Bag of Genres for Video Retrieval

Leonardo A. Duarte[1], Otávio A. B. Penatti[2], and Jurandy Almeida[1]

[1]Institute of Science and Technology, Federal University of São Paulo – UNIFESP

[2]Advanced Technologies, SAMSUNG Research Institute

Introduction

- Huge number of multimidia contents, like **videos**;
- YouTube receives arround **300 hours** of video per minute;
- Video **retrieval** is a challenging task;
- A single video may contain multiple categories (**genres**).

Literature

- There are two approaches to represent video content:
  1) **Spatio-temporal** methods [1];
  2) **Spatial-only** methods [2].
- A popular approach for spatial methods is **visual dictionaries**.

Bag of Genres

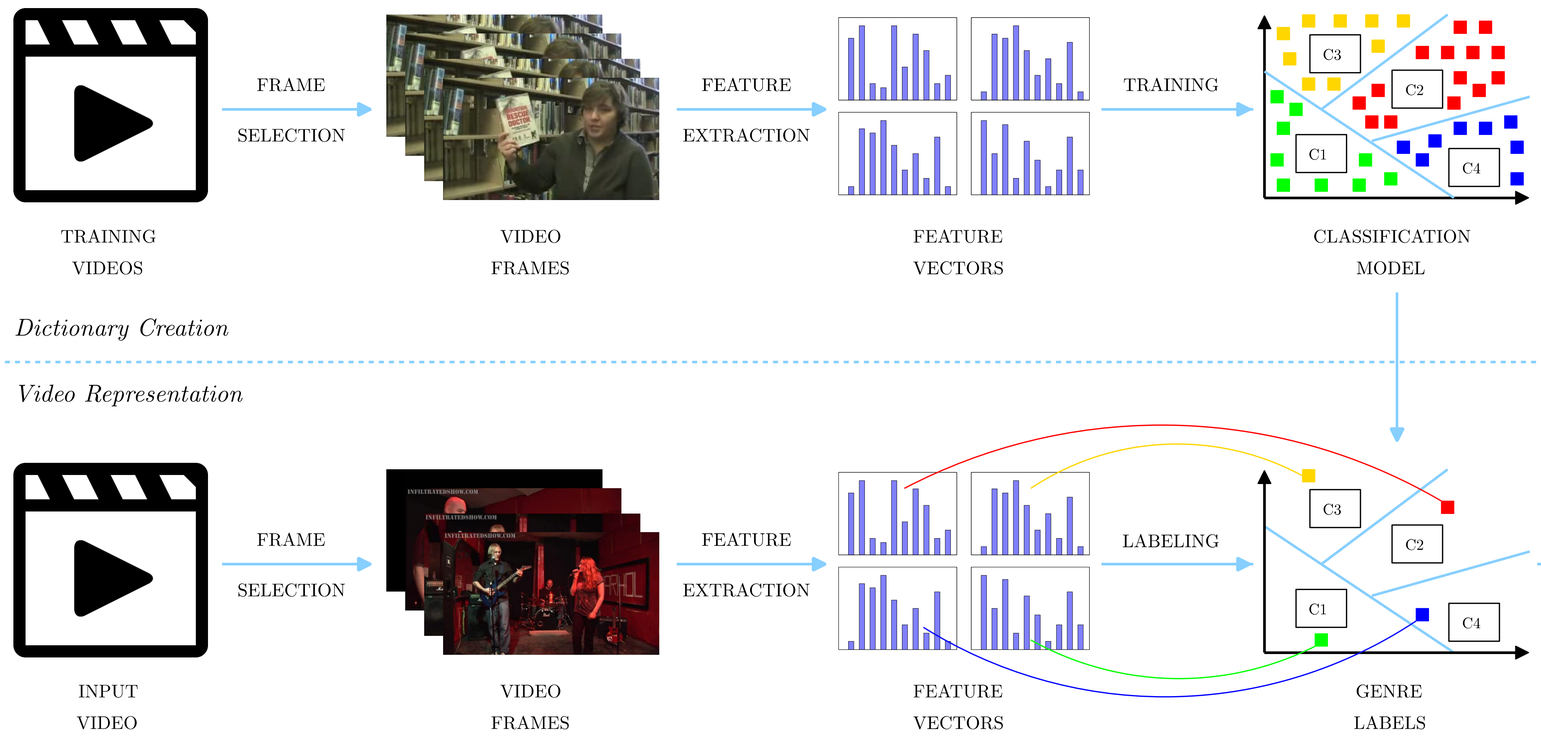

Figure 1. An overview of Bag-of-Genres model.

- **New**: visual words are **labeled regions** determined by a **classification model**;
- **Contribution**: **each dimension** of the feature space spanned by the model is associated to a **semantic concept**.

The diagram above is divided in two phases:

1) Dictionary creation;
2) Video representation.

Experimental Protocol

VIDEO GENRE RETRIEVAL

- **Genre Tagging Task** at MediaEval 2012 [3]:
  - **14,838 videos** (3.288 hours) collected from Blip.tv;
  - 5.288 videos (3.943.375 frames) for **training (36%)**;
  - 9.550 videos (7.273.996 frames) for **testing (74%)**;
  - **26 different genres** assigneded by Blip.tv;
- **7 content descriptors** of separated frames [4];
- **Balanced training** with **800 frames** of each genre;
- **200 queries** (5% of the dataset size) replicated **5 times**;
- Retrieval effectiveness **P10** and **MAP**.

VIDEO EVENT RETRIEVAL

- Event Video (**EVVE**) dataset [5]:
- **2,995 videos** (166 hours) collected from YouTube;
- **13 events** (categories);
- **620** (20%) of **query videos**;
- **2,375** (80%) of **reference vídeos**;
- Retrieval effectiveness **MAP**.

Experimental Results – Genre Retrieval

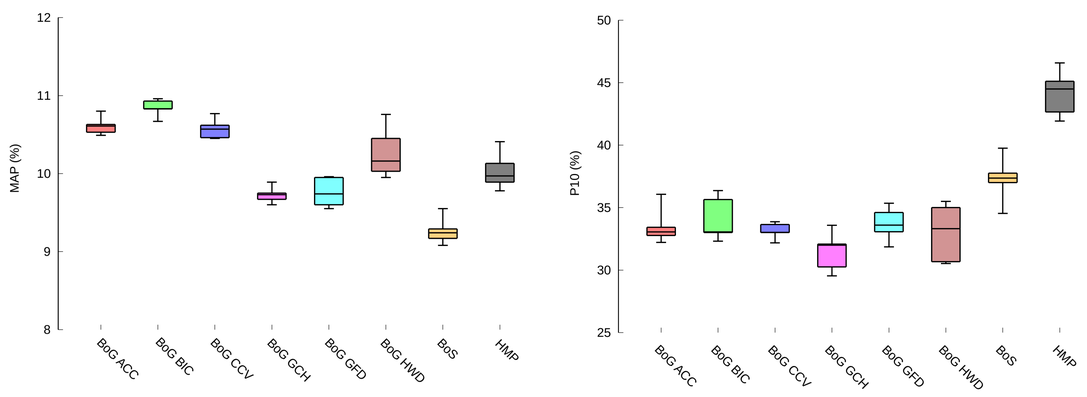

Figure 2. Results for video genre retrieval comparing BoG with the baselines in terms of MAP and P10.

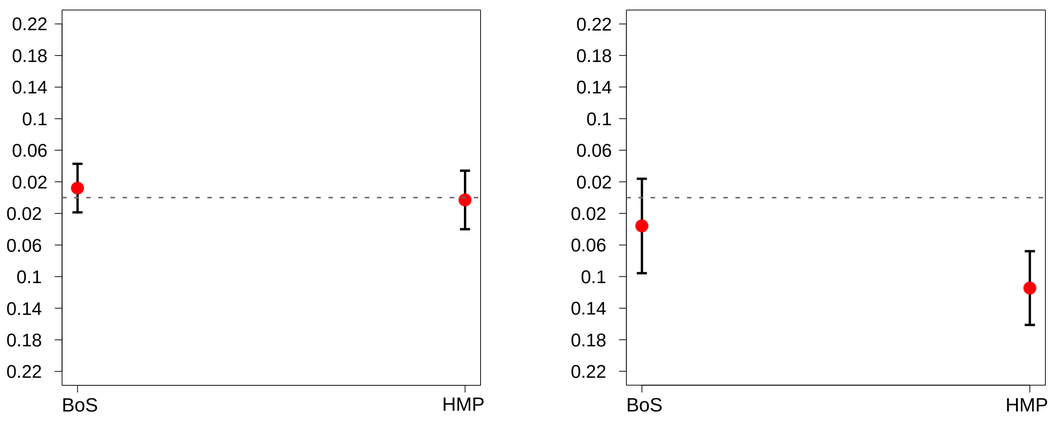

Figure 3. Paired t-test comparing the best BoG configuration and the baselines. BoG_{BIC} was similar to the baselines on MAP metric (left) but was outperformed by HMP on P10 (right).

Experimental Results – Event Retrieval

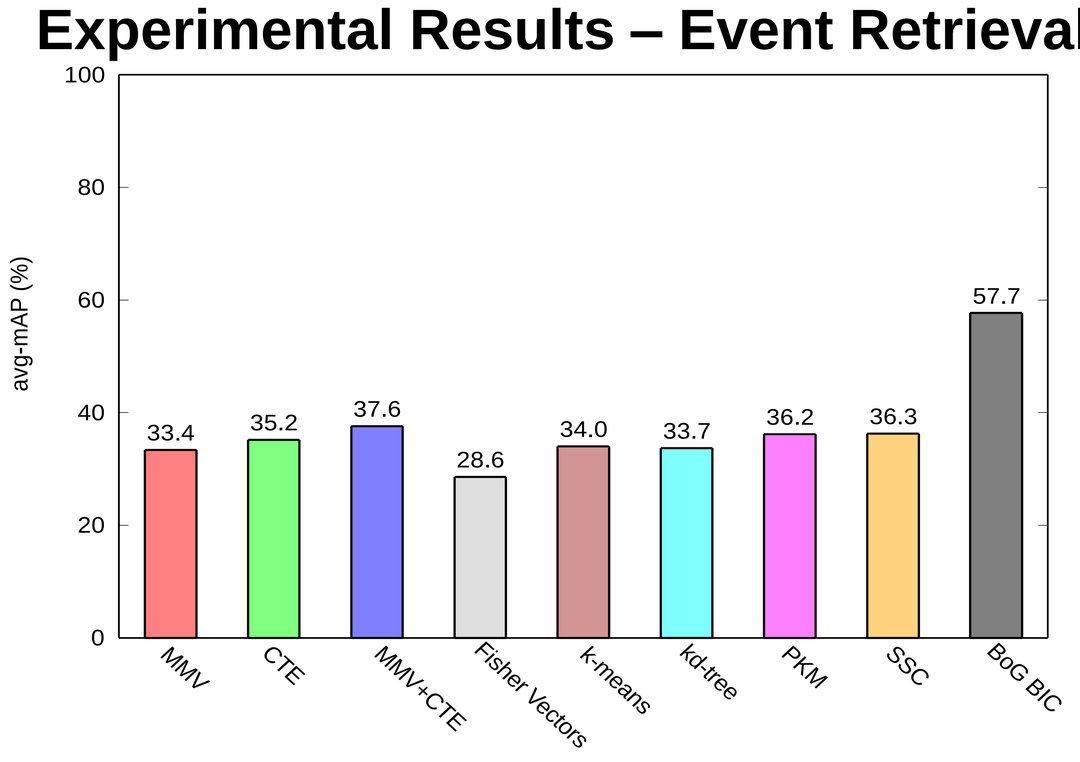

Figure 4. Performance of different methods for event retrieval on EVVE dataset. BoG_{BIC} outperformed all the baselines by a large margin.

Conclusions

- Approach performs **similar to state-of-the-art** methods on MediaEval dataset and it is the **state-of-the-art on EVVE dataset**;
- In the proposed visual dictionary, visual words are obtained by a **supervised classifier**; the method is **compact**; each dimension corresponds to a **semantic concept**;
- Future work includes the evaluation of other methods for **feature extraction** and **classification strategies**.

Bibliography

[1] J. Almeida, N. J. Leite, and R. S. Torres, “Comparison of vídeo sequences with histograms of motion patterns,” in *ICIP* 2011, pp. 3673–3676.
[2] O. A. B. Penatti, L. T. Li, J. Almeida, and R. da S. Torres, “A visual approach for video geocoding using bag-ofscenes,” in *ACM ICMR* 2012, pp. 1–8.
[3] S. Schmiedeke, C. Kofler, and I. Ferran´e, “Overview of mediaeval 2012 genre tagging task,” in *Working Notes Proceedings of the MediaEval 2012 Workshop*, 2012.
[4] O. A. B. Penatti, E. Valle, and R. S. Torres, “Comparative study of global color and texture descriptors for web image retrieval,” *JVCIR*, vol. 23, no. 2, pp. 359–380, 2012.
[5] J. Revaud, M. Douze, C. Schmid, and H. Jégou, "Event retrieval in large video collections with circulant temporal encoding," CVPR 2013, pp. 2459-2466.

Acknowledgments

The authors would like to thank CAPES, CNPq, and FAPESP (grant #2016/06441-7) for funding.