiszewski. “Experimental modeling the flow of character recognition results in video stream for document recognition” // International Conference on Machine Vision(ICMV 2018), 2019, V. 11041, 110411L, P. 1-6, DOI: 10.1117/12.2522970.
34. V. V. Arlazarov, O. A. Slavin, A. V. Uskov,I. M. Janiszewski, “Modelling the flow of character recognition results in video stream” // Bulletin of the South Ural State University, Series: Mathematical Modelling, Programming and Computer Software, V. 11, N 2, P. 14-28, 2018. DOI: 10.14529/mmp180202.
35. S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas., C. V. Jawahar, “Roadtext-1k: Text detection & recognition dataset for driving videos” // arXiv preprint 2005.09496, 2020.
36. Т. С. Чернов, Н. П. Разумный, А. С. Кожаринов, Д. П. Николаев, В. В. Арлазаров. “Оценка качества входных изображений в системах распознавания видеопотока” // Информационные технологии и вычислительные системы, 2017, № 4, С. 71-82.
37. K. Bulatov, D. Polevoy, “Reducing overconfidence in neural networks by dynamic variation of recognizer relevance” // European Conference on Modelling and Simulation (ECMS 2015), 2015, P. 488-491. DOI: 10.7148/2015-0488.
38. V. V. Arlazarov, K. Bulatov, T. Chernov, V. L. Arlazarov, “MIDV-500: A Dataset for Identity Document Analysis and Recognition on Mobile Devices in Video Stream” // Computer Optics, V. 43, N 5, P. 818-824, 2019. DOI: 10.18287/2412-6179-2019-43-5-818-824.
39. K. Bulatov, D. Matalov, V. V. Arlazarov, “MIDV-2019: Challenges of the Modern Mobile-Based Document OCR” // International Conference on Machine Vision(ICMV 2019), V. 11433, 114332N, P. 1-6, 2020. DOI: 10.1117/12.2558438.
40. L. Yujian, L. Bo. “A normalized Levenshtein distance metric” // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6), P. 1091-1095. DOI: 10.1109/TPAMI.2007.1078.
41. Y. S. Chernyshova, A. V. Sheshkus, V. V. Arlazarov, “Two-step CNN framework for text line recognition in camera-captured images” // IEEE Access, V. 8, P. 32587-32600, 2020. DOI: 10.1109/ACCESS.2020.2974051.

Арлазаров Владимир Викторович. Кандидат технических наук. Заведующий отделом Федерального исследовательского центра «Информатика и управление» Российской академии наук» Области исследований: искусственный интеллект, машинное обучение, системы распознавания, информационные технологии. E-mail: vva777@gmail.com

Methods for Combining Multiple Text Recognition Results

V. V. Arlazarov [I,II]

[I] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia
[II]Smart Engines Service LLC, Moscow, Russia

Abstract. The task of per-frame combination of text recognition results from multiple images is an important component of video stream document recognition systems. Currently there is no unified approach to solving this problem which would yield a high precision of text recognition. In this paper a comparative study is presented of known approaches to the combination of recognition results for identity document fields. It was demonstrated that different approaches are advantageous on different parts of the data sets, while a sepection of the potential best single result can still significantly outperform all the analyzed methods.

Keywords: text recognition, document analysis, video stream recognition, combination methods, OCR, image processing.

DOI 10.14357/20718594220309

References

1. S. C. Kosaraju, M. Masum, N. Z. Tsaku, P. Patel, T. Bayramoglu, G. Modgil, M. Kang. "DoT-Net: Document layout classification using texture-based CNN" // International Conference on Document Analysis and Recognition (ICDAR), 2019, P. 1029-1034. DOI: 10.1109/ICDAR.2019.00168.
2. D. He, D. Cohen, B. Price, D. Kifer, C. L. Giles. "Multi-scale multi-task FCN for semantic page segmentation and table detection" // International Conference on Document Analysis and Recognition (ICDAR), 2017, P. 254-261. DOI: 10.1109/ICDAR.2017.50.
3. F. Jia, C. Shi, Y. Wang, C. Wang, B. Xiao. "Grayscale-projection based optimal character segmentation for camera-captured faint text recognition" // International Conference on Document Analysis and Recognition, 2017, P. 1301-1306. DOI: 10.1109/ICDAR.2017.214.
4. J. Baek et al., "What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis" // IEEE/CVF International Conference on Computer Vision (ICCV), 2019, P. 4714-4722. DOI: 10.1109/ICCV.2019.00481.
5. H. Li, S. Wang, A. C. Kot. "Image recapture detection with convolutional and recurrent neural networks" // Electronic Imaging, 2017 (7): P. 87-91. DOI: 10.2352/ISSN.2470-1173.2017.7.MWSF-329.
6. N. Yusoff, L. Alamro. "Implementation of feature extraction algorithms for image tampering detection" // International Journal of Advanced Computer Research, 2019, 9(43), P. 197-211. DOI: 10.19101/IJACR.PID37.
7. D. Wemhoener, I. Z. Yalniz, R. Manmatha, "Creating an Improved Version Using Noisy OCR from Multiple Editions" // International Conference on Document Analysis and Recognition (ICDAR), 2013, P. 160-164. DOI: 10.1109/ICDAR.2013.39.
8. R. Wang, S. M. Pizer, J.-M. Frahm, "Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth" // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, P. 5555–5564.
9. J. Jeong, Y. H. Yoon, J. H. Park, "Reliable road scene interpretation based on itom with the integrated fusion of vehicle and lane tracker in dense traffic situation" // Sensors 20, 2020, Article No. 2457. DOI:10.3390/s20092457.
10. K. B. Bulatov, N. V. Fedotova and V. V. Arlazarov, "An approach to road scene text recognition with per-frame accumulation and dynamic stopping decision," // International Conference on Machine Vision(ICMV 2020), 2021, V. 11605, P. 116051S1-116051S10. DOI: 10.1117/12.2586912.
11. K. B. Bulatov, P. V. Bezmaternykh, D. P. Nikolaev and V. V. Arlazarov, "Towards a unified framework for identity documents analysis and recognition" // Computer Optics, V. 46, N. 3, P. 436-454, 2022. DOI: 10.18287/2412-6179-CO-1024.
12. D. V. Polevoy, K. B. Bulatov, N. S. Skoryukina, T. S. Chernov, V. V. Arlazarov, A. V. Sheshkus, "Key Aspects of Document Recognition Using Small Digital Cameras" // Vestnik RFFI, N. 4, P. 97-108, 2016. DOI: 10.22204/2410-4639-2016-092-04-97-108.
13. T. Kohonen, "Median strings" // Pattern Recognition Letters, V. 3, N. 5, 1985, P. 309-313. DOI: 10.1016/0167-8655(85)90061-3.
14. J. G. Fiscus. "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)." // IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, 1997, P. 347-354.
15. J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, "On combining classifiers" // IEEE Transactions on Pattern Analysis and Machine Intelligence, V. 20, N. 3, P. 226-239, 1998. DOI: 10.1109/34.667881.
16. A. B. Petrovsky. Methods of group classification of multi-featured objects (part 1) // Artificial intelligence and decision making. 2009. No 3. P. 3-14.
17. A. B. Petrovsky. Methods of group classification of multi-featured objects (part 2) // Artificial intelligence and decision making. 2009. No 4. P. 3-14.

18. D. V. Polevoy, M. A. Aliev, D. P. Nikolaev. “Choosing the best image of the document owner’s photograph in the video stream on the mobile device” // International Conference on Machine Vision (ICMV 2020), V. 11605, P. 1-9, 2021. DOI: 10.1117/12.2586939.
19. C. Zhanzhan, L. Jing, N. Yi, P. Shiliang, W. Fei, Z. Shuigeng. “You only recognize once: Towards fast video text spotting” // 27th ACM International Conference, 2019, P. 855-863. DOI: 10.1145/3343031.3351093.
20. V. L. Arlazarov, O. A. Slavin, V. V. Farsobina. “Algorithms for finding optional position of images upon summation” // Artificial intelligence and decision making, V. 2, P. 25-34, 2015.
21. M. Haris, G. Shakhnarovich, N. Ukita. “Recurrent backprojection network for video super-resolution” // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019, P. 3897-3906. DOI: 10.1109/CVPR.2019.00402.
22. K. Mehregan, A. Ahmadyfard, H. Khosravi. “Super-resolution of license-plates using frames of low-resolution video” // 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), 2019, P. 1-6. DOI: 10.1109/ICSPIS48872.2019.9066104.
23. C. Merino-Gracia, M. Mirmehdi. “Real-time text tracking in natural scenes” // IET Computer Vision, 2014, 8(6), P. 670-681. DOI: 10.1049/iet-cvi.2013.0217.
24. S. Zhang, P. Li, Y. Meng, L. Li, Q. Zhou, X. Fu. “A video deblurring algorithm based on motion vector and an encoder-decoder network” // IEEE Access, 2019, V. 7, P. 86778-86788. DOI: 10.1109/ACCESS.2019.2923759.
25. V. V. Myasnikov, E. A. Dmitriev. “The accuracy dependency investigation of simultaneous localization and mapping on the errors from mobile device sensors” // Computer Optics, 2019, V. 43, N. 3, P. 492-503. DOI: 10.18287/2412-6179-2019-43-3-492-503.
26. K. B. Bulatov, “Selecting optimal strategy for combining per-frame character recognition results in video stream” // Information technologies and computing systems, V. 3, 2017, P. 45-55.
27. R. Polikar. “Ensemble based systems in decision making” // IEEE Circuits and Systems Magazine, 2006, V. 6, N. 3, P. 21-45. DOI: 10.1109/MCAS.2006.1688199.
28. Z. H. Zhou “Ensemble methods: Foundations and algorithms”. New York: Chapman and Hall/CRC, 2012, ISBN: 978-1-4398-3003-1.
29. K. B. Bulatov, V. Y. Kirsanov, V. V. Arlazarov, D. P. Nikolaev, D. V. Polevoy, “Methods for integration of document text field recognition results in a videostream of a mobile device” // Vestnik RFFI, N 4, P. 109-115, 2016. DOI: 10.22204/2410-4639-2016-092-04-109-115.
30. T. I. Buldakova, O. A. Slavin, D. N. Putintsev, “Algorithms for integration of video stream recognition results of identity document fields” // Mezhdunarodnyy zhurnal prikladnykh i fundamentalnykh issledovaniy, N. 7, part 2, P. 172-175, 2017.
31. K. B. Bulatov, “A Method to Reduce Errors of String Recognition Based on Combination of Several Recognition Results with Per-Character Alternatives” // Bulletin of the South Ural State University, Series: Mathematical Modelling, Programming and Computer Software, V. 12, N 3, P. 74-88, 2019, DOI: 10.14529/mmp190307.
32. O. Petrova, K. Bulatov, V. L. Arlazarov, “Methods of weighted combination for text field recognition in a video stream” // International Conference on Machine Vision (ICMV 2019), V. 11433, 11433 2L, P. 1-7, 2020, DOI: 10.1117/12.2559378.
33. E. Andreeva, V. V. Arlazarov, O. Slavin, I. Janiszewski. “Experimental modeling the flow of character recognition results in video stream for document recognition” // International Conference on Machine Vision(ICMV 2018), V. 11041, 110411L, P. 1-6, 2019, DOI: 10.1117/12.2522970.
34. V. V. Arlazarov, O. A. Slavin, A. V. Uskov,I. M. Janiszewski, “Modelling the flow of character recognition results in video stream” // Bulletin of the South Ural State University, Series: Mathematical Modelling, Programming and Computer Software, V. 11, N 2, P. 14-28, 2018. DOI: 10.14529/mmp180202.
35. S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas., C. V. Jawahar, “Roadtext-1k: Text detection & recognition dataset for driving videos” // arXiv preprint 2005.09496, 2020.
36. T. S. Chernov, N. P. Razumnyy, A. S. Kozharinov, D. P. Nikolaev, V. V. Arlazarov, “Image quality assessment for video stream recognition systems.” // Infiormation technologies and computing systems, N 4, P. 71-82, 2017.
37. K. Bulatov, D. Polevoy, “Reducing overconfidence in neural networks by dynamic variation of recognizer relevance” // European Conference on Modelling and Simulation (ECMS 2015), 2015, P. 488-491. DOI: 10.7148/2015-0488.
38. V. V. Arlazarov, K. Bulatov, T. Chernov, V. L. Arlazarov, “MIDV-500: A Dataset for Identity Document Analysis and Recognition on Mobile Devices in Video Stream” // Computer Optics, V. 43, N 5, P. 818-824, 2019. DOI: 10.18287/2412-6179-2019-43-5-818-824.
39. K. Bulatov, D. Matalov, V. V. Arlazarov, “MIDV-2019: Challenges of the Modern Mobile-Based Document OCR” // International Conference on Machine Vision(ICMV 2019), V. 11433, 114332N, P. 1-6, 2020. DOI: 10.1117/12.2558438.
40. L. Yujian, L. Bo. “A normalized Levenshtein distance metric” // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6), P. 1091-1095. DOI: 10.1109/TPAMI.2007.1078.
41. Y. S. Chernyshova, A. V. Sheshkus, V. V. Arlazarov, “Two-step CNN framework for text line recognition in camera-captured images” // IEEE Access, V. 8, P. 32587-32600, 2020. DOI: 10.1109/ACCESS.2020.2974051.

Arlazarov Vladimir V. Candidate of technical sciences. Head of the Department, Federal Research Center "Computer Science and Control", the Russian Academy of Sciences. Research areas: artificial intelligence, machine learning, recognition systems, information technologies. E-mail: vva777@gmail.com