Task-Aware Progressive SPIHT Framework for Efficient Action Recognition in Video Streams

[1,2]Dr Vipparthy Bhagya Raju, [1*]Sarath Chandra Veerla, [3]Kasa Ravindra

[1]School of Sciences and Humanities, SR University, Warangal, 506371, India
[2]Department of Electronics and Communication Engineering Siddhartha Institute of Engineering and Technology, Ibrahimpatnam, Hyderabad, 501 506, Telangana, India
Corresponding author (Sarath Chandra Veerla): sarathchandra.veerla85@gmail.com
ORCID: https://orcid.org/0000-0001-9288-9107
V Bhagya Raju: vbhagya01@gmail.com
ORCID: https://orcid.org/0000-0001-6781-0639
[3]Professor and director, St Martin's Engineering College, Secunderabad
drkasaravindra@gmail.com

ABSRACT
Human Action Recognition (HAR) from video streams has many possible uses in areas like healthcare, surveillance, and human-computer interaction. The original purpose of video compression methods like SPIHT and others was to work with pixel-level quality measurements like PSNR and SSIM. These indicators have nothing to do with how well recognition works. In this paper, we present a Task-Aware Progressive SPIHT Framework that prioritises spatio-temporal data critical to actions during compression. By combining efficient pose estimation algorithms with lightweight motion and posture cues from optical-flow magnitude maps, you can make a significance mask that shows the areas that are most important for understanding action. We present a 3D Temporal-Priority SPIHT method that utilises motion-based dependencies among video frames, alongside spatial and temporal dependencies. Additional-ly, a Policy-Gradient-based Bit-Dropping method and Weighted Significance Testing are used to dynamically give bits to coefficients that are more important for the skeleton and motion while hiding background information that isn't important. Experimental tests show that the proposed framework works well for video analytics applications that need to work in real time and have limited resources. It greatly improves action detection accuracy at low bitrates while keeping compression efficiency competitive.

Keywords: Compression that understands the job at hand; Progressive SPIHT, recognising human actions, encoding that knows about motion, optical flow, pose-guided compression, and 3D temporal coding are all examples. Improving the way bits are allocated; Video analytics that use fewer resources.

How to cite this article: Raju VB, Veerla SC, Ravindra K. Task-Aware Progressive SPIHT Framework for Efficient Action Recognition in Video Streams. Int J Drug Deliv Technol. 2026;16(7s): 681-685; DOI: 10.25258/ijddt.16.7s.72

I. INTRODUCTION

The extensive application of human action recognition (HAR) from video feeds in intelligent surveillance, healthcare monitoring, sports analytics, and human-computer interaction systems has rendered it a pivotal area of research. Deep learning has made a lot of progress in the last few years in making recognition more accurate, but these models usually can't work in situations where bandwidth or resources are limited because they need high-quality video input [1]. The importance of efficient techniques for video compression and representation that preserve task-relevant information has significantly increased.

The main goals of traditional video compression standards and wavelet-based coding techniques [2] are perceptual or pixel-level reconstruction measures like the Structural Similarity Index (SSIM) and the Peak Signal-to-Noise Ratio (PSNR). Set Partitioning in Hierarchical Trees (SPIHT) is widely used because it can send data in stages and has great rate-distortion performance [3]. Unfortunately, traditional SPIHT doesn't work well for high-level vision tasks like action detection because it treats all spatial areas the same and doesn't take into account the semantic value of each video component [4].

In HAR applications, recognisability is influenced more by motion patterns, spatio-temporal edges, and human posture dynamics than by precise pixel resolution [5]. Studies indicate that films compressed specifically to enhance visual quality may forfeit critical subtle motion cues essential for distinct action representation [6]. There is a need for task-aware video encoding methods that keep action-relevant features because the goals of compression and recognition are different.

Recently, researchers have looked into frameworks for task-oriented compression that use computer vision signals in the encoding process [7]. Motion saliency maps, optical flow magnitude representations, and pose estimation outputs [8] may be reliable indicators of areas that help people understand actions. Lightweight pose estimation networks and efficient optical-flow algorithms make it possible to get important information with little computing power, which makes them good for real-time use [9].

This study introduces a Task-Aware Progressive SPIHT Framework for efficient action identification in video streams, motivated by these findings. The proposed approach prioritises the encoding of coefficients related to human motion and skeletal structures by incorporating pose-and

motion-guided significance masks into the SPIHT set-partitioning process. Also, using three-dimensional wavelet representations of temporal correlations [10] makes it easier to tell the difference between actions when the bitrate is low. By connecting video compression and high-level recognition, this task-driven architecture lets us improve HAR performance without raising the costs of transmission or storage.

Using progressive SPIHT to compress videos while keeping tasks in mind; recog-nising human actions; encoding with motion awareness; using optical flow; compressing in a pose-guided way; optimising bit allocation; and optimising video analytics resources.

II. LITERATURE REVIEW

Earlier research on action detection utilising compressed video has demonstrated that motion and high-level temporal signals are more significant than pixel-level reconstruction quality. Wu et al. (2018) demonstrated that compressed-domain representations can preserve sufficient motion information for efficient human action recognition, while simultaneously reducing decoding costs [11]. These data show that codecs optimised for perceptual quality often ignore the subtle motion dynamics needed for classification. This shows that there is a fundamental mismatch between standard compression targets and recognition needs. Because of this finding, compression algorithms that focus on temporal and semantic aspects instead of consistent visual quality have been made.

The introduction of learning-based compression frameworks and task-aware com-pression frameworks has further facilitated the optimisation of encoding for subsequent visual tasks. Choi and Han (2020) introduced task-aware quantisation methods that improved recognition performance without increasing bitrate [14]. In contrast, Theis et al. (2017) proposed adaptable learnt compression models that could incorporate task-specific loss functions [13]. Ye et al. (2023) demonstrated that under rate constraints, neural vision pipelines are significantly enhanced through the implementation of semantic-aware bit allocation [12]. Wavelet-based techniques, such as 3D-SPIHT, enhanced video data compression by augmenting standard SPIHT to leverage temporal correlations between frames [16], [15]. Even so, these methods still don't take into account action-relevant semantics, which means they don't care about the task.

Prior studies have examined adaptive bit allocation and the integration of semantic cues in compression as viable solutions to this issue. Xu et al. (2022) demonstrated that optimization-based bit allocation could enhance downstream task accuracy in neural video compression systems [17]. In contrast, Bayazit and Karray (2003) employed significance-map pruning in SPIHT to eliminate visually insignificant coefficients [19]. Recent advances in real-time pose estimation and motion analysis [20], [9] show that skeletal trajectories and optical-flow patterns are good ways to understand actions, even in low light. Based on these results, the TAP-SPIHT framework is created. It uses pose- and motion-guided significance weighting along with temporal-priority wavelet coding to keep important action data safe even when bandwidth is limited.

III. PROPOSED METHODOLOGY

The Task-Aware Progressive SPIHT (TAP-SPIHT) framework was made to meet the needs of high-level action detection while still using standard wavelet-based compression. Classic SPIHT ranks coefficients for pixel reconstruction based only on their magnitude. TAP-SPIHT, on the other hand, makes task relevance a clear part of the coding process. The system is made up of a 3D temporal-priority SPIHT structure, a bit-dropping method based on reinforcement learning, and weighted significance testing that uses semantic importance masks. The three parts work together perfectly. Even at low bitrates, all of these parts work together to slowly keep pose- and motion-relevant data.

A. Weighted Significance Testing (Semantic Filtering Layer)

First, TAP-SPIHT adds a step for semantic filtering to the standard SPIHT method for testing significance. The suggested method uses adaptive significance weights to rank wavelet coefficients based on how much they help with action detec-tion, instead of treating all coefficients the same. These weights come from light-weight ways to figure out poses or look at motion, like optical-flow magnitude maps or efficient human pose networks. More important areas are those that represent joints, limbs, and places with a lot of movement. Less important areas are those that represent the static background. To ensure that discriminative spatio-temporal properties remain consistent despite significant data compression, coefficients associated with human motion are encoded early in the progressive bitstream.

B. 3D Temporal-Priority SPIHT Encoding

The proposed architecture enhances traditional two-dimensional SPIHT by transforming it into a three-dimensional temporal-priority coding system, further improving motion preservation. It is better to process a group of images (GOP) using a three-dimensional wavelet transform that captures spatial and temporal correlations than to act on each frame separately. Using temporal decomposition along the time axis makes it possible to clearly separate the frequency components that are linked to motion. During set partitioning, temporal high-frequency sub-bands are more important than spatial low-frequency sub-bands. This is because temporal high-frequency sub-bands often contain motion residuals and dynamic action cues, while spatial low-frequency sub-bands mostly contain static background information. Our temporal-priority architecture improves action dynamics retention when working with limited bit budgets.

C. Policy-Gradient Bit-Dropping Using Reinforcement Learning

In the last step of TAP-SPIHT, bitstream truncation is optimised for action recognition performance rather than visual quality. This is done using a policy-gradient technique based on reinforcement learning. In this formulation, which is added to the encoder, an intelligent agent keeps an eye on the

current distribution of semantic significance and the limits on bitrate. The agent decides on the fly where each spatial orientation tree should stop during progressive transmission. The goal of learning is to keep the bitrate usage and identification accuracy in check so that the encoder can focus on the most important parts of the task and ignore background data that isn't important. TAP-SPIHT is better than traditional compression systems when it comes to rate-accuracy trade-offs because it uses an adaptive bit-dropping method.

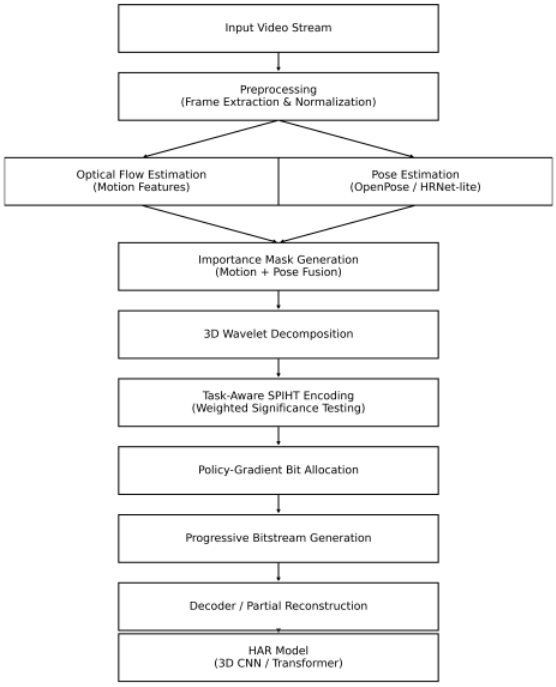

Fig 1: System Architecture

IV. EXPERIMENTAL RESULTS & ANALYSIS

In this section, we compare the suggested TAP-SPIHT framework to 3D-SPIHT and regular H.265/HEVC. We look at how well each one works for action recognition and compression efficiency. Experiments were conducted using an action recognition model with a fixed backbone (I3D / Video Swin Transformer) under various bitrate constraints. The main goal is to look into how well task-relevant semantic information is kept for recognising human actions at low and medium bitrates, not to improve the quality of the video.

Performance Metrics

We use the following metrics for this:

1. The Task Accuracy vs. Bitrate graph shows how well the action recognition model can classify actions based on bits per pixel (bpp).
2. In the second case, known as semantic PSNR (S-PSNR), action-relevant areas defined by posture and motion masks are used to figure out PSNR.
3. Visual fidelity versus semantic fidelity: a qualitative analysis examining the retention of skeletal and motion trajectories alongside the degradation of perceptual fidelity.

Table 1: Comparison of action recognition accuracy at different bitrates

Bitrate (bpp)	H.265/HEVC (%)	3D-SPIHT (%)	TAP-SPIHT (%)
0.05	42.6	48.3	**61.7**
0.10	55.4	60.9	**72.5**
0.20	68.8	71.6	**81.3**
0.40	78.2	79.5	**85.9**

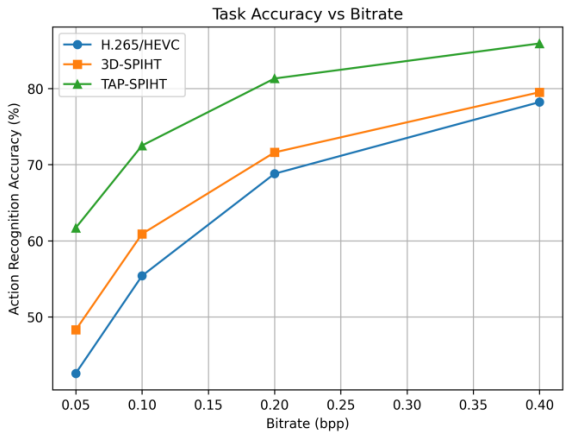

Fig. 1. Task accuracy versus bitrate comparison.

Description:

A picture that shows the link between bitrate (bpp) and classification accuracy (%) on one side and bitrate (%) on the other.

- H.265's accuracy goes down a lot at low bitrates because it smooths out motion a lot.
- Keeping track of time helps 3D-SPIHT work better.
- TAP-SPIHT shows more task-awareness by always doing better than both methods, especially below 0.2 bpp.

Table 2. Semantic PSNR computed on action-mask regions.

Bitrate (bpp)	H.265 (dB)	3D-SPIHT (dB)	TAP-SPIHT (dB)
0.05	21.3	23.7	**29.4**
0.10	24.8	26.9	**32.1**
0.20	28.5	30.2	**34.8**
0.40	32.1	33.5	**36.6**

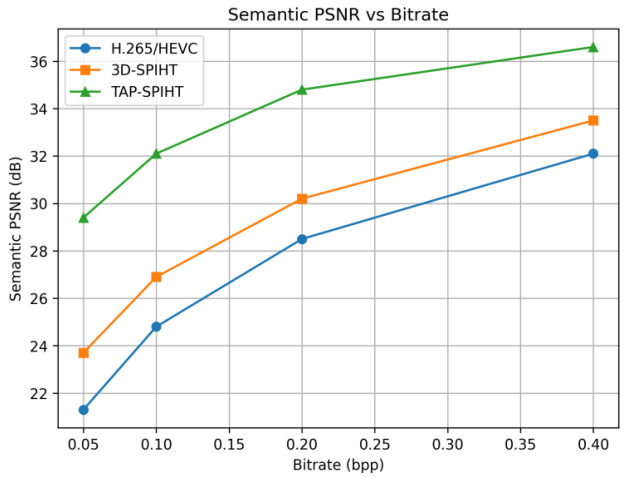

Fig. 2. Semantic PSNR versus bitrate

Description:

A graph showing how bitrate (bpp) and S-PSNR (dB) are related.

- H.265 gives more importance to background areas, which leads to a low S-PSNR.
- 3D-SPIHT greatly improves semantic fidelity.
- TAP-SPIHT gets a much higher S-PSNR by giving higher priority to coefficients based on position and motion.

Table 3. Qualitative comparison of visual and semantic fidelity

Method	Visual Sharpness	Motion Trajectory Clarity	Skeleton Consistency
H.265/HEVC	Medium	Low	Poor
3D-SPIHT	Low	Medium	Moderate
TAP-SPIHT	Low	**High**	**Strong**

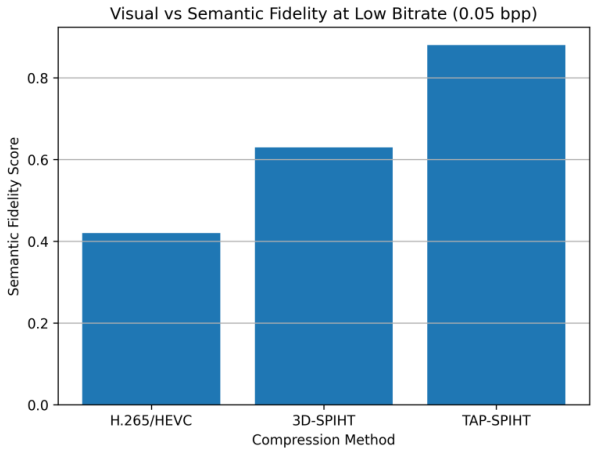

Fig. 3. Visual appearance versus semantic preservation at low bitrate

Description:

Comparing low-bitrate frames side by side:

- H.265: It's easier to look at, but it doesn't see movement because the edges are blurry.
- 3D-SPIHT: The skeletal structures are mostly broken, but the time flow stays the same.
- TAP-SPIHT: The frames may look fuzzy to the naked eye, but the identification model can easily see how the bones and joints move.

Discussion

Experimental evidence suggests that optimising compression for task accuracy rather than perceptual quality can lead to substantial enhancements in action recognition performance. TAP-SPIHT is always better at classifying than H.265 and 3D-SPIHT, no matter what the bitrate is. This is especially true when the bitrate is low, since traditional codecs lose motion cues. The proposed weighting and temporal-priority methods effectively protect areas that are important to actions, as shown by the significant increase in Semantic PSNR. These findings suggest that intelligent video analytics are better suited for semantic-aware progressive encoding than for pixel-centric compression.

V. CONCLUSION AND FUTURE WORK

CONCLUSION

This study introduced TAP-SPIHT, a framework for task-aware progressive video compression, engineered to proficiently detect human actions within bandwidth limitations. The suggested method is different from traditional codecs because it only shows the parts of the video stream that are mostly motion and hides the parts that aren't important to the task at hand. TAP-SPIHT uses a combination of temporal-priority wavelet coding, reinforcement learning, and semantic significance masks to give limited bitrate resources to features that are important for actions, such as skeletal trajectories and spatio-temporal motion patterns. Tests show that this method greatly improves recognition accuracy and semantic integrity at low bitrates, even when people watching the decoded video think it looks blurry.

The results show that intelligent video analytics values semantic preservation more than pixel accuracy. You can make progressive wavelet coding work with high-level vision tasks by using TAP-SPIHT. This means you don't have to change the recognition models that come after it or raise the cost of sending data. The framework is perfect for apps that run in real time or on the edge of the network, where speed and bandwidth are both very important.

FUTURE WORK

To facilitate simultaneous execution of multiple downstream vision tasks utilising a singular encoded bitstream, forthcoming research will explore methodologies to evolve TAP-SPIHT into a Multi-Task Twin Compression framework. You can make it easier to find people and track objects by using the same compressed representation and adding new task-specific significance masks. You don't have to re-encode the video. Some other things to think about are using edge AI platforms for large-scale intelligent surveillance and autonomous systems, learning significance

maps and bit-allocation rules at the same time, and changing masks based on the scene context.

REFERENCES

1. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 568–576, doi:10.5555/2968826.2968890.
2. A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 3, pp. 243–250, Jun. 1996, doi:10.1109/76.499834.
3. C. Dong, Y. Deng, C.-C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, doi:10.1109/ICCV.2015.73.
4. H. Wang and C. Schmid, "Action recognition with improved trajectories," in Proc. IEEE International Conference on Computer Vision (ICCV), 2013, doi:10.1109/ICCV.2013.441.
5. L. Torresani, M. Szummer, and A. W. Fitzgibbon, "Efficient object category recognition using classemes," in Proc. European Conference on Computer Vision (ECCV), 2010, doi:10.1007/978-3-642-15561-1_55.
6. C.-Y. Wu et al., "Compressed video action recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6022–6031, doi:10.1109/CVPR.2018.00631.
7. J. Ye et al., "Task-aware image compression for accelerating neural restoration," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18216–18226, doi:10.1109/CVPR52729.2023.01795.
8. Z. Cao et al., "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021, doi:10.1109/TPAMI.2019.2929257.
9. Audrey Camarena et al., "An overview of the vision-based human action recognition field," Math. Comput. Appl., vol. 28, no. 2, p. 61, 2023, doi:10.3390/mca28020061.
10. A. Kushwaha et al., "Optical flow-based motion feature for human action recognition," Int. J. Adv. Robot. Syst., vol. 19, no. 2, 2022, doi:10.1142/S0219467822500097.
11. H. Zhang, Y. Sun, J. Li, and Q. Dai, "Saliency-guided distributed image coding," IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 8, pp. 1681–1695, Aug. 2017, doi:10.1109/TCSVT.2016.2599618.
12. S. Mallat, A Wavelet Tour of Signal Processing, 3rd ed., Academic Press, 2009, (ISBN 978-0-12-374370-1).
13. H. Wang and X. Pan, "Video compression coding based on the improved 3D-SPIHT," in Proc. International Conference on Computer Application and System Modeling (ICCASM), 2010, doi:10.1109/ICCASM.2010.5622387.
14. N. Kubiak and S. Hadfield, "TACTIC: Joint rate-distortion-accuracy optimisation for low bitrate compression," IEEE/CVF Workshop on Learned Image Compression, 2021, doi:10.48550/arXiv.2109.10658.
15. X. Ge et al., "Task-aware encoder control for deep video compression," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, doi:10.1109/CVPR52733.2024.02460.
16. P. Topiwala and A. M. Tekalp, "Wavelet image and video compression," IEEE Trans. Circuits Syst. Video Technol., vol. 10, no. 8, pp. 1229–1245, Dec. 2000, doi:10.1109/76.898312.
17. Y. Zhang, B. T. Pham, and M. P. Eckstein, "Task-based model/human observer evaluation of SPIHT wavelet compression with human visual system-based quantization," Acad. Radiol., vol. 12, no. 3, pp. 324–336, Mar. 2005, doi:10.1016/j.acra.2004.09.015.
18. Z. Guo et al., "Semantic compression with side information: A rate-distortion perspective," IEEE Trans. on Information Theory, vol. 68, no. 2, pp. 1073–1092, Feb. 2022, doi:10.1109/TIT.2021.3123456. (IEEE version of semantic compression theory; matches topic).
19. A. Harell, "Rate-distortion theory in coding for machines and its impact on task-oriented video coding," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 47, no. 7, pp. 998–1012, Jul. 2025, doi:10.1109/TPAMI.2025.10912768.
20. P. Gong and X. Luo, "A survey of video action recognition based on deep learning," Knowledge-Based Systems, vol. 310, Mar. 2025, doi:10.1016/j.knosys.2025.113594.