

Optimal Crop-out for Photographing People During Sporting Activities

Anastasiia Lebedenko*

Abstract

The aim of this paper is to present a program that can process footage of people in motion to generate an output video of optimal dimensions, centered on the human and excluding excessive surroundings.

The solution uses Computer Vision techniques to accurately detect and track human's position in the input video, and then applies a cropping algorithm to generate the output video.

Currently, the program is capable of processing videos of a single person in motion, as well as multiple people in some partner sports. Its command-line interface enables user to customize the aspect ratio, graphic overlay and crop modes of the output video.

Cropping video frames reduces video size without compromising data quality, which is especially useful for training Machine Learning models, where large data sets are used frequently. Overall, the program offers a flexible and customizable solution for processing videos with people in motion, making it a valuable tool for researchers in various fields.

*xlebed11@stud.fit.vutbr.cz, Faculty of Information Technology, Brno University of Technology

1. Introduction

Computational time needed to train ML models is highly dependent on size of input data, with a direct linear relationship between the two [1]. While video data sets are crucial for effective training of models, designed to analyze human pose or movement, the surrounding context in these videos may introduce an unnecessary computational load. Removing the excessive area and centering just on the moving person can significantly reduce the amount of data for processing, leading to more efficient resource usage and faster computation time.

The goal of this work is to develop a tool to process video footage based on the requirements above. The ROI to be cropped out from original video must form an optimal bounding box around a moving person, optionally - multiple people and sporting equipment. Output video should contain optimally cropped frames, image quality of which must not be changed.

The manual approach to the solution is to record videos, zoomed in on the subject, so that little amount of surrounding area is captured. For static frames this can be achieved with a tripod setup, but for dynamic scene such solution is likely to be rather challenging. It

is also not fit for processing already existing footage.

Another manual solution is to use video editing software by marking the area around individuals in motion in every frame. This is also a highly time-consuming and labor-intensive process, which is not prone to human error and may not produce optimal results. On the other hand, such approach may be a better choice for low quality videos, where CV algorithms may fail to detect humans with sufficient confidence level.

State of the art software solution is Apple's Center Stage feature [2], available on certain compatible devices. It uses ML technology to recognise a user and keep them in centre view in real-time during video calls. The downside is that the feature requires a specific ultra-wide camera and cannot be applied on previously shot footage.

Solution, proposed in this paper, is a CLI program that can process multiple videos just by one command with modifiable parameters. The solution does not require user to have special skills or hardware in order to achieve the optimal crop.

2. Proposed Solution

The program is developed using the Python programming language and is divided into two modules: bounding box and crop-out box calculations. The intermediate results are stored in JSON-format and are shared between the modules to dismiss the need to process video more than once.

2.1 Bounding box calculation

The algorithm uses MediaPipe Pose Detection API to detect a maximum of 33 body landmarks. Maximal and minimal values of x and y coordinates are calculated for each set of landmarks. These 4 points define bounding box, as can be seen on figure 2.2 on the poster. Figure 3.1 displays bounding box overlay on the real video frame.

While pose landmarks are redundant to the task, they assist in more precise bounding box definition. Bounding boxes, calculated by other human detection solutions, were unstable from frame to frame and could include excessive area. Moreover, the chosen API succeeds in differentiating the most prominent person on the frame, which is useful for videos, where individual sport is performed with audience in the background. On the other hand, this mechanism is not fit for partner sports, as the set of landmarks will be calculated for just one person.

Detecting multiple people

In case of videos featuring dancing, boxing, and other partner sports, the MediaPipe Object Detection API is used. Detector output includes a name of object category i.e “human” and dimensions of the detected bounding box. The final bounding box is obtained by summing up the bounding boxes of the relevant classes (“human” is default, classes of sporting equipment are optional).

2.2 Crop-out box calculation

As the person moves, their dimensions and the corresponding bounding box can change. The crop-out box, based on which the final video frames are extracted, should have a stable size. The largest width and height value among all bounding boxes from the previous step 2.1 define size of an output video, in regards to chosen aspect ratio.

As per figure 2.4 on the poster, the crop-out box coordinates are calculated such that the bounding box is centered inside it. Figure 3.2 displays cut-out box overlay on the real video frame.

Edge cases

For frames like on the Figure 2.5a, when person is moving towards the edge of the frame, bounding box centering results in crop-out box values exceeding the dimensions of input video. In this case, centering constraint is not included in the calculations.

For frames to be cropped and extracted from input video, each frame has to have crop-out box values defined. As shown on Figure 2.5, if frame $n + 1$ is missing crop-out box coordinates, these values are iteratively propagated from frame n and vice versa. Same mechanism is applied in case bounding box is defined, but its dimensions differ more than by 30% from those of previous frame, which can be a sign of fault detection.

Stabilization

To ensure stable output video, crop out points' values are filtered using Savitzky-Golay filter from SciPy library. An array of each point's coordinate values in each frame is processed by `savgol_filter()` function. Figure 2.6 demonstrates the filtering effect.

2.3 Crop modes

Crop modes are special crop settings implemented in the program, appropriate for specific type of movements in the input footage. Yoga videos, where person remains on the same place, could be cropped by **fixed frame**, which signifies the border, inside which all action is happening. For movements, that are primarily up and down, or left to right, **one-direction** cropping mode eliminates frame fluctuations in the secondary direction. Default **two-direction** mode can be combined with **zoom** option: in footage, where person distances from the camera, such mode zooms the frame in and out, so that person appears to be in the same distance.

Acknowledgements

I would like to thank my supervisor prof. Ing. Adam Herout Ph.D. for his help and ideas.

References

- [1] Daniel Justus, John Brennan, Stephen Bonner, and Andrew Stephen McGough. Predicting the computational cost of deep learning models. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3873–3882, 2018.
- [2] Apple Inc. Use center stage on your ipad or studio display. <https://support.apple.com/en-us/HT212315>, 20.04.2023.