(19) Евразийское патентное ведомство

(11) 044785 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.09.29

(21) Номер заявки
202293486

(22) Дата подачи заявки
2022.12.27

(51) Int. Cl. ***G06T 7/10*** (2017.01)
G06F 18/00 (2023.01)
G06F 18/20 (2023.01)
G06V 10/00 (2022.01)
G06V 10/20 (2022.01)
G06V 10/25 (2022.01)
G06V 10/26 (2022.01)
G06V 10/96 (2022.01)
G06V 30/184 (2022.01)

(54) СПОСОБ И СИСТЕМА СЕГМЕНТАЦИИ СЦЕН ВИДЕОРЯДА

(31) **2022116268**
(32) **2022.06.16**
(33) **RU**
(43) **2023.09.27**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Лексутин Роман Валерьевич, Жилин Евгений Юрьевич (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-A1-20140037216
US-A1-20110013840
US-A1-20090109298
US-A1-20040233987
US-A1-20030035484
US-A1-20020176625

(57) Настоящее решение относится к области компьютерных технологий, в частности к способу и системе для сегментации сцен видеоряда. Техническим результатом является повышение точности определения контекстных сцен для сегментации видео, за счет параллельного анализа потоков данных, формирующих видео. Заявленный результат достигается за счет осуществления способа сегментации сцен видеоряда, выполняемого с помощью вычислительного устройства и содержащего этапы, на которых: получают входной видеоряд, содержащий видео и речевые данные; выполняют разделение входного видеоряда по трем потокам данных: изображения лиц, текстовые данные на основании транскрибированной речевой информации, и изображения контента, представленных в видеоряде; определяют признаки для каждого кадра видеоряда в каждом из упомянутых потоков данных; выполняют векторизацию упомянутых признаков в каждом из упомянутых потоков данных; осуществляют нормализацию векторных представлений, полученных в каждом потоке, и последующую конкатенацию нормализованных векторных представлений для получения общего набора признаков для каждого кадра видеоряда в виде единого вектора; определяют метрику расстояния для каждого потока данных, как косинусное расстояние между векторами для изображений лиц в видео, и как евклидово расстояние между векторами для текстовых данных и изображений контента; вычисляют на основании упомянутых метрик показатель общей метрики для упомянутого единого вектора, характеризующего каждый кадр видеоряда; выполняют сегментацию видеоряда на контекстно связанные сцены на основании сравнения получаемых единых векторных представлений кадров в векторном пространстве, при этом разделение выполняется на основании превышения порогового значения общей метрики векторных представлений единых векторов кадров видеоряда.

044785 B1

044785 B1

Область техники

Настоящее решение относится к области компьютерных технологий, в частности к способу и системе для сегментации сцен видеоряда.

Уровень техники

В различных процессах компаний (например - продажи, разработки продуктов, и т.п.) необходимо использование коммуникаций с контрагентами (клиентами, сотрудниками других подразделений и т.п.). В рамках этих коммуникаций создаются/передаются большие объемы неструктурированной/слабоструктурированной информации, которая используется для корректировки/изменений реализации соответствующих процессов.

Одним из наиболее распространенных способов коммуникации является онлайн встреча, в рамках которой используются различные каналы передачи данных - люди видят друг друга (видео связь), общаются с помощью аудио (может быть телефонная линия) и могут демонстрировать контент (презентации, демонстрации экрана и т.п.). Например, в процессах продаж, при коммуникациях с клиентами необходимо выявить потребность клиента и затем на основании выявленной потребности предложить соответствующие позиции из ассортимента товаров/услуг компании. Соответственно, необходимо в коммуникациях с клиентом определить промежуток времени и артефакты/объекты, относящиеся к потребностям клиента (например, описание/запрос коммерческого предложения и т.п.). Такой промежуток времени называется сценой. Определение сцен - это задача сегментации видео и другой не структурированной информации которой обмениваются стороны в процессе коммуникаций. По результатам анализа соответствующих сцен, могут совершаться определенные действия (фактически решая задачу классификации сцены по определенному набору действий/классов) - отправить покупателю соответствующие предложения из имеющегося в наличии ассортимента, рассчитать и предложить скидку на дополнительные товары и т.п. Для сегментации видео известным и доступным (включен во многие открытые библиотеки по работе с видео, например, opencv) подходом является разбиение видеоряда на сцены по переходу между кадрами (оценивая разницу между характеристиками последовательных кадров). Такой подход не учитывает контекстную составляющую соответствующих коммуникаций (смысловое содержание видео изображений, аудио, слайдов презентаций и т.п.) и не позволяет производить классификацию сцен для получения практических результатов/действий зависящих от контекста (т.е. смысла/содержания) сцен.

В патенте RU 2628192 C2 (Акционерное общество "Творческо-производственное объединение "Центральная киностудия детских и юношеских фильмов им. М. Горького", 15.08.2017) описано средство сегментации и классификации видео, но в качестве входных признаков используется только один канал передачи данных - видеоизображение. Данный подход не подходит для формата онлайн коммуникаций, в которых видеоизображение может быть статичным долгий промежуток времени, но при этом в аудио коммуникациях может обсуждаться и затрагиваться несколько тем, относящихся к разным контекстным сценам.

В статье A Local-to-Global Approach to Multi-modal Movie Scene Segmentation (https://arxiv.org/abs/2004.02678) описан фреймворк выделения контекстных сцен в фильмах, использующий мультимодальные характеристики каждого кадра (место, актерский состав, действие и аудио). Метод извлечения признаков в этом фреймворке является наиболее близким решением, но имеет отличия в той части, что для сегментации и классификации сцен используется подход на основе "обучения с учителем" (supervised) с помощью сети BNet, предназначенный именно под художественные фильмы. Использование supervised подхода невозможно в случае различной стилистики онлайн коммуникаций (в зависимости от назначения, стиля спикеров, используемого демонстрационного материала).

В заявленном решении для преодоления недостатков, присущих решениям, известным из уровня техники, предлагается подход, позволяющий выполнять сегментацию сцен с помощью классификации по трем каналам данных, формирующих видеоряд, с помощью моделей машинного обучения.

Сущность изобретения

Заявленное изобретение направлено на решение технической проблемы, заключающейся в создании эффективного способа сегментации видео, содержащего демонстрацию контента.

Техническим результатом является повышение точности определения контекстных сцен для сегментации видео, за счет параллельного анализа потоков данных, формирующих видео.

Заявленный результат достигается за счет осуществления способа сегментации сцен видеоряда, выполняемого с помощью вычислительного устройства и содержащего этапы, на которых:

получают входной видеоряд, содержащий видео и речевые данные;

выполняют разделение входного видеоряда по трем потокам данных: изображения лиц, текстовые данные на основании транскрибированной речевой информации, и изображения контента, представленных в видеоряде;

определяют признаки для каждого кадра видеоряда в каждом из упомянутых потоков данных;

выполняют векторизацию упомянутых признаков в каждом из упомянутых потоков данных;

осуществляют нормализацию векторных представлений, полученных в каждом потоке, и последующую конкатенацию нормализованных векторных представлений для получения общего набора признаков для каждого кадра видеоряда в виде единого вектора;

определяют метрику расстояния для каждого потока данных, как косинусное расстояние между векторами для изображений лиц в видео, и как евклидово расстояние между векторами для текстовых данных и изображений контента;

вычисляют на основании упомянутых метрик показатель общей метрики для упомянутого единого вектора, характеризующего каждый кадр видеоряда;

выполняют сегментацию видеоряда на контекстно связанные сцены на основании сравнения получаемых единых векторных представлений кадров в векторном пространстве, при этом разделение выполняется на основании превышения порогового значения общей метрики векторных представлений единых векторов кадров видеоряда.

В одном из частных примеров осуществления способа на основании изображений лиц формируют векторные представления, характеризующие по меньшей мере одно из: лицевые характеристики, пол, возраст, направление взгляда, эмоции. В другом частном примере осуществления способа дополнительно распознают жесты, отображаемые в видеоряде.

В другом частном примере осуществления способа дополнительно из речевых данных выделяют аудиохарактеристики голосов в видеоряде.

В другом частном примере осуществления способа аудиохарактеристики голосов включают в себя по меньшей мере одно из: тональность, интенсивность, форманты.

В другом частном примере осуществления способа демонстрируемый контент дополнительно подвергается OCR обработке для распознавания представленной информации.

Заявленное изобретение также осуществляется за счет системы сегментации сцен видеоряда, содержащая по меньшей мере один процессор и память, хранящую машиночитаемые инструкции, которые при их исполнении процессором реализуют вышеуказанный способ сегментации сцен видеоряда.

Краткое описание чертежей

Фиг. 1 иллюстрирует блок-схему заявленного способа.

Фиг. 2 иллюстрирует пример потоков данных видеоряда.

Фиг. 3 иллюстрирует пример формирования усредняющего нормализованного вектора для потока видео.

Фиг. 4 иллюстрирует общую схему вычислительного устройства.

Осуществление изобретения

На фиг. 1 представлена блок-схема выполнения заявленного способа (100) сегментации видеоряда. На первом этапе (101) исполняющее способ (100) устройство получает входные данные, которые представляют собой видео данные, в частности видеоряд, содержащий как изображения видео контента, так и демонстрацию сопутствующего контента в видеоряде, например, видео презентация. В качестве исполняющего устройства может применяться любой пригодный тип вычислительной техники, например, компьютер, сервер и т.п. Передача информации может осуществляться любым пригодным способ связи, например, с помощью вычислительной сети "Интернет", с помощью непосредственной загрузки данных в вычислительное устройство или любым другим известным принципом передачи цифровой информации.

Как показано на фиг. 2 на этапе (102) из полученного видеоряда выделяется три потока данных, в каждом из которых будет происходить вычисление соответствующих признаков:

видео данные (201);

изображения контента, представленных в видеоряде (202);

аудиопоток (203) и получаемые текстовые данные (2031) на основании транскрибированной речевой информации и их последующие векторные представления (2032).

Выделение потоков может выполняться с помощью известных в уровне техники подходов по выделению из кадра видеопотока контента, отображаемого в видео. Например из кадров видеоряда может выделяться область интереса (задаваемую определенной и фиксированной областью в кадре, например с помощью OpenCV алгоритмов) содержащая демонстрируемый контент. Из потока видеоданных выделяются изображения лиц людей, например, с помощью технологии распознавания лиц (алгоритмы Face recognition). Аудиопоток транскрибируется в текстовую форму для последующего анализа.

Полученные данные в каждом из потоков на этапе (103) обрабатываются для определения признаков в каждый момент времени (кадр видеоряда). В частности, для каждого потока (201-203) может устанавливаться временное окно, в котором будет происходить обработка информации (T1, T2, T3). Также, может определяться частота кадров для анализа кадров видеоданных (F1) и изображений контента (F3), представленного в видеоряде. Частота кадров - настраиваемый параметр, определяющие баланс между точностью и производительностью системы (рекомендуемое значение - 2 кадра в секунду, но не реже 1 кадра в 2 секунды).

Окно обработки информации в потоках данных выполняет две основные функции:

исправление ошибок и артефактов для видеоканалов (за счет последующего отброса аномальных значений в окне и усреднения изображений) - выявляются аномалии по значению прогноз-факт на основе определения ошибки (методом Upper Control Limit) в сравнении с прогнозом модели VAR (Vector Auto-Regression) с установкой параметра maxlag равным размеру окна обработки информации;

обеспечение возможности работы с речевыми признаками для аудио потока - использование алго-

ритмов речь-в-текст невозможно "в моменте" (речь всегда обрабатывается за определенный промежуток времени), поэтому для того, чтобы соотнести векторное представление смысла произнесенной речи с кадром видеоряда необходима обработка аудиопотока с движущимся окном T3.

Окна обработки (T1, T2, T3) - настраиваемые параметры, определяющие устойчивость (robustness) алгоритма (рекомендуемое значение для потока видео T1=100/F1 секунд, для контента T2=100/F3 секунд, для потока аудио T3-30 секунд). Для каждого потока данных определяются признаки, которые затем преобразуются в векторный вид (эмбеддинги на этапе (104) для их последующей обработки с помощью модели машинного обучения).

На каждый момент времени t (каждая секунда видео, аудио и т.п.) производится определение вектора признаков $\overline{p^t} = \begin{pmatrix} p_1^t \\ p_2^t \\ \cdots \\ p_n^t \end{pmatrix}$ в метрическом пространстве (P, d), где P - множество векторов, характеризующий контекст (включающий смысловое содержание видео и демонстрируемого контента, аудио и пр.) на данный момент времени, а d - метрика, определяющая расстояние между векторами из множества P. Например, для кадра формируется вектор признаков:

$V_{17}^1 =$

{

"roll": 3.5928382873535156,

"pitch": -3.403892993927002,

"yaw": 11.955580711364746,

"looks_aside": false,

....

"pos_percent_left": -0.02490421455938696,

"pos_percent_top ": 0.0029940119760478723,

"bad_position": false

}

А для транслируемого контента на этом кадре:

$V_{17}^3 =$

{

"duration": 62.879999999999995,

"text": "",

"similar_to_previous": false,

"slide_words": "",

...

"slide_sentences": "",

"num_words": 141

}

Описанное решение позволяет работать на любых числовых признаках, но в качестве опорного перечня выбираются следующие признаки:

видео - лицевые эмбеддинги (в том числе закодированные в них лицевых характеристики такие как направление взгляда, пол, возраст, эмоции), определение жестов, значения HSV;

аудио - перевод речи в текст и применение языковой модели для выделения эмбедингов предложений, аудиохарактеристики голосов всех людей, находящихся в анализируемом временном окне (тональность, интенсивность, форманты);

видео демонстрируемого контента - значения HSV, эмбеддинг изображения, детектирование (с помощью OCR) и получения эмбеддингов текста, описывающего соответствующий контент.

Далее для полученных векторов признаков в каждом из потоков (201-203) выполняется их нормализация и последующая конкатенация на этапе (105) для получения усредняющего вектора признаков для каждого потока данных. На фиг. 3 приведен пример получения усредняющего вектора (2011) для потока видео (201).

Для потока видеоизображений (201) производится отбрасывание аномальных векторов в рамках окна и усреднение оставшихся. Аналогично для потоков контента (202) и аудио (203) выполняется нормализация признаков векторов методом Max/Min Normalization в рамках групп признаков и их конкатенирование в единый вектор (так как конкатенирование отдельных эмбеддингов без нормализации повлияет на дальнейшее вычисление расстояния).

Например, итоговый единый вектор признаков для кадра №17 будет иметь следующий вид:

$$V_{17}=\begin{cases}V_{17}^{1}\\V_{17}^{2}=\\V_{17}^{3}\end{cases}$$

{

"roll": 0.5928382873535156,

"pitch": 0.403892993927002,

"yaw": 0.955580711364746,

"looks_aside": 0,

….

"pos_percent_left": 0.02490421455938696,

"pos_percent_top ": 0.0029940119760478723,

"bad_position": 0

…

"slide_sentences": 0,

"num_words": 0.01

}

Аналогично для каждого кадра видеоряда для соответствующего потока (202-203) также формируется усредненный вектор.

На основании полученного набора признаков для каждого потока данных (201-203) на этапе (106) определяется метрика расстояния, которая задает метрическое пространство, в котором каждый вектор описывает текущее состояние в момент времени. Под метрикой расстояния подразумевается числовая функция, удовлетворяющая требованиям/аксиомам определения расстояния в этом метрическом пространстве. Примерами такой метрики могут быть расстояние Хэмминга, евклидово расстояние, косинусное расстояние и т.д. Так как сравнение векторов в разных потоках данных может определяться разными метриками, то формально метрика d - это набор метрик (d1, d2, d3) и сравнение векторов выражается в применении отдельных метрик к разным потокам (201-203) и их последующее взвешенное усреднение (используя разные "веса" составляющих метрик d1, d2, d3).

Например, состав метрики d из набора метрик (d1, d2, d3) может иметь следующий вид:

метрика d1 для канала видео с изображениями лиц (с использованием предобученной архитектуры ResNet50 для получения эмбеддингов) представляет из себя косинусный коэффициент (косинусная близость) двух векторов;

метрика d2 для видео демонстрируемого контента (с использованием предобученной модели на базе ResNet50 для определения контекста сцены и модели GPT3 для получения эмбеддингов предложений) представляет из себя евклидовое расстояние между векторами;

метрика d3 для аудио канала (с использованием предобученной модели эмбеддингов на базе архитектуры BERT) представляет из себя евклидовое расстояние между векторами.

В результате выполнения предыдущих шагов для каждого времени T определяется вектор в метрическом пространстве (P, (d1,d2,d3)). Вектор признаков P меняется во времени по ходу видеоряда (так как вектор определяется для каждого кадра видеоряда демонстрируемого в момент времени t). То есть каждый момент видеоряда представляет из себя набор векторов, привязанных ко времени.

Далее на этапе (107) выполняется сегментация входного видеоряда на сцены за счет сравнения получаемых единых векторных представлений кадров. Для выявления данных и контекста сцен с последующей сегментацией используется вышеуказанное метрическое пространство (набор признаков и соответствующая метрика расстояния) и датасет с примерами сегментации информации в каналах коммуникаций (т.е. фактически, математическое описание отдельных областей в метрическом пространстве).

Так как векторы признаков для каждого кадра видеоряда не являются независимыми, а должны быть рассмотрены как последовательность, то в отличие от применения просто методов кластеризации, применяются методы кластеризации последовательности Optimal Sequential Grouping.

Для определения последовательностей, в которых разные кадры могут сильно отличаться, но при этом контекстно содержаться в одной сцене, алгоритм проходит в два этапа.

На первом этапе для каждых последовательных кадров производится сравнение по метрике d и в случае превышения порога чувствительности L производится предварительное разделение на s сегментов, которые определяются последовательностью временных "меток" segms=[$t^{(1)}$, $t^{(2)}$,..., $t^{(s)}$].

Пример результатов сравнения последовательных кадров по метрике d:

[..., 0.31, 0.25, 0.79, 0.12, 0.17, ... , 0.47, 0.85, 0.14]

Соответственно для этого примера границы сцен соответствуют кадрам с расстояниями от предыдущих равными 0.79 и 0.85.

Для полученных сегментов segms производится кластеризация Optimal Sequential Grouping с помощью решения оптимизационной задачи минимизации расстояния между центроидами векторов, входящих в сегменты segms по построенной матрице попарных расстояний между сегментами. В качестве алгоритма сегментации предлагается использование методов кластеризации и использование метода локтя для определения количества кластеров.

Для калибровки и выбора параметров, используемых для сегментирования в данном подходе может использоваться калибровочная выборка. Под калибровочной выборкой понимается размеченный датасет коммуникаций в видео, с разметкой сцен (сегментов) в виде меток начала и окончания сцены. В отличие от supervised подхода, для калибровки/выбора параметров при сегментации фильмов нужно не 21000 сцен, как это представлено в A Local-to-Global Approach to Multi-modal Movie Scene Segmentation, а всего 200 сцен.

При этом с помощью калибровочной выборки возможна оптимизация и выбор параметров, используемых для сегментации. Калибруемые параметры:

весовые коэффициенты w1, w2, w3 метрик d1,d2,d3 для расчёта метрики d;

порог чувствительности L.

Калибровка производится путем корректировки Cost функции, которая представляет сумму ошибок на калибровочной выборке (т.е. классическая задача минимизации ошибки).

Для отладки модели машинного обучения, применяемой для классификации сцен, может выполнять классификация по тэгам, которая производится на базе тех же самых единых векторов, которые получены на шаге (105), и сцен, полученных в ходе сегментации на этапе (107). Для классификации используется обучающая выборка извлекаемых объектов из сцен - тэгов. Для каждой сцены может быть несколько тэгов, то есть решается задача multilabel (множество меток) классификации.

Представлением сцены для классификации является усредненный вектор по всей сцене (под усреднением понимается вектор из сцены, который имеет наименьшее евклидово расстояние до среднего вектора). Так как вектор признаков уже подготовлен на этапе (105), то для классификации можно использовать не сложные deep learning end2end подходы, а производить обучение на обучающей выборке тэгов классическими методами машинного обучения (ML), например, с помощью метода градиентного бустинга. Предложенный подход может найти широкое применение в части эффективной автоматизированной сегментации видеоряда с помощью применяемых технологий и алгоритмов машинного обучения, которые за счет тренировки на соответствующих датасетах могут с высокой вероятностью классифицировать контекстно связанные сцены для их выделения из общего потока данных. Например, такое применение может быть полезно для эффективного разделения блоков презентаций или конференций, в части анализа демонстрируемого контента и сегментации на основании контекстно несвязанных блоков, что может потом передаваться в качестве сегментов во внешние системы демонстрирования контента, например, системы предоставления виде по запросу (video on demand) или т.п.

На фиг. 4 представлен общий вид вычислительной системы на базе вычислительного устройства (300), пригодного для выполнения способа (100). Устройство (300) может представлять собой, например, сервер или иной тип вычислительного устройства, который может применяться для реализации заявленного способа.

В общем случае вычислительное устройство (300) содержит объединенные общей шиной информационного обмена один или несколько процессоров (301), средства памяти, такие как ОЗУ (302) и ПЗУ (303), интерфейсы ввода/вывода (304), устройства ввода/вывода (305), и устройство для сетевого взаимодействия (306).

Процессор (301) (или несколько процессоров, многоядерный процессор) могут выбираться из ассортимента устройств, широко применяемых в текущее время, например, компаний Intel™, AMD™, Apple™, Samsung Exynos™, MediaTEK™, Qualcomm Snapdragon™ и т.п. В качестве процессора (301) может также применяться графический процессор, например, Nvidia, AMD, Graphcore и пр.

ОЗУ (302) представляет собой оперативную память и предназначено для хранения исполняемых процессором (301) машиночитаемых инструкций для выполнение необходимых операций по логической обработке данных. ОЗУ (302), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.).

ПЗУ (303) представляет собой одно или более устройств постоянного хранения данных, например, жесткий диск (HDD), твердотельный накопитель данных (SSD), флэш-память (EEPROM, NAND и т.п.), оптические носители информации (CD-R/RW, DVD-R/RW, BlueRay Disc, MD) и др.

Для организации работы компонентов устройства (300) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов B/B (304). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь: PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п.

Для обеспечения взаимодействия пользователя с вычислительным устройством (300) применяются различные средства (305) В/В информации, например, клавиатура, дисплей (монитор), сенсорный дисплей, тач-пад, джойстик, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п.

Средство сетевого взаимодействия (306) обеспечивает передачу данных устройством (300) посредством внутренней или внешней вычислительной сети, например, Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (306) может использоваться, но не ограничиваться: Ethernet карта, GSM модем, GPRS модем, LTE модем, 5G модем, модуль спутниковой связи, NFC модуль, Bluetooth и/или BLE модуль, Wi-Fi модуль и др.

Дополнительно могут применяться также средства спутниковой навигации в составе устройства (300), например, GPS, ГЛОНАСС, BeiDou, Galileo.

Представленные материалы заявки раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения, не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области техники.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ сегментации сцен видеоряда, выполняемый с помощью вычислительного устройства и содержащий этапы, на которых:

получают входной видеоряд, содержащий видео и речевые данные;

выполняют разделение входного видеоряда по трем потокам данных: изображения лиц, текстовые данные на основании транскрибированной речевой информации, и изображения контента, представленных в видеоряде;

определяют признаки для каждого кадра видеоряда в каждом из упомянутых потоков данных;

выполняют векторизацию упомянутых признаков в каждом из упомянутых потоков данных;

осуществляют нормализацию векторных представлений, полученных в каждом потоке, и последующую конкатенацию нормализованных векторных представлений для получения общего набора признаков для каждого кадра видеоряда в виде единого вектора;

определяют метрику расстояния для каждого потока данных, как косинусное расстояние между векторами для изображений лиц в видео, и как евклидово расстояние между векторами для текстовых данных и изображений контента;

вычисляют на основании упомянутых метрик показатель общей метрики для упомянутого единого вектора, характеризующего каждый кадр видеоряда;

выполняют сегментацию видеоряда на контекстно связанные сцены на основании сравнения получаемых единых векторных представлений кадров в векторном пространстве, при этом разделение выполняется на основании превышения порогового значения общей метрики векторных представлений единых векторов кадров видеоряда.

2. Способ по п.1, в котором на основании изображений лиц формируют векторные представления, характеризующие по меньшей мере одно из: лицевые характеристики, пол, возраст, направление взгляда, эмоции.

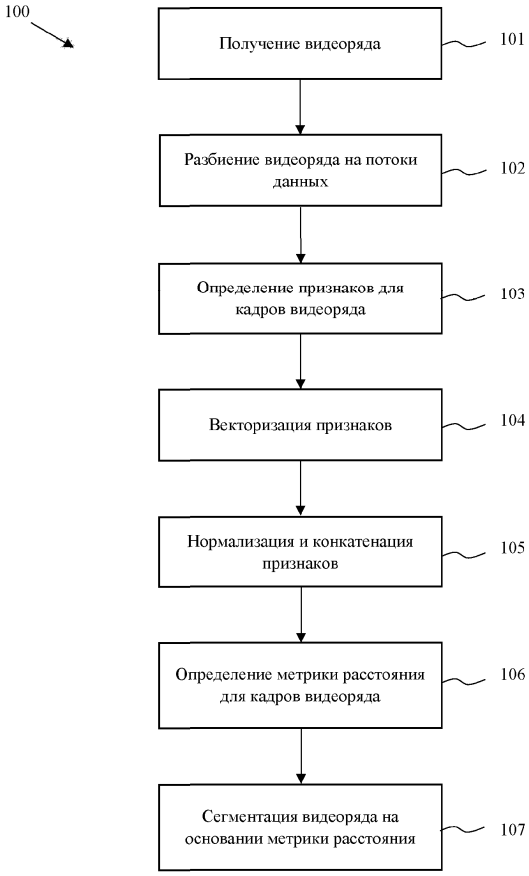
3. Способ по п.2, в котором дополнительно распознают жесты, отображаемые в видеоряде.

4. Способ по п.1, в котором дополнительно из речевых данных выделяют аудиохарактеристики голосов в видеоряде.

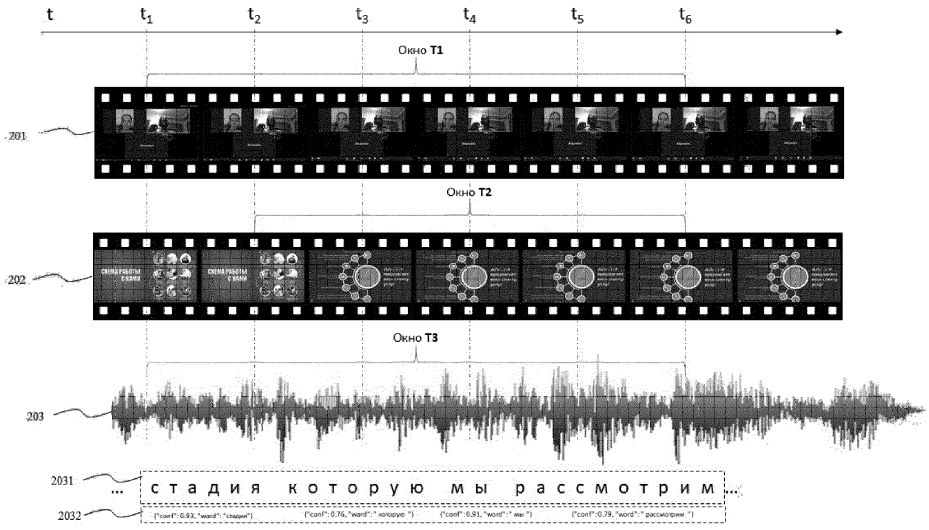
5. Способ по п.4, в котором аудиохарактеристики голосов включают в себя по меньшей мере одно из: тональность, интенсивность, форманты.

6. Способ по п.1, в котором демонстрируемый контент дополнительно подвергается OCR обработке для распознавания представленной информации.

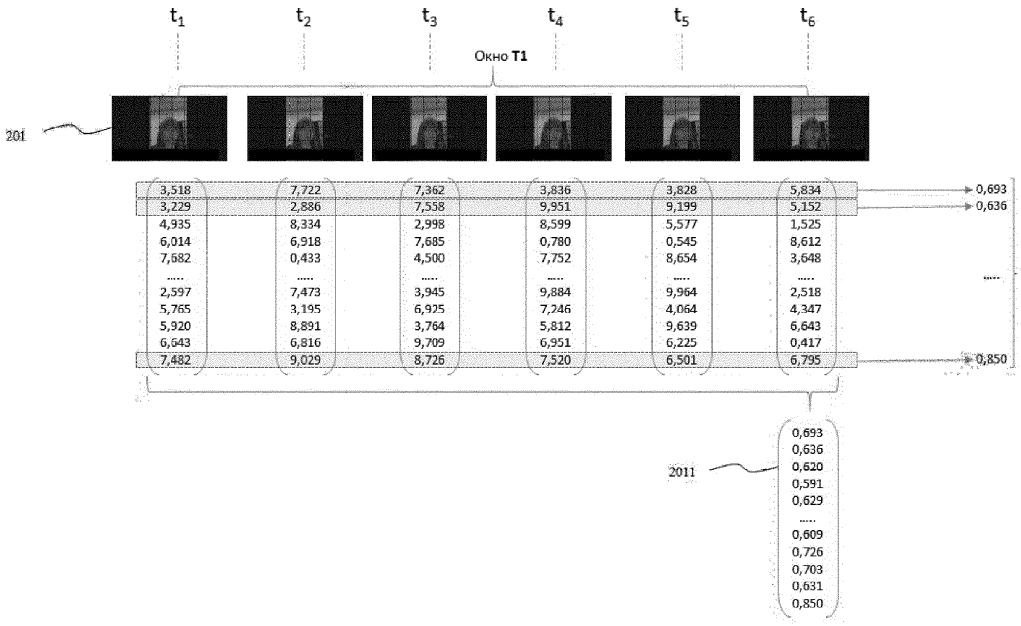
7. Система сегментации сцен видеоряда, содержащая по меньшей мере один процессор и память, хранящую машиночитаемые инструкции, которые при их исполнении процессором реализуют способ сегментации сцен видеоряда по любому из пп.1-6.


100
Получение видеоряда
101
Разбиение видеоряда на потоки данных
102
Определение признаков для кадров видеоряда
103
Векторизация признаков
104
Нормализация и конкатенация признаков
105
Определение метрики расстояния для кадров видеоряда
106
Сегментация видеоряда на основании метрики расстояния
107

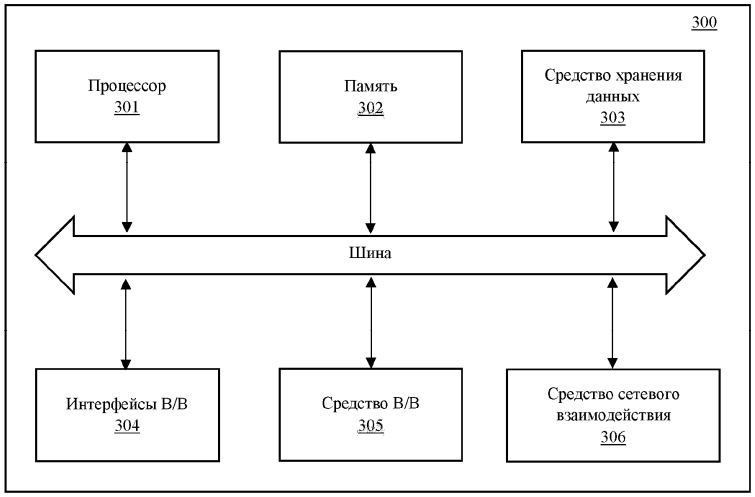
Фиг. 1


t
t1
t2
t3
t4
t5
t6
Окно T1
201
Окно T2
202
Окно T3
203
2031
стадия которую мы рассмотрим
2032

Фиг. 2


t1
t2
t3
t4
t5
t6
Окно T1
201
3,518 7,722 7,362 3,836 3,828 5,834 0,693
3,229 2,886 7,558 9,951 9,199 5,152 0,636
4,935 8,334 2,998 8,599 5,577 1,525
6,014 6,918 7,685 0,780 0,545 8,612
7,682 0,433 4,500 7,752 8,654 3,648
2,597 7,473 3,945 9,884 9,964 2,518
5,765 3,195 6,925 7,246 4,064 4,347
5,920 8,891 3,764 5,812 9,639 6,643
6,643 6,816 9,709 6,951 6,225 0,417
7,482 9,029 8,726 7,520 6,501 6,795 0,850
2011
0,693
0,636
0,620
0,591
0,629
0,609
0,726
0,703
0,631
0,850

Фиг. 3


300
Процессор
301
Память
302
Средство хранения данных
303
Шина
Интерфейсы В/В
304
Средство В/В
305
Средство сетевого взаимодействия
306

Фиг. 4