Towards Learning and Generating Audience Motion from Video

Kenny Chen
Cesium GS
Philadelphia, USA
kenny@cesium.com

Norman I. Badler
Cesium GS
Philadelphia, USA
norm@cesium.com

ABSTRACT

There has recently been an explosion of interest in creating large-scale shared virtual spaces for multiplayer content. However, rendering player-controllable avatars in real-time creates latency issues when scaling to thousands of players. We introduce a human audience video dataset to support applications in deep learning-based 2D video audience simulation, bypassing the need for background 3D virtual humans. This dataset consists of YouTube videos that depict audiences with diverse lighting conditions, color, dress, and movement patterns. We describe the dataset statistics, our implicit data collection strategy, and audience video extraction pipeline. We apply deep learning tasks on this data based on video prediction techniques, and propose a novel method for 2D audience simulations.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

audience augmentation, crowd generation, video datasets, video prediction, neural networks

ACM Reference Format:
Kenny Chen and Norman I. Badler. 2023. Towards Learning and Generating Audience Motion from Video. In *The ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA '23), August 04–06, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3606037.3606839

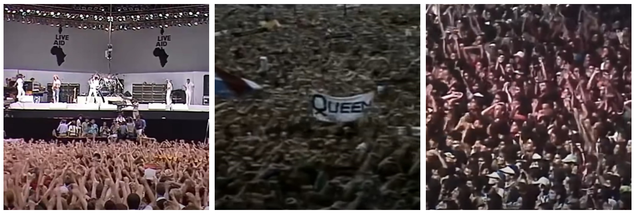

Figure 1: Example frames from our audience dataset from the "rock-concert" class. Frames contain a diverse range of audience and scene color, human and camera motion, audience size, and resolution. Images credits to YouTube creator raff809.

1 INTRODUCTION

While there have been several *crowd* video datasets [Grant and Flynn 2017; Waqar et al. 2022] for crowd analysis (counting and tracking) applications, there are no available large scale *audience* video datasets. While Durupınar et al. [2016] address audiences, they are not elaborated by type as in our work, but rather by psychological parameters. We observed that audiences exhibited distinct behavioral patterns that were unlike moving crowds, nor were they completely random. Accordingly, we hypothesized that we might be able to *automatically differentiate audience types to identify their context*, without overtly observing that context itself. Moreover, audience behavior over time is critically important to these distinctions. There has been much recent interest in massive multiplayer "Metaverse" events, which are currently unable to host thousands of simultaneous virtual attendees due to prohibitive rendering costs or network bandwidth requirements. For example, in hosting virtual events, Roblox divides live participants into manageable shards of 50-100 individual avatars, but would like to scale that number up into thousands [Morgan McGuire 2022]. Scalable methods to render realistic background audiences are therefore an important step in increasing the scope and immersion of these events.

Motivated by the inability of current systems to host thousands of simultaneous virtual attendees due to prohibitive rendering costs or network bandwidth requirements, we began to collect audience video data aimed toward conducting experiments for audience analysis and synthesis. We design a data collection pipeline for automatic extraction of audience videos, described in detail in Section 2. To demonstrate an application of this dataset, we propose a novel audience simulation method based on the video prediction literature, in which we train a video prediction generative model conditioned on 4 input frames to generate 10 frame predictions. This procedure is described in Section 4.

2 COLLECTING AN AUDIENCE DATASET FROM WEB VIDEOS

Our first goal is to collect a rich set of audience videos extracted from YouTube showing a diverse range of scenes and contexts. To achieve this, we implemented a data collection pipeline with three main steps. We describe steps (1)-(3) in more detail below. For a more extensive set of examples of videos in our dataset across all classes, please see our supplementary video.

2.1 Implicit Query for Raw Video Collection

Our data collection pipeline query strategy is influenced by Fouhey et al. [2017], which describes an *implicit* data collection strategy when querying online videos, whereby we search for "tennis match" or "concert" in which there almost certainly is some clip within

SCA '23, August 04–06, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0268-6/23/08.
https://doi.org/10.1145/3606037.3606839

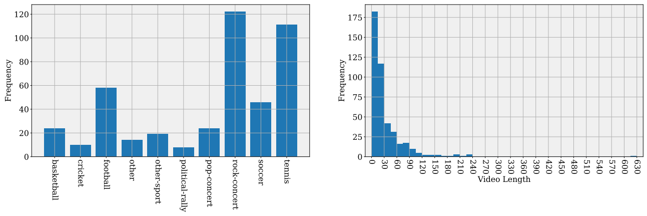

Figure 2: We display statistics of our human audience video dataset. Audience video label (left) and length in number of video frames (right) frequency are displayed as bar chart and histogram, respectively.

the video that pans to the audience. We extract these raw videos for further processing. Utilizing an *explicit* data collection strategy, using phrases such as "football audiences", leads to sparse and biased queries.

2.2 Audience Clip Extraction

After collecting this diverse set of raw videos, we extract clips containing audiences within them. We do this by running a crowd counting deep model, M-SegNet [Thanasutives et al. 2021], on our raw video dataset. We then ensure that the frame $\boldsymbol{F}$ contains an audience by computing the proportion of pixels with density above 0 density. We then accept a frame if this proportion is above δ, an empirical parameter. This is described by an inequality that sums the indicator variable representing the density $\mathbf{D}$ of a pixel at image location (x, y) being greater than 0, divided by the number of pixels, $w \cdot h$, in the frame. The frame must also contain a number of people, N, greater than 60, another parameter we determined empirically: $\frac{1}{w \cdot h} \sum_{(x,y) \in \boldsymbol{F}} 1\{\boldsymbol{D}(x, y) > 0\} > \delta, N > 60$.

2.3 Cleaning

We manually remove videos with significant proportion of the frame containing foreground or non-audience elements, such as frames focused on the event (i.e., on the tennis court) rather than the audience. Single frame examples collected after these 3 steps is shown in Figure 1.

2.4 Dataset Statistics

Our dataset contains 436 clips, 20,237 frames with an average of 46 frames per clip, minimum clip length of 15 frames, and maximum length of 640 frames. The average frame resolution is 1216x732. Frequency of each class and video length are displayed in Figure 2.

3 BENCHMARK CLASSIFICATION TASK

We use a 3D ResNet-50 [He et al. 2016] model trained on 4 input frames to classify each video into one of 10 classes. Our classification model achieves an accuracy of .75 and $\kappa = .61$. We use Cohen's κ due to the class label imbalance in our dataset.

4 AUDIENCE SYNTHESIS METHODOLOGY

We adapt our dataset to a video prediction task for audience behavior synthesis. While methods in video prediction have been traditionally applied in other contexts, we explore its novel application to audiences. Rendering massive crowds has been a problem in graphics for decades, and with the rise of massive multiplayer environments consisting of individual virtual avatars, the rendering costs of simulating thousands of humans with user-controlled movements becomes prohibitive. As such, our method renders realistic background audience video, conditioned on several input frames, at a similar speed regardless of audience size. We show in our supplementary video the result of training a conditional variational recurrent neural network (VRNN), using the method of Castrejon et al. [2019], with 4 input frames and 10 future predicted frames. We train our model on a single audience video scene in our dataset, containing 645 15-frame clips. We find that sampled results are plausible predictions of real audience videos, though high frequency regions are blurry. We hypothesize that it is difficult for networks to learn prediction results on our dataset due to the randomness of audience movements on an individual level and due to the density of humans in the scenes.

5 CONCLUSION AND FUTURE WORK

We presented our audience video dataset, which is the first of its kind used to categorize and reproduce non-locomotive crowds. We validate our hypothesis that we can identify scene context from audiences alone, with results showing that we can train a CNN on audience videos to retrieve class labels. We finally presented a novel methodology for conditional audience simulation by adapting video prediction work by Castrejon et al. [2019] to our dataset.

We expect this method to find application in the creation of background audiences in Metaverse venues. Populating virtual events and performance venues would enliven geospatial models and presumably pique participant interests in exploring what the synthetic audiences appear to be enjoying. By moving some of the 3D rigged model animation, rendering, and network computational burdens of audiences onto animated textures from 2D video synthesis from deep-learning models we can devote more compute bandwidth to enhancing the less numerous foreground actors and avatars.

REFERENCES

Lluis Castrejon, Nicolas Ballas, and Aaron Courville. 2019. Improved Conditional VRNNs for Video Prediction. In *The IEEE International Conference on Computer Vision (ICCV)*.

Funda Durupınar, Uğur Güdükbay, Aytek Aman, and Norman I. Badler. 2016. Psychological Parameters for Crowd Simulation: From Audiences to Mobs. *IEEE Transactions on Visualization and Computer Graphics* 22, 9 (2016), 2145–2159. https://doi.org/10.1109/TVCG.2015.2501801

David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. 2017. From Lifestyle Vlogs to Everyday Interactions. *CoRR* abs/1712.02310 (2017). arXiv:1712.02310 http://arxiv.org/abs/1712.02310

Jason M Grant and Patrick J Flynn. 2017. Crowd scene understanding from video: a survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 2 (2017), 1–23.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Morgan McGuire. 2022. Personal communication.

Pongpisit Thanasutives, Ken-ichi Fukui, Masayuki Numao, and Boonserm Kijsirikul. 2021. Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2382–2389.

Sahar Waqar, Usman Ghani Khan, M Hamza Waseem, and Samyan Qayyum. 2022. The utility of datasets in crowd modelling and analysis: a survey. *Multimedia Tools and Applications* (2022), 1–32.