УДК 004.932.75'1

АЛГОРИТМ ДЛЯ ЛОКАЛИЗАЦИИ ТЕКСТА НА ИЗОБРАЖЕНИЯХ СЛОЖНЫХ СЦЕН НА ОСНОВЕ РАЗРЕЖЕННОГО КОДИРОВАНИЯ

Т. А. Нанавова, С. И. Анищенко

Донской государственный технический университет, Ростов-на-Дону, Российская Федерация
ООО «11 бит», НИИ нейрокибернетики им. А. Б. Когана АБиБ ЮФУ,г. Ростов-на-Дону, Российская Федерация
stells318@gmail.com,
sergey@11bits.net

Рассматриваются вопросы локализации текста на основе разреженного кодирования. Задача данного исследования — разработка и оценка эффективности алгоритма получения оптимального разреженного базиса для надежного разделения текста и фона на изображениях сложных сцен.

Ключевые слова: обнаружение и локализация текста на видео, распознавание символов, машинное обучение, компьютерное зрение, разреженное кодирование.

UDC 004.932.75'1

TEXT LOCALIZATION ALGORITHMS ON COMPLEX SCENES IMAGES BASED ON SPARSE CODING

T. A. Nanavova, S. I. Anishchenko

Don State Technical University, Rostov-on-Don, Russian Federation
Ltd «11 bits», A. B. Kogan Research Institute for Neurocybernetics, Southern Federal University, Rostov-on-Don, Russian Federation

stells318@gmail.com,
sergey@11bits.net

The article considers issues of text localization based on sparse coding. The aim of this research is the development and evaluation of the effectiveness of algorithm to obtain optimal sparse basis for a reliable section of the text and images background on complex scenes.

Keywords: text detection and text localization on video frames, character recognition, machine learning, computer vision, sparse coding.

Введение. В настоящее время отмечается стремительное развитие цифровых видеоданных. В связи с этим возникает потребность в разработке новых и эффективных методов своевременного извлечения и индексирования информации из видеопотоков. Такая необходимость возникает, например, в следующих случаях:

— создание аннотационных видеороликов,
— извлечение машиночитаемых данных о спортивных соревнованиях в реальном времени,
— оцифровка показаний индикаторов систем с закрытым программным интерфейсом,
— навигация роботов,
— воспроизведение текстовой информации в аудиорежиме в помощь слабовидящим.

Классификация текста на видео. Текст на видео можно классифицировать следующим образом:

— графический текст, который непосредственно наносится при видеообработке или при монтаже;
— сценический текст (от англ. scene text), который содержится на предметах, снятых на видео.

Примерами же графического текста могут служить различные заголовки, субтитры, время и отметка расположения, имена людей и спортивные результаты. Тексты сцены размещаются, например, на попавших в кадр дорожных знаках, рекламных стендах, надписях на транспорте, одежде людей и т. д. На рис. 1 представлена стандартная схема системы извлечения текста из изображений или видео.

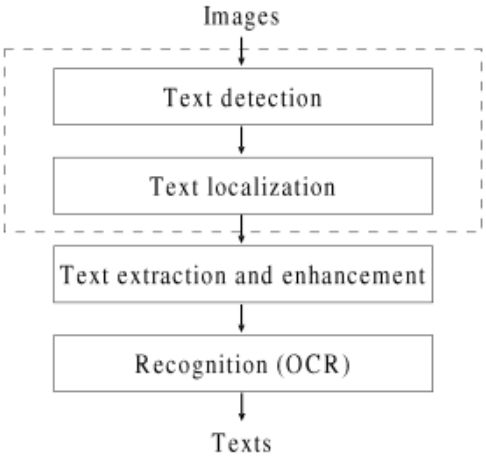

Рис. 1. Схема извлечения текста из изображений или видео [1]

Основными компонентами этой системы являются обнаружение и локализация текста. Видеоданные часто содержат изображения сложных сцен, на которых текст, содержащий важную информацию, занимает небольшую площадь (например, знаки ограничения скорости или счет футбольного матча). Поэтому для корректного распознавания текста на видео необходима его надежная локализация на кадрах.

Обзор существующих методов. Методы обнаружения текста разделяют на две группы:

— «сверху вниз» (вначале сцена рассматривается как единое целое, а затем пошагово выделяются области текста);

— «снизу вверх» (вначале выделяются компоненты изображения, а потом пошагово группируются в области текста).

Для обнаружения текста используются различные его признаки. Авторы работы [2] ориентируются на ширину штриха, предполагая, что для символов она изменяется в достаточно узком диапазоне. Детектор границ Кэнни [3] позволяет для каждого пикселя вычислить ширину штриха как расстояние между двумя ближайшими точками с противоположным по знаку градиентом. Затем находятся области, где ширина штриха характерна для текста и постоянна. Следует, однако, отметить недостатки этого подхода:

— ширина штриха на стыках букв часто бывает непостоянна;

— эффективность метода существенно зависит от результатов предварительной обработки изображения.

В работе Коатеса [4] для обнаружения текста предложен метод, базирующийся на обучении алгоритма извлечения признакового описания без учителя. Такой подход предполагает построение словаря признаков на основе обучающей выборки изображений слов с использованием кластеризации признакового пространства. В специфической области применения обнаружение текста может производиться при помощи пороговой или адаптивной бинаризации изображения. Это относится к тем случаям, когда априори известно, что текст характеризуется высоким уровнем контраста с фоном [5]. Обнаруженные строки разделяют на символы, используя вертикальный профиль, где локальные минимумы соответствуют пробелам между буквами. Затем из изображения символа извлекают вектор признаков. В качестве признаков могут использоваться моменты Цернике [5], полиномиальная аппроксимация контура символов или их фрагментов [6].

Актуальность работы. Цели и задачи. Многие методы локализации текста работают достаточно надежно. Однако в некоторых ситуациях современные алгоритмы не справляются с ло-

кализацией и распознаванием текста. Так, на рис. 2 представлен пример работы онлайн-сервиса компании *Microsoft*, который находится в свободном доступе [7].

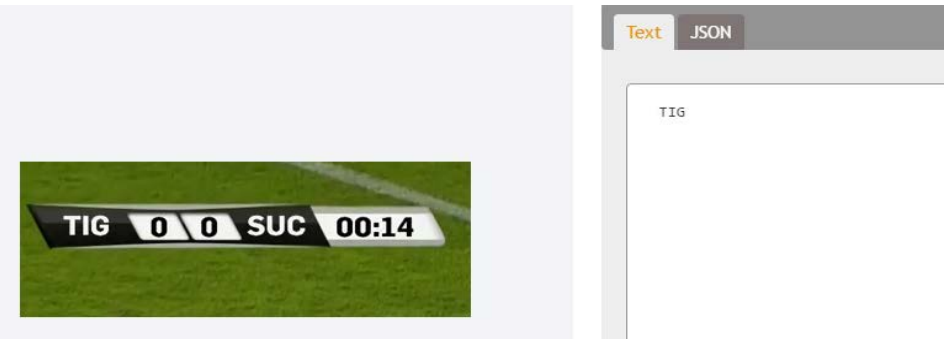

Рис. 2. Пример работы *Microsoft Computer Vision API*

Видно, что на изображении, которое представляет собой специально вырезанную полоску с футбольным счетом, найдено только три первых символа, но совершенно не распозналась самая важная информация — счет и время матча. К тому же на некоторых видеокадрах, если текст занимает небольшую площадь, данный сервис не дает вообще никаких результатов. Таким образом, на данный момент задача локализации текста на видео не решена в полном объеме и остается актуальной. Представленное исследование посвящено разработке нового эффективного метода для обнаружения и локализации текста на видео.

В рамках данной работы создавался алгоритм локализации текста на основе метода обучения признаков без учителя (*Unsupervised Features Learning*). Это подход базируется на методе разреженного кодирования [8]. Разреженное кодирование представляет собой класс методов обучения без учителя для более эффективного представления данных. Основной целью данного класса алгоритмов является нахождение набора базисных векторов φi таким образом, чтобы представить входной вектор x как линейную комбинацию базисных. Задача данной работы состоит в создании и исследовании алгоритма получения оптимального разреженного базиса для классификации текста и фона. Подход на основе разреженного кодирования эффективен при решении задач распознавания текста. Это показано в [4], где предложено признаковое описание областей изображения в 64-мерном пространстве.

Проведенное исследование. В отличие от работы Коатеса [4], в данном исследовании предлагается метод, позволяющий уменьшить количество базисных векторов и, соответственно, размерность признакового пространства, что способствует ускорению обработки данных и улучшению качества классификации. Для этого было проведено качественное исследование, состоявшее из 10 этапов.

1) Подготовка обучающей выборки, содержащей текстовые и фоновые области кадров видео.

2) Извлечение из изображений обучающей выборки области размером 8×8 пикселов и получение их линейной развертки. Таким образом будут получены векторы $X \in R^{64}$.

3) Нормализация каждого вектора x_i по яркости и контрастности путем вычитания среднего значения и деления на стандартное отклонение.

4) Визуализация полученных наборов векторов x_i для текста и фона. Для этого их размерность предварительно уменьшается (до трехмерных) с помощью метода главных компонент. Таким образом получатся 2 кластера для текста и фона. На рис. 3 можно увидеть визуализацию век-

торов x_i для областей изображений, содержащих текст, предварительно уменьшенной размерности с помощью метода главных компонент.

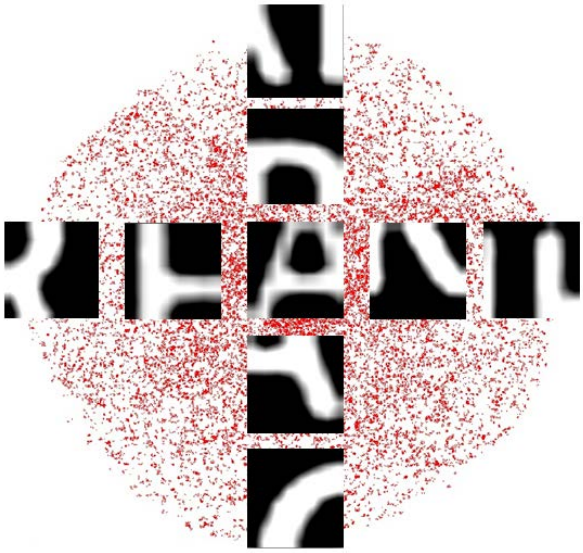

Рис. 3. Визуализация наборов векторов $x_i \in R^{64}$ для текста

Каждая точка представляет собой отдельный вектор, и здесь также можно увидеть визуализированные области изображения для некоторых векторов. Они представляют собой фрагменты букв (см. рис. 3). Перечисленные выше действия выполняются с целью обеспечения качественной оценки распределения.

5) Получение базиса для текста и фона. Для этого находятся центроиды полученных на шаге 4 кластеров с помощью метода К-средних. При применении метода К-средних число кластеров задано $n = 64$.

6) Уменьшение размерности конечного признакового описания путем исключения из текстового базиса элементов, которые по Евклидову расстоянию окажутся к фоновым ближе, чем заданный порог ($h = 5$). На рис. 4 продемонстрированы те базисные элементы для текстовых областей, которые остались после применения процедуры, описанной на шаге 6.

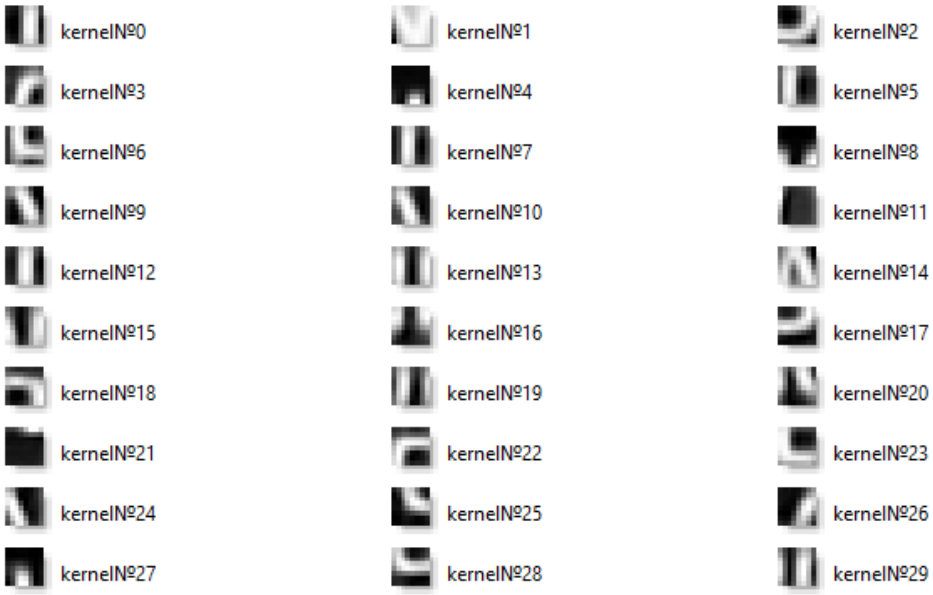

Рис. 4. Оставшиеся элементы базиса для текстовых областей

7) После этого каждый из оставшихся базисных векторов рассматривается как ядро для фильтрации всех кадров видео по отдельности. Признаковое описание области изображения получается путем свертки с каждым базисным вектором. То есть при наличии 30 базисных ядер можно получить 30-мерный признаковый вектор области изображения.

8) Получение вектора признаковых описаний $Y \in R^{30}$ в исходном кадре для каждой точки области с текстом и фоном при помощи подхода разреженного кодирования. Стоит отметить, что извлечение подобных признаковых описаний является простой задачей при использовании графических процессоров, поскольку вычисления легко распараллеливаются.

9) Визуализация полученных наборов векторов y_i для текста и фона (их размерность также предварительно уменьшается до трехмерных).

10) Качественный анализ признакового описания, полученного путем разложения по обученному базису для случая 64 и 30 базисных ядер.

Результаты представлены на рис. 5, 6 и 7 в виде визуализации всех полученных наборов векторов. Для текста использован красный цвет, для фона —синий.

На рис. 5 представлено изначальное признаковое пространство в оттенках серого. Это визуализированные в трехмерном пространстве, уменьшенные с помощью метода главных компонент исходные линейные развертки областей изображения размером 8×8 пикселов. Заметно, что это очень плохо разделимые облака точек (кластеры), и классифицировать их с помощью методов машинного обучения едва ли представляется возможным.

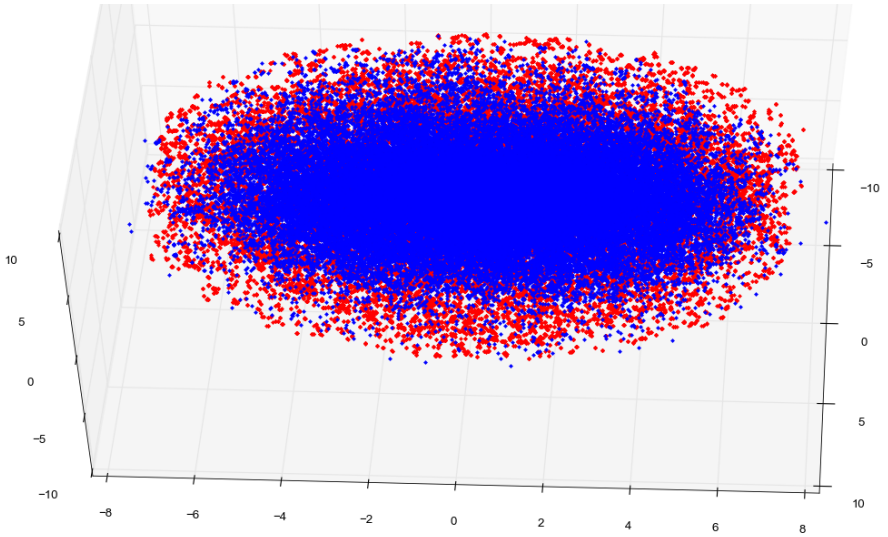

Рис. 5. Изначальное признаковое пространство (области 8×8) $V \in R^{64}$

На рис. 6 и 7 показаны визуализированные признаковые описания, полученные с помощью подхода разреженного кодирования в случаях с 64 и 30 базисными ядрами. Видно, что облака точек стали более различимыми, причем при уменьшении размерности более чем в 2 раза четкость границ не уменьшилась.

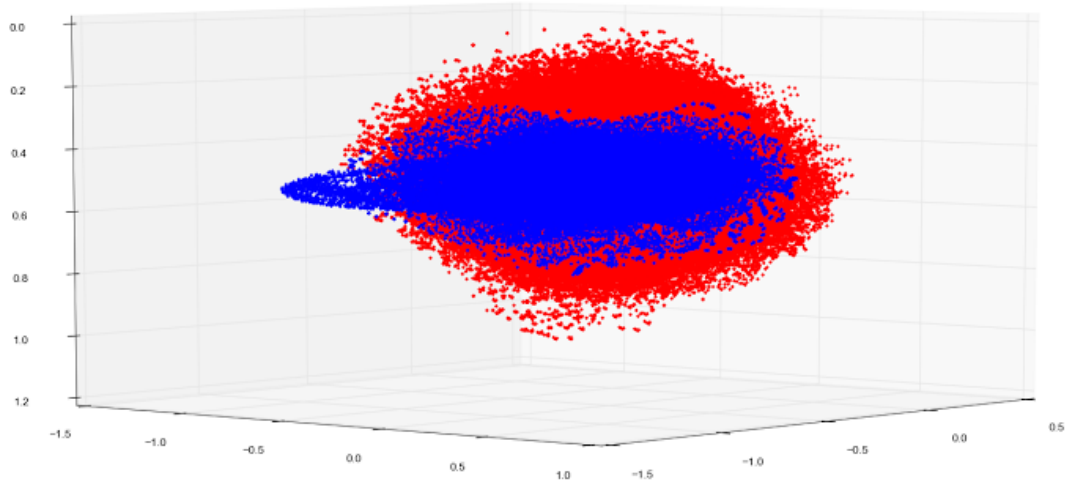

Рис. 6. Разреженные признаки $V \in R^{64}$

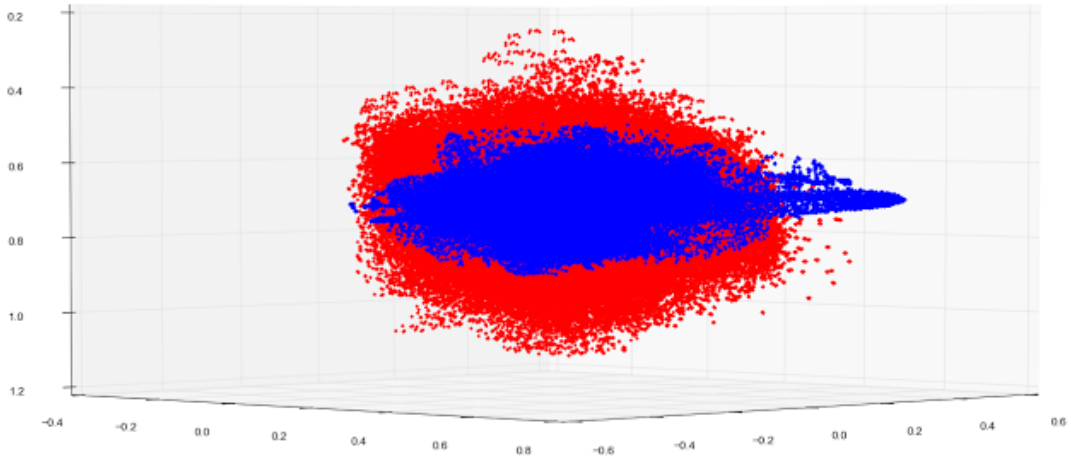

Рис. 7. Разреженные признаки $V \in R^{30}$

Выводы. Итак, подход на основе разреженного кодирования позволяет создавать разделимые признаковые описания, с помощью которых можно эффективно применять методы машинного обучения для выделения текста из фона. Признаковое описание, создаваемое с помощью подхода разреженного кодирования, получается относительно низкой размерности, благодаря чему достигается высокая вычислительная эффективность. Основное направление дальнейших исследований — обучение классификатора (искусственная нейронная сеть) для локализации текста.

Библиографический список

1. Jung, K. Text information extraction in images and video: A survey / K. Jung, K.-I. Kim, A.-K. Jain // Pattern Recognition. — 2004. — Vol. 37, № 5. — P. 977–997.

2. Epshtein, B. Detecting text in natural scenes with stroke width transform / B. Epshtein, E. Ofek, Y. Wexler // Computer Vision and Pattern Recognition (CVPR 2010) : IEEE Conference. — San Francisco, 2010. — P. 2963–2970.

3. Canny, J. A computational approach to edge detection / J. Canny // The IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1986. — Vol. PAMI-8, № 6. — P. 679–698.

4. Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning / A. Coates [et al.] // International Conference on Document Analysis and Recognition. — 2011. — P. 440–445.

5. Zhang, D. General and domain-specific techniques for detecting and recognizing superimposed text in video / D. Zhang, R.-K. Rajendran, Shih-Fu Chang // International Conference on Image Processing. — Rochester 2002. — P. I — 593.

6. Smith, R. An overview of the tesseract OCR engine / R. Smith // Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2007). — Curitiba, 2007. — P. 629–633.

7. Microsoft Computer Vision API. Extract rich information from images to categorize and process visual data—and protect your users from unwanted content [Электронный ресурс] / Microsoft. – Режим доступа: https://www.microsoft.com/cognitive-services/en-us/computer-vision-api (дата обращения: 15.04.16).

8. Efficient sparse coding algorithms / H. Lee [et al.] // Advances in neural information processing systems (NIPS 2006). — Vancouver, 2006. — P. 801–808.