EBU
OPERATING EUROVISION AND EURORADIO

TECHNICAL REVIEW

Q3 2013

VIDEO BASED OCR: A CASE STUDY OF REAL TIME IN-SCREEN SUBTITLE RECOGNITION

Внимание!
Данный перевод **НЕ** претендует на аутентичность
и может содержать отдельные неточности.
Оригинал документа на сайте https://tech.ebu.ch

OCR НА БАЗЕ ВИДЕО: ПРАКТИЧЕСКИЙ АНАЛИЗ РАСПОЗНАВАНИЯ СУБТИТРОВ С ЭКРАНА В РЕАЛЬНОМ ВРЕМЕНИ

MARCO SLIK, HANS JONGEBLOED & MARK VAN STAALDUINEN
NPO

РЕЗЮМЕ

NPO, Нидерландская общественная вещательная организация, реализовала решение оптического распознавания символов (OCR) для преобразования видимых на экране субтитров в речевой звук субтитров в реальном времени.

В Нидерландах телевизионные программы на иностранных языках субтитрируются. NPO предлагает услуги доступа для слабовидящих в форме речевых субтитров: слышимый текст на основе синтезированной речи. Чтобы это работало корректно, в прошлом были необходимы созданные заранее файлы субтитров или данные, встроенные внутри студийных сигналов. Теперь программы все чаще подаются с субтитрами, уже присутствующими в активном видео – так называемые «титр-копии» или «впаянные субтитры» на прежнем техническом жаргоне. Процент программ с речевыми субтитрами имеет тенденцию к падению, в то время как политическое требование услуг доступа для слабовидящих звучит все громче.

В поиске решения NPO объединила усилия с голландской исследовательской организацией TNO, поставщиком технологии почтового рынка Prime Vision, и с поставщиком эфирного оборудования NPO Ericsson Broadcast Services в работе над решением этой проблемы. С успехом…

ВВЕДЕНИЕ

NPO, Нидерландская общественная вещательная организация, реализовала решение для преобразования субтитров с экрана в данные скрытых субтитров в реальном времени. Это значит, что «впаянные» субтитры в видеоизображении конвертируются в отдельный поток данных, содержащий текст субтитров.

ПРАКТИЧЕСКИЙ АНАЛИЗ: РЕЧЕВЫЕ СУБТИТРЫ В НИДЕРЛАНДАХ

В Нидерландах около 430,000 человек имеет серьезные проблемы со зрением. В результате старения населения это число растет; сейчас около 76,000 человек являются слепыми или почти слепыми.

Для слабовидящих телевидение немного дополняет другие формы медиа потребления. Однако оно играет не менее важную роль, чем для видящей аудитории. Телевидение – важный способ быть в курсе текущих событий. Социальное значение совместного просмотра и социального взаимодействия, создаваемого программой, также не следует недооценивать. Несмотря на разные объемы и способы «просмотра» (этот термин используется самими слепыми), привычки слабовидящей аудитории в телепросмотре почти совпадают с привычками людей с нормальным зрением.

На голландском телевидении большинство программ или программных элементов на иностранных языках субтитрируется. Постсинхронизация / аудио дублирование применяется только в программах для детей до 12 лет. Для видящей аудитории полностью поддерживается ощущение оригинальной программы. Для слабослышащих идеальным способом телепросмотра является субтитрирование. Однако для слабовидящей аудитории более полезно дублирование.

Хотя это и не требуется по закону, NPO как общественная вещательная организация испытывает политическое давление на трансляцию программ по основным каналам NPO с потоками данных субтитров[1]; чтобы слабовидящие зрители могли слушать текст на голландском языке путем синтеза речи (т.е. автоматического преобразования текста в речь). Эта автоматизированная так называемая служба «речевого субтитрирования» работает в Нидерландах с 2001 г. Это экономичный метод обслуживания слабовидящей аудитории по сравнению с методом аудио описания, который применяется в других странах и требует ручной авторизации перед эфиром. При удовлетворении определенных условий речевое субтитрирование может достичь 100% охвата программ.

Конечный пользователь может принимать звук речевых субтитров следующими способами:

- В виде языковой дорожки DVB в цифровом распространении, что дает наложение звука речевых субтитров на программный стереозвук. Громкость звука программы аккуратно понижается во время присутствия текста субтитров; это так называемый подход «‘вещательного микширования». Этот сигнал подходит для пользователей, способных управлять относительно сложными (с точки зрения целевой аудитории) работой ТВ приставки. Это особенно полезно для большой группы дислексиков или зрителей с другими нарушениями способности к чтению.
- Через визуальные средства; например, отдельный приемный блок, декодирующий данные субтитров, передаваемые через страницу телетекста 889 или поток данных в интернете. Сегодня стоимость двух коммерческих визуальных средств возмещается через голландскую систему медицинского страхования. Эти визуальные средства также обеспечивают другие функции, например, речевые газеты, книги и журналы. Простота работы, включая обратную аудиосвязь, делает это решение самым походящим для слабовидящих «зрителей». Звук речевых субтитров подается из отдельного громкоговорителя, позволяя другим членам семьи слушать нормальный звук программы.

Разница в синхронизации между речью и субтитрами обычно менее нескольких ТВ кадров. В интернет-решении, где жесткая синхронизация с видео трансляцией теряется, применяется настраиваемая пользователем задержка для компенсации разницы в трактах распространения (спутниковом, наземном, кабельном и IPTV).

ПРОБЛЕМА

К сожалению, субтитры, необходимые для распространения данных и преобразования их в речь, имеются в виде отдельных файлов лишь для 70% программ. Этот процент уменьшается по следующим причинам:

[1] Для слабослышащих Dutch Media Act требует 95% видимого субтитрирования для речевых программ на голландском языке. Это осуществляется через Teletext стр. 888 и субтитры DVB, в дополнение к видимым субтитрам с переводом с иностранного языка. Политические требования к речевому субтитрированию (без определенных юридических оснований) имеют ту же основу, что и законы о субтитрировании для слабослышащих.

- В ежедневных выпусках новостей и прямых репортажах о текущих событиях субтитры все чаще впечатываются в видеосигнал при причине гибкости монтажа в последнюю минуту перед эфиром.
- Ручные операции с субтитрами в студийном производстве прямого эфира сводятся к минимуму из-за бюджетных ограничений. Программные элементы в файлах, произведенные в студии, сейчас должны быть полностью готовы к эфиру.
- Есть тенденция кроссмедийной передачи во время или до эфира. Зависимость от прямого эфира должна быть минимизирована, и желательна максимальная гибкость в создании версий для разных способов распространения.

РЕШЕНИЕ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ (OCR)

Решение, которое искала NPO – как автоматически распознавать текст субтитров внутри конечного видео продукта. NPO сотрудничала с TNO (крупнейшей независимой научно-исследовательской организацией в Нидерландах с 1932 г.) и Prime Vision (компанией, специализирующейся на OCR, главным образом в почтовой автоматизации) в работе над инструментом автоматического преобразования. Ericsson Broadcast Services, поставщик эфирного оборудования NPO, присоединился на этапе проверки концепции и ускорил практическую реализацию в системе воспроизведения. Требования на этапе определения проекта были сформулированы следующим образом:

- Система должна поддерживать видео в разрешении SD и HD.
- Должны быть включены рабочие процессы с файловыми и прямыми программами.
- Архитектура решения должна быть простой и эффективной. Следует по возможности избегать усложнения (цифрового) рабочего процесса.
- Поскольку ожидается, что OCR не сможет совпадать с «добрым старым» потоком данных на базе текста, преимущество должен иметь поток данных из оборудования субтитрирования, если он есть.
- 99% слов из OCR должны быть корректно конвертированы в текст.
- Следует конвертировать только голландские субтитры (исключая субтитры для языковых меньшинств, например, арабские, турецкие и т.д., которые потребуют лишних шрифтов и языковых наборов и существенно усложнят проект).
- Знаки пунктуации также должны считываться.
- Выходные данные системы OCR должны быть доступны в течение 300 мс после первого кадра, в котором появляется субтитр, для сохранения приемлемой синхронизации речи и субтитра во всей вещательной цепи.

Вследствие требований прямого эфира выбор был быстро сделан в пользу системы почти в реальном времени в эфирном видео потоке.

ВЫБОР ПАРТНЕРА ПО РЕАЛИЗАЦИИ

TNO начала искать партнера по бизнесу и реализации. Учитывая, что видео OCR довольно ново, TNO не ожидала найти готовый коммерческий продукт, удовлетворяющий целевым требованиям. Запрос на информацию десяти потенциальным поставщикам дал четыре ответа, из которых лишь два проявили должное понимание проблемы.

После дальнейшего запроса проектного предложения Prime Vision оказался самым подходящим. Компания имеет опыт разработки индивидуальных решений, понимания высококритичных операционных процессов, обработки больших объемов данных и имеет собственные средства разработки в Нидерландах. Это сокращает каналы связи, что является важным фактором успеха в проекте. Потенциальная трудность состояла в том, что они были новичками в передаче медиа потоков; этот риск был снижен убедительной демонстрацией первых результатов с характерным тестовым материалом SD.

Prime Vision имеет 60 лет опыта в чтении рукописного текста и печатных символов на почтовых отправлениях, бандеролях, в денежных операциях и т.п. Они предоставляют услуги почтовой автоматизации на базе OCR компаниям со всего мира. Помимо почтового рынка, Prime Vision профилирует себя как поставщик решений и в других новаторских областях. Пример – распознавание номерных знаков для автотранспортных приложений. Поэтому функции, требуемые NPO/TNO, вписывались в их проектное портфолио.

ОПРЕДЕЛЕНИЕ И ИЗМЕРЕНИЕ КАЧЕСТВА

По следующим тестовым изображениям на Рис. 1 – 6 видно, что хотя субтитры кажутся хорошо структурированными, эти примеры иллюстрируют вариантность внешнего вида и некоторые отличия в начертании шрифта, которые могут быть фатальными, если не уделить им должного внимания.

Рис. 1: Рядом с титрами, близко или внутри области субтитров

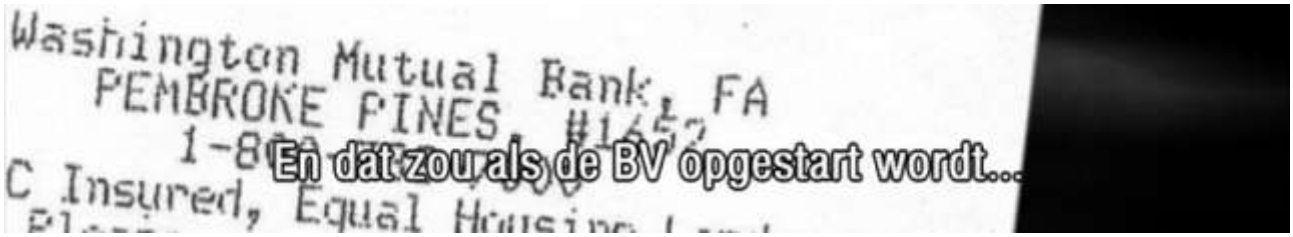

Рис. 2: Сложные пертурбации текста

Рис. 3: Разница в презентации (например, курсивом)

Рис. 4: Текст на иностранном языке; должен автоматически игнорироваться OCR

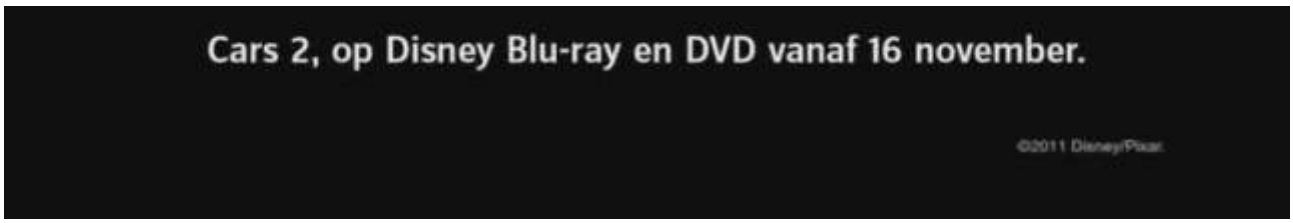

Рис. 5: Текст не субтитров внутри зоны субтитров, известный ложный результат

Рис. 6: Наложение символов, некорректная рамка

Предлагаемое решение OCR должно быть устойчивым к этим вариантностям.

Поэтому TNO разработала тест-план для измерения качества распознанных субтитров структурированным методом и безопасности процесса определения, должны ли быть выполнены требования.

Были определены два параметра качества:

- качество обнаружения: определение, является ли субтитром предполагаемый субтитр OCR.
- производительность распознавания: измерение корректного совпадения с шаблоном.

Измерение качества обнаружения не так банально, как кажется. Например, не каждый кадр содержит субтитр; но если субтитр присутствует, он всегда находится в ряде последовательных кадров. Наконец, данные субтитров должны экспортироваться только один раз. Поэтому важным требованием является обнаружение последовательности субтитров.

После корректного обнаружения субтитров следует сократить до минимума ошибки символов, чтобы получить приемлемый и понятный текст для речевого субтитрирования. После начального этапа разработки video-OCR группа тестеров сравнила партию захваченных кадров с результатами субтитров OCR в автоматизированной процедуре. Распознанные субтитры были проверены и скорректированы вручную. Это дало длинный список распознанных субтитров и их корректного вида. Эти субтитры были сравнены друг с другом и выявлены структурные ошибки.

Для этого использовался метод расстояния Левенштейна [1]. Он сравнивает две строки и определяет число поправок, необходимых для подгонки одной строки под другую. В данном случае определялось число поправок для изменения распознанного субтитра в корректную форму. Если оба субтитра были идентичны, число поправок было ноль. Если один символ был пропущен или написан неверно, то для устранения разницы требовалась одна поправка – добавление или замена, соответственно. Разумеется, чем меньше расстояние Левенштейна, тем лучше качество распознавания.

Тестирование качества на разных этапах развития не только доказало необходимость оптимизации, но и дало материал для определенного решения. В этой статье мы рассмотрим данный вопрос с точки зрения OCR.

OCR 101

OCR – технология распознавания по шаблону, способная читать машинописные или рукописные знаки. Она использует алгоритмы статистического совпадения шаблонов для изучения характеристик большого количества маркированных примеров (где «маркированный» означает: логическая пользовательская презентация символов предполагаемых групп пикселей изображения) и использует эти характеристики для распознавания. OCR широко применяется и является успешным методом автоматического преобразования отсканированных документов (изображений) в редактируемый текст.

Как эталон, человеческий мозг отлично распознает очень сложные формы шаблонов. Мы можем распознавать отклонения и учитывать контекст шаблонов в совпадении и маркировке. Это часть наших текущих сознательных и бессознательных естественных процессов обучения, начиная с рождения; например, распознавание лиц или объектов (идентификация, частично ли они закрыты, отвернуты или даже скрыты). Иногда наше обучение активно стимулируется в определенном направлении, например, обучение чтению на родном языке в начальной школе.

Компьютеры пока далеки от выполнения всех задач распознавания по шаблону, который доступен человеку. Однако OCR отлично работает в обусловленной среде (например, документы, почтовые отправления, бандероли); большинство символов в основном формате презентации распознаются (> 99% правильно) хорошо обученной системой OCR. Однако системы OCR не безотказны. Возможными причинами ошибок могут быть: почти похожие символы, шум изображения, низкое качество печати, неизвестные шрифты, символы, не входящие в сохраненный набор (например, греческие, математические символы, знаки валют и т.д.).

OCR, применяемая Prime Vision, включает ряд смарт-технологий, которые решают многие известные проблемы OCR:

1. Лучшая попытка распознавания делается по знакам пунктуации и диакритике (например, é, ë). Это нетривиально, т.к. знаки пунктуации и диакритика довольно маленькие (всего несколько пикселей и могут также быть фоновым шумом.
2. Статистические модели языка, обученные для голландского языка, используются для включения контекстного знания между символами, которые визуально похожи для OCR – например, ‘I’ (заглавная i) и ‘l’ (прописная L); ‘O’ и ‘o’ vs. ‘0’ (ноль). Типичный пример на голландском языке – отличие между ‘I’ в ‘Ik’ (значит ‘I, я), и ‘l’ в ‘elkaar’ (друг друга).
3. Символы не всегда четко разделены: типичные комбинации типа ‘f’ и ‘t’ (как в gift), ‘r’ и ‘t’ (как в porto) могут стоять так близко, что OCR распознает их как один символ (наиболее похожим знаком может быть ‘n’). Для решения этой проблемы обучающая последовательность OCR

была расширена и обучена так называемым «сросшимся символам», чтобы в случае «срастания» идентифицировать два корректных символа. Пример проблемы сросшихся символов показан на Рис. 7.

Рис. 7: Пример сросшихся символов

OCR НА БАЗЕ ВИДЕО ДЛЯ СУБТИТРОВ

Для применения OCR к видео нужны статичные, а не подвижные изображения. Следовательно, входящее видео разлагается на отдельные кадры через процесс ввода и регистрации кадров. Результат: файлы несжатых изображений с устранением чересстрочности в разрешении 1920 x 1080 пикселей и пиксельной глубиной 24 бит. Это дает объем памяти около 1 GB (гигабайт) на минуту или более 1 TB (терабайт) на сутки. Поэтому требования к внутренней памяти RAM и объему памяти и ограничения времени обработки субтитров были установлены для развития и внедрения финального решения. Изображения должны сохраняться на диск для регистрации, оценки и настройки. Чтобы не отставать от видео потока, требуется выдавать результат распознавания в течение 300 мс после первого кадра, содержащего субтитр. Поэтому алгоритмы должны были быть относительно быстрыми и не могли работать с обширным поиском в словаре или со сложными моделями языка.

Наряду с вышеупомянутыми технологическими аспектами были реализованы следующие этапы процесса для получения оптимальных результатов subtitle-OCR и для удовлетворения некоторых ключевых показателей производительности, которые следуют из трех вышеперечисленных проблем OCR:

4. Конфигурируемые зоны распознавания: фокусирование только на той области изображения, которая может содержать субтитр (т.е. нижняя треть в случае NPO; субтитры в верхней части изображения редки).
5. Методы обработки изображений: комбинирование множества кадров для обнаружения стабильных стационарных областей, где могут быть субтитры.
6. Интеллектуальные алгоритмы сегментации: отделение переднего плана (т.е. субтитра) от фона (видео изображения). Это особенно полезно, если видео фон также содержит текст. См. для примера изображение на Рис. 8.
7. Продвинутое обнаружение краев: находит, где находятся символы субтитров в видео изображении, пользуясь черной обводкой символов и кластеризацией цветных участков.
8. Модели языка N-gram: фильтрация неголландских субтитров (например, турецких в субтитрах на двух языках для программ, где полиция может просить публику помочь в поиске пропавших людей). Модель содержит статистику о языке; в данном примере типичное использование комбинаций символов было главной задачей процесса фильтрации.
   Решение о правильном написании на голландском языке выходного текста OCR основано на пороге подсчета модели языка score и на оптимальном пороге. Таким образом получается баланс между ложными признаниями и ложными отказами и решение основывается на оценке большого количе ства образцов субтитров.
   Модели языка также предотвращают бессмысленные результаты OCR, получившиеся из черно-белых шаблонов в изображении, которые могут вести к распознаванию нелогичного типа **‘l j ii J L l’.**

Рис. 8: Фрагменты видео могут содержать текст, создавая проблемы для OCR.

ФИНАЛЬНАЯ ПРОБЛЕМА СУБТИТРОВ

Так называемые «ложные результаты», текст в изображении, который был распознан, но не являлся (частью) субтитра, оказались самой серьезной проблемой в проекте. Представьте тикерную ленту, полосы, наложение титров с именами, элементы текста в рекламе, бегущие титры, анонсы передач, расписание программ и т.д.

Эта проблема в большой степени была решена следующими способами:

- Допуск только текста, стационарного в изображении (фильтрация тикерных лент).
- Допуск только корректно выровненного текста (например, только по центру изображения).
- Фильтрация результатов OCR с низкой долей уверенности из модели языка и/или вообще не содержащих нормальных слов (фильтрация текста низкого качества и прочего текста типа веб-ссылок и адресов email).
- Недопущение текста из одних заглавных букв (это бывает очень редко в нормальных субтитрах, но очень часто в титрах с именами).
- Настройка порога для чтения только (почти)-белых символов с черной обводкой; не весь текст субтитров везде имеет черную обводку, поэтому допускается некоторый пропуск черного. Однако это предотвращает чтение символов, которые не имеют никакой обводки и могут происходить из других типов текста на фоне.

Помимо независимой оценки производительности системы, TNO также посоветовала подход к решению технических проблем. Это было достигнуто путем глубокого анализа ошибок, обнаружения структур и последующей дискуссии с Prime Vision и NPO о том, как и до какого уровня следует обрабатывать структурные ошибки. Эти дискуссии стали возможны благодаря доверию, открытому и творческому отношению, быстрым каналам связи и стимулу к реальному решению проблемы. Самое важное, что они привели к нужному решению, где были удовлетворены требования и спецификации NPO:

- Общая скорость чтения, объем вовремя идентифицированных субтитров процессом OCR более 99%. С материалом субтитров, полностью соответствующим текущей спецификации NPO (т.е. корректный тип шрифта, размер, позиционирование, контур) – 100%.
- Пословная вероятность ошибок в распознанных субтитрах менее 1%.
- В качестве компромисса для максимально возможной скорости чтения программ некоторые ложные результаты были приняты только с непрограммным материалом. Настройка системы ограничивает число случаев, когда пользователи удивлены случайными недействительными субтитрами, которые читаются вслух между нормальными субтитрами в программе.

УРОКИ С ТОЧКИ ЗРЕНИЯ ВЕЩАТЕЛЕЙ

В проекте и в дальнейшей повседневной практике видео OCR стало ясно следующее:

- Мы считаем, что ТВ – это подвижные изображения … но статичный текст во многих формах является важной частью видеокоммуникации.
- Корректный визуальный дизайн – критический фактор успеха общей производительности OCR. Это может быть проблемой в творческом мире телевизионного маркетинга. Технология никогда не должна иметь приоритет перед творческим контентом, но они должны подстраиваться друг под друга.
- Необходимо внимание к субтитрированию в форме производственной политики и спецификации. Хорошее субтитрирование – непростая часть монтажного процесса; нужен определенный опыт. В Нидерландах мы наблюдаем замену традиционной передачи текстового файла субтитров в систему видеомонтажа заранее сформатированным прозрачным файлом QuickTime, который будет накладываться во время монтажа. Таким образом, гарантируется единообразная, сделанная специалистами презентация, которая дает гибкость монтажа в любом месте, в любое время и любым способом.

РЕЗЮМЕ И ЗАКЛЮЧЕНИЕ

ПРЯМОЙ ЭФИР: 100% СУБТИТРИРОВАННЫХ ПРОГРАММ ДОСТУПНО ДЛЯ СИНТЕЗА РЕЧИ

Начав пилотный проект в 2011 г., NPO внедрила решение OCR на всех трех общественных телевизионных премиум-каналах с марта 2012 г.

Решение вещательного микширования текста в речь на базе речевой технологии Nuance было разработано и предоставлено компанией Ericsson параллельно с этой работой.

Через несколько месяцев NPO пришла к выводу, что данные субтитров, сгенерированные системой OCR, имеют такое качество, что должны быть единственным источником речевых субтитров. Техни-

ческое оборудование для извлечения данных из файла субтитров или служебных данных, вставленных системами субтитрирования в удаленных студийных потоках или в эфире, было демонтировано, устранив таким образом единые точки отказа и повысив надежность в цепи воспроизведения.

Целевая аудитория весьма удовлетворена улучшением услуг. Жалобы сведены к минимуму. Тем не менее, развитие системы не прекратилось: выходные данные с эфира скоро будут использоваться в целях обогащения метаданных; а в рамках проекта замены системы воспроизведения NPO реализация протокола Softel 'Newfor' позволит OCR в реальном времени в системе Miranda iTX. Авторы желают, чтобы формат прямых субтитров, созданный в результате текущей работы над спецификацией EBU-TT part 3 [2], был уже доступен, но, как всегда, технологический прогресс идет шаг за шагом… или, для субтитров, знак за знаком...

ССЫЛКИ

[1] http://en.wkipedia.org/Wiki/Levenshtein_distance

[2] http://tech.ebu.ch/ebu-tt

БИОГРАФИИ АВТОРОВ

MARCO SLIK

Marco Slik – старший советник по политике, Research & Development, голландской общественной вещательной компании NPO.

После обширной карьеры в области управления вещательными проектами, консультирования и архитектурного дизайна для интеграторов систем и поставщиков услуг Marco с 2010 г. работает в распространении и вещании NPO. Его первый опыт был в области ТВ эфира и цифрового распространения, хотя он занимается также и звуком, и интернетом. Возглавляет проекты R&D и инфраструктурных изменений и консультирует по политике инноваций.

HANS JONGEBLOED

Hans Jongebloed работает в должности R&D консультанта в Prime Vision B.V., Нидерланды.

Hans – эксперт по распознаванию речи, моделей языка и распознавания по шаблону и работает в KPN Research, Dutchear, и Prime Vision соответственно с 1997 г. Имеет специальность по прикладной физике и искусственному интеллекту. Главные темы его работы – R&D инновация и дизайн новых решений проблем автоматизации новых абонентов в области речевой технологии и OCR.

MARK VAN STAALDUINEN

Mark van Staalduinen работает в области консультирования по медиа и сетевым услугами в TNO, Нидерланды.

После научной карьеры в технологии машинного зрения и распознавания по шаблону, применяемой в больших базах данных, Mark работает консультантом и менеджером инновационных проектов в TNO. Его задача – развитие передовых решений в области медиа и безопасности в сотрудничестве с различными партнерами из университетов, правительства и поставщиков коммерческой технологии.

Опубликовано European Broadcasting Union, Женева, Швейцария
ISSN: 1609-1469
Главный редактор: Simon Fell
Ответственный редактор: Eoghan O'Sullivan
E-mail: osullivan@ebu.ch

Ответственность за мнения, выраженные в данной статье, лежит исключительно на авторе(ах).