Human-centric saliency detection in 360° multimedia

Yi Zhang Deng-Ping Fan*

Abstract

Recent decades have witnessed a prosperous development of two-dimensional (2-D) saliency detection benchmarks. However, the lack of a well-established dataset representative of dynamic scenes with audio-visual information and high-quality annotations hinder the development of methods for saliency detection in 360° multimedia. In this work, we propose ***HCSDAV360****, a new benchmark dataset for* ***h****uman-****c****entric* ***s****aliency* ***d****etection in* ***a****udio-****v****isual* ***360****°. The dataset contains 69 4K-resolution 360° videos (59,853 frames), representing 65 backgrounds over five broad categories of human-centric daily scenes. Fixations per-frame, acquired by conducting an eye-tracking experiment upon all collected videos, are regarded as the guidance for salient human instance annotation. Within 10,033 uniformly sampled key frames, 22,225 salient human instances are pixel-wisely annotated at both the object-/instance-level, 8,584 sounding instances are also labeled with spherical bounding boxes. To further contribute the community a complete benchmark, (?) state-of-the-art (SOTA) saliency detection algorithms are fine-tuned and evaluated upon the proposed dataset, a comprehensive analysis is conducted based on both the benchmarking results and dataset attributes. In addition, we introduce a new 360° saliency detection method,* ***g****lobal-aware* ***t****angent image-based* ***n****etwork (****GTNet****), which outperforms all the benchmark models on the proposed dataset, thus providing insights for applying SOTA planar-based optimizations to spherical manifolds. The dataset and code will be made publicly available.*

1. Introduction

(Recent video salient object detection (VSOD) methods aim at segmenting salient object regions from two-dimensional (2-D) data and achieve good results on public benchmarks. However, when used for real-world multimedia applications, e.g. virtual reality (VR, or 360°) video segmentation, which requires the capability of learning features on dynamic sphere surface, a significant decline of their performances is initially discovered. Meanwhile,) (, including conversation, monologue, singing, instrument performance and miscellanea.) (With the various natural scenes and abundant annotations, the **HCSDAV360** could provide supports for many vision tasks, e:g:, video salient instance segmentation, fixation prediction, audio-visual learning and geometric learning, etc.)

2. Related Work

3. Proposed Dataset

3.1. Video Collection

69 360° videos with ambisonic sound captured at 4K.

3.2. Eye-tracking Experiments

Fixations
Scanpath

3.3. Data Annotation

Scene Category Labeling. Five categories based on the attributes of sounds.
Instance-/Object-Level Salient Persons Annotation. 10K key frames.
Sounding Object Labeling. All the sounding objects in the 10K key frames are labeled with bounding boxes.

3.4. Dataset Features and Statisitics.

(Special challenges for SOD in 360 compared with 2D's, reflected by examples from proposed dataset.)

3.5. Dataset Splits

4. Proposed Framework

5. Benchmark Experiments

5.1. Experimental Settings

Evaluation Metrics. (applying 2D SOD metrics on the equirectangular mesh.)
Benchmark Models and Protocols.

5.2. Performance Comparison

Performance of 2D video SOD Models.
Performance of proposed GTNet.
Runtime Analysis. (FPS)

[*] Corresponding Author: Deng-ping Fan

5.3. Ablation Study

(FCN-ResNet101 with ER / TI, GTNet)

6. Potential Applications

Human body part segmentation
Human Pose Estimation
Audio-visual Saliency prediction.
Ambisonic Sound Separation.
Sounding Object Localization.
Multi-modal Video Object Tracking and Segmentation.

7. Conclusion

8. Supplementary Materials

References