УДК 004.896

М.В. Ліпчанський
А.В. Зензеря

МЕТОД АВТОМАТИЧНОГО ВИЯВЛЕННЯ КЛЮЧОВИХ МОМЕНТІВ У КОРОТКИХ ВІДЕОЗАПИСАХ НА ОСНОВІ МУЛЬТИМОДАЛЬНОГО АНАЛІЗУ

Національний технічний університет "Харківський політехнічний інститут", м. Харків

Анотація

У роботі досліджено та розроблено метод автоматичного виявлення ключових моментів у коротких відеозаписах для платформ соціальних мереж. Запропоновано мультимодальний підхід на основі пізнього злиття незалежних оцінок п'яти аналітичних каналів: аудіо-, відео-, емоційного, кольорового та текстового. Для зведення числових показників застосовано мін-макс нормалізацію per-video, одновимірне гауссове згладжування для усунення поодиноких артефактних піків та алгоритм ковзного вікна змінної тривалості. Програмний комплекс реалізовано у вигляді мікросервісної архітектури на базі мов Go, Python із використанням брокера повідомлень RabbitMQ та СУБД MySQL. На основі користувацького опитування проведено порівняльне оцінювання системи на тестовому наборі відеозаписів різних жанрів. Результати дослідження підтвердили, що розроблений мультимодальний метод демонструє вищу стійкість до жанрової специфіки контенту та перевершує одномодальні базові лінії за середньою оцінкою та показником win rate.

Ключові слова: короткі відеозаписи, ключовий момент, мультимодальний аналіз, пізнє злиття, гауссове згладжування, ковзне вікно, мікросервісна архітектура.

Abstract

The paper investigates and develops a method for automated key moment detection in short video clips tailored for social media platforms. A multimodal approach is proposed based on the late fusion of independent scores from five analytical channels: audio, video, emotional, color, and textual. To aggregate numerical indicators, a per-video min-max normalization, one-dimensional Gaussian smoothing to eliminate isolated artifact peaks, and a sliding window algorithm of variable duration are applied. The software system is implemented as a microservice architecture using Go and Python, along with the RabbitMQ message broker and MySQL DBMS. Based on a user survey, a comparative and ablation evaluation of the system was conducted on a test dataset comprising video clips of various genres. The research results confirmed that the developed multimodal method demonstrates higher robustness to genre-specific content, outperforming single-modality baselines with an average rating and a win rate.

Keywords: short videos, key moment, multimodal analysis, late fusion, Gaussian smoothing, sliding window, microservice architecture.

Вступ

Сучасний сегмент соціальних медіа характеризується стрімким зростанням обсягів короткого відеоконтенту (тривалістю до 60 секунд), яскравими прикладами якого є платформи TikTok, Instagram Reels та YouTube Shorts. Велика щільність інформаційного потоку вимагає від користувачів та контент-мейкерів інструментів для швидкого пошуку, навігації та автоматичного виділення найважливіших фрагментів відео ("ключових моментів"). Складність вирішення цього завдання полягає у високій жанровій різноманітності контенту (розмовні відео, динамічні влоги, музичні кліпи, геймплей тощо), де інформаційне навантаження може розподілятися нерівномірно між візуальним рядом, мовою або звуковим супроводом. Традиційні одномодальні методи (наприклад, аналіз лише графічного контенту або текстових транскрипцій) часто виявляються неефективними, оскільки ігнорують комплексний контекст відео. Тому актуальним завданням є розробка стійких методів автоматичного виявлення ключових моментів на основі інтеграції інформації з кількох аналітичних каналів сприйняття [1].

Основна частина

У межах цього дослідження розроблено метод автоматичного виявлення ключових моментів, який базується на мультимодальному аналізі та стратегії пізнього злиття (late fusion). Запропонована архітектура обробки даних включає п’ять незалежних аналітичних каналів:

1) аудіоканал: фіксує зміну акустичних параметрів та гучності;

2) відеоканал (динаміка сцени): оцінює інтенсивність руху та частоту зміни планів;

3) емоційний канал: розпізнає мімічні прояви емоцій людей у кадрі;

4) кольоровий канал: аналізує яскравість та насиченість палітри зображення;

5) текстовий канал (NLP): виконує семантичний аналіз транскрибованого мовлення для пошуку логічних та емоційних кульмінацій.

Кожен із каналів генерує свій числовий вектор оцінок у часі. Оскільки шкали та розподіли значень відрізняються, на етапі передобробки застосовано per-video min-max нормалізацію, що приводить показники до єдиного діапазону [0; 1] відносно конкретного відеозапису. Для усунення випадкових шумів та поодиноких артефактних піків розраховані вектори піддаються одновимірному гауссовому згладжуванню [2].

Інтеграція даних (пізнє злиття) реалізована шляхом обчислення зваженої суми нормалізованих векторів, де вагові коефіцієнти каналів можуть адаптуватися під специфіку контенту. Для безпосереднього пошуку кульмінаційного інтервалу використовується алгоритм ковзного вікна змінної тривалості (2-3 секунди). Вікно зміщується по інтегральному вектору оцінок з фіксованим кроком, обчислюючи сумарну вагу фрагмента; інтервал із максимальним значенням визначається як головний ключовий момент відео [3].

Програмна реалізація запропонованого методу виконана у вигляді відмовостійкої мікросервісної архітектури. Для розробки сервісів використано мови Python (аналітичні модулі комп’ютерного зору та NLP) та Go (високопродуктивні сервіси оркестрації та агрегації). Асинхронну взаємодію та черги завдань організовано за допомогою брокера повідомлень RabbitMQ, а збереження результатів обробки у СУБД MySQL.

Для експериментальної перевірки методу було сформовано тестовий набір із 30 коротких відеозаписів різних жанрів. Оцінювання якості виявлення ключових моментів проводилося шляхом користувацького опитування, де респонденти оцінювали точність знайденого фрагмента за 5-бальною шкалою, а також порівнювали роботу мультимодальної системи з одномодальними базовими лініями.

Висновки

Проведені експерименти та дослідження підтвердили ефективність запропонованого підходу. Мультимодальний метод продемонстрував значно вищу стійкість до жанрових особливостей відео, ніж будь-яка з окремо взятих модальностей. Середня експертна оцінка точності визначення ключових моментів склала 4.12 з 5. У порівняльних тестах із базовими одномодальними рішеннями розроблена система досягла показника win rate у 68%. Практична реалізація у вигляді мікросервісів на базі Go, Python та RabbitMQ довела свою здатність до ефективного масштабування та паралельної обробки медіапотоків, що дозволяє інтегрувати розроблений метод у реальні високонавантажені платформи обміну відеоконтентом.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Apostolidis E., Adamantidou E., Metsai A., Mezaris V., Patras I. Video Summarization Using Deep Neural Networks: A Survey / Proceedings of the IEEE., 2021., Vol. 109, No. 11., P. 1838–1863.

2. Farneback G. Two-Frame Motion Estimation Based on Polynomial Expansion / Scandinavian Conference on Image Analysis., Springer, 2003., P. 363–370.

3. Loureiro D., Barbieri F., Neves L., Anke L., Camacho-Collados J. TimeLMs: Diachronic Language Model Trained on Twitter Time Data / ACL Anthology., 2022., P. 344–356.

Ліпчанський Максим Валентинович — к.т.н, доцент, доцент кафедри комп'ютерної інженерії та програмування, Національний технічний університет "Харківський політехнічний інститут", м. Харків, e-mail: Maksym.Lipchanskyi@khpi.edu.ua

Зензеря Арсеній Віталійович — студент групи КН-Н924б, Навчально-науковий інститут комп'ютерних наук та інформаційних технологій, Національний технічний університет "Харківський політехнічний інститут", м. Харків, e-mail: Arsenii.Zenzeria@cit.khpi.edu.ua

Maksym Lipchanskyi — PhD in Technical Sciences, Associate Professor, Associate Professor of the Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, e-mail: Maksym.Lipchanskyi@khpi.edu.ua

Arsenii Zenzeria — student of the group KN-N924b, Educational and Scientific Institute of Computer Science and Information Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, e-mail: Arsenii.Zenzeria@cit.khpi.edu.ua