Новые идеи: оркестровка видео, звука и метаданных

Александр Серов

В прежние годы кинокамерами владели немногие и демонстрировали свои съемки только узкому кругу лиц – семье, друзьям или сослуживцам. Процесс изготовления контента был трудоемок и затратен. К тому же кинопленки было практически невозможно копировать (в бытовых условиях, разумеется). Потом появились видеокамеры, позволившие упростить процесс создания движущихся изображений. Видео дало возможность расширить аудиторию – достаточно было раздобыть два видеомагнитофона, чтобы скопировать понравившуюся запись. Но аудитория все равно оставалась ограниченной.

Затем появился Интернет, а с ним и возможность выкладывать созданные видеозаписи на всеобщее обозрение. А затем произошло нечто и вовсе невероятное – появились смартфоны. И теперь видеокамеры есть почти у каждого в кармане. И каждый способен создать свой маленький телеканал – размещать свои записи, проводить прямые трансляции. Кроме этого, появление смартфонов позволило производить метаданные – сохранять координаты мест съемок, записывать движение камеры, условия съемок, обнаруживать на записи различные объекты и так далее. А в будущем развитие машинного обучения сделает производство метаданных еще более масштабным.

Возникло сразу несколько особенностей, связанных с избыточностью любительского контента. Например, если мероприятие популярно, на сайтах вроде YouTube можно обнаружить десятки, а то и сотни роликов или прямых трансляций, сделанных со смартфонов. Множество людей занимаются одним и тем же – производят контент, посвященный одному и тому же событию.

Существует специальный термин для любительского контента, записанного при помощи смартфонов или видеокамер, – UGC или user-generated content (контент, созданный пользователями или пользовательский контент).

К настоящему моменту только на одном Facebook ежедневно полмиллиарда человек просматривают пользовательское видео, а 84% пользователей публикуют свои собственные видеоролики. Скорость загрузок видео на Youtube составляет 72 ч/мин.

Можно ли как-то объединить этот контент, дав зрителю эффект погружения в происходящие события, возможность посмотреть на него сразу сотнями глаз? А использовать при этом метаданные? Например, при трансляции с концерта показывать синхронизированную информацию об исполнителе и музыкальном произведении, либо давать ссылку на такую информацию. Другая возможность – использовать любительские записи при производстве телепрограмм.

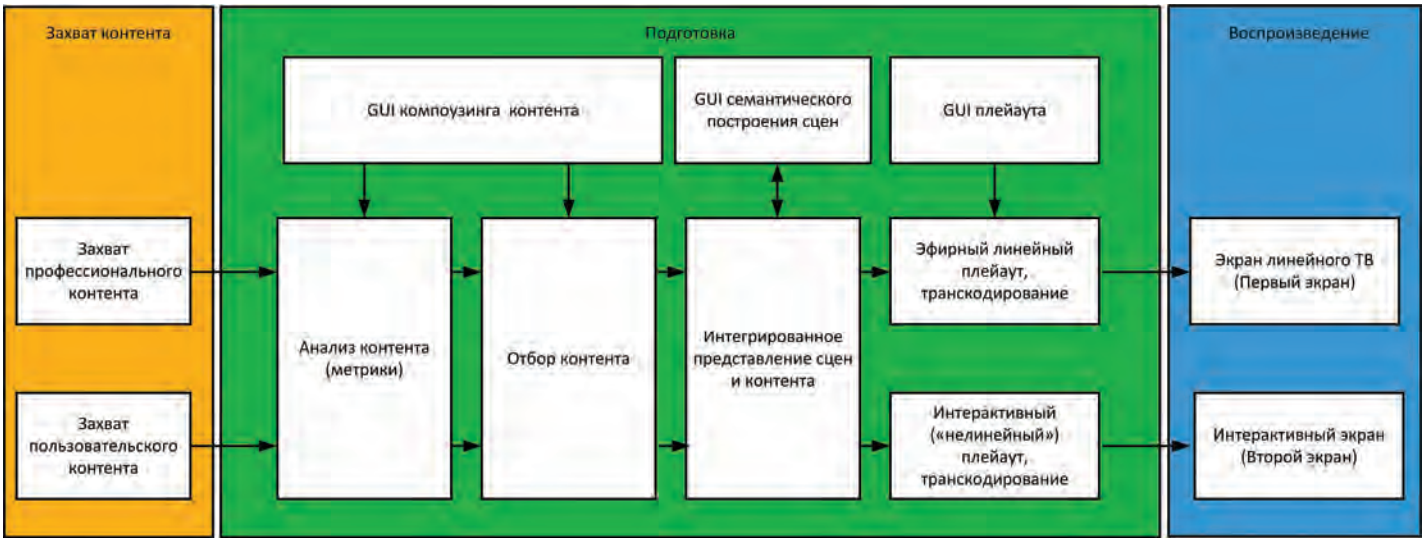
Одна из возможностей создания подобного «склеенного» контента – это формирование так называемого immersive video, что можно перевести как «видео с вовлечением в действие» или «видео с погружением». Immersive video создается уже давно при помощи специальных камер, дающих 360-градусный обзор. При просмотре пользователь имеет возможность при помощи мыши изменять ракурс, как бы поворачивая свою виртуальную голову. Подобный прием если не дает эффект присутствия на месте съемки, то, безусловно, обогащает восприятие. Однако 360-градусные панорамы создаются намеренно. А можно ли сделать что-то подобное без предварительной подготовки из любительских съемок со смартфонов, как упоминалось выше?

Здесь есть несколько сложностей, которые кажутся труднопреодолимыми. Для того чтобы правильно «собрать» видео, требуется знать положение точки съемки в пространстве. Кто-то снимает концерт из первого ряда, а кто-то – из середины зала. Съемки должны быть близки по формату. Кто-то снимает SD, а кто-то UHD. Анализировать технические параметры записи напрямую нельзя, поскольку UHD может быть снято камерой, которая не обеспечивает реальное качество UHD. Материалы должны быть приемлемы и сопоставимы по качеству. Кто-то снимает статичное изображение, кто-то делает панораму, а кто-то никак не может справиться с дрожанием рук. Таким образом, необходимы какие-то критерии, чтобы оценить как техническое, так и операторское качество материалов. В случае, если качество не соответствуют необходимому, должны быть предусмотрены алгоритмы гармонизации – какие-то съемки можно улучшить, а какие-то намеренно «ухудшить» и так далее.

Идея использовать UGC для того, чтобы создавать новые видео- и телевизионные материалы, нашла живой отклик в массах. Особенно она понравилась специалистам по рекламе, поскольку исследованиями было установлено, что «самодельный» контент вызывает большее доверие у зрителей.

Чтобы решать проблемы, обозначенные выше, появилось несколько проектов-стартапов. Один из крупнейших – европейский COGNITUS (cognitus-h2020.eu) был создан на грант Евросоюза и ставил перед собой цель создать набор средств (платформу), которые бы позволяли осуществлять конвергенцию UGC и вещательного контента. На сайте проекта сообщается, что были разработаны решения, позволяющие выполнять задачи увеличения формата изображения, преобразования скорости потоков, адаптации динамического диапазона изображения, стабилизации, преобразования ориентации кадра из портретного в альбомный формат. Для аудиодорожек UGC были разработаны

Съемка UGC для COGNITUS



Архитектура системы, предложенной ICoSOLE

алгоритмы коррекции шума ветра, отсечения звука (clipping) и синхронизации перекрывающихся потоков. Разработаны метрики интернет-маркетинга – оценки качества контента, методики создания объемного звука из большого числа аудиодорожек, полученных от разных зрителей.

Еще один проект, посвященный работе с пользовательским контентом, создан в виде консорциума нескольких известных предприятий, среди которых, например, Би-би-си. Консорциум называется ICoSOLE и целью своей ставит создание платформы, которая позволит комбинировать пользовательский контент с профессиональным контентом для проведения в прямом эфире трансляций различных зрелищных мероприятий. Консорциумом разрабатываются приложения для получения, обработки и распространения такого комбинированного контента.

Из схемы видно, что архитектура платформы состоит из трех больших и очевидных частей – получения информации (раздельно профессиональной и пользовательской), обработки и воспроизведения. Для воспроизведения планируется использование двух экранов – основного линейного и вспомогательного интерактивного. В качестве последнего может выступать, например, смартфон или планшетный компьютер. Самое важное содержится в блоке «Подготовка», где выполняются функции отбора контента и конструирования сцен из полученных «кирпичиков».

Два описанных выше проекта являются лишь частными случаями более общей задачи синхронизации различных информационных потоков: видео, аудио, метаданных. Для подобной синхронизации, предусматривающей получение информации и ее адаптацию, используется музыкальный термин orchestration (буквально: «оркестровка» или «гармонизация»). Задачи, для которых может использоваться оркестровка контента, выходят за границы телевизионных приложений. Это может быть утилитарное построение 360-градусных сцен из многих пользовательских фотографий для мониторинга, отслеживания перемещения лиц по фотографиям из разных источников, построения сцен виртуальной или дополненной реальности.

Как всегда, при появлении новых идей в их разработку инвестируют участники рынка и только позже подключаются разработчики стандартов. В случае с UGC это правило тоже выполняется. Упомянутые выше проекты являются коммерческими, и технологии, которые будут разработаны, – это специализированные (закрытые) технологии. Хотя COGNITUS заявляет, что часть решений будет распространена в виде кода open source.

На момент написания данной статьи полный стандарт оркестровки видеоданных отсутствует. Но некоторые его элементы существуют. В лидерах, как всегда, консорциум DVB, который еще в 2015 году опубликовал спецификацию ETSI TS 103 286 Companion Screens and Streams («Вспомогательные экраны и потоки»), состоящую из двух частей:

- Part 1: Concepts, roles and overall architecture – «Часть 1. Концепции, роли и общая архитектура»;
- Part 2: Content Identification and Media Synchronization – «Часть 2. Идентификация контента и синхронизация медиаинформации».

Как ясно из названия, в этом стандарте идет речь о создании экосистемы, в которой действуют первый и второй экраны, причем последний служит для показа синхронизированного дополнительного контента. Этим дополнительным контентом и может быть упоминавшийся выше UGC. Синхронизированный контент в данном случае означает, что вспомогательный контент синхронизирован с контентом на основном экране. Например, трансляция, ведущаяся со смартфона одного из зрителей во время футбольного матча, синхронизирована с основным видеоматериалом. То же касается и метаданных – положение игрока на футбольном поле может быть отмечено как на основном экране, так и на видео, которое транслируется с пользовательского устройства. ETSI TS 103 286 содержит много нюансов. Например, он учитывает ситуацию, в которой пользовательский контент и метаданные могут передаваться одновременно разным операторам, имеющим разные сценарии совмещения основного и дополнительного контента.

Развитием ETSI TS 103 286 в ближайшие годы может оказаться новый стандарт, опубликованный ISO в августе 2019 года под названием ISO/IEC 23001-13:2019 Information technology – MPEG systems technologies – Part 13: Media orchestration («Информационные технологии – Технологии систем MPEG – Часть 13: Оркестровка медиаинформации»).

Данный стандарт определяет способы координации между устройствами захвата и устройствами воспроизведения для различных видов информации: видео, аудио, метаданных в гетерогенной среде. Гетерогенной – то есть состоящей из устройств различного типа и действующих на различных сетях. Таким образом, в этом стандарте задача решается в более общем виде по сравнению с ETSI TS 103 286, хотя ISO/IEC 23001-13:2019 основан на принципах, появившихся в стандарте консорциума DVB.

Различные варианты применения оркестровки, а также используемые технические приемы будут рассмотрены в следующих статьях.