AutoShot: A Short Video Dataset and State-of-the-Art Shot Boundary Detection

Wentao Zhu[1], Xiufeng Xie[1], Wenxian Liu[1],
Jincan Deng[1], Debing Zhang[1], Zhangyang Wang[2], Ji Liu[1]
[1]Kuaishou Technology [2]University of Texas at Austin

1. SHOT Dataset

To accelerate the study of short video related shot boundary detection, we collect 853 short videos from one of the most widely used short video platforms. The dataset property comparison is listed in Table 1. The total number of frames is 960,794, close to one million frames. The frame-wise shot boundary annotation is a heavy task.

Inspired by the frame processing in TransNetV2 [1], we develop a video thumbnail image based annotation by resizing each frame to be 48×27 as shown in Fig. 1. The frame number is adaptively displayed in the upper left corner of each frame, which significantly reduces the annotator's efforts for frame number check. If the pixel value is dark in the frame number position, we display the frame number in light color. Otherwise, we display the frame number in dark color. We have three experts and an annotation team to complete all the annotation of 960,794 frames. The annotation team conducts the annotation for 459 short videos and obtains 6,111 shots totally. After the annotation, we conduct an expert inspection for the 6,111 annotations. We randomly choose 200 shot annotations, and find that 2% error rate for the 6,111 shot annotations exists. Considering the ambiguity of the shot definition and the annotation difference of annotators, we believe that the 2% error rate of the 6,111 shots is acceptable.

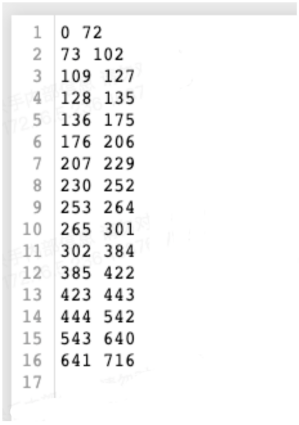
The quality of annotations on the test set directly affects the accurate evaluation of benchmark methods, *i.e.*, the quality of the short video dataset. To guarantee the high quality of the shot boundary annotations on the test set, we randomly choose 200 short videos from the rest of 394 short videos annotated by the three experts as the test set. The expert team is the production team who defines the shot transitions and writes the annotation guideline. Then the guideline and data are sent to contractors for the annotation using the thumbnail images, and QA auditing is conducted by the production team. Because of the importance of the test set, it is labeled by the production team completely. For the 200 test short videos, we employ two round annotations, where the first round yields 2,616 shots. In the second round, we conduct a rigorous check for the annotations. In addition to fixing a handful of false positive annotations, which cannot be corrected by our manually designed rules for the automated inspection, the second round totally yields 2,716 shots, *i.e.*, re-collecting 100 shots from false negatives.

Figure 1. A thumbnail image for one case from our collected SHOT dataset. The frame number is displayed in the upper left corner of each frame.

Dataset	BBC	RAI	ClipShots	SHOT
Complex Gradual Trans.	✗	✗	✗	✓
Virtual Scene	✗	✗	✗	✓
Ternary Video	✗	✗	✗	✓
Avg. Video Len. (s)	2944.8	590.5	236.5	39.5
Avg. Shot Len. (s)	6.57	5.65	15.34	2.59

Table 1. Comparison of different short boundary detection datasets.

In the dataset, for each short video, we have a text file with the same name as the video consisting of the annotations of the shot boundaries, *i.e.*, the beginning and the end frame number of each shot per line, as shown in the following Fig. 2. If there is no gradual transition, as in most shots, the beginning frame number of the next shot should be the next frame number of the end frame number of the current shot. For instance, the end frame number of the first shot is 72, and the beginning frame number of the second shot is 73. If there is a gradual transition, the beginning frame number is a clear transition, and the gap between the end frame number of the last shot and the beginning frame number of the current shot is the gradual transition. For in-



```
1  0 72
2  73 102
3  109 127
4  128 135
5  136 175
6  176 206
7  207 229
8  230 252
9  253 264
10 265 301
11 302 384
12 385 422
13 423 443
14 444 542
15 543 640
16 641 716
17
```

Figure 2. The shot boundary annotation text file. Each line denotes the beginning and the end frame number of one shot.

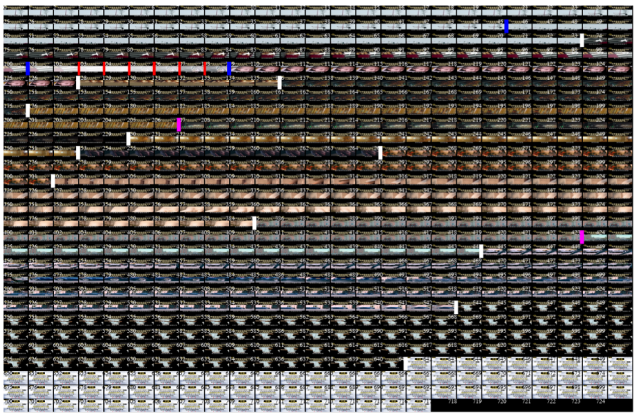


Figure 3. A thumbnail image for the same video as Figure 2. Light, red and pink colors mark the ground truth shot boundary annotation. The yellow circle denotes one gradual transition from frame 102 to frame 108.

stance, the shot in line 2 ends with frame number 102 as shown in Fig. 2, and the shot in line 3 begins with frame number 109. The frames between 102 and 108 are gradual transitions, as shown in the yellow circle of Fig. 3. Fig. 3 is a thumbnail image for one whole video same as Fig. 2, and the light, red and pink colors mark the ground truth shot boundary annotation.

To evaluate the shot boundary detection fairly, we follow the evaluation of ClipShots dataset in [1]. If the detected end frame number is within two frames of the annotation, the detected boundary is a true positive. For a gradual transition, if the detected end frame number is within the gradual transition, it is also a true positive. The metric evaluation script can be found in https://github.com/soCzech/TransNetV2/blob/master/training/metrics_utils.py#L26.

2. More Visual Comparison Results

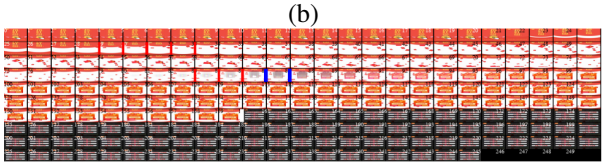
We add more figures in the supplementary to compare the prediction of AutoShot (blue, light, pink, cyan), TransNetV2 [1] (green, yellow, light, cyan) and ground truth (red, light, pink, yellow) in Fig. 4. The light color shows that both TransNetV2 and AutoShot have true positive on that frame, as shown in the video thumbnail image of Fig. 4 (a). The blue cuts in the Fig. 4 (b) show AutoShot successfully detects (within two frames of ground truth) minor shot transitions in the middle of the video. Fig. 4 (c) is a vertically ternary structured video, where the scene changes slightly and the minor shot transition is quite difficult to detect. The pink cut in Fig. 4 (d) shows AutoShot detects the shot transition, while the TransNetV2 fails. The blue color in frame one of Fig. 4 (d) shows AutoShot has a false positive detection, and it is visually reasonable that there is a gradual transition at the beginning of the video.

References

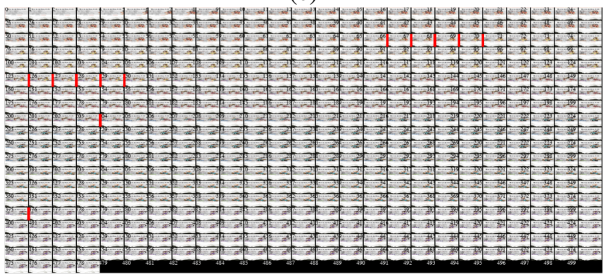
[1] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 1, 2

(a)

(b)



(c)



(d)

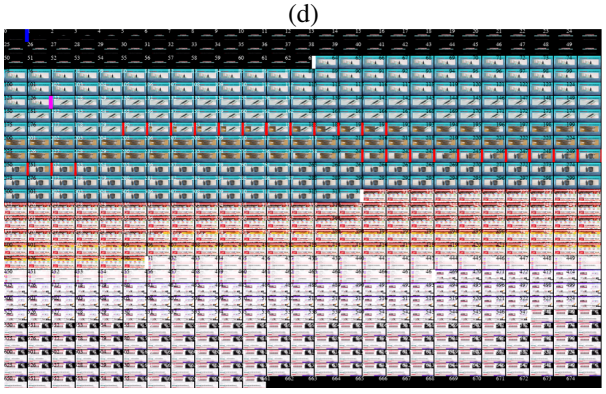


Figure 4. Visual comparison of shot boundaries of ground truth (red, light, pink, yellow), TransNetV2 (green, yellow, light, cyan), and AutoShot (blue, pink, light, cyan) on the collected SHOT dataset. AutoShot detects minor transition, shown in the pink color. The light color denotes that both the TransNetV2 and AutoShot successfully detects the shot boundary.