PLUMCOT at TRECVid Instance Search 2018

Contributors from LIMSI: Hervé Bredin, Benjamin Maurice, Ruiqing Yin, Camille Guinaudeau, Aman Behre, Claude Barras
Contributors from KIT: Vivek Sharma, Saquib Sarfraz

Briefly, list all the different sources of training data used in the creation of your system and its components.

Neural face embedding was trained with VGGFace2 dataset.
Neural place embedding was trained with Places365 dataset.
Neural speaker embedding was trained with VoxCeleb1 dataset.
SyncNet talking-detection was trained with VoxCeleb2 dataset.

Briefly, what approach or combination of approaches did you test in each of your submitted runs? (please use the run id from the overall results table NIST returns)

Run 1

For places, we computed “place feature” distances to the query. We used pyannote.video toolkit to perform temporal segmentation into scenes. The score of each is updated to be the best score with the scene it belongs to.

For fusion, we first remove all shots whose “place” score is below a threshold. Then, fusion is done by combining shot ranks in both (face and place) lists: 0.1 x face_rank + 0.9 x place_rank.

Run 2

Same as *Run 1* except we do not rely on temporal segmentation into scenes.

Run 3

We ran *syncnet* talking-face detection module and used the N best talking-faces for each query to train a speaker embedding for each query (using *pyannote.audio* toolkit).

Then, same as *Run 1* except face ranking is obtained by combining ranking of *Run 1* and ranking obtained by speaker embedding: 0.8 x face_rank + 0.2 x speaker_rank.

What if any significant differences (in terms of what measures) did you find among the runs?

Run 1 (0.230) >> Run 2 (0.096) shows that temporal segmentation into scenes helps a lot to improve place recognition.

Run 3 (0.221) < Run 1 (0.230) does not allow us to really conclude anything.

Since we were not aware of the existence of annotated development data, we could not really tune our face + speaker fusion.

Based on the results, can you estimate the relative contribution of each component of your system/approach to its effectiveness?

No. The fact that we are not provided with separate scores for the face recognition task and the place recognition task makes it very difficult to do.

Overall, what did you learn about runs/approaches and the research question(s) that motivated them?

Not much actually. We are not really interested in place recognition, but mostly in audiovisual person recognition (from face **and** voice). However, we could not really learn anything from this challenge with respect to this research question.