ПРОТОТИПИРОВАНИЕ ПОДГОТОВКИ ДАННЫХ ИЗ ВИДЕОРЯДА И СЕГМЕНТАЦИИ ДЛЯ ГЕНЕРАЦИИ 3D МОДЕЛИ ГОЛОВЫ ЧЕЛОВЕКА ПО 2D ИЗОБРАЖЕНИЯМ

Лукашевич М. М., Венгеренко В. В., Воронов А. А.
Кафедра информационных систем управления, Белорусский государственный университет
Объединенный институт проблем информатики Национальной Академии наук Беларуси
Минск, Республика Беларусь
E-mail: lukashevichmm@bsu.by, {vengerenko, voronov}@lsi.bas-net.by

Материал посвящен описанию разработки приложения для автоматического создания 3D модели головы человека на основе изображений, полученных с камеры. Приложение позволит исследовать модель во всех проекциях и найдет применение в медицине и судебной экспертизе.

Введение

Современные подходы к разработке алгоритмов для восстановления 3D модели лица и головы из одиночного или набора изображений позволяют автоматически создавать 3D модель головы человека и исследовать такую модель в различных проекциях. Решение этой задачи требует применения специализированных программных и технических средств, разработка которых – актуальное и перспективное направление научно-технической деятельности. Применяемые алгоритмы восстановления 3D модели, реализованные программно, определяют требования к необходимому аппаратному обеспечению (камерам видеонаблюдения) систем видеоаналитики, которые направлены на обеспечение мониторинга безопасности и полезны в судебной экспертизе или медицине.

I. Алгоритм восстановления 3D модели головы по 2D изображениям

Алгоритм восстановления 3D модели головы по 2D изображениям состоит из следующих шагов [1]:

1. Подготовка данных из видеопотока.
2. Выделение объектов интереса и отделение их от фона, сегментация изображений, дальнейшая обработка, связанная с удалением неинформативных артефактов и определение очертаний объектов интереса. В основе всех известных алгоритмов восстановления 3D модели лежит именно выделение объектов интереса, отделение их от фона и сегментация.
3. Построить трехмерную сетку и вписать в нее полученные изображения.
4. Текстурирование 3D модели.

Рассмотрим подробнее шаги 1–2 представленного алгоритма. Подготовка данных из видеопотока состоит из следующих шагов:

1. Вычисление основных характеристик видеопотока. Ширина, высота изображения и частота кадров (FPS).
2. Масштабирование изображения. Рассчитывается коэффициент уменьшения для приведения кадров к максимальному размеру 720 пикселей по большей стороне.
3. Выбор ключевых кадров. Выполняется анализ кадров по показателю резкости с помощью оператора Лапласа, который применяется к изображению после перевода его в полутон и бинаризации. Для каждой секунды видео определяется самый четкий кадр из диапазона 5–25 кадров. Это позволяет пропускать возможные артефакты в начале секунды.
4. Сохранение результатов. Оригинальные ключевые кадры сохраняются в соответствующий каталог «№секунды.png».
5. Определение пяти ключевых кадров на основе классификационной модели. Ключевые кадры из видеопотока – анфас, вид слева, справа, сзади, сверху.

II. Алгоритмы сегментации изображений головы 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 [2–8]. 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 изображения головы человека в разных ракурсах и эталонными масками сегментации. Модель SAM использовалась без дообучения с предварительно расстановкой опорных точек для сегментации. Ре-

зультаты тестирования показали, что наилучшую точность демонстрирует модель Segmenter (IoU = 0,9739, Dice = 0,9867, Accuracy = 0,9973).

Приведем описание алгоритма для нейросетевой архитектуры Segmenter.

1. Разбиение изображения на патчи. Входное изображение разбивается на фиксированные блоки (патчи) одинакового размера, например, 16x16 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'a через кросс-внимание, что позволяет decoder'у формировать детализированные маски, соответствующие реальной структуре изображения.
6. Генерация финальной карты сегментации. На выходе decoder'a получается набор масок, каждая из которых соответствует определённому объекту или классу. Эти маски объединяются в финальную карту сегментации, совпадающую по размерам с исходным изображением.

Пример результатов сегментации приведен на рис. 1.

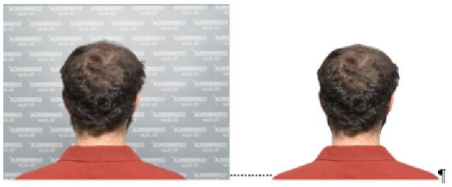

Рис. 1 – Примеры результатов сегментации

Заключение

В ходе работы были разработаны алгоритмы подготовки данных из видеопотока и удаления фона на изображениях головы человека, являющиеся важной частью решения задачи реконструкции 3D модели головы человека по серии фотографий, и выполнено их экспериментальное исследование. Выполнен анализ существующих подходов к 3D реконструкции головы человека: методов фотограмметрии, нейросетевых моделей и ручного моделирования. Установлено, что для условий мобильной съемки наиболее перспективным является использование методов фотограмметрии. Реализовано также 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 (бинарная сегментация) протестированы четыре современные архитектуры нейронных сетей: U-Net, PortraitNet, Segmenter, SAM. Результаты тестирования показали, что наилучшую точность демонстрирует модель Segmenter.

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".

Список литературы

1. Schönberger, J. L. Structure-from-Motion Revisited / J. L. Schönberger, J. M. Frahm // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 4104–4113. – DOI: 10.1109/CVPR.2016.445.
2. Wu, C. Towards Linear-Time Incremental Structure from Motion / C. Wu // Proceedings of the International Conference on 3D Vision (3DV). – 2013. – P. 127–134. – DOI: 10.1109/3DV.2013.25.
3. Yao, Y. MVSNet: Depth Inference for Unstructured Multi-View Stereo / Y. Yao, Z. Luo, S. Li, T. Fang, L. Quan // Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – P. 767–783. – DOI: 10.1007/978-3-030-01237-3_47.
4. Gu, X. Cascade Cost Volume for High-Resolution Multi-View Stereo / X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, P. Tan // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 3265–3274. – DOI: 10.1109/CVPR42600.2020.00257.
5. Choy, C. B. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction / C. B. Choy, D. Xu, J. Gwak, K. Chen, S. Savarese // Proceedings of the European Conference on Computer Vision (ECCV). – 2016. – P. 628–644. – DOI: 10.1007/978-3-319-46484-8_38.
6. Qi, C. R. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space / C. R. Qi, L. Yi, H. Su, L. J. Guibas // Advances in Neural Information Processing Systems (NIPS). – 2017. – P. 5099–5108.
7. Mildenhall, B. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis / B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng // Proceedings of the European Conference on Computer Vision (ECCV). – 2020. – P. 405–421. – DOI: 10.1007/ 978-3-030-58452-8_24.
8. Saito, S. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization / S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, H. Li // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2019. – P. 2304–2314. – DOI: 10.1109/ICCV.2019.00239.