УДК 621.391

ПРОТОТИПИРОВАНИЕ ПОДСИСТЕМЫ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЯ ГОЛОВЫ ОТ ФОНА В СОСТАВЕ СИСТЕМЫ ЦЕНТРАЛИЗОВАННОЙ ВИДЕОАНАЛИТИКИ НА БАЗЕ МАШИННОГО ЗРЕНИЯ БПЛА

М.М. ЛУКАШЕВИЧ[1], А.А. ВОРОНОВ[2], В.В. ВЕНГЕРЕНКО[2], М.М. ТАТУР[3]

1 – Белорусский государственный университет, Республика Беларусь
2 – Объединенный институт проблем информатики НАН Беларуси, Республика Беларусь
3 – Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь

Поступила в редакцию 6 апреля 2026

Аннотация. Выполнен обзор вариантов использования прототипа подсистемы сегментации изображения головы от фона в составе централизованной системы видеоаналитики с использованием БПЛА. Предложен алгоритм сегментации, использующий модификацию нейросетевой архитектуры. Показано, что алгоритм работает значительно быстрее и эффективнее, чем лучшие известные алгоритмы сегментации и стандартные нейросетевые архитектуры.

Ключевые слова: семантическая сегментация, видеоаналитика, БПЛА, дистанционное зондирование.

Введение

Мониторинг общественной безопасности предполагает наличие точной и оперативно обновляемой информации об объектах инфраструктуры. В настоящее время системы автоматической видеоаналитики достаточно распространены в мире, лидером по количеству камер видеонаблюдения является Китай, осуществивший теоретически полное покрытие всех крупных городов камерами видеорегистрации с продвинутой системой видеоаналитики, использующей идентификацию по FaceID и быстрый поиск по единой базе данных в 1 млрд. записей. Например, системы Hikvision iVMS-5200 (Китай) и QVR Pro (Тайвань).

Одним из лидеров по внедрению видеонаблюдения в DAS (Domain Awareness System) системы отдельных мегаполисов для обеспечения общественной безопасности являются США. Ниже перечислены некоторые из них: VIDEOMA INTELION, Milesight VMS Enterprise, Rhombus physical security platform, On-Net Surveillance Systems, Cisco MerakiMV Smart Cameras, Arcules Platform, VisionPoint, Verkada Video Surveillance & Management.

Третье место по количеству камер видеорегистрации принадлежит Москве с интеграцией в систему распознавания лиц и идентификации преступников. Основные разработчики необходимых функций – Ntechlab (поиск лица в кадре – детектирование), Visionlabs-Сбер (распознавание) и Tevian (идентификация – сверка с базами данных) или Сфера (при работе на транспорте).

Обработка видеоданных в системах видеонаблюдения позволяет расширить их функциональность. Решаются следующие задачи: детекция пешеходов; лиц людей с последующей идентификацией; автомобилей; номерных знаков автомобилей и номеров вагонов поездов с последующим распознаванием; детекция оставленных без присмотра вещей; слежение за объектом, в том числе и с использованием поворотных камер или камер дрона, с автоматическим управлением; анализ плотности скопления людей и автомобильного трафика, подсчет объектов интереса; целеуказание.

Решение каждой из перечисленных задач требует комплексного подхода, и включает ряд подзадач: предварительная обработка кадров видеопотока; выделение областей движения;

вычисление оптического потока; подавление тряски камеры (antishake); детекция объектов (лиц, людей, автомобилей и т.д.); слежение за объектом (трекинг); идентификация объекта.

Важным параметром алгоритмов обработки видеоданных, используемых в задачах видеонаблюдения, является их вычислительная сложность, так как обработка данных в таких системах ведется в реальном масштабе времени. Поэтому, важной характеристикой алгоритмов является возможность функционирования на недорогом оборудовании встроенных систем.

Для реализации быстрого распознавания лиц в системах видеонаблюдения применяются различные подходы к хранению данных и идентификации: использование единого центра обработки данных (ЦОД) или поиск в локальной базе данных (локальные системы хранения данных, напрямую подключенные к серверам для обработки видеопотоков в реальном времени с использованием ИИ-алгоритмов анализа). При использовании локальных баз данных минимизируются временные потери на сетевые коммуникации и возрастает автономность систем видеоаналитики, что важно в сфере безопасности на транспорте и торговле, в распределенной системе автоматизированного пограничного контроля.

Варианты развертывания

Видеоаналитика со стационарных камер не способна обеспечить отсутствие слепых зон при контроле безопасности, поэтому часто используют дроны, которые могут маневрировать и транслировать видеопоток с камер FHD по радиоканалу с достаточной пропускной способностью Далле, рассмотрим два варианта развертывания мобильных систем видеонаблюдения с применением беспилотного летательного аппарата (БПЛА) и использованием подсистемы сегментации. Оба они проилюстрированы на рис.1

Первый вариант для систем с централизованной обработкой. Основные системы видеоаналитики с централизованной обработкой, которые интегрированы с дронами следующие: Face-Six, которая лидирует в идентификации, DJI Analytics универсален по решаемым задачам, TRASSIR, Dedrone [1-6]. Видеопоток от дрона передаётся в облачный/вычислительный центр, где выполняется полнофункциональная обработка: стабилизация, детекция лиц (YOLOv8 и более ресурсоемкие нейросетевые архитектуры), сегментация, трекинг, калибровка камер, стереосопоставление, 3-D-реконструкция и идентификация. Такой подход обеспечивает более высокую точность и полноту анализа, но требует стабильной сетевой связи и больших вычислительных ресурсов. Сетевые задержки и пропускная способность остаются критичными ограничениями, особенно при использовании 4G/LTE; 5G пока недоступен в некоторых регионах

Одним из вариантов использования подсистемы сегментации лица человека на фотографиях и восстановления 3D модели для последующей идентификации по FaceID является описанный выше способ трансляции с дрона видеопотока в высоком разрешении на сервер, обладающий необходимыми вычислительными возможностями для сложной обработки и анализа видео в реальном времени. При анализе основной аппаратно-программной инфраструктуры предлагаемого варианта использования подсистемы сегментации в составе системы централизованной видеоаналитики, следует обратить внимание на используемые кодеки сжатия, варианты организации радиоканала и коммуникационной сети для передачи видеопотока с дрона на сервер. В одном из стандартов, регламентирующих процесс передачи видеопотока с дронов военного назначения определяются уровни сжатия MISM-L9H/L10H для Full HD изображений равные соответственно 80:1 и 20:1. Камеры дронов получают изображение Full HD 1920×1080×(50p) и 1920×1080×(30p). Для сокращения требований к радиоканалу применяются уровни сжатия из стандарта MISM-L10H и 10-разрядное представление для одного пикселя, так для передачи видеопотока Full HD 1920×1080×(30p) требуется пропускная способность 50 Мбит/с, а минимальный канал необходимый для передачи видеопотока, декодируемого при помощи MPEG-2 со степенью сжатия от 5 до 10 составляет от 34 до 100 Мбит/с. Сжатие потока Full HD с коэффициентами до 45:1 и 80:1 соответствует уровню MISM-L9 стандарта. При сжатии на уровне MISM-L9H достаточно радиолинии с пропускной способностью всего 20 Мбит/с для передачи такого видеопотока. В стандарте определены содержание и формат данных, получаемых с радаров обнаружения движущихся целей на фоне земной поверхности (GMTI – Ground Moving Target Indicator). В зависимости от пропускной способности каналов связи описанный в стандарте формат GMTI позволяет передать только

информацию о движущихся целях либо еще и сопутствующие радиолокационные изображения высокого разрешения (рис.1) [5]. Данные транслируются пакетами размером 65535 байт, что позволяет передать информацию о 80 отметках целей и их траекториях, сопровождаемых с высоким разрешением, или о 4300 целях и их трассах, сопровождаемых в обычном режиме. Всего на борту дрона может быть 256 источников данных (основных и вспомогательных) и генерировать они могут данные различных типов:64 сенсоры, формирующие изображение, так же данные от других бортовых датчиков: навигационного оборудования, бортовых систем управления и системы управления видеокамерой на борту дрона. Учитывая мировой опыт на борту дрона эффективно использовать видеокамеры высокой четкости с прогрессивной разверткой и квадратной формой пикселя, например, Full HD 1920×1080×(50p), чтобы не загружать для пересылки видеопотока радиоканал, имеющий ограниченную пропускную способность, целесообразно сжимать видео и использовать коэффициенты сжатия из стандарта соответствуюшие уровням MISM-L9H/L10H.

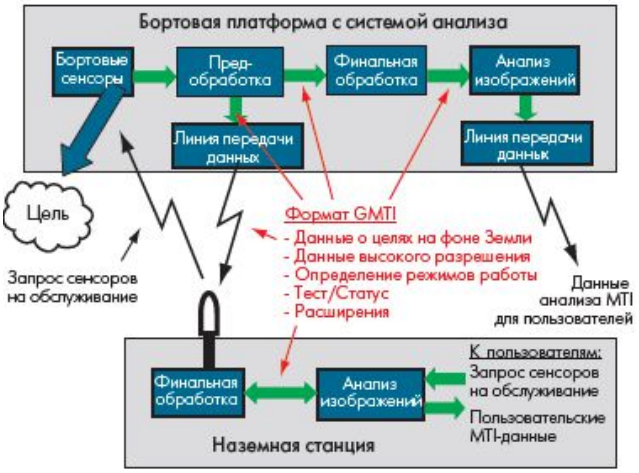

Рис. 1. Архитектура системы передачи данных об объектах интереса с дрона

В условиях ограниченной пропускной способности радиоканала в мобильных системах наблюдения на практике передача видеопотока невозможна без сжатия. Для сжатия отдельных кадров из которых и состоит видеопоток в настоящее время широко используются кодеки стандартов JPEG (Joint Photographic Experts Group), JPEG2000, основанные на алгоритмах энтропийного кодирования Хаффмана и арифметического; стандартов GIF (Graphics Interchange Format), TIFF (Tagged Image File Format), основанные на алгоритме LZW (Lempel-Ziv-Welch); методы и алгоритмы RLE (Run Length Encoding), LZ (Lempel-Ziv), усредненного блочного кодирования, EZW (Embedded Zerotrees of Wavelet transforms), SPIHT (Set Partitioning In Hierarchical Trees), SPECK (Set Partitioned Embedded bloCK) и ряд других [6]. Однако эти кодеки, методы и алгоритмы ориентированы, как правило, на сжатие отдельных изображений, учитывают только кодовую, межэлементную избыточность

Когда имеем дело с видеопотоком используются кодеры, которые могут работать в двух основных режимах: внутрикадрового и межкадрового прогнозирования. Внутрикадровое прогнозирование используется для кодирования опорных кадров. Межкадровое прогнозирование основано на блочной компенсации движения и используется для формирования прогнозных кадров на основе опорных кадров или других прогнозных кадров. Видеодекодер, осуществляет обратное преобразование и сглаживание значений восстановленных пикселей на границах блоков для вычисления ошибки прогнозирования, последующего ее кодирования и передачи. Основные отличия кодеков MPEG-4, H.264, H.265 заключаются в использовании различных преобразований и диапазонах изменения размеров блоков. В видеокодеке MPEG-4 применяется дискретное косинусное преобразование. Видеокодек H.264 использует дискретное вейвлет-преобразование. В видеокодеке H.265 существенно расширен диапазон изменения размеров блоков при блочной компенсации движения. При пиксельной компенсации движения, применяемой в MPEG-2, MPEG-4, H.264, H.265 прогнозный кадр формируется в результате

смещения всех пикселей опорного кадра на одинаковую величину [7]. Это позволяет компенсировать поворот, тангаж и смещение.

Данные передаются по радиолинии пакетами и для каждого пакета вычисляется контрольная сумма CRC. В зонах покрытия сотовой связью используется мобильная сеть. Если нет устойчивого покрытия мобильной связью, то используются специальные радиосети в том числе с ретранстляторами на базе БПЛА обеспечивающие надежную передачу информации с вероятностью ошибки на уровне от 10^{-7} до 10^{-6}, что для FullHD кадра равно одному пикселю на кадр. Для случая возникновения ошибок применяется помехоустойчивое кодирование, которому необходима минимум 30 % избыточность за счет добавления служебной информации при кодировании [8]. Помехоустойчивое кодирование обеспечивает восстановление передаваемой информации при ее искажениях в канале передачи. Если сжатие ориентировано на уменьшение избыточности передаваемой информации, то помехоустойчивое кодирование, наоборот, основано на внесении в передаваемую информацию избыточности, позволяющей приемной стороне обнаружить и исправить ошибки, возникающие в канале передачи. Избыточность в передаваемую информацию, в простейшем случае, вноситься за счет добавления проверочных и корректирующих символов, значения которых выбираются в соответствии с алгоритмом помехоустойчивого кодирования. В более сложном случае осуществляется преобразование передаваемой информации. Различные методы помехоустойчивого кодирования обеспечивают различную корректирующую способность и различную избыточность. В общем случае, избыточность увеличивается пропорционально увеличению корректирующей способности. Поэтому, при использовании помехоустойчивого кодирования необходимо предусмотреть соответствующее увеличение коэффициента сжатия при эффективном кодировании для того, чтобы сохранить требуемую пропускную способность канала передачи. Широкое распространение для помехоустойчивого кодирования получили коды Рида-Соломона (для борьбы с ошибками внутри модулей), табличные низкоплотные коды (LDPC-коды, позволяющие ускорить декодирование), БЧХ-коды (для борьбы со случайными ошибками) и их модификации [9]. Далее приведены примеры технологий, обеспечивающих передачу видеопатока необходимого качества, так для технологии Wi-Fi, например, скорости передачи от 1 Гбит/с до 54 Мбит/с при радиусе зоны уверенного приема 300 м на открытой местности [10] Для технологии WiMAX скорость передачи 75 Мбит/с при радиусе зоны уверенного приема 25 км [10], а в мобильной версии со скоростью 40 Мбит/с и меньшим радиусом зоны уверенного приема 5 км. При использовании сети мобильной связи 4G(LTE) можно передавать Full HD видеопоток с задержкой, если передача идет по сети 5G передача идет без пауз, но сеть 5G массово развернута только в Китае, США, Южной Корее, ОАЭ и 80 странах Азии, Америки и Европы. Первый вариант использования с развертыванием подсистемы сегментации изображений головы и восстановления 3D модели головы в ЦОД и трансляции качественного видеопатока практически реализуем с большей вероятностью.

Второй вариант развертывания подсистемы сегментации и распознавания лиц во встроенном исполнении на базе Raspberry Pi 5 использующий специализированные облегченные(квантованные) нейронные сети для детектирования (нейронная сеть YOLOv8n квантованная) и идентификации (FaceNet квантованная) соответственно [11]. В данном исполнении видеопоток обрабатывается на борту дрона, что позволяет принимать решения в реальном времени даже в условиях ограниченной связи, соответственно не требуется высокая пропускная способность сети передачи данных. Выполняемые вычисления не должны быть требовательны к ресурсам и адаптированы для запуска на встроеных вычислителях, что делает данный вариант жизнеспособным для мобильного использования, но с ограниченным качеством распознавания, и соответственно такими же ограничениями по применению сценариев 3D моделирования головы человека и идентификации в реальном времени. В тоже время нейронные сети семейства YOLO в последнее время приобрели популярность среди разработчиков и доминируют в решениях для детектирования: улучшилось выделение признаков и разномасштабное обнаружение, появилась возможность применения в автономном вождении, наблюдении и в робототехнике [12] Чтобы решить задачу идентификации используются нейронные сети с низкой вычислительной сложностью, компактное хранилище данных, и высокую точность распознавания лиц [13] для изображений с изменяющимся качеством при

неблагоприятных погодных условиях. В статье [14] описаны мультимодальные системы, использующие несколько каналов для входных данных различной природы.

Аппаратно-программная база для второго варианта развертывания подсистемы сегментации. Аппаратная основа системы с обработкой видео непосредственно на встроенной системе дрона: Raspberry Pi 5 для выполнения вычислений, камера - источник видеопотока. Основа программной части: квантованная для запуска на Raspberry Pi 5 нейронная сеть YOLOv8n, предварительно обученная на наборе данных COCO для обнаружения лица человека, и также квантованная нейросеть DeepFace (FaceNet) для распознавания лица и поиска в базе данных существующих фото с использованием метрики косинусного подобия. Как можно заметить в предложенном варианте развертывания актуальными для практического использования являются те методы и алгоритмы детекции и идентификации лица человека, которые не требуют больших вычислительных мощностей.

Структура системы видеоаналитики

Структура системы видеоаналитики в которой могут быть использованы предлагаемые алгоритмы сегментации головы человека от фона для последующей реконструкции 3D модели головы человека в реальном времени и развертывания вычислений в ЦОД представлена на рис.2, так как 3D моделирование и поиск нужного ракурса для получения фото вычислительно сложные задачи.

Рис. 2. Архитектура системы видеоаналитики при передаче видеопотока об объектах интереса с дрона в центр обработки данных

Блок получения видео предназначен для получения последовательности кадров видеоизображения. Источником изображения может служить Любой стандартный источник видеопотока (windows video device); видеофайл в формате AVI; графический файл в форматах PNG, JPG, BMP; текстовый файл, содержащий список файлов изображений.

Блок предварительной обработки и стабилизации изображения предназначен для улучшения качества поступающих изображений. В основе данного блока лежат алгоритмы медианной фильтрации для устранения шума. Необходимо устранения возможного дрожания изображения вследствие вибраций камеры с помощью мультимасштабного анализа. Блок детекции лиц и последующей сегментации от фона (выделения оконтуренных объектов) на изображении.

Блок трекинга лиц, вычисление оптического потока предназначен для вычисления областей на изображении, где присутствует визуальное движение, вычисляется усредненная сумма векторов потока по всей области интереса, которая характеризует глобальное перемещение паттерна.

Блок калибровки камер системы предназначен для вычисления фундаментальной матрицы преобразования в соответствии с эпиполярной геометрией. В качестве исходных данных используются точки, указанные пользователем вручную. Вычисленная калибровка сохраняется в файле и может быть многократно использована, но только в том случае, если не изменилось взаимное расположение камер. Калибровка рассчитывается для каждой камеры в системе.

Блок стереосопоставления формирует на выходе стереолица и их 3D модели и выполняет идентификацию лиц непосредственно по их их 3D моделям, либо определяет оптимальный ракурс предназначеный для выбора наиболее качественного изображения лица с точки зрения идентификации. Для этого анализируются показатели качества лиц как в многоракурсном домене (стереолицо), так и во временном домене (история наблюдения за лицом).

Методы сегментации лица человека

Методы, применяемые для решения задач сегментации, классификации и идентификация. Сегментация лица человека от фона может использоваться в различных приложениях: например, восстановления 3D модели головы и лица [15], идентификации на основе использования фотографий построенной 3D модели лица и головы с одной стороны и фотографий от системы видеонаблюдения, что и предлагается использовать в экспериментальной системе видеоаналитики, архитектура которой описана ниже и представляет конвейер обработчиков данных.

Методы обнаружения изображения головы на статических кадрах, полученных из видеопотока, можно разделить на следующие:

Эмпирические методы, основанные на знаниях о внешнем виде головы. В таких методах используется набор правил, выполнение которых может привести к решению рассматриваемой задачи. Примерами таких правил могут быть для случая фронтального изображения головы, следующие: лицо имеет симметричную форму, центральная часть лица имеет практически не изменяющуюся текстуру. Как правило, они заключаются в установлении простых закономерностей, например, для фронтального изображения головы: лицо человека обычно симметрично, глаза, нос и рот, как правило отличаются по яркости и т.д. В качестве примеров можно назвать иерархические и проекционные методы [15].

Признаковые методы основаны на поиске отличительных особенностей лица на изображении и сопоставлении найденных признаков для восстановления позиции лица в целом. В качестве отличительных признаков могут рассматриваться контуры лица текстуры, цвет кожи, форма.

Методы моделирования предполагают использование некоторой модели лица, которое сопоставляется с представленным изображением. Модели могут состоять из контуров, характерных точек. В литературе встречаются подходы с жесткими и деформируемыми моделями.

Методы внешнего вида опираются на инструментарий распознавания образов. При этом классификатор настраивается на решение двух классовой задачи распознавания. Общее у методов, основанных на внешнем виде лица, состоит в алгоритме получения тестовых областей на изображении. Так как изображение лица может находиться в любом месте кадра и иметь любой размер, то для рассмотрения всех возможных вариантов необходимо перебрать все области на изображении при всех подходящих масштабах.

Каждая такая область рассматривается как тестовый образ и подается на вход классификатора. В качестве методов классификации могут использоваться: факторный анализ (FA – Factor Analysis); линейный дискриминантный анализ (LDA – Linear Discriminant Analysis); скрытые Марковские модели (HMM - Hidden Markov Models); активные модели внешнего вида (AAM – Active Appearance Models); метод опорных векторов (SVM – Support Vector Machines); разреженная сеть просеивающих элементов (SNoW - Sparse Network of Winnows); собственные лица (Eigenfaces); искусственные нейронные сети (NN – Neural Networks). Кроме того, нейросетевые методы могут использоваться для определения ориентации изображения лица, а также для каскадирования слабых классификаторов [16].

Последние два метода рассматриваются сегодня как лучшие с точки зрения высокого процента верного обнаружения лиц и низкого процента ложных детекций. Нейросети давно с успехом применяются для решения поиска локальных экстремумов является базовой операцией для множества задач обработки изображений.

Нейросетевые методы позволяют добиться высокого процента верного обнаружения лиц и низкого процента ложных детекций. Нейросети давно с успехом применяются для решения многих задач в области машинного зрения. Недостатком нейросетевых методов является их

вычислительная сложность. Они не применимы в чистом виде для решения задач детекции лиц в реальном времени. Для решения этой проблемы предпринимаются различные шаги. Например, в [16] для обеспечения работы детектора лиц на базе сверточной нейронной сети в реальном времени используется специализированный нейросетевой процессор.

Алгоритм сегментации изображений головы человека от фона

Алгоритмы удаления фона на изображениях головы человека используют такие архитектуры глубоких сверточных нейронных сетей, как U-Net, DeepLab-V3+ и Mask R-CNN, которые показывают высокое качество сегментации, а для работы на мобильных устройствах и в реальном времени предложены оптимизированные архитектуры: LiteUNet, MODNet и PortraitNet. Они сочетают в себе приемлемое качество сегментации и высокую скорость обработки, что делает их подходящими для внедрения в мобильные потребительские устройства. Одним из наиболее перспективных направлений является использование Transformer-архитектур, таких как Segmenter, Trans-UNet, Swin-Unet, которые особенно эффективны в обработке информации [17–23]. В ходе реализации проекта разработаны и экспериментально исследованы алгоритмы сегментации при помощи нейросетевых архитектур UNet, PortraitNet, Segmenter, SAM. Модели U-Net, PortraitNet, Segmenter, SAM были обучены на наборе данных, представленном исходными изображениями головы человека в разных ракурсах и эталонными масками сегментации. Модель SAM использовалась без дообучения с предварительно расстановкой опорных точек для сегментации. Результаты тестирования показали, что наилучшую точность демонстрирует модель Segmenter (*IoU* = 0,9739, *Dice*(*F1*) = 0,9867, *Accuracy* = 0,9973), табл 1.

При вычислениях метрик точности *IoU* (Intersection over Union) или *Jaccard Index* – отношение пересечения (предсказанной и истинной масок) к их объединению и *Accuracy* (точность) – доля правильно классифицированных пикселей и *Dice* коэффициента (*F1* метрика) приняты следующие обозначения: *TP* (True Positive) - пиксели, правильно предсказанные как объект; *TN* (True Negative) – пиксели, правильно предсказанные как фон; *FP* (False Positive) – пиксели, ошибочно предсказанные как объект (ложные срабатывания).

Dice коэффициент вычисляется по формуле (1), *IoU* и *Accuracy* по формулам (2, 3)

$$Dice = \frac{2TP}{2TP + FP + FN}, \tag{1}$$

$$IoU = \frac{TP}{TP + FP + FN}, \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

Приведем описание алгоритма для нейросетевой архитектуры Segmenter.

1. Разбиение изображения на патчи. Входное изображение разбивается на фиксированные блоки (патчи) одинакового размера, например, 16×16 пикселей. Каждый патч преобразуется в вектор признаков с помощью линейного слоя.

2. Добавление позиционных эмбеддингов. Каждому визуальному токену (вектору патча) добавляется позиционный эмбеддинг, который указывает на его местоположение в исходном изображении. Это позволяет модели учитывать пространственную структуру изображения.

3. Обработка через Transformer-encoder. Последовательность токенов подается в Transformer-encoder, где каждый токен взаимодействует со всеми остальными через механизм внимания (self-attention). На этом этапе модель выделяет глобальные признаки и устанавливает связи между различными частями изображения.

4. Инициализация масочных токенов. Decoder использует набор обучаемых масочных токенов (learnable mask tokens), каждый из которых отвечает за формирование определённой семантической маски (например, для каждого класса объекта).

5. Восстановление масок через кросс-внимание. Масочные токены взаимодействуют с закодированными признаками из encoder'а через кросс-внимание, что позволяет decoder'у формировать детализированные маски, соответствующие реальной структуре изображения.

6. Генерация финальной карты сегментации. На выходе decoder'а получается набор масок, каждая из которых соответствует определённому объекту или классу. Эти маски объединяются в финальную карту сегментации, совпадающую по размерам с исходным изображением.

Табл. 1. **Результат тестирования моделей нейронных сетей сегментации**

Архитектура	IoU	Dice	Accuracy
U-net	0,9307	0,9641	0,9927
PortraitNet	0,9159	0,9560	0,9914
Segmenter	0,9739	0,9867	0,9973
SAM	0,6634	0,7931	0,9597

Заключение

Доклад посвящен рассмотрению возможных вариантов использования(развертывания) подсистемы сегментации изображения головы челвека от фона на фотографиях и выбору перспективных алгоритмов для программной реализации в подсистеме. Описан разработанный алгоритм подготовки данных из видеопотока (удаления фона на изображениях головы человека), являющийся важной частью решения задачи реконструкции 3D модели головы человека по видеопотоку, выполнено экспериментальное исследование алгоритмов удаления фона. Выполнен анализ существующих подходов к семантической сегментации изображений головы человека на фотографиях для последующей 3D реконструкции модели головы человека. Установлено, что для условий мобильной съемки с борта БПЛА наиболее перспективным является извлечение ключевых кадров, их масштабирование с сохранением пропорций и оценка резкости с использованием оператора Лапласа, что позволяет эффективно отбирать кадры для дальнейшей сегментации. Для решения задачи удаления фона (бинарной сегментации) протестированы современные архитектуры нейронных сетей: U-Net, PortraitNet, Segmenter, SAM. Модели U-Net, PortraitNet и Segmenter, которые были обучены на наборах данных, представленных исходными изображения головы человека в разных ракурсах и эталонными масками сегментации. Модель SAM использовалась без дообучения с предварительной расстановкой опорных точек для сегментации. Результаты тестирования показали, что наилучшую точность демонстрирует модель Segmenter, а вариант использования или развертывания подсистемы сегментации с практической точки зрения наиболее перспективен с использованием значительных вычислительных мощностей центра обработки данных.

Работа выполнена в рамках договора с БРФФИ Ф25ТУРГ-001 от 05 марта 2025 г., тема "Приложение для генерации компьютерной 3D модели головы человека и ее сравнение с изображениями, полученными с камеры".

PROTOTYPING A HEAD-BACKGROUND SEGMENTATION SUBSYSTEM AS PART OF A CENTRALIZED VIDEO ANALYTICS SYSTEM BASED ON UAV MACHINE VISION

M.M. LUKASHEVICH, A.A. VORONOV, V.V. VENGERENKO, M.M. TATUR

Abstract. The article is devoted to not only processing aerial photographs of human face for the task of safety monitoring, but also to the review of possible use cases for a prototype head-background segmentation subsystem within a centralized UAV-based video analytics system and development of algorithms for processing and segmentation of photoimages of human face for the task of safety monitoring. Segmentation algorithm using a modified neural network architecture is proposed. It is shown to be significantly faster and more efficient than the best known segmentation algorithms and state of art neural network architectures.

Keywords: semantic segmentation, video analytics, UAVs, incident detection, classification.

Список литературы

1. Ye Z., Lan C. [et al.] // Journal of Visual Communication and Image Representation. 2023. Vol. 95, P. 103903–103910.
2. Yin N., Liu C. [et al.] // IEEE Transactions on Multimedia. 2024. Vol. 26. P. 7812–7822.
3. Zhou G., Xu Q. [et al.] // IEEE Transactions on Geoscience and Remote Sensing. 2025. Vol. 63. P. 1–13.
4. Wang Y., Yang Y. [et al.] // Proc. of European Conference on Computer Vision. 2020. P. 651–664.
5. Слюсар В. // Электроника НТБ Часть 1. 2009. № 5. С. 68–73.
6. Lazzaroni F., Leonardi R. [et al.] // IEEE Signal Processing Letters. 2003. Vol. 10. P. 28–31.
7. Deniz O., Bueno G. [et al.] // Pattern Recognition. Vol. 44. P. 2887–2901.
8. Цветков В. // Доклады БГУИР. 2019. №. 3(121). С. 25–36.
9. Sullivan G., Ohm J. [et al.] // IEEE Transactions on Circuits and Systems for Video Technology. 2012. Vol. 22. P.1649–1668
10. Вишневский В., Портной С. [и др.] Энциклопедия WiMax. Путь 4G. М., 2009.
11. Yang Y., He H. [et al.] // Algorithms. 2025. Vol. 18. P. 453–460.
12. Nagpal S., Etnubarhi N. [et al.] // Proc. of Innovations in Intelligent Systems and Applications Conference (ASYU). 2024. P. 1–6
13. Ren Z., Liu X. [et al.] // Journal of Imaging. 2025. Vol. 11. P. 24-30.
14. Wu W., Chen X. [et al.] // Proc. of International Joint Conference on Neural Networks (IJCNN). IEEE. 2024. P. 1–8.
15. Вежневец В. // Математические методы распознавания образов (ММРО-10): 10-я Всероссийская конференция. 2001. С. 179–180.
16. McRaven J., Scheutz M. [et al.] // IEEE international workshop Cellular Neural Networks and their Applications (CNNA 2004). 2004. P. 196-201.
17. Wu C. // Proc. of the International Conference on 3D Vision (3DV). 2013. P. 127–134.
18. Yao Y., Luo Z. [et al.] // Proc. of the European Conference on Computer Vision (ECCV). 2018. P. 767–783.
19. Gu X., Fan Z. [et al.] // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 3265–3274.
20. Choy C., Xu D. [et al.] // Proc. of the European Conference on Computer Vision (ECCV). 2016. P. 628–644.
21. Qi C., Yi L. [et al.] // Proc. of conf. Advances in Neural Information Processing Systems (NIPS). 2017. P.50 99–5108.
22. Mildenhall B., Srinivasan P. [et al.] // Proc. of the European Conference on Computer Vision (ECCV). 2020. P. 405–421.
23. Saito S., Huang Z. [et al.] // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. P. 2304–2314.