MIC_TJ at TRECVID 2014

Lei Wang, Yun Yi, Bowen Zhang, Fengkuangtian Zhu, Bo Xiao, Tianyao Sun, Hanli Wang
Department of Computer Science and Technology, Tongji University,
201804 Shanghai, P. R. China
{110_wangleixx, 13yiyun, 102310, 2zhufengkuangtian, 1314xiaobo, 1333783_sty,
hanliwang}@tongji.edu.cn

Abstract

Our MIC_TJ team (Multimedia and Intelligent Computing Lab at Tongji University) participated in the Instance Search (INS) task and the Multimedia Event Detection (MED) task at TRECVID 2014 [1]. In this paper, we mainly present the framework and approaches used in our systems. For the INS task, we submit a speed up system with a GPU cluster, while in the MED task, we adopt the classic Bag-of-Words (BoW) framework with trajectory based features and audio feature. This paper presents the methods and findings for INS and MED task. For the INS task, we submitted 13 runs, and all of the training data are extracted from “BBC Eastenders”. Regarding the MED task, we submitted 1 run. The training data of this run is extracted from the datasets of 000Ex, 010Ex and 100Ex for different sub-tasks, respectively. The corresponding runs for INS and MED tasks are summarized below.

INS 2014	**MED 2014**
F_D_MIC_TJ_1: Using ASMK, Hessian-Affine detectors with both ColorSIFT and RootSIFT descriptors being united together.	MIC_TJ: Using MFCC, Salient Trajectory with HOG, HOF and MBH.
F_A/B/C/D_MIC_TJ_2: Using ASMK with Hessian-Affine detectors and RootSIFT descriptors.	
F_A/B/C/D_MIC_TJ_3: Using ASMK with Hessian-Affine detectors and ColorSIFT descriptors.	
F_A/B/C/D_MIC_TJ_4: Using the asymmetrical method with Hessian-Affine detectors and RootSIFT descriptors.	

For the INS 2014 task, the highest score of these 13 runs is 0.146 (F_D_MIC_TJ_1). It indicates that our approaches used in this task are reasonable while still requiring further improvement. Additionally, we also use the MapReduce framework as well as multiple Graphics Processing Units (GPUs) to accelerate large-scale data processing which significantly improve the training and searching efficiency. As far as the MED 2014 task is concerned, we find that the combination of audio feature and video features are important to system's performance. Thus, it is critical to design an appropriate fusion technique to fuse different features.

1. Instance Search (INS) Task

1.1 Introduction

The Instance Search (INS) task is to retrieve a series of video shots (these video shots consist of a series of sequential video frames that are similar or partly similar to each other in content) which most likely contain a specific entity from a collection of test video clips [2]. For the INS task at TRECVID 2014, the description of a master shot reference and several topics (*i.e.*, queries) are already available.

Regarding our implementation, different sets of examples for a topic are used. For set A, only example 1 is used for searching. While for the set B/C/D, multiple query images are employed which give more information to the system, thus higher performances can be achieved. In the set B/C/D, the local features extracted from multiple query images of the same kind are combined together to form a new query.

Moreover, we employ the Aggregated Selective Match Kernel (ASMK) [3] algorithm and the asymmetrical method [4] for the INS 2014 task. A single late fusion on different features is added at last. In order to deal with such a huge amount of data, both advanced algorithms, powerful computing technologies and platforms are required. The MapReduce [5] framework is a parallel programming model aiming for cloud computing and an associated implementation for processing and generating large-scale data is originally proposed by Google. Meanwhile, the Graphics Processing Units (GPU) is a powerful technology proposed by NVIDIA. Comparing with CPU, GPU owns higher computational ability [6] and can significantly improve the processing speed. Our team combines the advantages of both MapReduce model 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.

1.2 Framework Overview

The overview of the proposed system is shown in Fig. 1. 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 MapReduce framework, combined with multi-GPUs in each node to cooperate with CPUs.

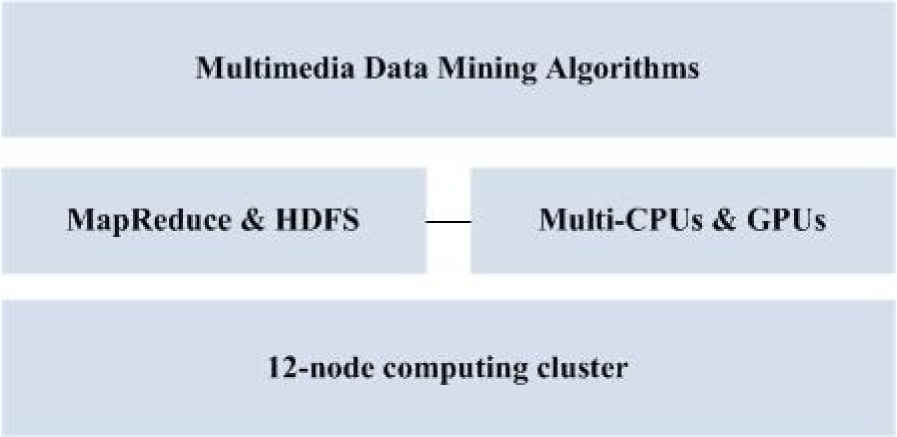

Fig.1. Architecture of the proposed multimedia data mining system. (1) The lowest level is a computing cluster constructed by 12 machines. (2) The middle level is a combination of the MapReduce framework with Hadoop Distributed File System (HDFS) [7]. (3) The highest level is the algorithm level, which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

1.3 Key Frames Extraction

Considering the INS 2014 task, we extract the key frames from the video collection based on the available master shot reference files.

1.4 Local Features Extraction

In the proposed system, we use the Hessian-Affine detector [8], SIFT descriptor [9], RootSIFT [10] descriptor and ColorSIFT [11] descriptor.

1.5 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 to generate the visual vocabulary dictionary with the dictionary size equal to 10k. Moreover, we use all the descriptors extracted from the key frames for clustering.

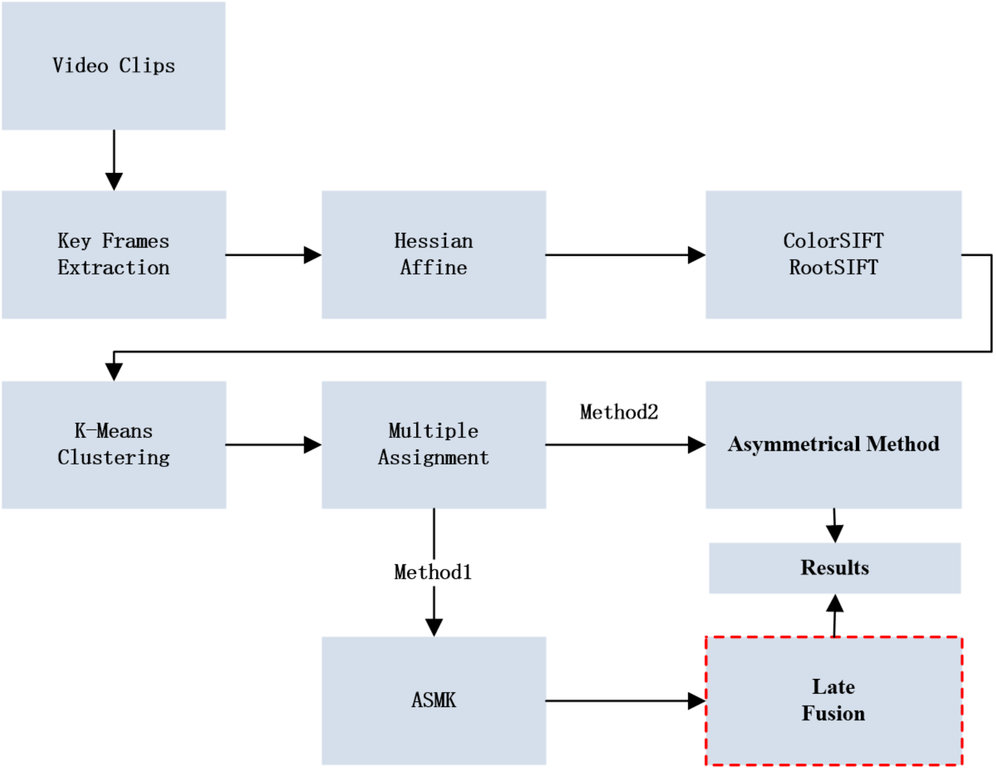

Fig. 2. Framework of our system in INS 2014.

1.6 Late Fusion

In order to improve the performance of the final results, multiple features are combined together. In our systems, the original results achieved by RootSIFT and ColorSIFT are mixed using a single linear fusion as

$$\text{FinalScore} = a*\text{ScoreRootSIFT}+b*\text{ScoreColorSIFT},$$

where a=2 and b=1 in our settings. The top-1000 results in the final list are then returned.

2. Multimedia Event Detection (MED) Task

2.1 Overview

For the 2014 TRECVID Multimedia Event Detection (MED) task, the sub evaluation is processed. Our submitted runs included Pre-Specified and Ad-Hoc event collections. For each collection, we submitted 3 exemplar conditions, including 000Ex, 010Ex and 100Ex.

In our system, we use motion feature, static visual feature and audio feature to describe events with the fisher vector [12] model being utilized to aggregate these features. Both early and late fusion techniques are used to combine the different low-level features. Figure 3 is the schematic overview of the proposed system.

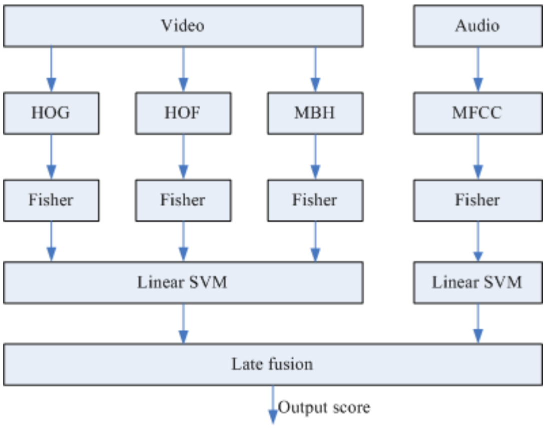

Fig. 3 Overview of the system for MED 2014.

2.2 System Description

2.2.1 Audio Feature

Due to auditory clues in some events, such as beekeeping, fixing a musical instrument and tuning a musical instrument, features of audio segments are considered in our system. Hence, we adopt the famous Mel-Frequency Cepstral Coeffcients (MFCC) algorithm for audio feature extraction and description. The window time for MFCC is 32 ms and there is 50% overlap between two adjacent windows. To fully utilize the discrimination ability of MFCC, we append delta and double-delta of 20 dimensions MFCC vector to generate a 60-dimension MFCC vector.

In order to use a single vector to represent a whole audio file, we adopt the classic Bag-of-Words (BOW) framework, where the fisher vector and Gaussian mixture model (GMM) are used. The clustering number of GMM is set to 500, therefore the dimension of fisher vector for each audio is 60000.

2.2.2 Visual Features

Visual information is captured by using trajectory features. Firstly, salient trajectories are generated to track human actions at multiple spatial scales. Then, camera motion elimination is utilized to further improve the robustness of the trajectories. To depict human motions accurately and efficiently, the Histogram of Oriented Gradient (HOG) [13], Histogram of Optical Flow (HOF) [14] and Motion Boundary Histogram (MBH) [15] are employed. The dimensions of these three descriptors are 96 for HOG, 108 for HOF and 192 for MBH, respectively.

After the extraction of descriptors, these feature vectors are normalized with the signed square root and L1 normalization, and then, PCA is individually applied to each of these three feature vectors for dimension reduction. After the feature descriptors are extracted, the fisher vector model is applied to construct a codebook for each descriptor. We compute one fisher vector over the complete video, and apply the signed square root and L2 normalization which is able to improve the recognition performance in combination with linear Support Vector Machine (SVM).

2.3 Fusion and Classification

Before training the related SVM classifier, an early fusion strategy is used to concatenate the trajectory based visual features, including HOG, HOF and MBH. To combine the audio and visual features, we employ a late fusion strategy, which linearly combines the classifier scores computed from the audio and visual features. As far as the classification is concerned, the linear SVM is employed in this work. In our implementation, the standard LIBSVM [16] is used with the penalty parameter *C* equal to 100.

3 Conclusion

In the INS task at TRECVID 2014, we propose a hybrid CPU+GPU cloud computing platform to deal with large-scale multimedia data mining algorithms. And we shall step forward to improve our algorithm to reach a better result in the future. Regarding the MED task, the BoW framework with fisher vector is applied and the trajectory visual features and MFCC audio feature are employed to build up a MED system. In the future, we will improve our system by using different kinds of features and more advanced learning methods.

4 References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quénot, "TRECVID 2014 -- An overview of the goals, tasks, data, evaluation mechanisms and metrics", *Proceedings of TRECVID 2014*, 2014.

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

[3] G. Tolias, Y. Avrithis, and H. Jegou, "To aggregate or not to aggregate: selective match kernels for image search", in *ICCV'13*, pp.1401-1408, Dec. 2013.

[4] C.-Z. Zhu, H. Jegou, and S. Satoh, "Query-adaptive asymmetrical dissimilarities for visual object retrieval", in *ICCV'13*, pp.1705-1712, Dec. 2013.

[5] 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.

[6] B. He and N. K. Govindaraju, "Mars: A MapReduce framework on graphics processors", in *PACT'08*, pp. 260-269, Oct. 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] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve

object retrieval”, in *CVPR’12*, pp. 2911-2918, Jun. 2012.

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, “Evaluating color descriptors for object and scene recognition”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582-1596, Sept. 2010.

[12] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the fisher kernel for large-scale image classification”, in *ECCV’10*, pp. 143-156, Sept. 2010.

[13] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, in *CVPR’05*, pp. 886-893, Jun. 2005.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies”, in *CVPR’08*, pp. 1-8, Jun. 2008.

[15] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance”, in *ECCV’06*, pp. 428-441, Sept. 2006.

[16] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, Apr. 2011.