City-Identification of Flickr videos using semantic acoustic features

Benjamin Elizalde - Carnegie Mellon University

Outline

1. Task
2. Approach
3. Experiments
4. Results
5. Conclusion

City-identification of videos

- Aims to determine the likelihood of a video belonging to a set of cities.
- Our approach focuses only on the audio track.

Outline

1. Task
2. Approach
3. Experiments
4. Results
5. Conclusion

Approach to City-identification of videos

- Expresses the relationship between a taxonomy of urban sounds and the city-soundtracks.
- Computes and used semantic acoustic features to show evidence of the relationship.
- Contrasts to only using frequency analysis of the city-soundtrack.

Our sounds and cities

- The 10 urban sounds:
  - air conditioner, car horn, children playing, dog bark, engine idling, gun-shot, jackhammer, siren, drilling, and street music.

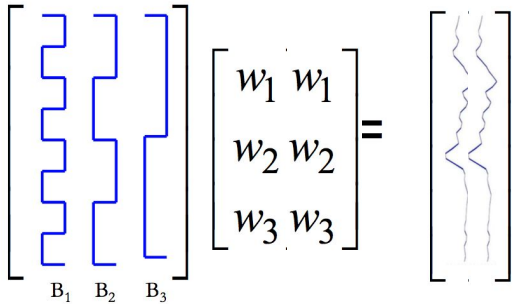
- The 18 cities consists of :
  - Bangkok, Barcelona, Beijing, Berlin, Chicago, Houston, London, Los Angeles, Moscow, New York, Paris, Prague, Rio, Rome, San Francisco, Seoul, Sydney, Tokyo.

A combination of sounds to approximate the city-soundtrack

$$\widehat{Signal} \approx Bases \times Weights = B_1W_1 + B_2W_2 + ... + B_nW_n$$

$$City - \widehat{soundtrack} \approx carW_1 + sirenW_2 + ... + drillingW_{10}$$

$$Weights = pinv(Bases) \times Signal$$



A combination of sounds to approximate the city-soundtrack

$$\widehat{Signal} \approx Bases \times Weights = B_1W_1 + B_2W_2 + ... + B_nW_n$$

$$City - \widehat{soundtrack} \approx carW_1 + sirenW_2 + ... + drillingW_{10}$$

- The linear combination and the weight matrix can be used as the acoustic features.

A combination of sounds to approximate the city-soundtrack

$$\widehat{Signal} \approx Bases \times Weights = B_1W_1 + B_2W_2 + ... + B_nW_n$$

$$City - \widehat{soundtrack} \approx carW_1 + sirenW_2 + ... + drillingW_{10}$$

- The linear combination and the weight matrix can be used as the acoustic features.
- The weight matrix carries the semantic evidence, indicating the presence of a given sound in a city-soundtrack.

A combination of sounds to approximate the city soundtrack

$$\widehat{Signal} \approx Bases \times Weights = B_1W_1 + B_2W_2 + ... + B_nW_n$$

$$City - \widehat{soundtrack} \approx carW_1 + sirenW_2 + ... + drillingW_{10}$$

- The linear combination and the weight matrix can be used as the acoustic features.
- The weight matrix carries the semantic evidence, indicating the presence of a given sound in a city-soundtrack.
- Successful examples of sound retrieval were achieved using the weight matrix i.e. sirens in a Berlin video.

Outline

1. Task
2. Approach
3. Experiments
4. Results
5. Conclusion

End-to-end pipeline for city-identification

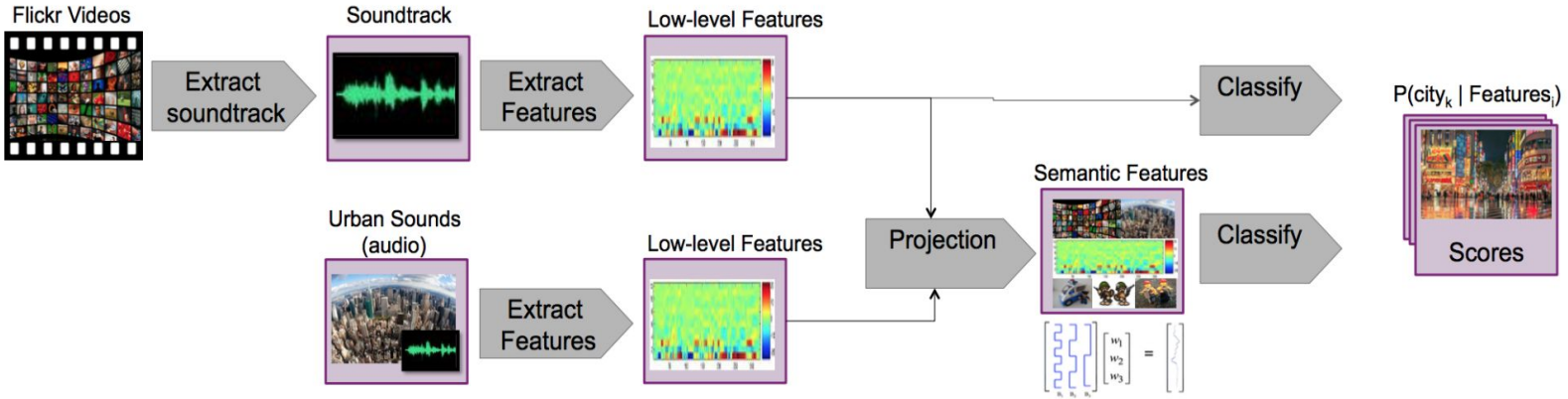


Fig. 1. On the top row the system uses low-level features. The soundtrack is extracted from the videos, then low-level features are computed and a classifier was trained and used to identify the videos. On the lower row, the system uses the semantic features. We computed low-level features of the urban sounds and together with the low-level features of the city soundtracks we computed the semantic features. These semantic features were used to train a classifier and perform city identification on the videos.

Outline

1. Task
2. Approach
3. Experiments
4. Results
5. Conclusion

Our approach outperforms the state-of-the-art

Fig. 2. The MLP outperformed the RF, which is suggested by a better use of the temporal information in the features. For the low-level features labeled as *Statistical Features*, the EER is 29.5% for the RF and 24.3% for the MLP. For the semantic features we have the *Weights Matrix* type with EER 29.3% and 24.6%; and the *Linear Combination* type with EER of 25.6% and 24.2%.

**Statistical Features are statistics derived from MFCCs, such as mean, variance, kurtosis, etc.*

More bases help and extend the semantic evidence

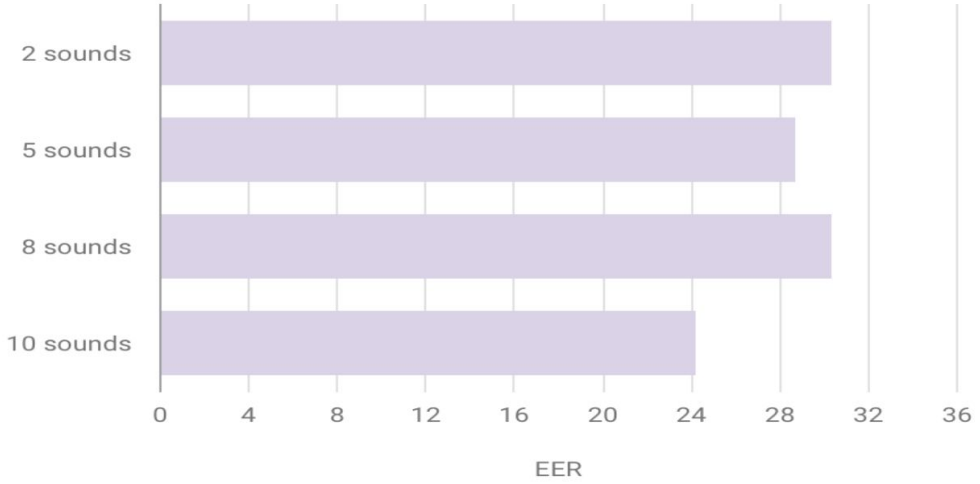
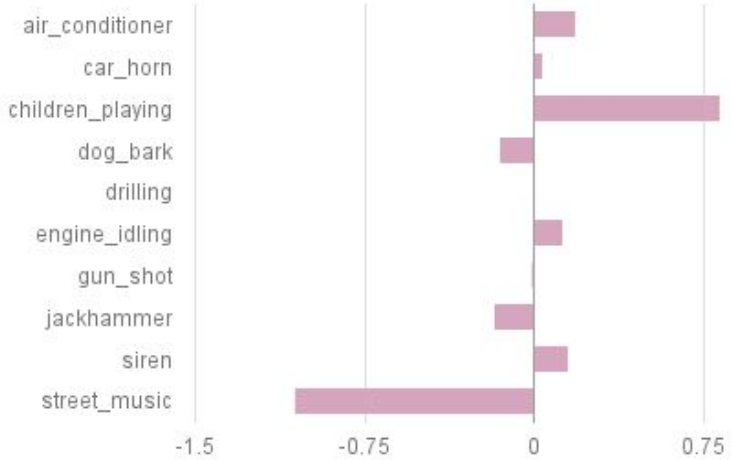


Fig. 4. The average EER for city-identification improves with the numbers of sound bases. Experiment suggests that more bases could improve city-identification as well as extending the semantic evidence.

Retrieval result: children playing and siren in Rome



Outline

1. Task
2. Approach
3. Experiments
4. Results
5. Conclusion

Audio can help city-identification of videos

1. City soundscapes contain information that aids its identification and geolocation.

2. Our method not only aids city-identification but also provides evidence.

3. More bases/sounds could improve our results and extend our evidence.


Q&A
bmartin1@andrew.cmu.edu