TEMPORAL AND MULTIMODAL BEHAVIORAL ANALYSIS FOR IMPROVING REAL-TIME VIDEO DETECTION

Mingturayeva Dildora

Annotatsiya. *Ushbu tezis real vaqt rejimidagi videoaniqlash tizimlarida inson xatti-harakatlarini temporal va multimodal tahlil qilish yondashuvlarini ko‘rib chiqadi. Zamonaviy tadqiqotlarda harakatlarni vaqt bo‘yicha aniqlash, inson tana holatini baholash, yuz xatti-harakatlarini tahlil qilish, videodagi hissiy holatlarni aniqlash va turli ma’lumotlarni birlashtirish uchun chuqur o‘rganish usullaridan keng foydalanilmoqda. Bitta kadrga asoslangan tahlil xatti-harakatning vaqt davomida rivojlanishini to‘liq ifodalamaydi; temporal modellashtirish ketma-ket kadrlar orasidagi bog‘liqlikni, multimodal tahlil esa yuz, tana va harakat kabi o‘zaro to‘ldiruvchi belgilarni birgalikda hisobga oladi. Tezisda mavjud yondashuvlar tahlil qilinadi hamda real vaqt rejimida ishlovchi video tizimlarida xatti-harakat belgilarini vaqt bo‘yicha va ko‘p modalli tarzda ifodalashga asoslangan konseptual sxema taklif etiladi.*

Kalit so‘zlar: *real vaqt rejimidagi videoaniqlash, xatti-harakatlar tahlili, vaqt bo‘yicha tahlil, ko‘p modalli tahlil, kompyuter ko‘rishi, sun’iy intellekt, inson harakatlarini tanib olish.*

Аннотация. *В тезисе рассматриваются подходы к временному и мультимодальному анализу поведения человека, направленные на повышение надежности систем видеоаналитики в реальном времени. Современные исследования широко используют методы глубокого обучения для обнаружения действий во временной последовательности, оценки позы человека, анализа мимики, распознавания эмоций и объединения взаимодополняющих признаков различных модальностей. Анализ отдельного кадра не всегда позволяет отразить развитие поведения во времени, тогда как временное моделирование учитывает последовательность кадров и изменения поведенческих признаков. Мультимодальный анализ позволяет совместно использовать взаимодополняющие признаки лица, позы и движения. В тезисе рассматриваются современные методы и предлагается концептуальная схема временного и мультимодального представления поведенческих признаков.*

Ключевые слова: *видеоаналитика реального времени, анализ поведения, временной анализ, мультимодальный анализ, компьютерное зрение, искусственный интеллект, распознавание действий.*

Abstract. *This thesis examines temporal and multimodal behavioral analysis approaches for improving real-time video detection systems. Recent studies use deep learning, temporal action detection, human pose estimation, facial behavior analysis, video emotion recognition, and multimodal fusion for understanding human actions. Frame-level analysis may fail to represent how behavior develops over time, whereas temporal modeling captures dependencies across observations. Multimodal analysis allows complementary cues, including facial appearance, body pose, and motion, to be represented jointly. This thesis reviews these approaches and proposes a conceptual framework in which behavioral features are extracted from video, represented temporally, fused across modalities, and used for a final behavioral decision. The framework provides a methodological basis for subsequent experimental research on reliable real-time video detection.*

Keywords: *real-time video detection, behavioral analysis, temporal analysis, multimodal analysis, computer vision, artificial intelligence, human action recognition.*

1. INTRODUCTION

Real-time video analytics has developed rapidly with advances in computer vision and deep learning. Modern systems can detect people, recognize actions, estimate human pose, and analyze facial or emotional behavior. Human action recognition is applied in surveillance, human-computer interaction, robotics, healthcare, and anomaly detection [1,3,4]. However, continuous video interpretation is more difficult than classifying isolated images.

A key limitation of frame-level analysis is that one frame describes only a moment. Many behaviors are defined by change, duration, repetition, or transition. Temporal action detection therefore identifies action intervals in untrimmed video and explicitly models their temporal structure [1,2]. Multimodal analysis provides a second complementary direction by combining facial, pose, motion, audio, or other signals [7–10].

2. EXISTING RESEARCH

2.1. Visual behavioral analysis and action recognition

Human action recognition transforms video into visual representations and classifies activities. Recent work describes a progression from conventional approaches to CNN-, RNN-, and Transformer-based architectures [3]. For continuous video,

temporal action detection includes anchor-based, boundary-based, and query-based approaches, as well as fully supervised, weakly supervised, and unsupervised learning [2]. These methods demonstrate that temporal localization is distinct from frame-level classification.

2.2. Facial behavior

Facial behavior can be represented through landmarks, local features, and Facial Action Units. Recent work surveys machine and deep learning methods for facial expression and Action Unit recognition [6]. CNN and attention-based models can learn visual patterns, but facial signals remain context-dependent and should be interpreted as one component of broader behavior.

2.3. Body pose and motion

Human pose estimation converts video into structured body-joint representations useful for action recognition. Recent surveys identify depth ambiguity, occlusion, data scarcity, and computational cost as important challenges [4,5]. Pose and motion can complement facial features by describing body configuration and movement dynamics over time.

2.4. Temporal behavior modelling

Temporal modeling represents dependencies across frames. RNNs, temporal convolutional methods, attention mechanisms, and Transformers have been used to model short- and long-range temporal relationships [1–3]. Conceptually, frame-level analysis asks 'what is visible now?', whereas temporal analysis asks 'how does the behavior develop over time?' This distinction is important when isolated observations are ambiguous.

2.5. Multimodal fusion

Multimodal fusion combines complementary information from different sources. Reviews describe early, intermediate, and late fusion strategies and emphasize the importance of synchronization, modality quality, and learned representations [8–10]. For video behavior analysis, face, pose, and motion may be treated as complementary components of one behavioral representation.

3. PROPOSED CONCEPTUAL FRAMEWORK

Based on the reviewed literature, a conceptual framework for real-time behavioral video analysis is proposed:

Video stream → Human detection → Behavioral feature extraction → Temporal representation → Multimodal fusion → Behavioral pattern detection → Decision

Behavioral feature extraction includes facial, pose, and motion representations. These features are organized as sequences so that order, duration, repetition, and

transition can be modeled. Multimodal fusion then integrates the complementary representations before a behavioral decision. The exact fusion strategy should be selected empirically because early, intermediate, and late fusion have different computational and representational properties [8–10].

The framework is conceptual and does not claim experimentally validated gains. Its purpose is to define a coherent methodological basis for subsequent algorithm development and evaluation.

4. CHALLENGES AND EVALUATION

Real-time temporal and multimodal analysis faces several practical challenges: individual variability, occlusion, illumination, camera angle, noisy data, temporal complexity, computational cost, dataset bias, and annotation limitations. Video emotion surveys also identify synchronization, generalization, interpretability, and dataset bias as persistent problems [7,9,10]. Pose-estimation reviews emphasize occlusion and data scarcity [4,5].

Evaluation should therefore go beyond a single accuracy value. A future experimental comparison should include: (1) a frame-level or single-modal baseline; (2) a temporal model; (3) a multimodal model; and (4) a temporal-plus-multimodal model. Precision, recall, F1-score, false-positive rate, and inference time are particularly relevant. False-alarm reduction is reserved for the subsequent scientific article and the wider PhD study; at the thesis stage, the contribution is the methodological framework and evaluation logic.

5. CONCLUSION

The literature demonstrates substantial progress in temporal action detection, human pose estimation, facial behavior analysis, video emotion recognition, and multimodal fusion [1–10]. These areas provide complementary components for continuous behavioral video analysis: temporal modeling captures how behavior develops, while multimodal fusion integrates facial, pose, and motion information.

The proposed framework organizes these components into a sequence from video input and behavioral feature extraction to temporal representation, multimodal fusion, and behavioral decision. It provides a methodological basis for subsequent experimental research on reliable real-time video detection and, at a later stage, false-alarm reduction.

REFERENCES

1. Vahdani E., Tian Y. Deep Learning-Based Action Detection in Untrimmed Videos: A Survey // IEEE Transactions on Pattern Analysis and Machine

Intelligence. — 2023. — Vol. 45, № 4. — P. 4302–4320. DOI: 10.1109/TPAMI.2022.3193611.

2. Hu K. et al. Overview of temporal action detection based on deep learning // Artificial Intelligence Review. — 2024. DOI: 10.1007/s10462-023-10650-w.

3. Alomar K., Aysel H. I., Cai X. CNNs, RNNs and Transformers in human action recognition: a survey and a hybrid model // Artificial Intelligence Review. — 2025. — Vol. 58. — 387. DOI: 10.1007/s10462-025-11388-3.

4. Liu Y., Qiu C., Zhang Z. Deep learning for 3D human pose estimation and mesh recovery: A survey // Neurocomputing. — 2024. — Vol. 596. — 128049. DOI: 10.1016/j.neucom.2024.128049.

5. Neupane R. B., Li K., Boka T. F. A survey on deep 3D human pose estimation // Artificial Intelligence Review. — 2025. — Vol. 58. — 24. DOI: 10.1007/s10462-024-11019-3.

6. Priego M. et al. Machine and deep learning in facial expression recognition: a survey based on facial action units // Scientific Reports. — 2026. — Vol. 16. — 28377. DOI: 10.1038/s41598-026-53853-1.

7. Hallur S., Gavade A. B., Gavade P. A survey on video emotion recognition: Segmentation, classification, and explainable AI techniques // Signal Processing: Image Communication. — 2025. — Vol. 141. — 117442. DOI: 10.1016/j.image.2025.117442.

8. Aruna G. et al. Survey on multimodal approaches to emotion recognition // Neurocomputing. — 2023. — Vol. 556. — 126693. DOI: 10.1016/j.neucom.2023.126693.

9. Geetha A. V., Mala T., Priyanka D., Uma E. Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions // Information Fusion. — 2024. — Vol. 105. — 102218. DOI: 10.1016/j.inffus.2023.102218.

10. Hallur S., Gavade A. Video emotion analysis: Segmentation, classification, and explainability techniques in review // The Franklin Open. — 2026. — Vol. 15. — 100604. DOI: 10.1016/j.fraope.2026.100604.