Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

V-SENSE

TOWARDS GENERATING AMBISONICS USING AUDIO-VISUAL CUE FOR VIRTUAL REALITY

Aakanksha Rana*, Cagri Ozcinar*, Aljosa Smolic

Problem And Objective

- Automatic spatial audio estimation based on audio-visual cue.
- 360 Audio-Visual Dataset (360AVD) which contains 265 video clips with a well-annotated ground-truth providing the sound direction and location.
- Propose evaluation criteria: 360 SSD and 360 OvErr.

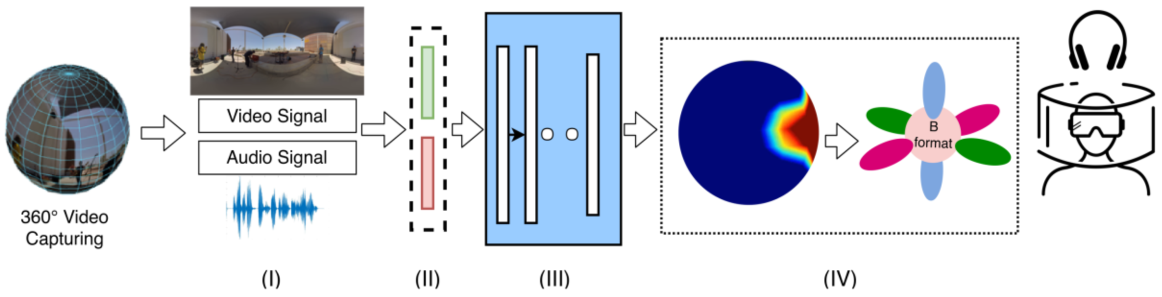
360 Video

Cubemap Projection

Equirectangular Projection

Proposed System

- Stage I: Representation, where audio and visual signals are pre-processed.
  Visual Signal is transformed into equirectangular or cubical format.
- Stage II: Feature Embedding, where we used :
  - VGG-19 network to compute feature maps from 15 frames and average them to obtain one feature map.
  - Extract the 128-dimensional audio representation, using a pre-trained VGGish network.
- Stage III: Prediction Module, to predict the 3D volumetric maps.
  - SsM Module [1]--- 3 conv layers $S_p^{SsM} = f(\sigma(\mathcal{L}^T conv_l)),$
  - ATT Module [2]--- Uses attention module
    $$S_p^{Att} = f(softmax(\omega \cdot \rho(l_v) + l_a)),$$
- Stage IV: Ambisonics Encoding, (B format).



Prediction Pipeline.

Metrics

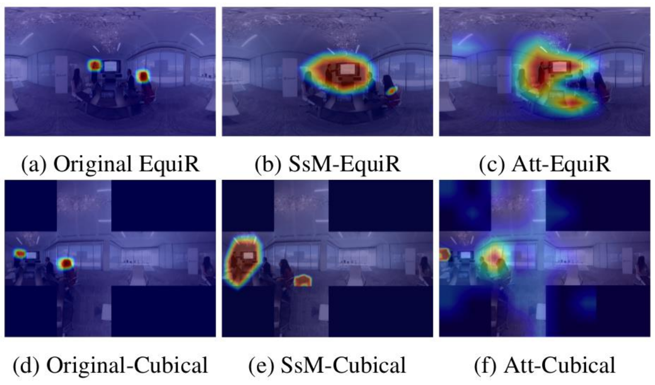
- 360- SSD: Euclidean distance between the centre of the predicted i-th sound source, and the centre of ground truth i-th sound source.
  All distances are normalized, and the probability spheres have radius 0.5
- 360- OvErr: Ratio of an intersection of the predicted and ground truth probability volumes to the union.

Evaluations

- 265 Omnidirectional videos.

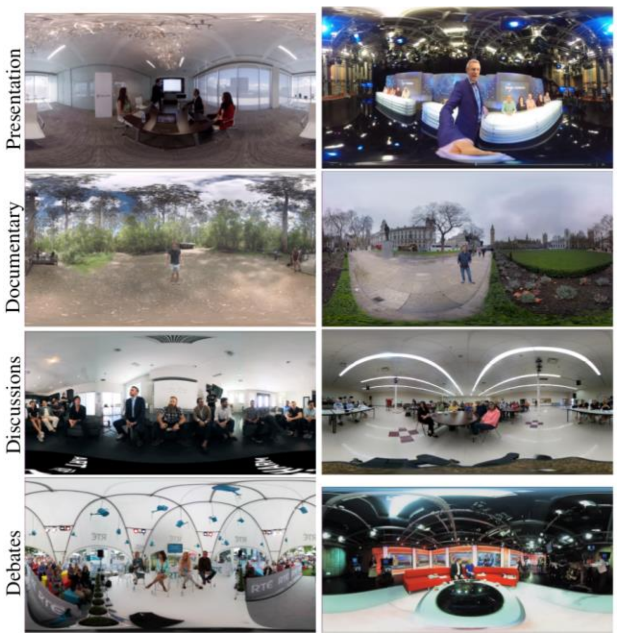
Models	360-SSD			360-OvErr		
	ϵ=0.6	0.5	0.4	0.6	0.5	0.4
SsM-Cubical	**0.71 ± 0.04**	0.72 ± 0.08	0.74 ± 0.06	**0.71 ± 0.06**	0.77 ± 0.05	0.82 ± 0.04
SsM-EquiR	0.75 ± 0.06	0.77 ± 0.09	0.79 ± 0.07	0.78 ± 0.07	0.84 ± 0.06	0.88 ± 0.08
Att-Cubical	0.72 ± 0.05	0.73 ± 0.05	0.74 ± 0.04	0.72 ± 0.05	0.74 ± 0.08	0.78 ± 0.08
Att-EquiR	0.76 ± 0.04	0.77 ± 0.08	0.78 ± 0.06	0.84 ± 0.06	0.85 ± 0.06	0.86 ± 0.06

Quantitative Results on 360AVD Dataset. The scores are averaged on 265 ODVs for all models.


(a) Original EquiR (b) SsM-EquiR (c) Att-EquiR
(d) Original-Cubical (e) SsM-Cubical (f) Att-Cubical

Dataset

- 265 Omnidirectional Video clips.
- Annotated sound source and direction.
- Each clip is 10 secs.
- Categories: presentation, documentary, debates and casual discussions.
- Data : https://github.com/V-Sense/360AudioVisual



REFERENCES

[1] A. Owens et al., “Audio-visual scene analysis with self-supervised multisensory features,”ECCV, 2018.

[2] T. Yapeng et. al., “Audio-visual event localization in unconstrained videos,” ECCV, 2018

2019 IEEE International Conference on Acoustics, Speech and Signal Proccessing

Contact: ranaaa@tcd.ie

v-sense.scss.tcd.ie