H2V4Sports: Real-Time Horizontal-to-Vertical Video Converter for Sports Lives via Fast Object Detection and Tracking

Yi Han
YANGSHIPIN Co., Ltd.
hanyi@yangshipin.cn

Kaidong Li
YANGSHIPIN Co., Ltd.
likaidong@yangshipin.cn

Zihan Song
Department of Computer Science and Technology, Tsinghua University
songzhan19@gmail.com

Wei Feng
Department of Computer Science and Technology, Tsinghua University
fw22@mails.tsinghua.edu.cn

Xiang Cao
YANGSHIPIN Co., Ltd.
caoxiang@yangshipin.cn

Shida Guo
YANGSHIPIN Co., Ltd.
guoshida@yangshipin.cn

Xin Wang
Department of Computer Science and Technology, BNRist, Tsinghua University
xin_wang@tsinghua.edu.cn

Xuguang Duan
Department of Computer Science and Technology, Tsinghua University
dxg18@mails.tsinghua.edu.cn

Wenwu Zhu
Department of Computer Science and Technology, BNRist, Tsinghua University
wwzhu@tsinghua.edu.cn

ABSTRACT

We present H2V4Sports, a real-time horizontal-to-vertical video converter specifically designed for sports live broadcasts. With the increasing demand of smartphone users who prefer to watch sports events on their vertical screens anywhere, anytime, our platform provides a seamless viewing experience. We achieve this by fine-tuning and pruning an object detector and tracker, which enables us to provide real-time, accurate key-object tracking results despite the complexity of sports scenes. Additionally, we propose a video virtual director platform that captures the most informative vertical zones from horizontal video live frames using various director logic for a smooth frame-to-frame transition. We have successfully demonstrated our platform in two popular sports: basketball and diving, and the results indicate that our technology delivers high-quality vertical scenes that are beneficial for smartphone users and other vertical scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; **Tracking**; *Computer vision*; *Interest point and salient region detections.*

KEYWORDS

Real-Time Video Editing; object detection; broadcasting platform

ACM Reference Format:
Yi Han, Kaidong Li, Zihan Song, Wei Feng, Xiang Cao, Shida Guo, Xin Wang, Xuguang Duan, and Wenwu Zhu. 2023. H2V4Sports: Real-Time Horizontal-to-Vertical Video Converter for Sports Lives via Fast Object Detection and Tracking . In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3581783.3612669

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612669

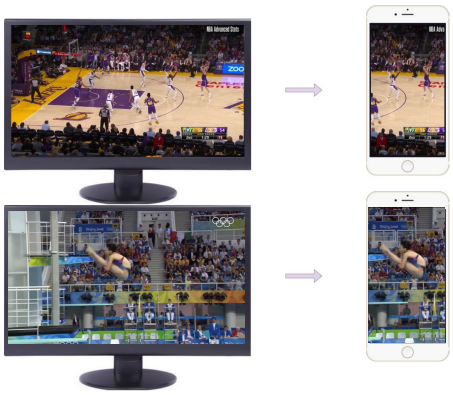

Figure 1: Examples of video H2V for basketball and diving

1 INTRODUCTION

With the increasing number of smartphone users, there has been a growing trend of people watching videos on their vertical screens. This is especially true for sports fans who want to watch their favorite competitions live on their smartphones immediately. However, existing solutions for video conversion are mainly focused on converting static videos into vertical ones, which do not cater to the real-time and sports-specific needs of viewers. To address the aforementioned issues, we propose the H2V4Sports platform, which is designed specifically for sports live broadcasts and provides real-time video conversion as well as an easy-to-use terminal interface and web interface for video platforms. We apply real-time object detection model YOLO, as well as object tracking strategies such as Kalman filter and single object tracking model Mixformer[2], to locate the key parts in the horizontal video frame to generate the vertical frame accordingly. Figure 1 illustrates the functionality of our H2V4Sports system.

2 SYSTEM ARCHITECTURE

The core architecture of the proposed H2V4Sports system is shown in Figure 2. The system utilizes an object detector and an object tracker to locate key objects in horizontal video frames. Then the system performs object matching and smoothing on adjacent frames to smooth the key-objects and filter out noisy detections. Finally, a director logic is designed to generate vertical shots from the location of key objects, from which the vertical video is rendered out.

2.1 Object Detecting and Tracking

The core of video H2V lies in locating the key areas, which are determined by the key objects in sports videos, such as basketballs in basketball games and divers in diving competitions. Therefore, we use the classic one-stage object detector YOLOv3[4] to detect the key objects, such as basketballs and divers, in the video. However, sports videos are of complex scenes, and situations such as the occlusion of the basketball by players and divers entering the water can cause the detector to fail. In addition, irrelevant objects in the scene can also cause confusion of the object detector, leading to false detection. Therefore, we utilize an object tracker Mixformer[2], which introduces attention mechanisms to search for the area that best matches the target area in the following frame to achieve tracking. By combining detection and tracking, we can accurately and stably locate the key objects in the video.

Key objects in different sports games have their own unique features. Therefore, to locate the key objects more accurately, we build a small sport-specific dataset, which contains about 20 10s-long videos that were annotated frame by frame with the bounding boxes of basketballs and divers, respectively, and fine-tuned the YOLOv3 model on the datasets by keeping the other parts of the model unchanged and modifying the classifcation head to output only one single category. Testing results indicate that the fine-tuned detection model significantly improved the accuracy of locating key objects.

2.2 Object Matching and Frame Smoothing

In sports scenes, there are usually multiple key objects, so H2V4Sports system is not just about locating the targets, but also about matching different key objects between frames. We used the classic multi-object tracking algorithm SORT[1] based on Kalman filter and Hungarian algorithm to track and match multiple targets between frames. As for frame smoothing, we utilize the following momentum-based camera smoothing method:

$$c_t = \beta_t c_{t-1} + (1 - \beta_t) d_t, \tag{1}$$

where c_t represents the center of the vertical shot for frame t, and d_t stands for the center of the key object in frame t. β_t is a smoothing parameter which can be calculated as follows:

$$\beta_t = 1 - \frac{2\,|d_t - c_{t-1}|}{w}, \tag{2}$$

where w is the width of the vertical shot. This shot movement method ensures that the target remains close to the center of the shot without the shot shaking severely.

2.3 The Director Logic

The role of the director is to generate corresponding vertical shots based on the position of the key objects given by the algorithm above. The director logic we use is as follows: a) If no object is detected, an empty shot will be generated in the center area of the horizontal video. b) If a single object is detected, a single shot will be generated with that object at the center. c) If there are multiple key objects, several largest targets will be selected according to the area of their bounding boxes and corresponding single shots are generated. Then, based on the shape of the objects, they are spliced horizontally or vertically to form a multi-shot.

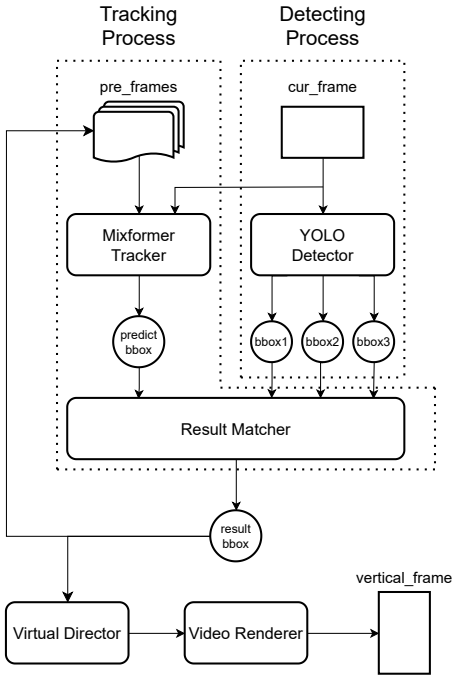

Figure 2: H2V4Sports system architecture

2.4 Real-time System

In order to meet the real-time requirements of live broadcasting in sports events, we have accelerated our system by using a Ray-based[3] multiprocess framework to perform parallel inferences on the detection model and tracking model in different processes. According to our test, our system achieves a processing speed of 24fps on NVIDIA GeForce RTX 3090, which is sufficient to be used for inference on live broadcasting.

3 CONCLUSION

We design and present a real-time system, H2V4Sports, to transform two typical kinds of live broadcasting sports, basketball and diving, from horizontal video to vertical video. Test results show that our system delivers high-quality vertical videos in real-time, which basely meets the requirement of smartphone users to watch sports lives on their vertical screens conveniently.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China No. 2020AAA0106300, NSFC (No. 62250008, 62222209, 62102222), BNRist Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

REFERENCES

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. 3464–3468. https://doi.org/10.1109/ICIP.2016.7533003

[2] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13608–13618.

[3] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. 2018. Ray: A distributed framework for emerging {AI} applications. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*. 561–577.

[4] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).