ЭНЕРГОЭФФЕКТИВНЫЙ АЛГОРИТМ АВТОМАТИЧЕСКОЙ ЗАМЕНЫ ФОНА ДЛЯ ВИДЕОКОНФЕРЕНЦИЙ

Пирштук Д. И.
Кафедра дискретной математики и алгоритмики, Факультет прикладной математики и информатики,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: PirshtukDI@bsu.by

В работе предложен алгоритм отделения человека от фона на видео на основе быстрой полносверточной сегментацинной нейронной сети и специальной процедуры обнаружения движения сегментируемого объекта в кадре. Предлагаемый алгоритм потребляет около 3-6% CPU при обработке HD-видеопотока скоростью 30 кадров/сек и может может использоваться для замены виртуального фона в видеоконференциях и других случаях дополненной реальности без существенного увеличения энергопотребления, разрядки аккумулятора ноутбука или мобильного телефона.

Введение

Для эффективной удаленной коммуникации между людьми важны не только передача аудио, но и видео, так как большинство людей являются визуалами, которые в процессе диалога любят использовать жесты и хотят видеть эмоции и жесты собеседников. Поэтому организаторы видеоконференций настоятельно просят участников включать камеру во время созвонов. Но с другой стороны включенная камера при созвоне из дома создает конфликт между рабочим и личным пространством.

Замена фона с использование высококонтрасного однотонного экрана (хромакея) хоть и является хорошим аппаратным решением, но очевидно неудобно для домашнего использования. Поэтому является актуальной задача программно отделить человека от фона в обычном видеопотоке с веб-камеры и заменить фон на виртуальный. Обработка видеопотока в реальном времени в облаке затруднительна из-за затрат, высоких задержек при передаче данных через Интернет и проблем конфиденциальности данных. Решение должно работать именно на клиентском устройстве, причем функционал виртуальной замены фона во время видеозвонка должен потреблять минимальное количество дополнительных вычислительных ресурсов, так как прием-передача и отображение нескольких видеопотоков уже и так является достаточно ресурсоемкой задачей. Более того, в отличие от настольных компьютеров, многие ноутбуки и смартфоны не могут долго работать с высокой производительностью. Либо процессор перегревается и снижает тактовую частоту, либо быстро разряжается аккумулятор.

Предлагаемое решение состоит из двух компонент – модуля обнаружения движений в кадре и нейронной сети для очень быстрой сегментации человека в кадре, и потребляет около около 3-6% CPU.

I. Быстрая сегментация человека

Для отделения человека от фона предлагается использовать асимметричную полносверточную U-Net-подобную [1] архитектуру типа Encoder-Decoder с использованием MobileNetV3Small-0.75x [2] в качестве кодировщика и легковесным декодером. Для объединения признаков использовалась модификация модуля Squeeze-and-excitation [3] в качестве модуля внимания. В качестве функции потерь использовалась функция потерь, описанная в [4].

Первоначальное обучение нейросети выполнялось в разрешении 256×256 с последующим дообучением под рабочее разрешение 256×144. Заметим, что это ровно 1/5 от HD-разрешения 1280×720, наиболее часто используемого в видеозвонках, что позволяет эффективно уменьшать размер кадра без дополнительной интерполяции.

Предварительное обучение выполнялось на смеси разных изображений: фотографии людей в различных условиях в полный и частичный рост со всевозможными перекрытиями и другими объектами в кадре, профессиональные портреты, селфи, снятые на камеру мобильного телефона, все возможные интерьеры без человека и др. Но заметим, что сверточные сети не инвариатны к изменению масштаба объектов и перекрытиям. Частично эта проблема может быть компенсирована увеличением количества фильтров на слоях и модулем контекстной аггрегации. Для достижения максимального качества сегментации при минимальном количестве вычислений вместо этого мы выполнили дообучение на размеченных кадрах видеозвонков для адаптации сети под сегментацию человека, сидящего чаще всего за столом при различных освещениях и жестикулирующего.

Для оценки качества использовалась метрика Intersection-over-Union (IoU) между размеченной и предсказанной масками $\frac{TP}{TP+FP+FN}$, где TP, FP и FN – количество истинно-положительных, ложно-положительных и ложно-отрицательных пикселей соотвественно.

Предложенная архитектура требует около 59 миллионов операций с плавающей точкой для обработки одного кадра, что занимает около 4 мс на обработку одного кадра в один поток на ноутбуке с процессором Intel Core i7–7500U на фреймворке Tensorflow Lite с использованием XNNPACK делегата [5]. Обученная нейросеть достигла точности сегментации по метрике IoU в 96,04% на нашем тестовом датасете. Пример сегментации на кадре из тестового видео приведен на рисунке 1.

II. Определение движения в кадре

Заметим, что движения рта, изменения выражения лица, движения глаз не влияют на границу между между человеком и фоном, а изменятся граница в следующих случаях:

1. Выход части тела за пределы кадра.
2. Появление части тела в кадре.
3. Движение частей тела человека, например, жесты или кивки головой.

Предлагается следующая процедура:

1. Выделим все граничные пиксели между человеком и фоном путем усредняющего размытия предсказанной сегментационной маски с ядром 11 и включеним в границу пиксели со значениями маски 0,2–0,8. Количество пикселей на границе обозначим через N.
2. Возьмем в качестве области интереса все пиксели границы, а также полоски по 5 пикселя периметра кадра.
3. Будем считать, что пиксель нового кадра подозрительный, если он принадлежит области интереса и изменился в оттенках серого на 10% диапазона 0–255. Пусть M – количество подозрительный пикселей.
4. Не пересчитываем сегментационную маски, если $M/(N+1) < 2{,}5\%$ и количество непрерывно пропущенных кадров меньше 30.

Процедура применяется к уменьшенному до разрешения 256×144 изображению. Константы подобраны по датасету видеозвонков и обусловлены в том числе тем, что веб-камеры большинства потребительских устройств достаточно сильно шумят.

Эксперименты показывают, что люди удовлетворены качеством замены фона, если IoU больше 95%. Алгоритм допускает пропуск пересчета сегментацинной макси, если метрика IoU для нового кадра со старой маской не может быть более, чем на 2,5% хуже IoU для изначальной пары. Так как среднее IoU нашей сегментационной нейронной сети больше 96%, то если сегментация на исходной паре имела высокое значение метрики IoU, то и для нового кадра со старой маской при выполнении критерия пропуска значение IoU будет удовлетворительным. При этом на случай, когда изначальная пара была случайно отсегментирована сильно хуже среднего ожидаемого IoU и отслеживаемая граница неверная, предусмотрен принудительный пересчет хотя бы раз в 30 кадров.

Данная процедура позволяет пропускать на видеозвонках в среднем около 60% кадров, снижая среднее потребление CPU до 3–6%.

Заключение

Предложен алгоритм отделения человека от фона на видео на основе нейронной сети и критерия обновления сегментационной маски со среднем потреблением на обычных CPU потребительких устройств в 3–6% при обработке видеопотока 30 кадров/сек. Предложение может быть также использовано для других задач улучшения видеозвонков, таких как сглаживание кожи, перекраска губ, перекраска волос, и в других сегментационных задачах обработки видеопотока в реальном времени.

1. Ronneberger, O. U-net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer , T. Brox. // Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. – Springer International Publishing, 2015. – P. 234–241.
2. Howard, A. Searching for MobileNetV3 / A. Howard [et al.] // Proceedings of the IEEE/CVF international conference on computer vision. – 2019. – P. 1314–1324.
3. Hu, J. Squeeze-and-excitation networks / J. Hu, L. Shen, G. Sun // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – P. 7132–7141.
4. Zhang, S. PortraitNet: Real-time portrait segmentation network for mobile device / S. Zhang [et al.] // Computers & Graphics. – 2019. – Vol. 80. – P. 104–113.
5. Accelerating TensorFlow Lite with XNNPACK Integration [Electronic resource] – Mode of access: https://blog.tensorflow.org/2020/07/accelerating-tensorflow-lite-xnnpack-integration.html. – Date of access: 16.10.2023.

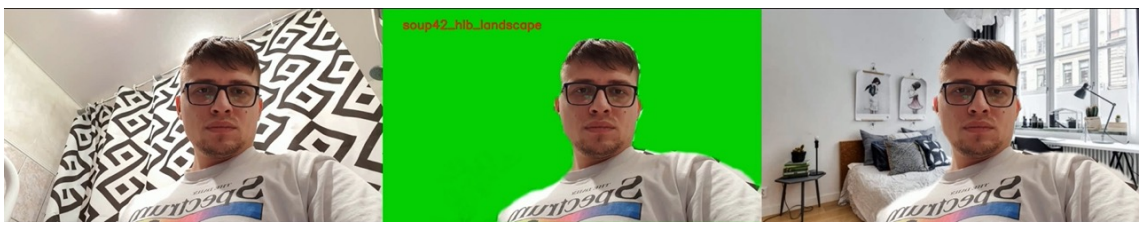

Рис. 1 – Пример замены фона на видео. Слева направо: исходный кадр, результат сегментации, виртуальная замена фона