Разработка нейросетевого алгоритма оценки карт визуального внимания для 360-градусных видео

Плошкин Александр Игоревич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: alexander.ploshkin@.graphics.cs.msu.ru

Научный руководитель *— Ерофеев Михаил Викторович*

360-градусное видео — это видео, предназначенное для просмотра в шлемах виртуальной реальности (VR-шлемах). Поскольку во время просмотра зритель видит лишь небольшую его часть, оно должно иметь более высокое разрешение, чем видео, предназаначенное для просмотра на экране. Стандартом в индустрии для 360-градусных видео сейчас считается разрешение 4K или 4096×2160 пикселей, но для человеческого глаза этого недостаточно — из-за близкого расстояния от экрана до роговицы пиксели по-прежнему различимы, поэтому в будущем разрешение экранов и, соответственно, видео будет повышаться. Видео в высоком разрешении требует больше памяти для хранения и создаёт сильную нагрузку на сеть при передаче. Одним из способов снижения объёма видео являются методы контекстного сжатия видео [1], которые учитывают области визуального внимания человека, в результате чего кадры сжимаются неравномерно. Для эффективного контекстного сжатия необходимо наиболее точно оценить карту визуального внимания человека, то есть определить области, на которые более вероятно будет смотреть человек при просмотре видео.

Целью работы является создание нейросетевого алгоритма оценки карты визуального внимания человека для 360-градусных видео. На вход подаётся 360-градусное видео в виде равнопромежуточной проекции, на выходе — карта визуального внимания человека для каждого кадра входного видео.

К алгоритму предъявляются следующие требования:

1. скорость работы: >1 FPS на Nvidia GeForce GTX1080;

2. высокая точность работы в смысле общепринятых метрик, таких как Normalized Scanpath Saliency (NSS) и расстояние Кульбака–Лейблера (KL);

3. временная стабильность: карты внимания для соседних кадров не могут значительно отличаться друг от друга.

В качестве основной нейросетевой модели, извлекающей высокоуровневые признаки изображения, была выбрана ResNeXt-50 [2]; для учёта временных зависимостей между кадрами была использована модель ConvLSTM [3].

Равнопромежуточная проекция сферы имеет геометрические эффекты, которые операция свёртки, лежащая в основе подавляющего большинства нейросетевых моделей для обработки изображений, не может адекватно обработать. Например, размеры объектов меняются в зависимости от расстояния до экваториальной линии, а их форма сильно искажается. Поэтому для обработки кадров 360-градусных видео в этой работе используется кубическая проекция: каждый кадр видео проецируется на поверхность куба, в таком случае краевые эффекты на гранях гораздо менее заметны. Затем изображения с граней полученной кубической проекции обрабатываются нейросетевой моделью независимо, и итоговая карта внимания получается обратным проецированием куба на плоскость.

В качестве обучающей выборки был использован набор видео с записанными движениями глаз при просмотре, предоставленный ИППИ РАН им. А. А. Харкевича. Всего он включает в себя 20 видео, в просмотре которых участвовало 90 человек.

Литература

1. Lyudvichenko V et al. Improving Video Compression With Deep Visual-AttentionModels // In Proceedings of the 2019 International Conference on Intelligent Medicine and Image Processing, Firenze, Italy, 2019, P. 88–94.
2. Xie S et al. Aggregated residual transformations for deep neural networks // In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, P. 1492–1500.
3. Xingjian S. H. I. et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting // In Advances in neural information processing systems, 2015, P. 802–810.