ZoomGate: Scale-Aware Action Recognition Across Mixed Zoom Levels

Kseniia Buzko[1] David Clausi[1] John Zelek[1] Yuhao Chen[1]
[1]Vision and Image Processing Group, Systems Design Engineering, University of Waterloo
{kbuzko, dclausi, jzelek, yuhao.chen1}@uwaterloo.ca

Abstract

Human action recognition and facial-expression analysis in cinematic footage remain challenging because most systems ignore camera-dependent changes in visible detail. Close-ups, medium shots, and long shots contain fundamentally different expressive cues, yet most models treat scale as irrelevant. This paper addresses this gap by introducing ZoomGate, a unified, scale-aware pipeline for human behaviour understanding across mixed zoom levels. Using a movie-trailer dataset with frame-level zoom annotations, we train image backbones to classify camera scale and route video segments to view-specific recognition modules tailored for facial emotions, micro-gestures, upper-body actions, full-body motion, or hand-only gestures. For close-up segments, a multimodal Gemini-based analysis produces structured, temporally aligned descriptions of emotion dynamics and articulatory behaviour. Experiments demonstrate that scale-conditioned processing yields more coherent and interpretable predictions than scale-agnostic baselines. ZoomGate provides a principled foundation for building computer-vision systems and AI characters that adjust behaviour naturally with camera distance.

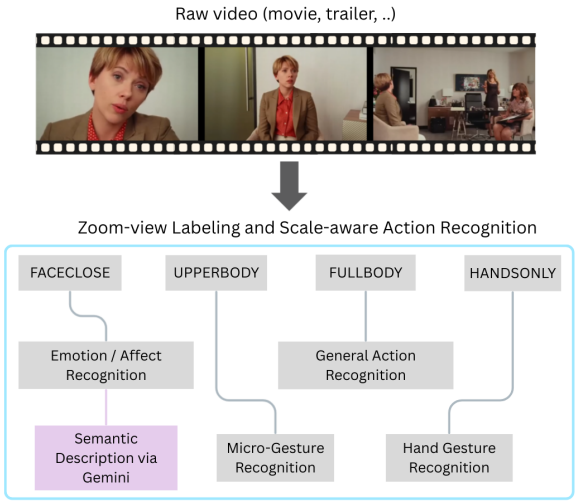

Figure 1. **Overview of the ZoomGate pipeline.** Raw video is assigned to zoom-view categories and routed into view-specific models: emotion analysis for close-ups, micro-gesture recognition for upper-body shots, full-body action recognition for long shots, and hand-gesture analysis for hand-only crops. A Gemini-based module produces rich semantic descriptions of facial dynamics in FACECLOSE segments, enabling animation-ready expression metadata.

1. Introduction

AI-driven characters in games, virtual production, and social platforms have become increasingly expressive, yet they still often do not act naturally. Even when animation is physically plausible, performances frequently feel inconsistent across camera changes: a gesture that appears subtle in a close-up may look inert in a long shot, while an animation crafted for a wide shot may appear exaggerated when framed tightly, reflecting well-documented findings that animated motion is interpreted differently depending on how it is presented to the viewer. Current avatar-generation and character-behaviour systems tend to overlook this mismatch between animation and camera scale, treating expressive control as camera-agnostic.

We argue that scale-aware perception is a missing component in modern character pipelines. Cinematic practice explicitly distinguishes close-ups, medium shots, and long shots, each carrying different expectations for facial detail, gesture magnitude, and emotional readability [4]. However, most learning-based systems lack any signal about view scale and therefore cannot learn how expressions and actions should adjust as the camera moves.

To address this gap, we introduce ZoomGate (Figure 1), a unified pipeline for view-scale human action recognition (HAR) and expressive facial analysis. Built from high-quality movie trailers, ZoomGate begins with frame-level zoom-view annotation aligned to person tracks, enabling robust training of scale classifiers based on ResNet-18 [11], ConvNeXt-Tiny [12], and ViT-B/16 [8]. These zoom predictions then drive a second stage of scale-aware action recognition, in which each zoom regime (FACECLOSE

(FC), UPPERBODY (UB), FULLBODY (FB), HANDSONLY (HO)) is routed to a specialized HAR module tailored to the level of visible detail. Finally, close-up segments are processed through a Gemini-based multimodal analysis module that produces structured descriptions of emotion dynamics, gaze shifts, mouth-shape trajectories, and muscle activations. By integrating these stages, ZoomGate provides a principled way for models to understand how human behaviour should vary, enabling future AI characters to adapt their expressive behaviour naturally as humans do.

2. Dataset

Our experiments are conducted on a semi-supervised movie-trailer dataset with frame-level zoom-view labels and a derived subset for human action recognition (HAR). For zoom-view classification, we annotate 19,864 frames sampled from eight trailers with five categories: FACECLOSE (FC), UPPERBODY (UB), FULLBODY (FB), HANDSONLY (HO), and NOPARTS (NP), where NOPARTS indicates that no valid body region is visible. The resulting distribution reflects typical cinematic framing: medium shots (UPPERBODY) dominate, while rare, highly localized views such as HANDSONLY are under-represented. These frame-level labels are used to train scale classifiers directly from RGB images.

For HAR, we construct short clips centered on person tracks extracted from the same trailers. Since no meaningful action can be inferred when no body parts are visible, the NOPARTS category is excluded from the HAR subset. Table 1 summarizes the combined statistics, reporting per-label counts for zoom-view frames as well as the number of clips and frames in the HAR subset.

Table 1. **Combined statistics of the Zoom-view (frame-level) and Human Action Recognition (HAR) clip datasets.** "NP" (NOPARTS) appears only in the zoom-view corpus, because frames without visible human parts cannot support reliable action labels.

Label	Zoom Frames	HAR Clips	HAR Frames
FC	989	18	995
HO	61	3	101
FB	1642	34	1957
UB	9748	266	43300
NP	7424	–	–
Total	**19864**	**319**	**46284**

3. Methodology

ZoomGate consists of three sequential stages, each operating at a different level of visual granularity. First, we predict the zoom-view category of every frame (Section 3.1), which determines the visible body scale. These scale-stable segments then enable scale-aware action recognition (Section 3.2), where specialized HAR modules are routed according to the detected view. Finally, close-up face segments are enriched with semantic facial descriptions obtained from Gemini (Section 3.3), providing fine-grained emotional and articulatory information for downstream animation tasks.

Figure 2. **Illustrative examples of the five zoom-view categories used in ZoomGate.** The frames are drawn from cinematic footage and serve as visual reference for the scale taxonomy.

3.1. Zoom-View Classification

Our first stage predicts the camera scale for each frame. Given a single RGB frame, the goal is to classify it into one of five zoom-view categories: FACECLOSE, UPPERBODY, FULLBODY, HANDSONLY, and NOPARTS. These labels later gate which downstream modules are applied, such as emotion analysis for FACECLOSE and full-body actions for FULLBODY.

We benchmark three standard image backbones, all initialized from ImageNet-1K [7]. The first is ResNet-18 [11], a classical residual CNN widely used for mid-scale visual recognition tasks. The second is ConvNeXt-Tiny [12], a modern ConvNet with Transformer-inspired architectural elements. The third model is ViT-B/16 [8], a vision transformer operating on 16×16 image patches. For all architectures, we replace the final classification layer with a five-way linear head to predict the zoom-view category.

To avoid memorizing particular movies, we use the following split: five trailers for training, one for validation, and two for testing. Frames are resized to 224×224 and augmented with RandomResizedCrop, ColorJitter, hor-

Table 2. **Zoom-view classification accuracy.** ConvNeXt-Tiny achieves the best performance among the evaluated backbones.

Model	Val. Acc.	Test Acc.
ResNet-18	93.35%	59.59%
ConvNeXt-Tiny	93.35%	**63.13%**
ViT-B/16	93.74%	43.81%

izontal flipping, and standard normalization, trained on 20 epochs. Because UPPERBODY and NOPARTS dominate the dataset, we apply a weighted cross-entropy loss with inverse-frequency class weights.

Table 2 summarizes the main result: ConvNeXt-Tiny provides the strongest baseline for zoom-view classification, while ViT-B/16 underperforms in this relatively small-data regime.

3.2. Scale-Aware Action Recognition

Before applying the view-specific HAR modules, we first construct coherent person-centric clips. Each video is decomposed into frame-level detections using YOLOv8 [17], chosen for its strong balance of speed and accuracy. To maintain temporal identity under fast motion and occlusion, detections are linked using the BoT-SORT tracker [1], producing stable tracklets rather than isolated bounding boxes. For every tracked individual, we extract anatomical cues using MediaPipe [13], including facial regions, body pose keypoints, and hand keypoints. Because per-frame labels can be noisy due to detector jitter, we apply a temporal smoothing stage to merge short runs and bridge small gaps, yielding stable zoom-view segments from which the final action-recognition clips are formed.

Once these temporally consistent zoom-view segments are available, we attach an additional layer of human action recognition (HAR) that is specialized to each view category. For HANDSONLY clips, we apply X-CLIP [14], a multimodal vision–language transformer capable of zero-shot gesture recognition from video–text pairs. FACECLOSE clips are processed with DeepFace [16], applied to a few sampled frames and aggregated by majority vote for robust facial-emotion estimation. UPPERBODY segments use the CONFIDANT architecture [3], originally designed for micro-gesture analysis, which we repurpose for cinematic medium shots. FULLBODY clips are processed with a pretrained STGCN++ [9] model operating on body-pose sequences, benefiting from Kinetics-400 [5] pretraining. Together, these specialized recognizers enrich each zoom-view segment with semantically appropriate actions and emotions. Figure 3 shows typical examples in which the zoom classifier selects a view scale and the corresponding HAR module predicts a scale-appropriate label.

3.3. Semantic Facial Descriptions via Gemini

Close-up facial footage carries subtle expressive cues that are difficult to capture using standard supervised labels. To enrich the metadata attached to FACECLOSE segments, we incorporate a multimodal semantic analysis stage using the Gemini 2.5 Flash model [6]. The objective is to obtain structured, machine-readable descriptions of emotion dynamics, gaze shifts, mouth-shape trajectories, and muscle activations—signals directly relevant for controllable avatar generation and blendshape-based animation.

FACECLOSE clip is passed to Gemini under a constrained JSON schema. The model produces (i) a global affect summary, (ii) compact descriptions of facial components, and (iii) a per-frame temporal timeline capturing changes in emotion, gaze direction, mouth configuration, and brow tension.

To ensure consistency and downstream usability, we fix the output structure through the following prompt:

```
Return ONLY a valid JSON object with:

{
  "global_state": {
    "dominant_emotion": "string",
    "valence": "float (-1 to 1)",
    "arousal": "float (0 to 1)",
    "emotion_change": "stable | escalating |
    de-escalating | mixed"
  },
  "face_description": {
    "eye_shape": "string",
    "gaze_behavior": "string",
    "brow_tension": "string",
    "mouth_behavior": "string",
    "notable_muscles": ["token1", "token2"]
  },
  "timeline": [{
    "frame_index": "int",
    "time_position": "float (sec)",
    "emotion": "string",
    "valence": "float",
    "arousal": "float",
    "gaze_direction": "string",
    "mouth_shape": "string",
    "brow_tension": "string"
  }]
}
```

This schema mirrors concepts widely used in animation, such as gaze vectors [15], viseme-based mouth shapes [2], and FACS-related muscle units [10], making the result directly compatible with blendshape regressors, talking-head models, or emotion-driven motion synthesis pipelines.

Figure 4 shows an example where Gemini tracks the evolution from sadness to a pained smile, recovering meaningful changes in valence, arousal, and muscle engagement. Such temporally aligned descriptors provide a richer supervisory signal than categorical labels alone and complete the ZoomGate pipeline with high-level, animation-ready facial semantics.

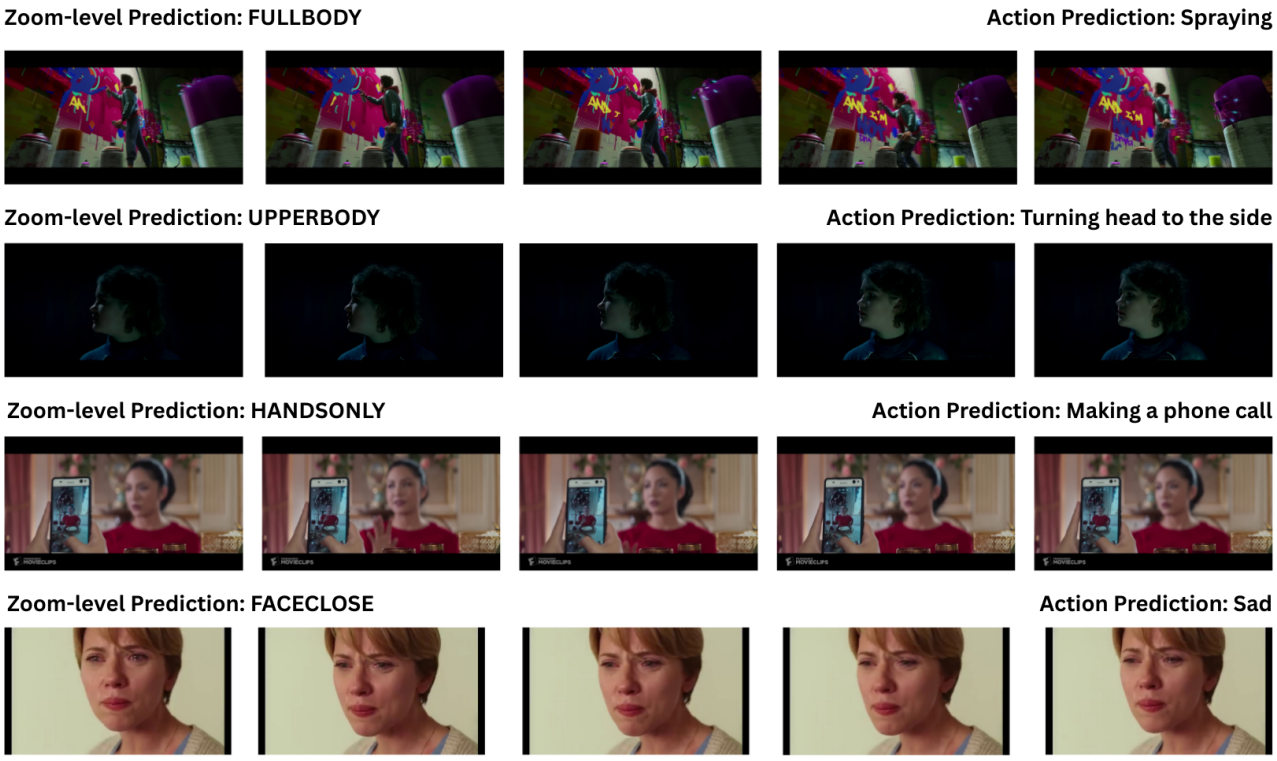

Figure 3. **Qualitative examples of scale-aware action recognition.** Each row shows frames from a zoom-view segment (FULLBODY, UPPERBODY, HANDSONLY, FACECLOSE) together with the predicted action. ZoomGate first predicts the zoom level, then routes the clip to a view-specific HAR module, yielding interpretable, scale-consistent action labels.

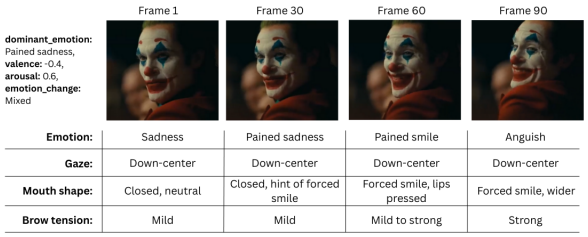

Figure 4. **Gemini-based semantic analysis of a FACECLOSE segment.** The model produces structured JSON describing global affect, facial-component behavior, and a per-frame timeline of emotional and articulatory states.

4. Conclusion

ZoomGate presents a unified framework for understanding cinematic footage through the lens of view scale, enabling downstream action recognition and semantic facial analysis that adapt naturally to the camera's distance from the actor. By combining a zoom-view classifier, scale-aware human action recognition, and a Gemini-based semantic description module, our pipeline forms a lightweight data engine for extracting expressive, animation-ready signals from real film material. The resulting representations (zoom classes, action labels, and temporally detailed facial descriptors) offer a principled foundation for building AI-driven characters that can behave, gesture, and emote coherently across mixed camera scales.

Despite demonstrating the value of scale-aware perception, several limitations remain. Our dataset is still relatively small and biased toward trailer-style cinematography; broader coverage of genres and shot styles would improve generalization. Additionally, HANDSONLY remains severely under-represented. Future work should integrate end-to-end, jointly trained models, expand multimodal supervision like audio-driven emotion cues, and explore how scale-dependent semantics can directly condition generative avatar models. Ultimately, we view ZoomGate as a step toward characters that not only move realistically, but also modulate their behaviour with the same sensitivity to framing that human actors and cinematographers instinctively apply.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. *arXiv*

Label	Zoom Frames	HAR Clips	HAR Frames
FACECLOSE	989	18	995
HANDSONLY	61	3	101
FULLBODY	1,642	34	1,957
UPPERBODY	9,748	266	43,300
NOPARTS	7,424	–	–
Total	**19,864**	**319**	**46,284**

Table 3. Zoom-view frame distribution and Human Action Recognition (HAR) clip statistics.

preprint arXiv:2206.14651, 2022. 3

[2] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of ACM SIGGRAPH*, pages 353–360, 1997. 3

[3] Kseniia Buzko. From far-field dynamics to close-up confidence: Action recognition across varying camera distances. Master's thesis, University of Waterloo, Waterloo, Canada, 2025. Systems Design Engineering. 3

[4] Katalin Eva Bálint, Janine Nadine Blessing, and Brendan Rooney. Shot scale matters: The effect of close-up frequency on mental state attribution in film viewers. *Poetics*, 83:101480, 2020. 1

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 3

[6] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 2

[9] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7351–7354, 2022. 3

[10] Paul Ekman and Wallace V. Friesen. Facial action coding system (facs). https://doi.org/10.1037/t27734-000, 1978. Database record. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 2

[12] Zhuang Liu, Hanzi Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 1, 2

[13] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. 3

[14] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 638–647, New York, NY, USA, 2022. Association for Computing Machinery. 3

[15] Christopher Peters and Carol O'Sullivan. Animating gaze for virtual characters. In *Proceedings of the ACM Symposium on Applied Perception*, pages 23–30, 2003. 3

[16] Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107, 2024. 3

[17] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. 3