SkeletonVis: Interactive Visualization for Understanding Adversarial Attacks on Human Action Recognition Models

Haekyu Park[1], **Zijie J. Wang**[1], **Nilaksh Das**[1], **Anindya S. Paul**[2], **Pruthvi Perumalla**[1], **Zhiyan Zhou**[1], **Duen Horng Chau**[1]

[1]Georgia Institute of Technology,
[2]Intel
{haekyu, jayw, nilakshdas, pp329, zzhou406, polo}@gatech.edu,
anindya.s.paul@intel.com

Abstract

Skeleton-based human action recognition technologies are increasingly used in video based applications, such as home robotics, healthcare on aging population, and surveillance. However, such models are vulnerable to adversarial attacks, raising serious concerns for their use in safety-critical applications. To develop an effective defense against attacks, it is essential to understand how such attacks mislead the pose detection models into making incorrect predictions. We present SKELETONVIS, the first interactive system that visualizes how the attacks work on the models to enhance human understanding of attacks.

Introduction

Skeleton-based human action recognition technologies have been widely used in many video understanding based applications, such as home robotics, eldercare, and surveillance (Yan, Xiong, and Lin 2018; Choutas et al. 2018; Saggese et al. 2019). These models have demonstrated promising prediction capabilities by analyzing the joints of human bodies (Yan, Xiong, and Lin 2018). However, such techniques are vulnerable to adversarial attacks; an attacker can add visually imperceptible perturbations to an input and fool the models into arriving at incorrect predictions (Liu, Akhtar, and Mian 2019; Freitas et al. 2020). This jeopardizes the use of human action recognition technologies in safety-critical applications, such as autonomous driving. To make the models more robust against attacks, it is essential to understand how attacks inflict harm on the models. Interactive visual analytics systems have great potential in helping users gain insights into adversarial attacks on deep learning models (Das et al. 2020a,b; Park, Hohman, and Chau 2019). We present SKELETONVIS (Figure 1), the first interactive visualization tool for understanding how adversarial attacks manipulate the input to induce incorrect prediction.

Demonstrating SKELETONVIS

SKELETONVIS takes as inputs a benign video clip (e.g., *person lunging*) and its "attacked" version (e.g., *exercising with a gym ball*) generated by an attack through perturbing pixels in benign clips (Figure 1A). SKELETONVIS visualizes the detected human joints in the benign and attacked videos, and highlight how they may misalign and contribute to wrong predictions. Furthermore, SKELETONVIS provides quantitative measurements across video frames to quantify the abnormal signals introduced in the attacked frames, which helps users more easily hone in on the specific frames exploited by the attacks (Figure 1B).

Extracting Human Joints to Predict Action

SKELETONVIS supports end-to-end skeleton-based classification:

- It first detects 17 human joints in each video frame, using Detectron2's body joint R-CNN model (Wu et al. 2019), providing a spatial confidence map for each joint.
- It then combines the resulting spatial joint information from all video frames to infer the human action performed using ST-GCN (Yan, Xiong, and Lin 2018), a state-of-the-art model for skeleton-based action detection.

Adversarial attacks are performed on the combined end-to-end model.

Skeleton View: Explaining How Attacks Manipulate Detection of Human Joints

Skeleton View (Figure 1A) visually explains how the attacks are manipulating the human joints to cause misclassification, via a *Comparison View* (Figure 1A-1) and a *Split View* (Figure 1A-2).

Comparison View enables *spatial* comparison of how the detected joints are located differently in benign and adversarial frames by overlapping the two sets of joints. For example, by revealing that the left foot is detected in unexpected locations in adversarial frames, users can more easily glean insights into why the pose detection model misclassifies the video (e.g., person lunging) as an incorrect action (e.g., exercising with an exercise ball).

Split View compares *movement* of the human joints in the benign and adversarial clips. For a clip, we visualize the trajectory of a human joint as follows. Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T$ be the sequence of the coordinates of a human joint in a clip. We display the movement as in Figure 1A-2, where joints from earlier frames (i.e., $\mathbf{x}_1, \mathbf{x}_2, ...$) are visualized as more transparent dots, and those from frames (i.e., $\mathbf{x}_T, \mathbf{x}_{T-1}, ...$) are visualized as more opaque dots. The joint trajectories and

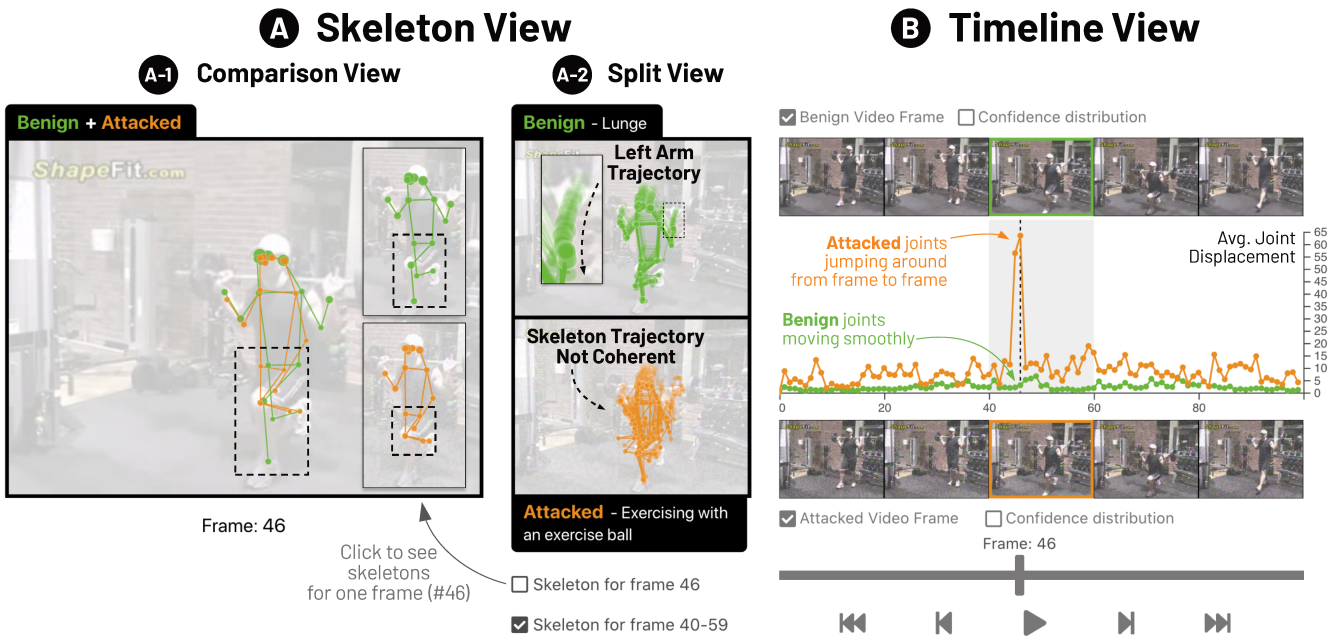

Figure 1: The interface of SKELETONVIS, visualizing how the *Fast Gradient Method* manipulates the left foot joints detected by the Detectron2 Keypoint R-CNN model. (A) The *Skeleton View* shows the joints perturbed to unexpected locations. (B) *Timeline View* reveals the attacked joints spuriously jumping around from one frame to the next, leading to a "spike" in the average joint displacement across attacked frames. These manipulations finally sway the ST-GCN action detection model into misclassifying the attacked frames as "exercising with exercise ball," instead of the correct "lunge" classification.

frames of the benign video are shown at the top, and those of the adversarial videos are shown at the bottom.

Timeline View: Visualizing Abnormal Signals from Adversarial Attacks

To help users discover attacked frames, SKELETONVIS quantifies and visualizes the pose detection models' responses to benign and attacked videos in the *Timeline View* (Figure 1B), enabling users to interactively compare and more easily pinpoint frames that they need to be wary of.

At the top of *Timeline View*, users can examine a *video segment* (i.e., a range of frames) in detail, by clicking on the segment's thumbnail. This highlights the thumbnail and its corresponding average joint displacement values in a line chart, where the horizontal axis represents frames, and the vertical axis represents the displacement from one frame to the next.

At the bottom of *Timeline View*, users can control which frame or segment to analyze through familiar controls used by typical video players. The interactive timeline consists of a *frame slider* and *time buttons*. Moving the knob on the *frame slider* sets the selected frame in *Comparison View* (Figure 1A-1) and the segment in *Split View* (Figure 1A-2). Pressing the "Play" button pauses or resumes the video playback. The "Fast backward" and "Fast forward" buttons at both ends bring users to the very beginning and the very end of the clip respectively. The "Forward" and "Backward" buttons step forward or backward by one frame.

Engaging the Audience

Our demonstration will focus on how even a basic adversarial attack, the Fast Gradient Method (Goodfellow, Shlens, and Szegedy 2014), is able to manipulate a Detectron2 Keypoint R-CNN model into making incorrect predictions of human joint positions. As described earlier, the attack will be crafted by combining the human joint detection model with the ST-GCN skeleton-based action detection model. SKELETONVIS will help the users better understand how the attack fools the downstream ST-GCN model into incorrectly classifying the action being performed, by directly perturbing pixels on an example video from the UCF-101 action classification dataset (Soomro, Zamir, and Shah 2012). For example, the audience of our demonstration will witness how the position of the detected feet is manipulated by the attack by very slightly perturbing the video pixels, that finally leads the model to misclassify "lunge" action as "exercising with gym ball" (Figure 1).

Acknowledgments

This work was supported in part by Defense Advanced Research Projects Agency (DARPA). Use, duplication, or disclosure is subject to the restrictions as stated in Agreement number HR00112030001 between the Government and the Performer. This work was supported in part by NSF grants IIS-1563816, CNS-1704701, and gifts from Intel, NVIDIA, Google, Symantec, eBay, Amazon.

References

Choutas, V.; Weinzaepfel, P.; Revaud, J.; and Schmid, C. 2018. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7024–7033.

Das, N.; Park, H.; Wang, Z. J.; Hohman, F.; Firstman, R.; Rogers, E.; and Chau, D. H. 2020a. Bluff: Interactively Deciphering Adversarial Attacks on Deep Neural Networks. In *IEEE Visualization Conference (VIS)*. IEEE.

Das, N.; Park, H.; Wang, Z. J.; Hohman, F.; Firstman, R.; Rogers, E.; and Chau, D. H. 2020b. Massif: Interactive Interpretation of Adversarial Attacks on Deep Learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7.

Freitas, S.; Chen, S.-T.; Wang, Z. J.; and Chau, D. H. 2020. Unmask: Adversarial detection and defense through robust feature alignment. *IEEE BigData* .

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

Liu, J.; Akhtar, N.; and Mian, A. 2019. Adversarial Attack on Skeleton-based Human Action Recognition. *arXiv preprint arXiv:1909.06500* .

Park, H.; Hohman, F.; and Chau, D. H. 2019. NeuralDivergence: Exploring and Understanding Neural Networks by Comparing Activation Distributions. In *Poster, Pacific Visualization Symposium (PacificVis)*. IEEE. URL http://haekyu.com/neural-divergence/.

Saggese, A.; Strisciuglio, N.; Vento, M.; and Petkov, N. 2019. Learning skeleton representations for human action recognition. *Pattern Recognition Letters* 118: 23–31.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* .

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI* .