TRECVid 2013 Experiments at Dublin City University

Zhenxing Zhang, Rami Albatal, Cathal Gurrin, and Alan F. Smeaton

INSIGHT Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland
{zzhang, ralbatal, cgurrin, asmeaton}@computing.dcu.ie

Abstract. In a move away from previous years' participation in TRECVid ([1] [2] [3]), this year our team focused on the instance search task (INS). We improved our system from last year by applying large vocabulary quantization, soft assignment of visual words, spatial verifications and query expansion. Overall, four automatic runs have been submitted for evaluation. In this paper, we present first our system, then we discuss the results and findings of our experiments.

1 Introduction

This paper describes the fourth participation of the iAD (Information Access Disruptions) Centre at the TRECVid workshop [4] in 2013 [5]. iAD is a research centre partially funded by the Norwegian Research Council. It is directed by the Microsoft Development Center Norway (MDCN) in collaboration with Accenture, and various universities including Cornell University, Dublin City University, BI Norwegian School of Management and the Universities in Tromso (UiT), Trondheim (NTNU) and Oslo (UiO). Given the researchers' varying expertise in video search and analysis, the consortium's efforts were coordinated by the group from Dublin City University. The main purpose of the iAD project is to research information access technologies, so we focus our research on developing and evaluating extreme precision search and recommendation solutions that support effective multi-modal access to multimedia content. This year, we kept working on the visual similarity search over large video collections and developed a particular object retrieval system to participate in the INS task.

Our system is built upon our experience from the participation in previous years, however new technologies, like large vocabulary quantization, soft assignment of visual words, spatial verifications and query expansion, have been employed to maximize the performance which will be served as the "state of the art" solution for our further experiments.

This paper is structured as follows: first in section 2, we describe the implementation of our proposed system for the instance search task; then section 3 focuses on experiment setting and presents the results. Our contribution in this year's experiment and future work are then summarized in Section 4; and finally section 5 concludes this paper.

2 System Implementation

This section overviews the system we developed for this year's TRECVid Instance Search Task. Our system employed the most successful solution for visual object retrieval, which is based on the bag of visual words approach [6]. The proposed instance search system includes several main components, and this sections presents the baseline components for all the submitted runs.

2.1 Baseline Components

Figure 1 gives an overview of the workflow of our system.

Fig. 1: Basic workflow of DCU-iAD INS system

Keyframe and Feature Extraction In order to reduce the computing complexity, a common first step in content based video analysis is to segment a video into elementary shots, each of which contains a scene happening in continues time, or same scene. In this experiment shot boundaries have been given in the data sources, so we extract the "most middle" frame to represent the content of that shot.

For each keyframe, the affine invariant interest regions are detected and scale invariant features are extracted. More specifically, we use the ColorDescriptor library[7] and select the Harris-Laplace detector and the SIFT descriptor. At the end of this step, we extracted about 480,000 keyframes and 0.96 billion SIFT descriptors.

Feature Quantization The purpose of this step is to aggregate a set of local descriptors into one vector and to represent each keyframe with a fixed-sized feature vector. This work often involves generating a visual vocabulary by using clustering algorithms, such as k-means, and mapping local descriptors into the visual words (i.e. cluster centres) of this vocabulary[6]. As shown in the work of [8], a large size vocabulary with more discriminative power is necessary for large scale object retrieval. However the time complexity of k-means will increase dramatically because in each iteration there are $k * N$ euclidean distance calculations. So we adopted Approximate k-means algorithm [8] which replaces the exact distance computation by an approximate nearest neighbour search method.

Search algorithm The focus of this step is to rank the keyframes according to their visual similarity to the query topics. We used the well-studied vector space model in information retrieval combined with a standard *tf-idf* weighting scheme. In this approach, the query topic and each keyframe are represented by a high-dimensional, sparse vector, the relevance score is calculated using the normalized L_2 distance. The open source library Lucene[9] is employed to accomplish the work.

Multi Images for Query Topic Experiments reported in the work of [10] proved that the performance of an object retrieval system can be improved significantly by issuing multiple images queries. In our experiment there are 4 images for each query topic and each of them has a mask to specify the item of interest. So we extracted the interests area, generated the bag of visual words representation, then summed up all the words into one query vector and finally we query the database for results using our search algorithm.

Spatial Verification and Re-ranking Until now an initial ranked result has been generated according to the occurrences of visual words and their statistical weighting (*tf-idf model*). Geometric information about the query object has been ignored during the step of feature quantization. As shown in [6] [8] [11], retrieval performance can be improved by verifying the geometric consistency between matched keypoints from the query images and retrieved images. The relevance scores can be recalculated from the verified inlier number by estimating affine homographies between the query image and initial top ranked results. The reason that only a few top-ranked results (20 in our experiments) have been verified is that estimating the best affine transform could be a time-consuming task due to the RANSAC algorithm [12].

2.2 Advanced Components

Compared to our last year's participation, we made some improvements to maximize the performance of our system.

Soft Assignment Ideally image features detected from the same object should be the same, however the variabilities of imaging conditions lead to changes in the feature descriptions, and clustering technologies can be used to overcome some of the variations. Recalling the feature quantization step in the

bag of words image representation algorithm, two image features are considered identical if they are clustered to the same visual word. Suppose two image features are located on both sides of a cluster boundary, they would be assigned to different visual words no matter how close they are in the feature space. In order to give a more precise measurement of the relationship between image features, we describe an image feature by the nearest n visual words instead of only one visual word. Additionally, a weighting has been calculated and assigned to each visual word according its distance to the image feature. We set $n = 3$ to keep a reasonable computing time.

Query Expansion Query expansion is a standard methodology for improving retrieval performance in text retrieval. In this method, a number of top-ranked documents from the initial results are added to the search query and allow the retrieval system to find more challenging documents by using relevant terms that are not included in the original query. The work of [11] illustrated that query expansion can be applied in visual object retrieval system. In our experiment, we chose a simple but effective approach called average query expansion, to automatically enhance our search performance. After the initial search with the original query, we first apply a strong spatial constraint between the query image and the top 20 results to filter out any false positive results which usually introduce noisy data and ruin the expansion. Then only the image features from the region of interested in the verified image are used to construct a richer search query by averaging the number of visual words from those image. Then more matched images would be retrieved by re-querying the database using the expanded query model. It is important to note that the performance of query expansion depend heavily on the initial returned results.

3 Experiments and Results

In this section, our experimental settings are presented and the final results are discussed.

During the experiment, we mainly test two technologies: Soft assignment in image representation and Query expansion for re-ranking. We submitted 4 automatic runs for the Instance Search Task of TRECVid 2013. The short description and their results of four automatic runs are shown in Table 1.

In general, all 4 runs contain the basic components: *a*) One keyframe is extracted for each shot and only SIFT descriptors are used for feature extraction. *b*) A visual vocagulary of one million clusters (visual words) is generated from feature quantization. *c*) The vector space model combined with *tf-idf* weighting scheme is selected as the scoring function.

The main difference between the runs is that Soft assignment and Query expansion technologies have been applied. Run 4 served as the baseline in this year's experiment. In run 1 and 3, query expansion is used to extend the retrieval results. Then soft assignment has been applied in Run 1 and 2.

Figure 2 illustrates our best results compared to best and median performance over all runs. Our system yielded better results in 10 topics than the

Table 1: iAD-DCU on INS Task of TRECVid 2013

Run ID	mAP	Description
F_N0_iAD_DCU_1	0.198	Soft assignment in visual words representation, Spatial verification and Automatic query expansion
F_N0_iAD_DCU_2	0.193	Soft assignment visual words representation and Spatial verification.
F_N0_iAD_DCU_3	**0.216**	Hard assignment in visual words representation, Spatial verification and Automatic query expansion.
F_N0_iAD_DCU_4	0.191	Hard assignment visual words representation, and Spatial verification.

average results, and most especially we almost achieved the best in search topic 9069. In the rest of the topics our system did not perform very well, and the reasons will be given in section 4.

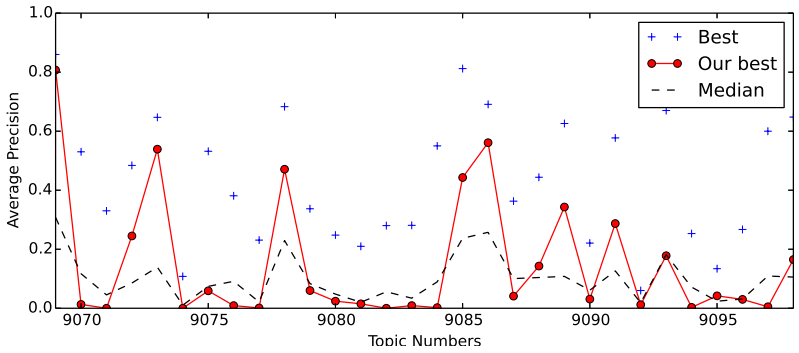
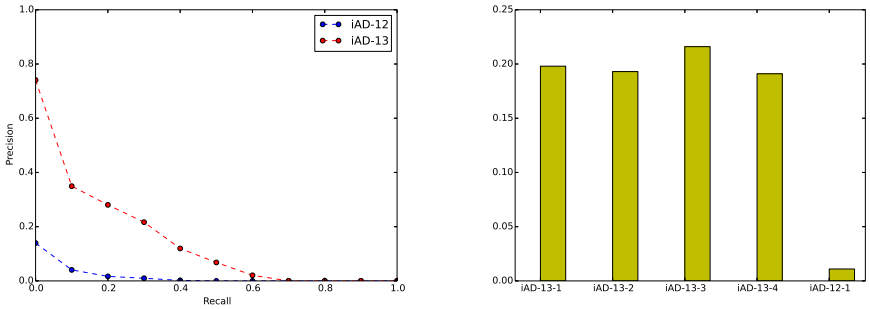


Fig. 2: Average Precision by topics for our best run and mean average precision over all runs in INS task of TRECVid 2013

4 Discussion and Future work

Contribution As shown in figure 3, our performance improved significantly both in precision and recall, compared to our last year's participation [3]. The main contributions this year come from three aspects:

1. the first one is higher dimensional vocabulary with the soft assignment scheme that give more discriminative features in image representation;
2. then the second is the spatial verification and the re-ranking technologies that removed some false positive results and improved the precision;



(a) Precision-Recall Curve (b) mean Average Precision

Fig. 3: Performance Comparison between iAD-12 and iAD-13

3. the third and most important one is the multiple images query and query expansion technologies that consistently enhance the search query and significantly improved the recall performance. For example, figure 4 displays the precision-recall curve which reveal the success of our system in improving recall.

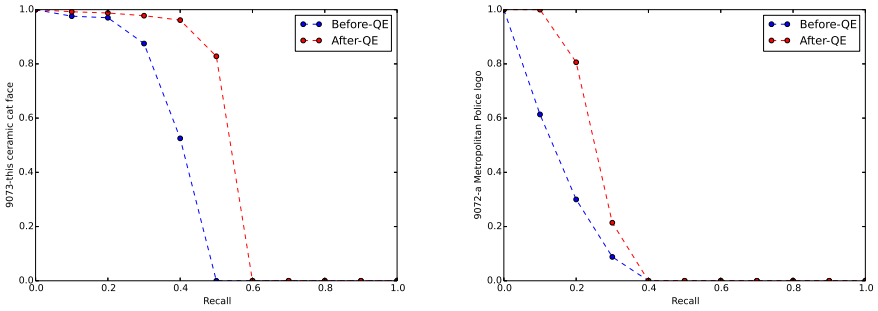


Fig. 4: Performance Comparison before and after Automatic query expansion

Problems and Solutions Even though we made a big step in this year in terms of progressing our our technologies, we still have some problems which need to be addressed.

Firstly due to the locking of rich descriptors (only SIFT in our case), some tasks were extremely difficult for our retrieval system compared to others. For instance, our system performed badly in topics about the logo (9071, 9082, 9087), sculpture (9076) , smooth object (9083, 9094), and person and animal (9077,

9084, 9096). The solution to this problem would be to employ a robust and rich combination of local and global visual feature descriptors.

The second problem with some of the queries is where there is an interesting object located in a complex background scene, which will lead to misunderstanding of the search topics. Take search topic 9092 (A man wearing white shirt and standing in a bar) for example, the results returned by our system (Figure 5) focused on the same bar scene instead of the man. The possible solution is to recruit discriminative learning technologies for object detection [13] and train a discriminative model which can learn from negative examples.

Fig. 5: Example of misunderstanding of the search topics

The third problem comes from practical experience. Our efficiency for data processing has to be improved in future. In this year's experiment, it took three week to finish the 1 million visual vocabulary generation, which made it very hard for us to test more possible approaches and add more feature to our system. To overcome this, we will try cloud-based computing resources in future years.

5 Conclusions

This year, our team focused on one task: Instance Search Task. We chose the most successful approaches based on text retrieval and bag of visual word framework, and then add two advanced components, soft-assignment and query expansion, to maximize our retrieval performance. The results from the official evaluation campaign indicate that a significant improvement has been made by our group in this year compared to previous years. We talked about the components of our proposed system in section 2, and then presented an overview of our experiment and results in section 3. We discussed our performance in this year's experiment and point out the problem of our currently system in section 4 which also light the direction of our future work.

Acknowledgments

This research was supported by the Norwegian Research Council (CRI number: 174867) and Science Foundation Ireland under Grant No.: SFI/12/RC/2289 IN-

SIGHT). We would like to thank our contacts at Internet Archive for assisting us in downloading the data corpus.

References

1. Foley, C., Guo, J., Scott, D., Ferguson, P., Wilkins, P., McCusker, K., Diaz Sesmero, E., Gurrin, C., Smeaton, A.F., Giro-i Nieto, X., Marques, F., McGuinness, K., O'Connor, N.E.: TRECVid 2010 Experiments at Dublin City University. In: TRECVid 2010. (12 2010)
2. Scott, D., Guo, J., Foley, C., Hopfgartner, F., Gurrin, C., Smeaton, A.F.: TRECVid 2011 Experiments at Dublin City University. In: TRECVid 2011. (12 2011)
3. Guo, J., Zhang, Z., Scott, D., Hopfgartner, F., Gurrin, C., Smeaton, A.F.: TRECVid 2012 Experiments at Dublin City University. In: TRECVid 2012. (11 2012)
4. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, New York, NY, USA, ACM Press (2006) 321–330
5. Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Kraaij, W., Smeaton, A.F., Quenot, G.: Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2013, NIST, USA (2013)
6. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision. Volume 2. (October 2003) 1470–1477
7. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1582–1596
8. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)
9. Lucene: The Lucene search engine (2005)
10. Arandjelović, R., Zisserman, A.: Multiple queries for large scale specific object retrieval. In: British Machine Vision Conference. (2012)
11. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: IEEE International Conference on Computer Vision. (2007)
12. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press, ISBN: 0521540518 (2004)
13. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV. (2011)