CAUVIS-IME-USP at TRECVID 2011: instance search

Arnaldo Câmara Lara and Roberto Hirata Jr.

Rua do Matão, 1010, Cidade Universitária, São Paulo, Brazil

Institute of Mathematics and Statistics - University of Sao Paulo

E-mail of corresponding author: alara@vision.ime.usp.br

Structured Abstract

In this paper, we describe the participation and results of CAUVIS-IME-USP at TRECVID 2011 in instance search pilot task.

Briefly, what approach or combination of approaches did you test in each of your submitted runs?

- CAUVIS_USP: PHOW descriptor and the bag-of-word approach.

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

We submitted just one run.

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

The best performance of our run was in the two queries of type location that we obtained score above the median of runs. In both locations, the texture of the scenes is the most important characteristic. One of the descriptors we are using is PHOW and it is a very suitable to describe texture. Other important factor is that for both queries the mask are big, including almost all the image. Our method did not work well when the interested instance is small.

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

PHOW descriptor and the bag-of-word approach is the state-of-art in object recognition but it did not work well in instance search task of TRECVID.

Introduction

The pilot task of TRECVID instance search, introduced in 2010, consists of searching for a specific person, object or place given an image and a mask of the interested object [1]. This problem can be used in real applications like semantic indexing, search in personal video collections, detect events of surveillance among others. In 2011, the task continues to be a pilot task. Accuracy in the localization of instances and time spent to run the queries are important issues, but they are not used for evaluation purposes for now.

The task has as input a database with thousands of video shots, 25 queries, each query with until 6 visual examples of frames and a mask that delimits the instance object. The proposed method must return as output 1000 video shots most likely to contain the instance object.

Method

To build a visual dictionary, we extracted a PHOW descriptor [3] in a dense grid of points from each visual sample of each query. Each sampled point has 5 pixels of distance from the next point vertically and horizontally. After this, 600000 descriptors are randomly selected and k-means is used to cluster the descriptors in 300 distinct centers, the visual words, building the visual dictionary [2]. Finally, the interested mask is applied in the image of the visual sample delimiting just the object instance. The model of the instance is built as a frequency histogram of the visual words.

For each video shot of the database, we extracted the central frame as a representative frame the video shot. The frames are described as a frequency histogram of the visual words. For each model of instance, it is calculated the chi-squared distance from each frame of video shots. The distances are sorted and the 1000 shortest distances are returned.

Analysis of Results and Conclusion

We did not use a different approach for each type of query (person, object or location) and our method using PHOW descriptor obtained a reasonable score just for the location queries: 9024 and 9029. For 9024, the media MAP (mean average precision) was 0.241 and our result was 0.268. For 9029 query, the media MAP was 0.184 and our result was 0.276. In these two location queries, the texture of the scenes is adequate to applying a PHOW descriptor to characterize it.

PHOW descriptor and bag-of-words approach showed good results in object recognition as reported in literature [4] but it was not enough to obtain good results in TRECVID instance task. For people instance search, it is necessary to use face detection and recognition methods and for objects, shape and color of the objects should be characterized in the model too. With the experience accumulated in this participation, we are planning to participate in the next edition of TRECVID using different approaches.

References

1. Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722
2. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, Visual categorization with bags of keypoints, ECCV International Workshop on Statistical Learning in Computer Vision (Prague, Czech Republic), 2004.
3. A. Bosch, A. Zisserman, and X. Munoz, Image classi_cation using random forests and ferns, 11th International Conference on Computer Vision (Rio de Janeiro, Brazil), october 2007, pp. 1-8.
4. J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, Locality-constrained linear coding for image classi_cation, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on (San Francisco, USA), IEEE Computer Society, june 2010, pp. 3360{3367.