*Artificial Intelligence, Social Computing and Wearable Technologies,* Vol. 113, 2023, 88–96
https://doi.org/10.54941/ahfe1004182

# Automated Generation of Synthetic Person Activity Data for Training AI Models

**Dominik Breck, Max Schlosser, Rico Thomanek, Christian Roschke, Marc Ritter, and Matthias Vodel**

University of Applied Sciences Mittweida, Mittweida, 09648, Germany

## ABSTRACT

This paper presents a workflow for automated generation and annotation of synthetic video data. This includes the provision and animation of person models of different ethnicities as well as the filming of the characters in a virtual environment. By generating an exemplary dataset of 90,000 primitively designed videos consisting of 13 activity classes, it could be shown that the workflow is suitable for the field of synthetic video generation for activities. The dataset called AAR-P01 is downloadable for free[1]. It can be assumed that the workflow can also be used for other generation methods, such as synthetic data for training object classifiers, due to its flexible, modular structure.

**Keywords:** Synthetic data generation, Video activity detection, Image and video analysis, Deep learning

## INTRODUCTION

Today's digitalized world is subject to constant change, which continuously provides new approaches to solving existing problems (Elstner et al., 2016). For example, new approaches in the field of artificial intelligence are being used more and more in a wide variety of business and research areas. For example, the field of machine video analysis is also strongly influenced by the use of artificial intelligence (Lichtenthaler, 2021). Various approaches and algorithms can be used to analyze video footage in order to solve specific problems in various areas of everyday life, such as general security (Ahmed and Echi, 2021) (Fontes et al., 2022).

A special area is the recognition of activities performed in images or videos of people. As in other fields, there are general problems in the application of artificial intelligence to activity recognition. One such problem is the huge amount of data needed to train such methods. This limitation is relevant to the extent that the data sets used and the underlying video material must first be laboriously collected and annotated. Another resulting problem is the potential bias due to a lack of diversity in the data (Djeffal, 2020).

[1] https://github.com/Schloool/activity-animation-recorder

These issues are not trivial to solve and have only been addressed to a limited extent. One approach is to provide synthetic data, which is artificially generated and has become increasingly important, especially in deep learning (Nikolenko, 2021). For video data, it has been shown that models based on synthetic data or hybrid ingestion of such data sometimes outperform methods based on real data alone (Kim et al., 2022). Virtual methods that minimize distracting elements in the background show particularly good results in tuning (Ballout et al., 2020). At the same time, virtual environments are not limited by the number of cameras to be used, allowing synthetic methods to additionally represent many otherwise underrepresented perspectives (Varol et al., 2021). Several real-time-based applications have already been used to generate synthetic video datasets for specific areas of activity detection (Hwang et al., 2021) (de Souza et al., 2017). Such approaches tend to focus on scenarios that are difficult to extend to new domains. Thus, the lack of flexibility limits not only the breadth of the dataset, but also the ability to feed it into machine learning processes.

As a new solution for synthetic video generation, a workflow that can be used to generate large datasets in the field of human activity recognition is presented. The focus is on the development of the Activity Animation Recorder (AAR) application: a real-time software that combines different aspects of current research and provides tools for optimized data generation and much easier subsequent processing. We focus on several advantages of the workflow:

- **Flexibility to integrate with existing machine learning workflows:** Components of the workflow can be interchanged as needed. In addition, datasets generated by the application can be used for both initial training and model tuning.
- **Size of datasets:** Generated datasets have tremendous versatility in their environmental design and viewports displayed. The use of characters from a wide range of ethnicities also prevents bias in the generated videos. In addition, extensive metadata is provided for each recorded clip, with a focus on categorizing and describing the activities used.
- **Usability:** The interface of the real-time application offers numerous possibilities to generate annotated datasets even without prior knowledge in the areas of computer science, game engines or machine learning. In order to represent a broad spectrum of human activities, environmental factors such as camera settings, scene objects, background or lighting can be modified according to the user's needs.

## METHODOLOGY

The methodology presented includes a workflow for generating synthetic video with rich annotations in a fully automated process. The workflow uses human activity as an example for generating large amounts of data.

Different types of data may be needed to perform machine learning processes. In the case of activity recognition of everyday human activities, data sets are needed that represent examples of the activities to be recognized along with annotations of which activities are performed in the clip. The structure and type of the target data set can be very different and also depends on the machine learning processes that will be used later. Examples include, but are not limited to, providing an image-based video or frame-based dataset, or using a node-based dataset that only maps data to bone position data during an activity. Ultimately, the goal of the process described here is a dataset that unifies a variety of information, annotations, and metadata. Thus, video data and the nodes and bone points generated from them will be persisted, as well as metadata on the cameras used and annotations on the videos generated.

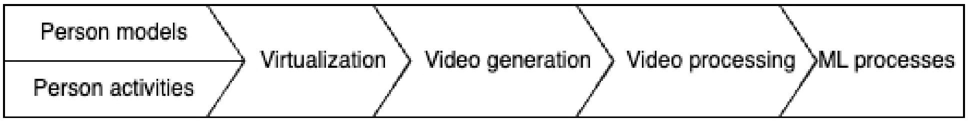

**Figure 1:** Process steps of the targeted workflow.

To achieve this goal, the objective here is to create a chain of individual tools that can be linked together to form an automated pipeline. To do this, a closer look at the general process must be taken. As can be seen in Figure 1, this process can be divided into six steps, five of which are considered in this workflow. First of all, person models are required as a starting point for the actions to be performed. These actions have to exist as person activity animation files. These components are then brought together in the third component of the visualization. Here the goal is that the activities are performed by one or more persons in combination with object interactions. In the fourth step, videos can be generated, which are further processed in the fifth step and made usable for machine learning processes. This general process is now transferred to a software level.

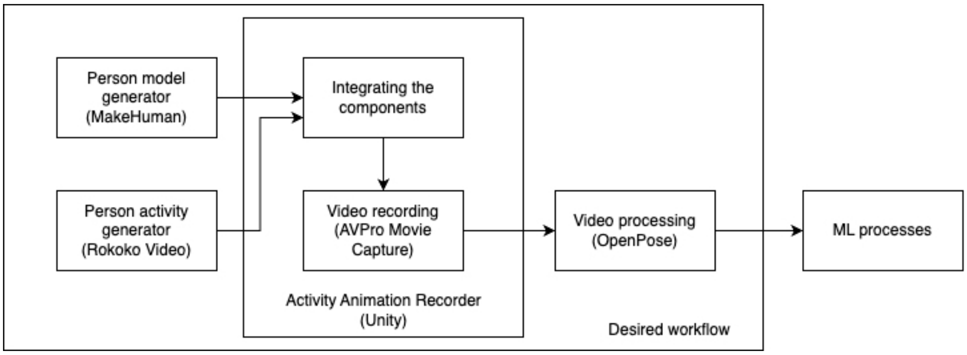
Person model generator (MakeHuman)
Person activity generator (Rokoko Video)
Integrating the components
Video recording (AVPro Movie Capture)
Activity Animation Recorder (Unity)
Video processing (OpenPose)
Desired workflow
ML processes

**Figure 2:** Outline of the components of the workflow.

For the components just described in the process, individual tools are now selected as examples and assigned to the process steps. Due to the modularity of the system, the selected components are only examples for an initial implementation and can be replaced by alternative components. The system components selected here as examples are shown in Figure 2 and will be examined in detail below.

### MakeHuman Plugin for Character Model Generation

The goal of the model generation process step is to create person models that can be provided with activity animations. MakeHuman[2] can be used for this purpose. By default, this tool allows the creation of person models by manual manipulation of buttons within the application. However, the process described here requires a large number of different person models. Such an export of several randomly generated person models is not provided by the application by default, but can be developed independently and thus integrated using the integrated plugin support. In this case, a large number of random-looking person models are automatically generated by specifying the desired number and exported as an FBX file and stored in the file system.

### Animation Generation With Camera-Controlled Motion Capturing

In addition to the required person models, person activities are provided in the second step of the process. This is done using Rokoko Video[3]. This software tool makes it possible to generate animations from recordings of people. In this way, the required person activities can be performed and recorded by a camera. The recordings are then imported into Rokoko Video and processed into animations.

### Activity Animation Recorder

The AAR software allows the recording of virtual scenarios in which human-related activities are performed. The application was developed using the Unity game engine. The application is based on 3D models and associated humanoid animations, which can be played and virtually recorded from different perspectives within the application.

In order to capture as many perspectives as possible, up to 500 virtual cameras are arranged in a hemispherical pattern around the virtual person being filmed. To save time, multiple cameras are recorded in parallel. At the same time, the virtual time in the application is accelerated. After recording, the accelerated video is slowed down to match the real time of the animation. The Unity plugin AVPro Movie Capture[4] is used for the outgoing recording from the cameras. Metadata is stored for each recorded video clip, which can include information about the type of activity being performed, camera settings used, changes in the character's bone points based on the animation, and data about the ethnicity or physical characteristics of the person being filmed. The application provides numerous configuration options for setting up the capture and generating metadata for each video clip.

The application's adaptability to different scenarios was ensured by a modular animation system. Users can customize animations as needed and

[2] http://www.makehumancommunity.org/

[3] https://www.rokoko.com/products/video

[4] https://renderheads.com/products/avpro-movie-capture/

configure the application for specific use cases. For example, animations can be configured to move characters in one direction at a given speed. Interaction between characters and other objects is also possible in two ways. First, it is possible to attach specific virtual objects to bone points on a model. This allows a character to hold a cell phone, for example. The second way is to interact directly with objects or characters: During animation, the subject can turn or move to another scene element while also playing independent animations. For example, processes such as talking to another person or getting into a car can be simulated. The application also supports scene setup at multiple points. By directly using the Unity game engine, the virtual environment can be designed according to one's own ideas with a detailed arrangement of 3D objects. Simplified methods of scene customization are also supported. These include selecting a skybox, randomly placing 3D models in the scene, and adjusting lighting. The setup and capture process is visualized in Figure 3.

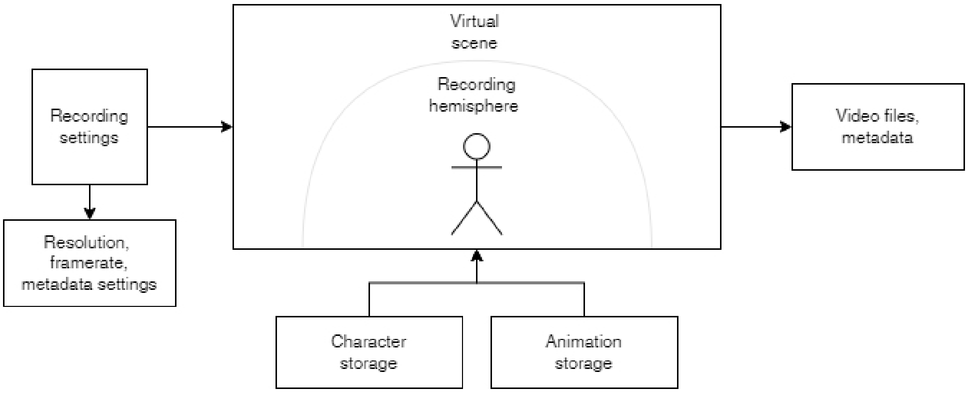

**Figure 3:** Exemplary representation of the components of the application AAR.

In addition to the aforementioned aspects, user-friendliness and ease of use for developers were also taken into account during development. The application allows users to control the recording via a graphical user interface or via the command line, depending on their individual needs. Among other things, parameters have been implemented to define the scope and format of stored metadata or to perform automatic train-test splitting of generated videos. These aspects simplify the integration of the generated data into machine learning procedures and make it possible to generate large data sets even without explicit knowledge of game engines or video generation.

## Video Processing

As a final step in the process chain, further processing steps are proposed for the videos. These can be used to extend the dataset itself by slight modifications or to derive new metadata for video clips.

Simple steps to modify video data include flipping clips and inserting visual artifacts that simulate anomalies found in real-world recordings. In addition, processes that are relevant to the analysis of activity data can be inserted.

For example, the extraction of bone points in the viewport of the image using OpenPose (Cao et al., 2017) or by approximation in three-dimensional point space using AlphaPose (Fang et al., 2022) is proposed.

## RESULTS

To illustrate the applicability of the workflow, an example dataset has been generated. The simplest configurations are used, as the focus is only on the applicability of the components implemented so far. The last step of video processing has been omitted to focus on the general generation of synthetic video material.

In preparation for dataset generation, a determination was made for various parameters. A total of 13 activities were selected, which exemplify a subset of the Deep Intermodal Video Analytics (DIVA) annotation definitions of the MEVA dataset (Corona et al., 2021). For each activity, three to seven clips were recorded using Rokoko software and incorporated into AAR. Additionally, using the written MakeHuman plugin, ten models were generated that exemplified different stereotypical ethnicities for both male and female characters. In order to fulfill the highest possible level of neutrality for the shots, all models wear completely white clothing and discard various physical traits. At the same time, the scene is provided with a static gray background that creates a high contrast to the characters. Depending on the activity, different objects were also deposited for the interaction of the animation. Basic parameters for recording in AAR were also set for this exemplary dataset. Each animation is captured per character by three newly generated camera hemispheres with radii of 6, 10 and 15 virtual length units. Each radius generates a set of 50 cameras that are evenly distributed over five latitudes. This results in a total number of 1,500 shots for one animation clip of an activity, resulting in a total of 90,000 files for all shots. A maximum size of three parallel camera recordings is used for generating the clips. In addition, all metadata supported by the application is stored for the clips.

The time to generate all animations has been 11,27 hours, which corresponds to an average duration of 0.45 seconds for one animation clip recording with all metadata. The system used in this process uses an i9-12900k Intel processor on a 3.19 GHz clock rate and an NVIDIA RTX 3080 graphics card. Comparatively, the mean time to annotate one single image in MOTS is 0.5 seconds (Voigtlaender et al., 2019). The effort for datasets such as VIRAT (Oh et al., 2011), which were partially or entirely manually annotated, is significantly higher.

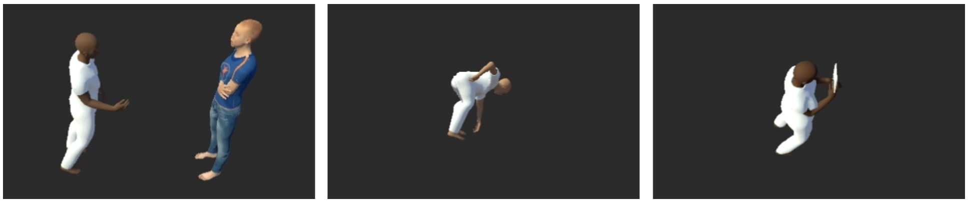

**Figure 4:** Examples of activity viewports recorded by AAR.

The total size of the data set is 13.1 gigabytes. All clips have a resolution of 1600x900 pixels recorded at 30 frames per seconds. The length of a clip was trimmed to a constant 60 frames, which corresponds to a real-time length of two seconds. The videos are organized into individual folders according to their included activities. Figure 4 shows selected viewports recorded for different activities in the dataset. For each animation, a subfolder is created for the associated metadata, which is all in XML data format. The metadata for a recording includes a reference to the clip name and the position and rotation of the associated camera as well as information on the corresponding character. In addition, for each animation, the translation and rotation of the character model's bone points is serialized for each individual frame of a recording run.

**Table 1.** Overview of the data provided by the dataset.

| Type of data | Data included |
|---|---|
| Video clips | 90,000 clips from 13 activity classes |
| Metadata | Various metadata for each recording |
| Bone point information | Temporal change of translation and rotation of all bone points for an animation |

The dataset that is the first prototypical generation by the workflow and from AAR is tagged with the identifier AAR-P01. A summary of all data provided is given in Table 1.

## DISCUSSION

Based on the exemplary generated dataset, it could be shown that the presented proof-of-concept of the workflow can already be used to generate large datasets of synthetic videos. It should be noted that the recordings are based on settings with minimalistic models, animations, and environmental details consisting of primitive geometric objects. By using purposefully produced animations as well as detailed scene objects, it is expected that the synthetic data will appear much more realistic. It also needs to be investigated whether the use of primitive objects has an impact on the quality of trained models.

It can be deduced that the application could be used for different scenarios. Even though the focus for this work was on the representation of person activities, other application areas in the field of machine learning could also benefit from synthetic data of the workflow. For example, it would be conceivable to support object recognition or tracking methods using three-dimensional images from the application. Relevance to multi-camera systems, such as those used to recognize specific people or objects in an environment, is also conceivable.

For the future, several points can also be derived on the basis of which the application AAR can be improved. First and foremost is the enhancement with new user-friendly functions that enable the insertion of further complex animations or entire process chains. The focus is also on extending the variability of the recordings, for example by providing the scene with more options for automated randomization of objects. Last but not least, a practical test of the dataset using a classification procedure is pending. Here, the inference should again be drawn to existing datasets and findings from procedures that use synthetic data, in order to be able to assess the quality of the workflow even better.

## CONCLUSION

In this work, the problem for flexibly generating annotated video data used for training activity recognition models was addressed. For this purpose, a prototypical generic approach for generating large datasets of synthetic videos was presented. By implementing the workflow, the exemplary dataset AAR-P01 mapping 13 activity classes using primitive scene settings was generated, containing a total of 90,000 fully annotated recordings. Further studies could explore the use of the exemplar dataset for machine learning processes as well as extending the application for use in other scenarios.

## REFERENCES

Ahmed, A. A., and Echi, M. (2021). Hawk-eye: An ai-powered threat detector for intelligent surveillance cameras. IEEE Access, 9, 63283-63293.

Ballout, M., Tuqan, M., Asmar, D., Shammas, E., and Sakr, G. (2020). The benefits of synthetic data for action categorization. In 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8.

Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299.

Corona, K., Osterdahl, K., Collins, R., and Hoogs, A. (2021). Meva: A large-scale multiview, multimodal video dataset for activity detection. In Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1060–1068.

Djeffal, C. (2020). Künstliche Intelligenz. In: Klenk, T., Nullmeier, F., Wewer, G. (eds) Handbuch Digitalisierung in Staat und Verwaltung. Springer VS, Wiesbaden.

Elstner, Steffen. Feld, Lars P. and Schmidt, Christoph M. (2016). Bedingt abwehrbereit: Deutschland im digitalen Wandel. Sachverständigenrat zur Begutachtung der Gesamtwirtschaftlichen Entwicklung.

Fang, H. S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., and Lu, C. (2022). Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Fontes, C., Hohma, E., Corrigan, C. C., and Lütge, C. (2022). AI-powered public surveillance systems: why we (might) need them and how we want them. Technology in Society, 71, 102137.

Hwang, H., Jang, C., Park, G., Cho, J., and Kim, I. J. (2021). Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. IEEE Access.

Kim, Y. W., Mishra, S., Jin, S., Panda, R., Kuehne, H., Karlinsky, L., and Feris, R. (2022). How Transferable are Video Representations Based on Synthetic Data?. Advances in Neural Information Processing Systems, 35. pp. 35710–35723.

Nikolenko, S. I. (2021). Synthetic data for deep learning (Vol. 174). Springer Nature.

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C. C., Lee, J. T., and Desai, M. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In CVPR 2011. pp. 3153–3160

Roberto de Souza, C., Gaidon, A., Cabon, Y., and Manuel Lopez, A. (2017). Procedural generation of videos to train deep action recognition networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4757–4767.

Ulrich Lichtenthaler and Springer Fachmedien Wiesbaden Gmbh (2021). Künstliche Intelligenz erfolgreich umsetzen Praxisbeispiele für integrierte Intelligenz. Wiesbaden Springer Fachmedien Wiesbaden Gmbh Springer Gabler.

Varol, G., Laptev, I., Schmid, C., and Zisserman, A. (2021). Synthetic humans for action recognition from unseen viewpoints. International Journal of Computer Vision, 129(7). pp. 2264–2287.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., and Leibe, B. (2019). Mots: Multi-object tracking and segmentation. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 7942–7951.