MIC_TJ at TRECVID 2013: Instance Search Task

Hanli Wang, Lei Wang, Bo Xiao, Kuangtian Zhufeng
Department of Computer Science and Technology, Tongji University,
201804 Shanghai, P. R. China
{hanliwang, 110_wangleixx, 1314xiaobo, 2zhufengkuangtian}@tongji.edu.cn

Abstract

The MIC_TJ team (Multimedia and Intelligent Computing Lab at Tongji University) participated in the Instance Search (INS) task at TRECVID 2013[1]. In this paper, we will mainly present the framework and approaches used in our system. For the INS task, we submitted two runs: (1) Bag-of-Feature (BoF) baseline; and (2) BoF with spatial re-ranking. A summary of our submissions is shown below.

- F_NO_MIC_TJ_1: Basic approach using BoF [2] model and a global geometric verification using RANSAC [3] is added to re-rank the top-n results. Cosine distance is employed to estimate the similarity between video frames.
- F_NO_MIC_TJ_2: Only the basic approach using BoF model is applied to represent and retrieve key frames, with the cosine distance being used to estimate the similarity between frames.

The results of both runs are not satisfactory due to the simple algorithms we used. However, the acceleration ideas and technologies using Map-Reduce [4] framework with multiple Graphics Processing Units (GPUs) to deal with large scale multimedia data are worth considering.

1. Introduction

Instance Search (INS) task is to retrieve a set of shots (a shot consists of a series of sequential frames that are similar to each other in content) which most likely contain a specific entity from a collection of test video clips [5]. For the INS task at TRECVID 2013, the description of a master shot reference and several topics (i.e., queries) are already available.

As mentioned above, this task is about multimedia data mining and a large scale of video data or image data need to be analyzed. In order to deal with such a huge amount of data, advanced algorithms and powerful computing technologies or platforms are both essential and required.

The Map-Reduce [4] framework is a parallel programming model aiming for cloud computing, and an associated implementation for processing and generating large scale data was originally proposed by Google. It is simple and powerful, and highly abstracts the process of complex parallelization. In addition, the Map-Reduce framework provides automatic task/data management, inter-machine communication and fault tolerance. On the other hand, the Graphics Processing Unit (GPU) is another very powerful technology proposed by NVIDIA. As compared with CPU, GPU owns higher computational ability [6] and can significantly improve the processing speed.

Our team combines the advantages of both cloud computing and GPUs to build a novel parallel computing system. The proposed system can greatly shorten the processing time while keeping a fundamental accuracy for the multimedia task. Moreover, this is the first time for our team to

participate in the INS task, our goal is mainly to study and test our ideas as well as systems on processing large scale multimedia data.

2. Framework Overview

The overview of the proposed system is shown in Figs. 1-2. The system is built based on a CPU+GPU cluster with 12-node computers including 12 CPUs (Intel Core i5-3470) and 24 GPUs (GeForce GTX 660). It is designed and implemented using the Map-Reduce framework, combined with multi-GPUs in each node to cooperate with CPUs.

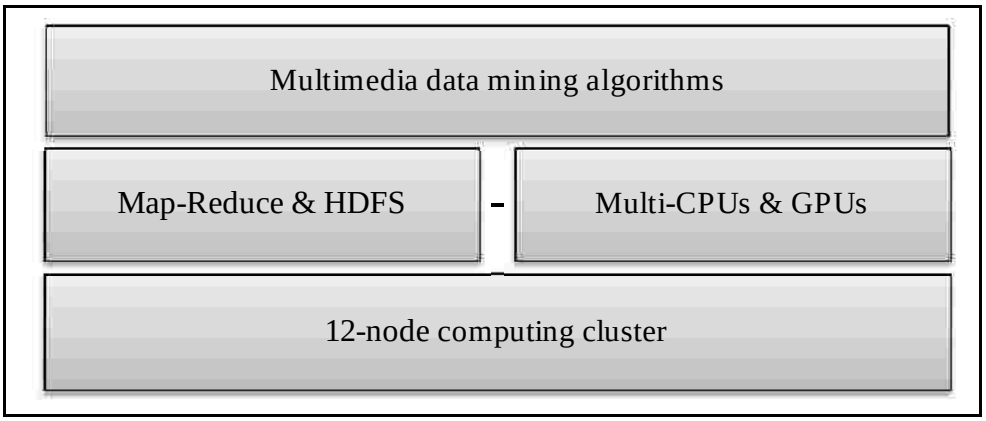

Fig. 1. Structure of the proposed multimedia data mining system. (1) The lowest level is a computing cluster constructed by 12 machines. (2) The middle level is a combination of Map-Reduce framework with Hadoop Distributed File System (HDFS) [7]. (3) The highest level is the algorithm level, which realizes all the programs related to the task.

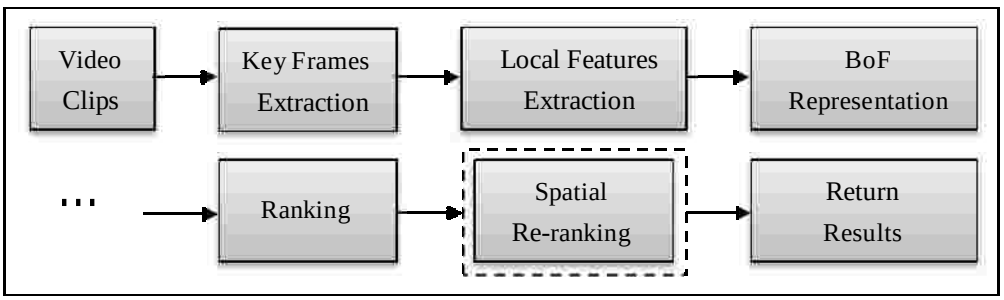

Fig. 2. Framework of our instance search system.

2.1. Key Frames Extraction

Considering the INS task, we firstly extract the key frames from the video collection based on the provided master shot reference files. For each shot, if the length of the shot is shorter than 1 second, we only take the first and the last frame of this shot as key frames; otherwise, we take a frame every one second from the start to the end of this shot.

2.2. Local Features Extraction

In the proposed system, we use the Hessian-Affine detector [8] and the SIFT [9] descriptor following the research of [10].

2.3. Vocabulary

Owing to the good performance of the proposed system in dealing with large scale data, we do not use any improved clustering method and just employ the flat K-Means clustering method to generate the vocabulary dictionary. Moreover, we use all the descriptors extracted from the key frames for clustering.

2.4. Ranking and Spatial Re-ranking

At the ranking stage, we use the cosine distance to evaluate the similarity of two frames, which gives the score of 1.0 if the two images are completely same. After the initial ranking, a fast spatial re-ranking algorithm [11] using RANSAC [3] is employed as a post-processing to re-rank the top-100 images. At last (just in F_NO_MIC_TJ_1), the system returns the final ranked list including the result of 1000 video shots.

3. Conclusion

In the INS task at TRECVID 2013, we propose a hybrid CPU+GPU cloud computing platform to deal with large scale multimedia data. In the future, we will put more attention on the algorithms while keeping optimizing our framework.

4. References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. Smeaton and G. Quenot, "TRECVID 2013 - An overview of the goals, tasks, data, evaluation mechanisms and metrics", TRECVID, 2013.

[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos", in *IEEE ICCV'03*, Vol. 2, pp. 1470-1477, Oct. 2003.

[3] M. A. Fischler and R. C. Bolles, "Random sample consensus", *Comm. ACM*, Vol. 24, No. 6, pp. 381-395, 1981.

[4] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters", *Communications of the ACM - 50th anniversary issue: 1958-2008*, Vol. 51, No. 1, pp. 107-113, Jan. 2008.

[5] A. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVID", in *MIR'06*, pp. 321-330, 2006.

[6] B. He and N. K. Govindaraju, "Mars: A MapReduce framework on graphics processors", in *PACT'08*, pp. 260-269, 2008.

[7] Apache Hadoop, http://hadoop.apache.org/

[8] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors", *Int. J. Comput. Vision*, Vol. 60, No. 1, pp. 63-86, Jan. 2004.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *Int. J. Comput. Vision*, Vol. 60, No. 2, pp. 91-110, Jan. 2004.

[10] H. Jegou, M. Douze and C. Schmid, "Improving bag-of-features for large scale image search", *Int. J. Comput. Vision*, Vol. 87, No. 3, pp. 316-336, May. 2010.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching", in *IEEE CVPR'07*, pp. 1-8, Jun. 2007.