Временная локализация технических действий в видеозаписях сумо с использованием различных представлений видеоданных

Г. И. Балтаев*, А. П. Рыжов‡

Исследуется временная локализация двух приёмов в полных видеозаписях сумо. Для ActionFormer сравниваются признаки YOLO-Pose, RGB-I3D и VideoMAE. Совместное дообучение VideoMAE-H и временных адаптеров даёт лучший результат: mAP 83.57% при tIoU 0.50 и 57.43% в среднем по пяти порогам.

Ключевые слова: временная локализация действий, сумо, компьютерное зрение, ActionFormer, VideoMAE, AdaTAD.

1. Введение

Методы интеллектуального анализа данных успешно применяются в спорте высших достижений: обзор описывает обработку изображений и сигналов, моделирование и планирование [3]. Задачи и специфика вида спорта определяют выбор технологий. Одна из распространённых задач видеоанализа — временная локализация значимых событий, то есть определение типа и границ действия в непрерывной записи.

Цель работы — сопоставить представления видеоданных и способы их адаптации при общем детекторе ActionFormer [1], оценивая распознавание и точность временных границ для выбора наилучшего инструмента для временной локализации приемов в сумо.

Для видео $V = (I_t)_{t=1}^{T}$, где I_t — кадр с номером t, а T — общее число кадров, требуется найти обнаружения

$$\widehat{Y} = \{(s_i, e_i, c_i, p_i)\}_{i=1}^{N},$$

Балтаев Глеб Иркинович — студент факультета ВМК МГУ имени М. В. Ломоносова, e-mail: st02220065@gse.cs.msu.ru.

Baltaev Gleb Irkinovich — Student, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University.

‡*Рыжов Александр Павлович* — д. т. н., профессор факультета ВМК МГУ имени М. В. Ломоносова, e-mail: ryjov@cs.msu.ru.

Ryzhov Alexander Pavlovich — Professor, Doctor of Engineering, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

где $[s_i, e_i)$ — интервал, c_i — класс действия, p_i — уверенность. Целевые классы — `cosinage` и `csuciminage`; остальные кадры относятся к фону.

2. Данные и методы

Используются восемь записей: четыре обучающие с повторениями приёмов и изменением ориентации борцов и четыре тестовые с полноценными схватками. В обучающих записях около 40000 кадров, в тестовых — около 10000; съёмка ведётся двумя камерами. Размечены интервалы целевых действий. Для ActionFormer сравниваются три представления.

YOLO-Pose. Сравниваются исходные и нормализованные относительно общей области координаты ключевых точек. Проверяются окна 200–500 кадров с разными перекрытиями и смесь фрагментов 50–1000 кадров.

RGB-I3D. Предобученная на Kinetics-400 модель ResNet-50 извлекает признаки размерности 2048. Сравниваются исходный и дообученный на видео сумо кодировщики; детектор обучается на сохранённых признаках.

VideoMAE и AdaTAD. Сравниваются VideoMAE-S и VideoMAE-H с адаптацией через TIA (Temporal-Informative Adapter, темпорально-информативный адаптер) [2]: обучение только адаптеров, совместное обучение адаптеров и кодировщика и полное дообучение без адаптеров. VideoMAE-S обучался совместно с обучением адаптера. Применяется Soft-NMS.

3. Полученные результаты

Для каждого класса вычисляется AP (Average Precision, средняя точность) — площадь под интерполированной кривой «точность–полнота»; mAP (mean Average Precision) — среднее AP по двум классам. Обнаружение засчитывается при совпадении класса и достаточном временном пересечении:

$$\mathrm{tIoU}(S_p, S_g) = \frac{|S_p \cap S_g|}{|S_p \cup S_g|}.$$

Здесь S_p и S_g — предсказанный и размеченный интервалы, $|S|$ — длительность в кадрах; tIoU — temporal intersection over union. Сопоставление однозначное. Среднее mAP считается по порогам $\mathcal{T} = \{0.50, 0.60, 0.70, 0.80, 0.90\}$ и обозначается mAP_5.

Для нормализованных YOLO-Pose-признаков лучшее окно — 500 кадров с перекрытием 25 кадров (табл. 1): mAP_5 равна 48.76% против 42.40% на исходных координатах. Обучение на смеси длин даёт 40.52%.

Таблица 1: Результаты с общим детектором ActionFormer, %

Представление и режим обучения	mAP@0.50	mAP_5
YOLO-Pose, исходные координаты	65.42	42.40
YOLO-Pose, норм., окно 500 кадров	70.18	48.76
YOLO-Pose, норм., разные длины	66.75	40.52
RGB-I3D, фиксированный кодировщик	52.16	31.94
RGB-I3D, предварительное дообучение	60.49	31.20
VideoMAE-S	69.07	33.23
VideoMAE-H, только адаптеры	82.89	53.72
VideoMAE-H, веса и адаптеры	**83.57**	**57.43**
VideoMAE-H, без адаптеров	59.34	32.09

Поскольку меняются признаки и фрагментация, эффект нельзя приписать только нормализации.

Дообучение RGB-I3D повышает mAP@0.50 с 52.16 до 60.49%, но снижает mAP_5 с 31.94 до 31.20%: точность границ при строгих порогах не улучшается.

Лучшая конфигурация — совместное дообучение VideoMAE-H и адаптеров: $mAP_5 = 57.43\%$, на 8.67 пункта выше YOLO-Pose и на 3.71 пункта выше обучения только адаптеров. mAP при порогах 0.50–0.90 составляет 83.57, 81.11, 72.52, 42.05 и 7.88%, что показывает сохраняющуюся проблему точного определения границ.

4. Заключение

Лучшее качество показывает совместное дообучение VideoMAE-H с адаптерами, тогда как YOLO-Pose остаётся более компактной альтернативой. Дальнейшая работа связана с уточнением границ и различением сходных приёмов.

Список литературы

[1] C.-L. Zhang, J. Wu, Y. Li, “ActionFormer: Localizing Moments of Actions with Transformers”, *Computer Vision – ECCV*, 2022, 492–510. DOI: 10.1007/978-3-031-19772-7_29.

[2] S. Liu, C.-L. Zhang, C. Zhao, B. Ghanem, “End-to-End Temporal Action Detection with 1B Parameters Across 1000 Frames”, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, 18591–18601. DOI: 10.1109/CVPR52733.2024.01759.

[3] F. Hammes, A. Hagg, A. Asteroth, D. Link, “Artificial Intelligence in Elite Sports—A Narrative Review of Success Stories and Challenges”, *Frontiers in Sports and Active Living*, **4** (2022), 861466. DOI: 10.3389/fspor.2022.861466 .

Статья поступила 8 октября 2026 г.

Temporal Localization of Technical Actions in Sumo Videos Using Different Video Data Representations

G. I. Baltaev, A. P. Ryzhov

We study temporal localization of two actions in untrimmed sumo videos, comparing YOLO-Pose, RGB-I3D, and VideoMAE features for ActionFormer. Joint fine-tuning of VideoMAE-H and temporal adapters yields the best result: 83.57% mAP at tIoU 0.50 and 57.43% averaged over five thresholds.

Keywords: temporal action detection, sumo, computer vision, ActionFormer, VideoMAE, AdaTAD.

References

[1] C.-L. Zhang, J. Wu, Y. Li, “ActionFormer: Localizing Moments of Actions with Transformers”, *Computer Vision – ECCV*, 2022, 492–510. DOI: 10.1007/978-3-031-19772-7_29.

[2] S. Liu, C.-L. Zhang, C. Zhao, B. Ghanem, “End-to-End Temporal Action Detection with 1B Parameters Across 1000 Frames”, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, 18591–18601. DOI: 10.1109/CVPR52733.2024.01759.

[3] F. Hammes, A. Hagg, A. Asteroth, D. Link, “Artificial Intelligence in Elite Sports—A Narrative Review of Success Stories and Challenges”, *Frontiers in Sports and Active Living*, **4** (2022), 861466. DOI: 10.3389/fspor.2022.861466.

Received on October 8, 2026