Информатика и вычислительная техника
Informatics and computer engineering

Научная статья
УДК 004.855
DOI: 10.14529/ctcr260201

АЛГОРИТМ СЕГМЕНТАЦИИ СЦЕН ВИДЕОЛЕКЦИЙ НА ОСНОВЕ СРАВНЕНИЯ ВИЗУАЛЬНЫХ ЭМБЕДДИНГОВ КАДРОВ

М.Е. Исмагулов[1], *m_ismagulov@ugrasu.ru, https://orcid.org/0009-0007-3280-5259*
А.В. Мельников[1, 2], *melnikovav@uriit.ru, https://orcid.org/0000-0002-1073-7108*
[1] *Югорский государственный университет, Ханты-Мансийск, Россия*
[2] *Югорский научно-исследовательский институт информационных технологий, Ханты-Мансийск, Россия*

Аннотация. В условиях роста объема учебных материалов в формате видеолекций актуальной является задача их автоматического преобразования в письменный формат, который обеспечивает в ряде случаев лучшее усвоение. Использование для решения этой задачи ручной разметки видеолекций характеризуется высокой трудоемкостью, что обуславливает необходимость разработки алгоритмических методов разграничения видеолекций на смысловые фрагменты на основе анализа визуальной информации. **Цель исследования:** разработка алгоритма сегментации видеолекций на сцены, основанного на сравнении визуальных эмбеддингов кадров. Предлагаемый подход направлен на выявление границ временных интервалов видеолекции, внутри которых сохраняется устойчивость визуального содержания, что позволяет интерпретировать такие интервалы, как сцены, соответствующие логически завершённым фрагментам изложения учебного материала. **Материалы и методы.** В современных исследованиях задачи автоматизированной обработки видеолекций часто решаются с использованием мультимодальных больших языковых моделей, способных учитывать взаимосвязи между аудиальной и визуальной информацией. Вместе с тем применение подобных моделей сопровождается рядом ограничений, связанных с интерпретируемостью результатов, вычислительной сложностью и требованиями к объему обучающих данных. В рамках исследования используется метод многомодельной обработки видеоданных, основанный на раздельном анализе модальностей видеолекции с применением специализированных моделей. Такой подход позволяет учитывать особенности каждого типа данных и повышать точность обработки. Для анализа визуальной информации применяются модели трансформерных эмбеддингов, в частности DINOv2 и CLIP, обеспечивающие получение устойчивых и семантически информативных представлений кадров, используемых для их последующего сравнения и выявления границ сцен. **Результаты.** В результате проведенного исследования был разработан многоступенчатый алгоритм разграничения видеолекций на сцены, основанный на использовании визуальных эмбеддингов кадров. Лучшим результатом обладают алгоритмы, основанные на модели DINOv2. Качество разграничения оценивалось путем сравнения предсказанных границ сцен с эталонной разметкой на основе метрик precision, recall и F1-score. **Заключение.** Значения указанных метрик подтверждают эффективность предложенного алгоритма при решении задачи автоматического разграничения видеолекций на сцены.

Ключевые слова: видеолекция, разграничение сцен, сегментация видеоданных, визуальные эмбеддинги, многомодельная обработка данных, трансформерные модели, анализ визуального содержания, автоматизированная обработка образовательного контента

Для цитирования: Исмагулов М.Е., Мельников А.В. Алгоритм сегментации сцен видеолекций на основе сравнения визуальных эмбеддингов кадров // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». 2026. Т. 26, № 2. С. 5–15. DOI: 10.14529/ctcr260201

Original article
DOI: 10.14529/ctcr260201

AN ALGORITHM FOR VIDEO LECTURE SCENE SEGMENTATION BASED ON VISUAL FRAME EMBEDDING COMPARISON

M.E. Ismagulov[1], *m_ismagulov@ugrasu.ru, https://orcid.org/0009-0007-3280-5259*
A.V. Melnikov[1, 2], *melnikovav@uriit.ru, https://orcid.org/0000-0002-1073-7108*
[1] *Yugra State University, Khanty-Mansiysk, Russia*
[2] *Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia*

Abstract. With the rapid growth of educational content in the form of video lectures, the task of their automatic transformation into a written format often providing better comprehension has become increasingly relevant. Manual annotation of video lectures is highly labor-intensive, which necessitates the development of algorithmic methods for segmenting video lectures into semantically meaningful fragments based on visual analysis. **Objective.** To develop an algorithm for segmenting video lectures into scenes based on the comparison of visual embeddings of frames. The proposed approach is aimed at identifying temporal boundaries within a video lecture where visual content remains stable, allowing such intervals to be interpreted as scenes corresponding to logically complete fragments of instructional material. **Materials and Methods.** In modern research, automated processing of video lectures is often addressed using multimodal large language models capable of capturing relationships between audio and visual information. However, the use of such models is associated with limitations related to interpretability, computational complexity, and data requirements. In this study, a multi-model video processing approach is employed, based on the separate analysis of lecture modalities using specialized models. This approach enables more accurate processing by accounting for the specific characteristics of each data type. For visual analysis, transformer-based embedding models, specifically DINOv2 and CLIP, are used to obtain stable and semantically informative representations of frames, which are then compared to detect scene boundaries. **Results.** As a result of the study, a multi-stage algorithm for segmenting video lectures into scenes based on visual frame embeddings was developed. The best performance was achieved by methods based on the DINOv2 model. The segmentation quality was evaluated by comparing predicted scene boundaries with ground truth annotations using precision, recall, and F1-score metrics. **Conclusion.** The obtained metric values confirm the effectiveness of the proposed algorithm for automatic segmentation of video lectures into scenes.

Keywords: video lecture, scene segmentation, video data segmentation, visual embeddings, multi-model data processing, transformer models, visual content analysis, automated processing of educational content

For citation: Ismagulov M.E., Melnikov A.V. An algorithm for video lecture scene segmentation based on visual frame embedding comparison. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*. 2026;26(2):5–15. (In Russ.) DOI: 10.14529/ctcr260201

Введение

Задача автоматизированного мультимодального преобразования видеолекции в текстовый документ является актуальной в условиях роста объема цифрового образовательного контента и распространения дистанционного обучения [1, 2]. Видеолекция представляет собой сложный информационный объект, включающий аудио, визуальные материалы (слайды, записи на доске, демонстрации) и их временную динамику [3]. Существенная часть содержания распределена между различными модальностями и не может быть полноценно извлечена средствами одномодального анализа [4]. Существующие подходы, ориентированные преимущественно на автоматическую транскрипцию речи, не обеспечивают согласованную интеграцию аудиальных и визуальных данных, что приводит к потере семантической связности формируемого текста [5].

Практическая значимость исследования заключается в возможности применения предложенного метода для автоматизированной генерации структурированных текстовых представлений онлайн-курсов, подготовки учебно-методических материалов, а также протоколирования вебинаров, конференций и иных научно-образовательных мероприятий [6]. С учетом достигнутого уровня развития методов и моделей машинного и глубокого обучения, продемонстрировавших высокую эффективность при решении задач обработки речи, изображений и их совместного ана-

лиза, актуальной становится задача построения прикладных систем, интегрирующих данные модели в единый алгоритмический контур и обеспечивающих их согласованное функционирование в рамках конвейера [7].

В работе конвейера одной из ключевых задач является разделение видеолекции на сцены – временные интервалы, характеризующиеся семантической однородностью и отражающие завершенный этап изложения учебного материала [8]. Корректная сегментация на сцены обеспечивает локализацию смысловых фрагментов и формирует основу для последующего структурирования и иерархической организации текстового документа [9, 10].

В рамках исследования планируется обработка трех форматов видеолекций, на основе которых будет формироваться соответствующий набор данных. Выбор форматов основан на наиболее распространенных типах образовательных видеороликов, доступных на открытых видеохостингах:

- видеолекция «Лектор и сопровождающая презентация» – формат, в котором примерно 1/3 кадра занимает лектор, а оставшаяся часть отведена презентационным или иллюстративным материалам, отображающим краткое содержание речи лектора в данный момент времени;
- видеолекция «Презентация и закадровый голос» – один из наиболее популярных форматов на видеохостингах, поскольку он прост в реализации и не требует сложного оборудования для съемки. Такой формат основан на концепции «скринкаста» – записи рабочего стола компьютера с помощью специализированных программ;
- видеолекция «Лектор и маркерная или меловая доска» – формат, характерный для записи очных лекций, где основное внимание уделяется преподавателю и его работе с доской.

В статье представлена разработка алгоритмов сегментации видеолекций на сцены, адаптированных к трем различным по структурной организации форматам представления материала. Предложенные алгоритмы включают методы, основанные на анализе визуальных эмбеддингов [10, 11], применении адаптивных пороговых критериев [12], сравнении гистограмм изображений [13, 14], использовании перцептивных хешей [15] и других инструментов количественной оценки межкадровых различий, что позволяет учитывать как низкоуровневые визуальные признаки, так и более абстрактные семантические изменения сцены.

1. Математическая постановка задачи разграничения сцен видеолекции

Формализация данных. Пусть видео состоит из N кадров:

$V = \{f_1, f_2, \dots, f_N\}, f_i \in \mathbb{R}^{H\times W\times C}$,

где H, W, C – высота, ширина и количество каналов (обычно RGB). Каждый кадр можно преобразовать в вектор признаков через цветовые гистограммы или эмбеддинги.

Таким образом, видео представляется как последовательность признаков:

$X = \{x_1, x_2, \dots, x_N\}, x_i \in \mathbb{R}^d$.

Определение задачи. Разграничение на сцены – это поиск индексов $S = \{s_1, s_2, \dots, s_M\}$, где каждый s_j – граница сцены (кадр, после которого начинается новая сцена). Цель: найти такие S, что кадры внутри сцены максимально похожи друг на друга, а кадры между сценами максимально различны.

Тогда функции сходства можно определить отдельно для каждого типа:

$d_h(i,j) = \text{dist}_h(h_i, h_j), \quad d_e(i,j) = \text{dist}_e(e_i, e_j)$.

Для гистограмм в данном исследовании используется метрика «расстояния χ^2»:

$$d_h(i,j) = \frac{1}{2}\sum_{k=1}^{K}\frac{\left(h_i^k - h_j^k\right)^2}{h_i^k + h_j^k}.$$

Для визуальных эмбеддингов в данном исследовании – евклидово расстояние или косинусное сходство:

$d_e(i,j) = \| e_i - e_j \|_2^2$ или $d_e(i,j) = 1 - \cos(e_i, e_j)$.

Критерий сцены. Сцена изменяется там, где сходство резко падает:

$s_j = \arg\max_i d(i, i+1)$

или с порогом θ:

$s_j = \{i : d(i, i+1) > \theta\}$.

В качестве порога различимости используется адаптивный порог MAD (Median Absolute Deviation), применение которого позволяет сделать алгоритм более устойчивым к выбросам и шуму в кадрах. То есть задача сводится к определению порога различия или нахождению локальных максимумов функции различий.

Пусть у нас есть последовательность различий между соседними кадрами:

$D = \{d_{\text{total}}(1,2), d_{\text{total}}(2,3), \dots, d_{\text{total}}(N-1, N)\}$.

Тогда медианная абсолютная девиация вычисляется как

$\text{MAD} = \text{median}(|\ D_i - \text{median}(D)\ |)$.

Порог можно задать как кратное MAD:

$\theta = \text{median}(D) + k \cdot \text{MAD}, k \geq 1,$

где k – коэффициент, регулирующий чувствительность детектора границ. При этом малое k значит больше границ, чувствительно к мелким изменениям, если большое k, то менее чувствительно, выделяются только резкие изменения.

Тогда формула границы сцены принимает вид:

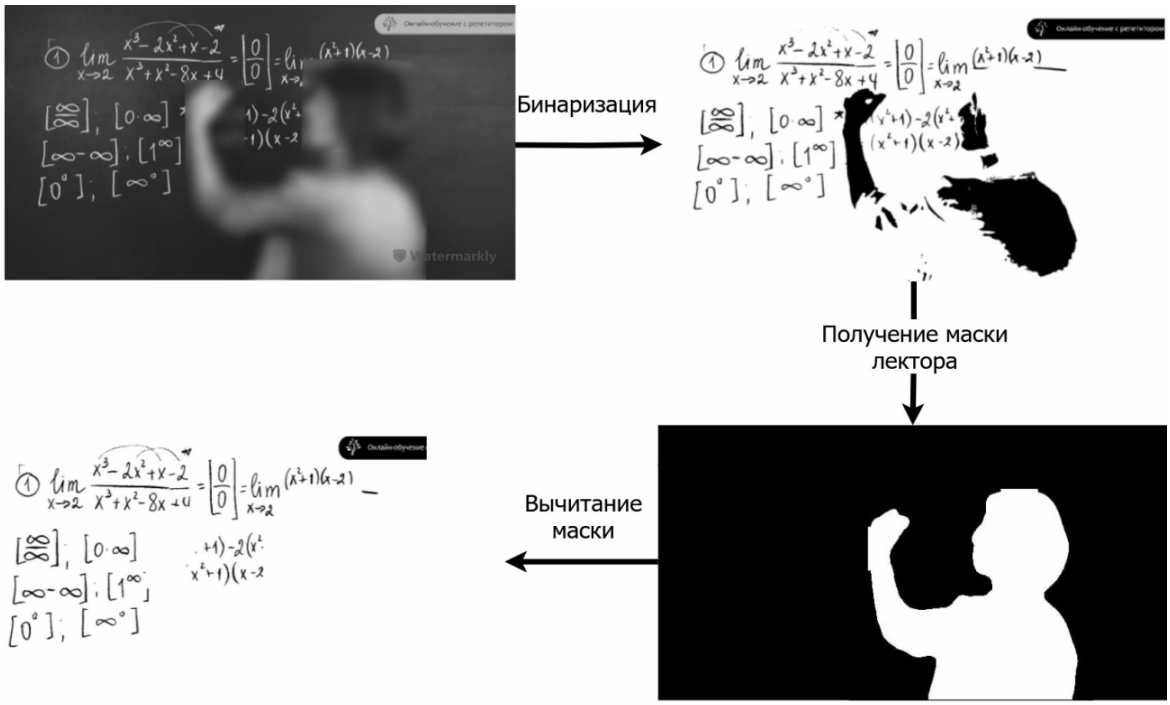
$s_j = \{i: d_{\text{total}}(i, i+1) > \text{median}(D) + k \cdot \text{MAD}\}$.

В рамках исследования коэффициент k взят по самому низкому порогу, так как это нивелируется особенностями многоступенчатого алгоритма, что будет подробно изложено в следующих разделах.

2. Методы и модели решения задачи

В качестве методологической основы исследования использованы алгоритмы, методы и модели машинного и глубокого обучения в сочетании с методом абстрагирования данных, направленным на выделение устойчивых компонент визуального содержания видеолекции.

Абстрагирование реализовано посредством исключения динамически изменяемых объектов сцены (в частности, фигуры лектора) с целью формирования целостного и инвариантного представления ключевой информационной области – доски или презентационного материала (рис. 1). Это позволило минимизировать влияние несущественных изменений и снизить внутрисценовую дисперсию признаков.



Источник: видео TutorOnline (https://youtu.be/I9TSR9rrwxQ), авторские права принадлежат TutorOnline

Рис. 1. Схема процесса абстрагирования лектора
Fig. 1. Lecturer abstraction process scheme

Для извлечения семантически значимых признаков кадров применялись предобученные модели визуальных трансформеров – DINOv2 и CLIP, формирующие эмбеддинги в латентном пространстве высокой размерности. Обнаружение границ сцен осуществлялось посредством анализа расстояний между эмбеддингами соседних кадров, что позволило формализовать задачу разграничения как задачу детекции значимых изменений распределения визуальных признаков во времени.

3. Описание конвейеров, адаптированных под разные форматы видеолекции

Архитектура конвейера обработки видеолекций носит адаптивный характер и конфигурируется в зависимости от структурного формата исходного видеоматериала. Для различных типов видеолекций ядро конвейера включает различный набор моделей и алгоритмических процедур.

Для видеолекции формата «Лектор и презентация» конфигурация конвейера обработки определяется совокупностью этапов, представленных на рис. 2.

Рис. 2. Алгоритм разграничения сцен для видеолекции «Лектор и презентация»
Fig. 2. Scene segmentation algorithm for the “Lecturer and presentation” video lecture

На первом этапе применяется модель детекции объектов YOLOv8 Small, обеспечивающая локализацию области лектора в кадре с целью последующего абстрагирования. Далее в качестве ядра конвейера используется алгоритм сравнения гистограмм изображений из библиотеки PySceneDetect с пониженным пороговым коэффициентом, что позволяет выполнить грубое первичное определение границ сцен. На втором этапе применяется алгоритм перцептивного хеширования pHash, предназначенный для выявления и объединения дублирующих сцен, а также выбора уникального репрезентативного кадра.

Алгоритм обработки видеолекций типа «Презентация и закадровый голос» структурно идентичен конвейеру, применяемому для формата «Лектор и презентация», за исключением этапа сегментации и последующего абстрагирования лектора (рис. 3). Отсутствие визуального присутствия докладчика в кадре исключает необходимость использования модели сегментации, что упрощает вычислительную схему и снижает совокупную сложность обработки без изменения логики последующих этапов детектирования и уточнения границ сцен.

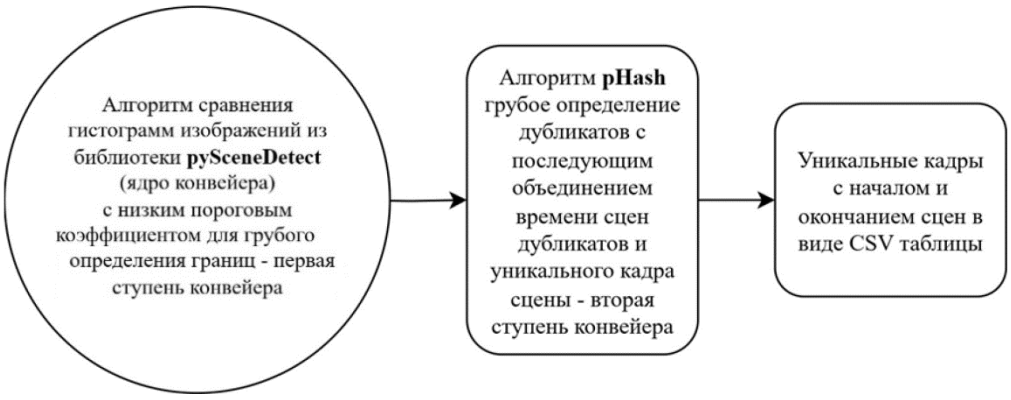


Рис. 3. Алгоритм разграничения сцен для видеолекции «Презентация и закадровый голос»
Fig. 3. Scene segmentation algorithm for the “Presentation and voice-over” video lecture

В рамках предложенного конвейера обработки видеоданных на первом этапе применяется модель сегментации YOLOv8 Seg Small для выделения контурной маски лектора с целью последующего абстрагирования его области в кадре (рис. 4). Далее осуществляется грубое детектирование границ сцен на основе визуальных эмбеддингов DINOv2 с использованием адаптивного порога, вычисляемого по метрике MAD. На следующем этапе выполняется уточнение сцен за счет сглаживания сигнала отношения эмбеддингов методом медианной фильтрации, что позволяет подавить локальные флуктуации и стабилизировать процедуру пороговой сегментации. Завершающим шагом является дедупликация кадров алгоритмом перцептивного хеширования (pHash), в результате чего формируется таблица уникальных сцен с указанием временных интервалов их начала и окончания в формате CSV.

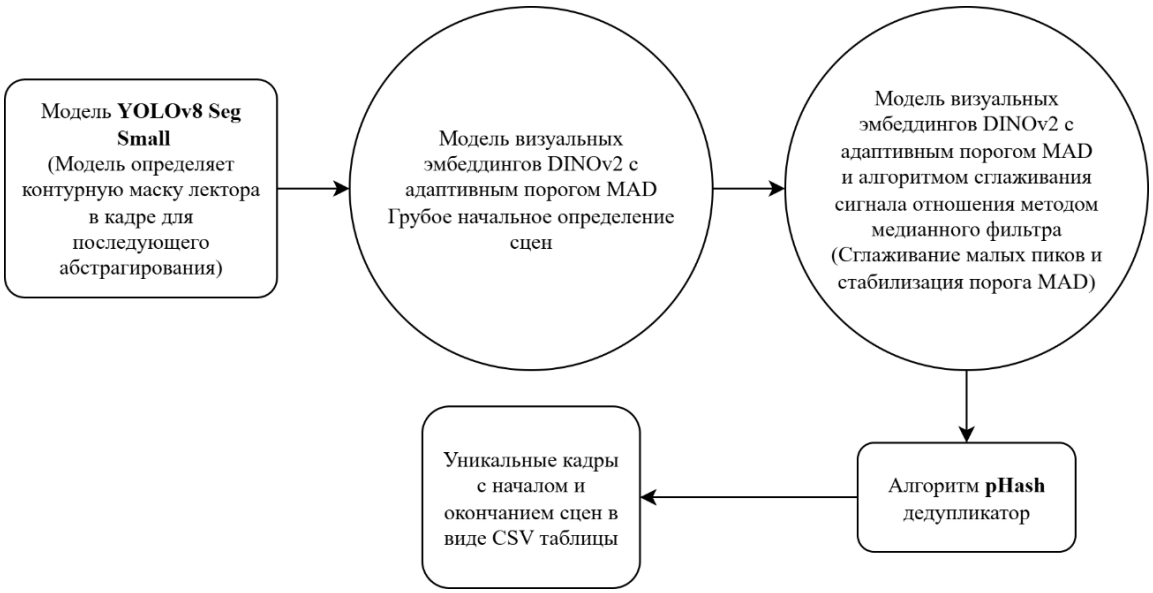


Рис. 4. Алгоритм разграничения сцен для видеолекции «Лектор и доска» на основе визуальных эмбеддингов модели DINOv2
Fig. 4. Scene segmentation algorithm for the "Lecturer and board" video lecture based on visual embeddings of the DINOv2 model

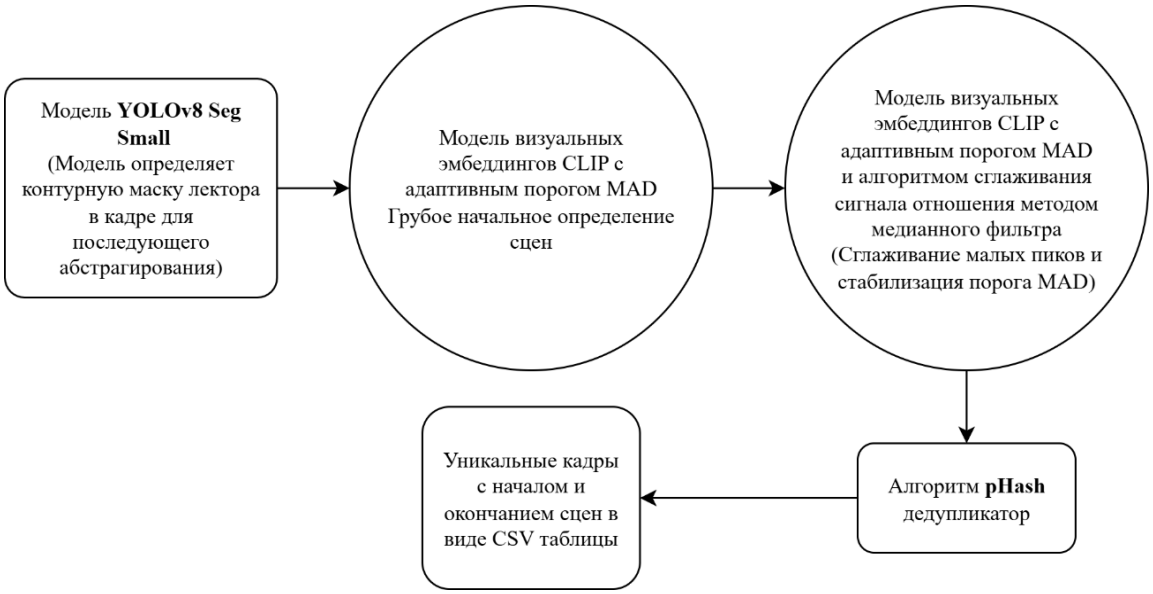


Рис. 5. Алгоритм разграничения сцен для видеолекции «Лектор и доска» на основе визуальных эмбеддингов модели CLIP
Fig. 5. Scene segmentation algorithm for the "Lecturer and board" video lecture based on visual embeddings of the CLIP Model

Конвейер с использованием модели CLIP реализован в структурно идентичной конфигурации по отношению к варианту на основе DINOv2 (рис. 5). Сохранение неизменной архитектуры обработки, включая методы расчета сигнала сходства, адаптивного порога (MAD) и процедуры сглаживания, необходимо для обеспечения корректности сравнительного анализа. Такой подход позволяет изолировать влияние типа визуальных эмбеддингов на итоговые метрики качества сегментации.

4. Набор данных

В качестве экспериментального корпуса были отобраны видеоматериалы нескольких онлайн-курсов, размещенных на платформах массовых открытых онлайн-курсов – Coursera и Stepik, а также на открытых видеохостингах, включая RuTube и VK Video. Совокупный объем выборки составляет 20 видеозаписей, из которых 10 относятся к формату «Лектор и презентация» со средним хронометражем порядка 15 мин, 8 – к формату «Лектор и доска» со средним хронометражем около 30 мин, и 2 – к формату «Презентация и закадровый голос» продолжительностью свыше 1 ч. Такой состав выборки обеспечивает вариативность визуальной структуры видеоряда и позволяет провести сопоставимый анализ эффективности алгоритмов для различных типов лекционного контента.

5. Результаты и обсуждение

Анализ сигналов изменения динамики кадров показывает, что при отсутствии медианной фильтрации (рис. 6) наблюдаются выраженные высокоамплитудные пики и повышенная частота их возникновения. Это приводит к росту оценки дисперсии сигнала и, как следствие, к увеличению адаптивного порога, вычисляемого по метрике MAD. Включение медианного фильтра (рис. 7) подавляет локальные выбросы и снижает влияние кратковременных флуктуаций, что приводит к стабилизации сигнала и формированию более выраженной квазипрямоугольной структуры на участках смены сцен. В результате кластеры, соответствующие границам сцен, становятся более компактными и однородными, а значение порога MAD снижается, повышая чувствительность алгоритма к устойчивым структурным изменениям видеоряда при одновременном подавлении шумовых переходов.

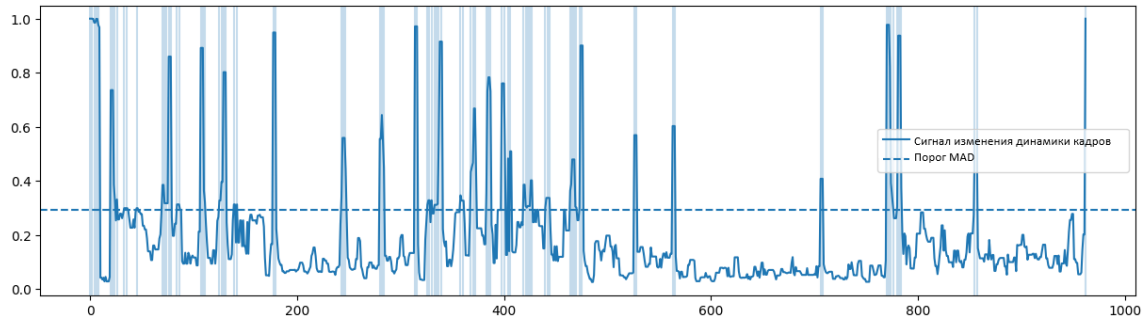


Рис. 6. Сигнал визуальных эмбеддингов DINOv2 без медианного сглаживания, с более высоким порогом MAD
Fig. 6. DINOv2 visual embedding signal without median smoothing, with a higher MAD threshold

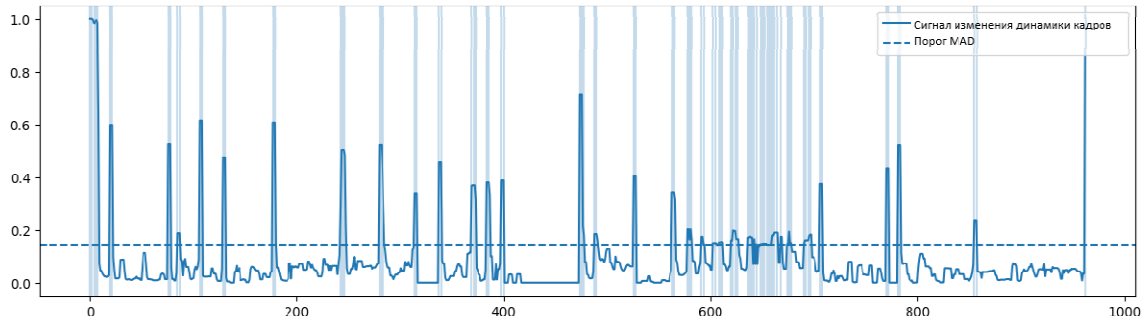


Рис. 7. Сигнал визуальных эмбеддингов DINOv2 с наличием медианного сглаживания, с более низким порогом MAD
Fig. 7. DINOv2 visual embedding signal with median smoothing, with a lower MAD threshold

Для эмбеддингов CLIP также наблюдается эффект стабилизации сигнала при применении медианной фильтрации, что подтверждается прямым сравнением графиков (рис. 8, 9). Однако в отличие от DINOv2, подавление мелких пиков выражено слабее. Это обусловлено большей чувствительностью эмбеддингов CLIP к локальным визуальным и семантическим изменениям (текст на слайде, мелкие объекты, незначительные перестройки композиции), что приводит к сохранению части высокочастотных флуктуаций даже после сглаживания [16]. В результате вклад выбросов в оценку разброса остается выше, а снижение адаптивного порога MAD оказывается менее выраженным по сравнению с вариантом на основе DINOv2 [17].

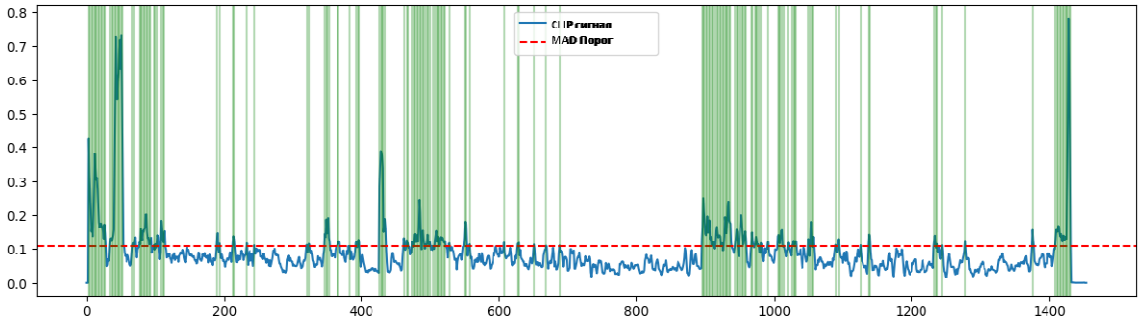


Рис. 8. Сигнал визуальных эмбеддингов CLIP без медианного сглаживания, с более высоким порогом MAD
Fig. 8. CLIP visual embedding signal without median smoothing, with a higher MAD threshold

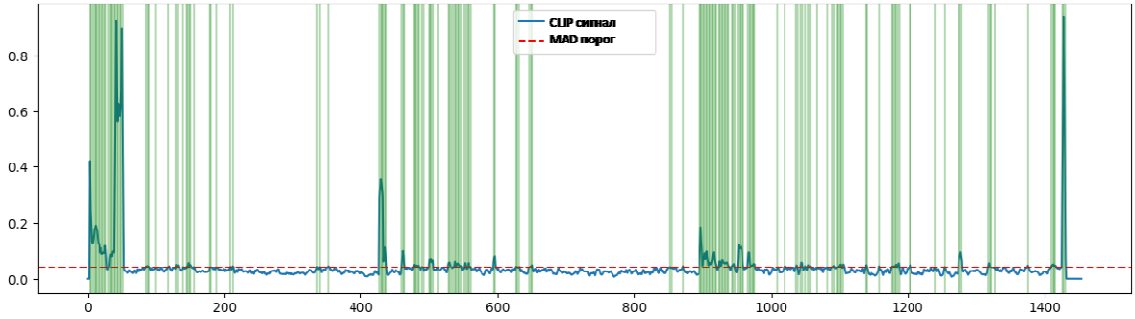


Рис. 9. Сигнал визуальных эмбеддингов CLIP с наличием медианного сглаживания, с более низким порогом MAD
Fig. 9. CLIP visual embedding signal with median smoothing, with a lower MAD threshold

Сравнительный анализ распределения границ сцен показывает, что для DINOv2 (голубые вертикальные маркеры) характерно формирование более устойчивых и компактных кластеров без выраженных разрывов. Границы сцен группируются в плотные интервалы, соответствующие устойчивым структурным изменениям видеоряда. В случае CLIP (зеленые маркеры) наблюдается большая фрагментация: внутри одного и того же перехода фиксируются дополнительные срабатывания, что приводит к разреженности кластеров и появлению промежуточных разрывов. Данное различие указывает на более высокую устойчивость DINOv2 к локальным и семантически несущественным вариациям изображения, что положительно сказывается на стабильности детектирования сцен.

Для оценки качества детектирования сцен использовались стандартные метрики precision, recall и F1-score. Сравнительный анализ проводился на основе сопоставления предсказанных моделью границ сцен с вручную размеченной эталонной разметкой. Для нивелирования неточности ручной разметки была введена разница времени, равная 2 с. Истинно положительными считались детекции, совпадающие с размеченными переходами в пределах заданного временного допуска; ложноположительные и ложноотрицательные срабатывания определялись соответственно как избыточные или пропущенные границы. Такой подход позволяет количественно оценить как точность локализации переходов, так и полноту их обнаружения.

Результаты количественной оценки демонстрируют принципиально различный характер поведения моделей.

Для конвейера обработки видеолекций формата «Лектор и презентация» получены следующие значения показателей качества: $precision = 0{,}948$, $recall = 0{,}987$ и $F1 - score = 0{,}967$. Достигнутые результаты указывают на высокую корректность выделения временных границ сцен, что объясняется сравнительно простой и структурно однородной визуальной композицией видеоматериала данного типа.

Для формата «Презентация и закадровый голос» зафиксированы значения $precision = 0{,}963$, $recall = 0{,}977$ и $F1 - score = 0{,}971$. Максимальные показатели качества в данном случае обусловлены упрощённой архитектурой конвейера, поскольку отсутствует необходимость обработки и сегментации образа лектора, что снижает вариативность визуальных данных и упрощает задачу детектирования сцен.

Для CLIP при $\Delta t = 2{,}0$ с получены следующие значения:

$precision = 0{,}214$; $recall = 0{,}949$; $F1 - score = 0{,}349$.

Высокое значение *recall* указывает на то, что модель практически не пропускает реальные границы сцен. Однако крайне низкий *precision* свидетельствует о значительном числе ложноположительных срабатываний. Это означает, что модель чрезмерно чувствительна к локальным визуальным и семантическим изменениям, вследствие чего происходит избыточное дробление сцен. На практике это проявляется в формировании большого количества коротких фрагментов и мелких кластеров кадров.

Для DINOv2 при том же допуске $\Delta t = 2{,}0$ с получены:

$precision = 0{,}653$; $recall = 0{,}821$; $F1 - score = 0{,}727$.

Здесь наблюдается более сбалансированное соотношение метрик. Незначительное снижение *recall* компенсируется существенным ростом *precision*, что приводит к двукратному увеличению меры *F*1 по сравнению с CLIP. Практически это означает, что модель реже генерирует ложные границы, формируя более устойчивые и укрупненные кластеры сцен, соответствующие реальным структурным переходам видеоряда.

Заключение

В ходе исследования разработаны и экспериментально апробированы конвейеры детектирования сцен для различных форматов видеолекций, основанные на использовании визуальных эмбеддингов CLIP и DINOv2 с адаптивной пороговой сегментацией (MAD) и процедурой медианного сглаживания сигнала. Проведенный аналитический и численный анализ показал, что применение медианной фильтрации снижает влияние локальных выбросов и стабилизирует сигнал изменения эмбеддингов, что положительно влияет на устойчивость выделения границ сцен. Сравнительная оценка по метрикам *precision*, *recall* и *F*1 продемонстрировала принципиальные различия моделей: при высокой полноте CLIP характеризуется значительным числом ложноположительных срабатываний и избыточным дроблением сцен ($F1 - score = 0{,}349$), тогда как DINOv2 обеспечивает более сбалансированное соотношение точности и полноты ($F1 - score = 0{,}727$), формируя устойчивые и структурно согласованные кластеры сцен.

К достоинствам предложенного решения относится модульная архитектура конвейера, инвариантность к формату видеолекции, адаптивность пороговой процедуры и возможность корректного сравнительного анализа различных типов эмбеддингов в идентичных условиях. Практическая значимость работы заключается в повышении качества автоматической структуризации видеоконтента, что может быть использовано при формировании конспектов лекций, индексировании образовательных видеоматериалов и построении систем мультимодального поиска. Перспективы дальнейших исследований связаны с гибридизацией эмбеддингов и интеграцией сценной сегментации в более сложные мультимодальные пайплайны анализа видеоданных.

Список литературы/References

1. Коробкин Д.М. Система автоматического субтитрирования видеофайлов // Системный анализ в науке и образовании: сетевое научное издание. 2022. № 2. С. 23–27. [Korobkin D.M. System of automatic subtitling of video files. *System Analysis in Science and Education*. 2022;(2):23–27. (In Russ)]. Available at: http://sanse.ru/download/469.

2. Market.us. Digital Education Content Market Size: CAGR of 26.1%. Available at: https://market.us/report/digital-education-content-market/ (accessed 25 February 2026).

3. Lee D.W., Ahuja C., Liang P.P., Natu S., Morency L.-P. Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in Educational Videos. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France, 2023. P. 20030–20041. DOI: 10.1109/ICCV51070.2023.01838

4. Mayer R.E. *Multimedia Learning*. 3rd ed. Cambridge: Cambridge University Press, 2020. 416 p.

5. Khosravi H., Denny P., Moore S., Stamper J. Learnersourcing in the Age of AI: Student, Educator and Machine Partnerships for Content Creation. *Computers and Education: Artificial Intelligence*. 2023;5:100151. DOI: 10.1016/j.caeai.2023.100151

6. Freisinger S., Schneider F.P., Herygers A., Georges M., Bocklet T., Riedhammer K. Unsupervised Multilingual Topic Segmentation of Video Lectures: What can Hierarchical Labels tell us about the Performance? In*: Conference: SLaTE 2023 (9th Workshop on Speech and Language Technology in Education)*. Dublin, Ireland, 2023. DOI: 10.21437/SLaTE.2023-27

7. Kubala F., Colbath S., Liu D., Srivastava A., Makhoul J. Integrated Technologies for Indexing Spoken Language. *Communications of the ACM*. 2000;43(2):48–56. DOI: 10.1145/328236.328146

8. Tseng S.M., Yeh Z.T., Wu C.Y., Chang J.B., Norouzi M. Video Scene Detection Using Transformer Encoding Linker Network (TELNet). *Sensors*. 2023;23(16):7050. DOI: 10.3390/s23167050

9. Chen S., Nie X., Fan D., Zhang D., Bhat V., Hamid R. Automatically Identifying Scene Boundaries in Movies and TV Shows. *Amazon Science*. Available at: https://www.amazon.science/blog/automatically-identifying-scene-boundaries-in-movies-and-tv-shows (accessed 26 February 2026).

10. Mun J., Shin M., Han G., Lee S., Ha S.J., Lee J., Kim E.S. Boundary-Aware Self-Supervised Learning for Video Scene Segmentation. *ArXiv preprint: arXiv:2201.05277*. 2022. Available at: https://arxiv.org/abs/2201.05277 (accessed 26 February 2026).

11. Berman N., Botach A., Ben-Baruch E., Hakimi S.H., Gendler A., Naiman I., Yosef E., Kviatkovsky I. Scene-VLM: Multimodal Video Scene Segmentation via Vision-Language Models. *ArXiv preprint: arXiv:2512.21778*. 2025. Available at: https://arxiv.org/abs/2512.21778 (accessed 26 February 2026).

12. Raja Suguna M., Kalaivani A., Anusuya S. The Detection of Video Shot Transitions Based on Primary Segments Using the Adaptive Threshold of Colour-Based Histogram Differences and Candidate Segments Using the SURF Feature Descriptor. *Symmetry*. 2022;14(10):2041. DOI: 10.3390/sym14102041

13. Kar T., Kanungo P., Mohanty S.N. et al. Video Shot-Boundary Detection: Issues, Challenges and Solutions. *Artificial Intelligence Review*. 2024;57:104. DOI: 10.1007/s10462-024-10742-1

14. Zhou S., Wu X., Qi Y. et al. Video Shot Boundary Detection Based on Multi-Level Features Collaboration. *Signal, Image and Video Processing*. 2021;15:627–635. DOI: 10.1007/s11760-020-01785-2

15. PySceneDetect. Detection Algorithms. *PySceneDetect Documentation*. Available at: https://www.scenedetect.com/docs/latest/api/detectors.html (accessed 26 February 2026).

16. Lavoie M.A., Mahmoud A., Zaimi A. et al. CLIP Is Shortsighted: Paying Attention Beyond the First Sentence. *ArXiv preprint: arXiv:2602.22419*. 2026. Available at: https://arxiv.org/abs/2602.22419 (accessed 2 March 2026).

17. Oquab M., Darcet T., Moutakanni T. et al. DINOv2: Learning Robust Visual Features without Supervision. *ArXiv preprint: arXiv:2304.07193*. 2024. Available at: https://arxiv.org/abs/2304.07193 (accessed 2 March 2026).

Информация об авторах

Исмагулов Милан Ерикович, аспирант 3-го года обучения по направлению 2.3.1 «Системный анализ, управление и обработка информации, статистика», Инженерная школа цифровых технологий, Югорский государственный университет, Ханты-Мансийск, Россия; m_ismagulov@ugrasu.ru.

Мельников Андрей Витальевич, д-р техн. наук, проф., Инженерная школа цифровых технологий, Югорский государственный университет, Ханты-Мансийск, Россия; директор, Югорский научно-исследовательский институт информационных технологий, Ханты-Мансийск, Россия; melnikovav@uriit.ru.

Information about the authors

Milan E. Ismagulov, 3rd year postgraduate student in the field of 2.3.1 "Systems analysis, management and information processing, statistics", Engineering School of Digital Technologies, Yugra State University, Khanty-Mansiysk, Russia; m_ismagulov@ugrasu.ru.

Andrey V. Melnikov, Dr. Sci. (Eng.), Prof., Engineering School of Digital Technologies, Yugra State University, Khanty-Mansiysk, Russia; Director, Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia; melnikovav@uriit.ru.

Вклад авторов: авторы принимали участие во всех этапах подготовки статьи.

Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors participated in all stages of the preparation of the manuscript.

The authors declare no conflicts of interests.

Статья поступила в редакцию 03.03.2026
The article was submitted 03.03.2026