TRECVid 2012 Experiments at Dublin City University

Jinlin Guo, Zhenxing Zhang, David Scott, Frank Hopfgartner, Rami Albatal, Cathal Gurrin, and Alan F. Smeaton

CLARITY: Centre for Sensor Web Technologies
Glasnevin, Dublin 9, Ireland
{jguo,zzhang,dscott,fhopfgartner,ralbatal,cgurrin,
asmeaton}@computing.dcu.ie

Abstract. Following previous participations in TRECVid, this year, the DCU-IAD team participated in four tasks of TRECVid 2012: Instance Search (INS), Interactive Known-Item Search (KIS), Multimedia Event Detection (MED) and Multimedia Event Recounting (MER).

1 Introduction

This paper describes the third (after [1] [2]) participation of the iAD (Information Access Disruptions) Centre at TRECVid. iAD is a research centre partially funded by the Norwegian Research Council. It is directed by Microsoft Development Center Norway (MDCN) in collaboration with Accenture, and various universities: Cornell University, Dublin City University, BI Norwegian School of Management and the universities in Tromso (UiT), Trondheim (NTNU) and Oslo (UiO). Given the researchers' expertise in video search and analysis, the consortium's efforts were coordinated by the group from Dublin City University.

Since iAD is about researching information access technology, we focused this year on a wide range of tasks within TRECVid, including Instance Search (INS) Interactive Known-Item Search (KIS), Multimedia Event Detection (MED) and Multimedia Event Recounting (MER). For some tasks, such as KIS, we build upon our experience gained from the participation in last year's KIS task, where we asked novice users to use a tablet-based graphical user interface to evaluate different display methodologies for KIS interaction. For the MED task, we follow last year's work, but extend it by including the audio stream of the videos. In the MER task, we performed a template-based textual recounting framework using related semantic concepts and objects. Finally, in the INS task, we continued our research on scale visual instance searching over large video collections by incorporating a bag of visual word approach.

This paper is structured as follows: In Section 2, we describe our approach in the instance search task. Section 3 focuses on the known-item search task. Our efforts in the multimedia event detection and recounting tasks are summarized in Sections 4 and 5. Section 6 concludes this paper.

2 Instance Search

This year was our first time to participate in the instance search task at TRECVid. We submitted one automatic run for evaluation in which a vector space model based on high dimensional bag of visual words (BoVW) representation was applied. The method was inspired by the work of Philbin et al. [3]. In the remainder of this section, we outline the implementation of our approach in detail.

2.1 Preprocessing

Firstly, we extracted one keyframe per second from the full dataset, which resulted in approximately 800,000 keyframes for the whole collection. In order to reduce computation, we then identified duplicated images based on the MEPG-7 Color Layout feature. Finally, the amount of keyframes are reduced by roughly 50%, resulting in a higher density of discriminative keyframes.

2.2 Feature Extraction

For each keyframe, we detect the affine-invariant Harris-Laplace regions [4]. These regions are the stable areas that are invariant to viewpoint, illumination and scale changes. For each keyframe with size of 640*480, around 2,000 interesting regions were detected. Based on each interest region, we then generated a 384-dimension RGBSift descriptor [5]. This feature extraction process resulted in 0.79×10^9 descriptors for 74,955 video clips in total.

2.3 BoVW Representation

Next, we generated a BoVW representation for each keyframe by creating a 0.5 million dimensional vocabulary. We randomly sampled 30.8 million descriptors for clustering the visual vocabulary. Due to time and space reasons, we then adopted the approximate k-means [3] – an alternation to the original k-means algorithm to generate this vocabulary. By applying the approximate nearest neighbor method, the cluster centroid assignment time for each point got reduced, hence significantly increasing the overall speed of computation. After creating the vocabulary, every keyframe was quantified to a full text document with corresponding visual word terms.

2.4 Searching Algorithm

A weighted vector model with 0.5 million dimensional spaces was created using the standard tf-idf weighting scheme. We converted every query into a vector in the same vector space. Similarity between the query and documents was measured by their cosine similarity. We built an inverted index to allow efficient retrieval in very large data collections. The high-performance, full-featured search engine library Lucene [6] was used to perform the real-time searching.

2.5 Results

Figure 1 shows our instance search task evaluation results in comparison to the median and best results of all submissions. Totally, there are 21 topics listed in the figure. As the figure indicates, the average performance of all systems achieved no more than 0.2 in average precision for most of the queries. The best results are much better, and achieved 0.5 for nearly half of the queries. Excluding three topics, our results are very close to the median.

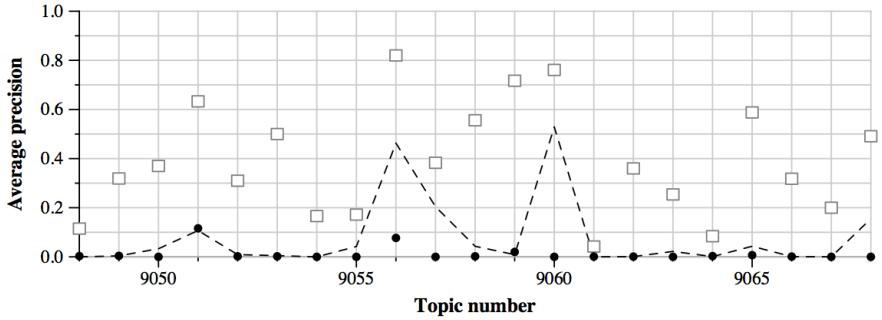

Fig. 1. iAD INS Task Results (•) iAD Run (- -)Median (□)Best by Topic

3 Known-Item Search

For this set of experiments we reuse our TRECVid prototype developed for both TRECVid 2010 and 2011, an iPad application which employs an interface designed for lean-back style interaction, which presents content in a simple and intuitive manner to the user. For the first year we utilize only the meta-data in developing our system but utilize a clustering technique based on opponentSIFT features and a k-means algorithm to group the content returned from the text ranked list. We developed two systems, one which uses single keyframe representations as shown in our TRECVid 2011 experiments and a second system which utilizes a multiple keyframe approach. We calculate this multiple keyframe representation based on MPEG-7 visual features. For each keyframe we build a set of dissimilar frames in the video and build the representation based on the intersection of these sets iteration based on the most dissimilar.

3.1 Users

Similar to previous years, we asked novice video searchers to participate in our evaluation; this time from DCU as opposed to from a partner institution. All participants were new to the creation and testing of video search engines. We

recruited eight subjects to test for this experiment, all members are over the age of 25 with the majority being graduates. Most of the users would regard themselves as heavy internet users, using services such as Google or YouTube quite frequently. An overview over the participants is given in Table 1.

Participant Profile		
Age:	25 and Younger	0
	Older than 25	8
Web Search	Regular	7
(inc Video)	Infrequent	1
Handheld	Never	1
Usage	Infrequent	5
	Regular	2
Education	Undergraduate/No Degree	1
	Graduate	7
Gender	male	3
	female	5

Table 1. Participants Profile TRECVid 2012

3.2 Experiment

Eight novice users attained from adjacent research groups form the participants who are assigned twelve tasks each, 6 topics on single keyframe and 6 on multiple keyframe, the distribution of which is shown in Table 2. Participants were provided with instructions on how to use the system and a set of training topics which were used to familiarize the participants with the system. Finally, users were given a survey form which they had to complete at each stage of the experiment, pre-experiment to capture demographic and usage data, post-experiment to capture their overall perception of the test and a survey at the end of each of the 6 assigned topics to capture immediate feedback.

We again employed a clustering technique to group similar content, k-means with k, the amount of cluster centers, set to 100 having been defined from previous experiments. Both systems used the output provided by this visual clustering, therefore, the only element we tested was the single vs multiple keyframe representation. Due to the use of clustering we cannot represent the multiple keyframes in the traditional (storyboard) approach, instead each video keyframe is inserted into their related cluster: In this way, each video can be represented in more than a single cluster increasing the likelihood of finding the known-item. We limit the chances of duplicates by allowing only one keyframe per video in each cluster representation.

	User 1	**User 2**	**User 3**	**User 4**	**User 5**	**User 6**	**User 7**	**User 8**
Topic 1:	a	b					a	b
Topic 2:	a	b					a	b
Topic 3:	a	b					a	b
Topic 4:	a	b					a	b
Topic 5:	a	b					a	b
Topic 6:	a	b					a	b
Topic 7:	b	a			a	b		
Topic 8:	b	a			a	b		
Topic 9:	b	a			a	b		
Topic 10:	b	a			a	b		
Topic 11:	b	a			a	b		
Topic 12:	b	a			a	b		
Topic 13:			a	b			b	a
Topic 14:			a	b			b	a
Topic 15:			a	b			b	a
Topic 16:			a	b			b	a
Topic 17:			a	b			b	a
Topic 18:			a	b			b	a
Topic 19:			b	a	b	a		
Topic 20:			b	a	b	a		
Topic 21:			b	a	b	a		
Topic 22:			b	a	b	a		
Topic 23:			b	a	b	a		
Topic 24:			b	a	b	a		

Table 2. Table outlining the topic distribution over our eight participating users, (a) denotes the Single-Keyframe system (b) denotes the Multi-Keyframe system

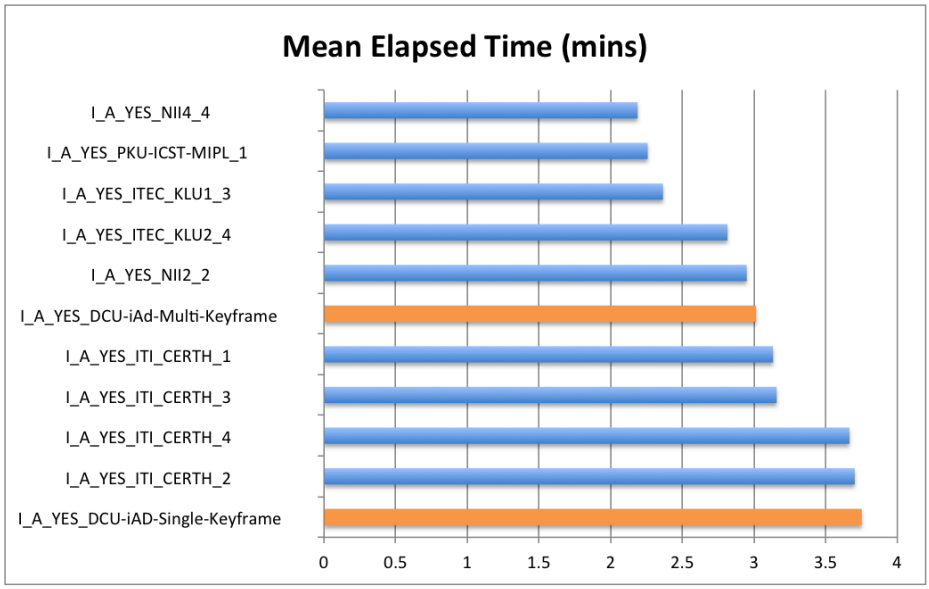

Fig. 2. Mean Elapsed Time of teams participating in TRECVid known-Item Task, our submission highlighted in Orange

3.3 Results

In this set of experiments we decided to not include the use of visual classification. Consequently, we only found ten of the twenty four topics on the single keyframe system, with a further two found in the multiple keyframe system, this is a stark contrast from last year's experiments where we attained the best results finding fourteen of the known-items on a single keyframe representation system.

With regard to Mean Elapsed Time our multiple keyframe approach out performs the single keyframe representation by almost a minute, see Figure 2. In terms of Mean Inverted Rank (see Figure 3), we also observe that users of the multiple keyframe system perform better finding more known-items. Overall, our results rank in about average for the multi-keyframe representation to the bottom for the single keyframe representation. We see from this that multiple keyframes appear to perform significantly better than the single keyframe representations and that the classifiers do appear help, given the decrease in performance this year when classifiers were not included.

4 Multimedia Event Detection

Following the work of last year [2], we still perform the event detection as a fusion of multimodal sources and consideration as a machine learning problem in MED task of TRECVid 2012. In the Pre-Specific (PS) task, we consider event detection as a fusion of multimodal sources, including low-level features

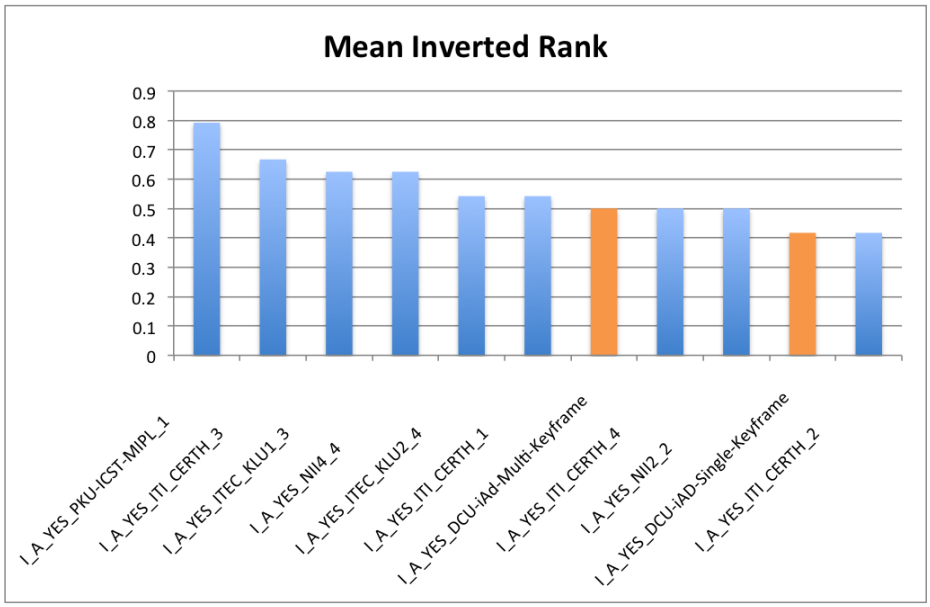

Fig. 3. Mean Inverted Rank of teams participating in TRECVid known-Item Task, our submission highlighted in Orange

and high-level semantic concepts. However, only low-level features are used in the Ad Hoc (AH) task. The flowchart for the PS task is shown in Figure 4

4.1 Low-Level Features Extraction

We consider both visual and audio features, and extract static and motion features in this work. Visual feature includes OpponentSIFT [5] feature extracted on keyframe level, 3D Histogram of Gradient (HOG3D) [7] extracted on pre-computed shot level. MFCC audio features are extracted on two scales, 3-second level and entire clip level.

The pre-defined semantic concepts are from the event kits provided by the organizers. In total, 52 visual concepts including objects, person, scene and human action and 7 audio concepts are selected for test events following the method in [8].

4.2 Representation Construction And Event Classification

After extracting the audio-visual low-level features, higher level feature representations are constructed at the video level. For OpponentSIFT and HoG3D Bag-of-visual-words (BoVW) representation, K-means are applied for constructing the visual vocabularies with 1024 words, respectively. For MFCC, GMM with

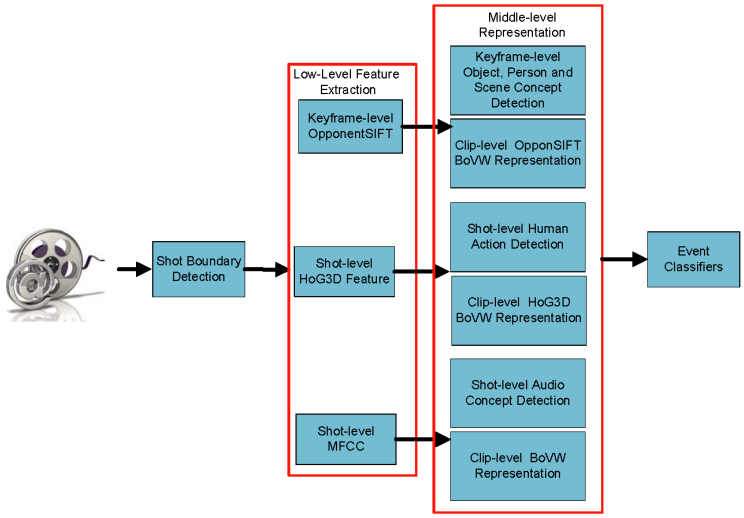

Fig. 4. The flowchart for the PS MED task

100 components is used on the entire clip level. Further, following the work in [9], 9 components are used on the 3-second level for audio concepts detection.

We use the *Maximum* operator for fusing the results of the concept detectors [8]. In order to train the concept detectors, we aggregated the keyframes extracted from the positive video samples as the training samples, so as to reduce manual annotation labour.

When training the event detectors, SVMs with χ^2 kernel are used as classifiers with 2207 (52+7+1024+1024+100) dimensional representation as feature input for the PS task, and with 2148 (1024+1024+100) dimensional for the AH task. This year, we only submit one primary run for PS and AH task respectively.

4.3 Results and Analysis

In order to protect the progress test set, limited evaluation results are released. Here, we report the results based on Missed Detection (PMiss) errors and False Alarm (PFa) errors. Comparison of our results with the Best, Median and Worst results released for PS task and AH task are shown in Table 3 and Table 4.

Table 3. Comparison our results with the Best, Median and Worst results released for PS task

	Best	**Median**	**Worst**	**Ours**
PFa	0.0009	0.0269	0.1556	0.1246
PMiss	0.2113	0.3537	0.8980	0.5788

Table 4. Comparison our results with the Best, Median and Worst results released for AH task

	Best	**Median**	**Worst**	**Ours**
PFa	0.0000	0.0327	0.6702	0.3952
PMiss	0.2004	0.3134	1.000	0.4004

For both the PS and AH task, our run reports worse false alarm (PFa) than most other runs, which may attribute to the performance of the detection framework. Furthermore, in order to get higher recall, we also reported a detection threshold (DT) which would introduce larger PFa, but at a lower rate of missed detection (PMiss).

Comparing our results for PS and AH task, we can conclude that introducing semantic concepts for event detection results in a significant reduction in the false alarm error rate, even the performance of semantic concept detection is far from perfect. This suggests that there are significant opportunities for improvements available.

5 Multimedia Event Recounting

This year, TRECVid initiated a new task called Multimedia Event Recounting (MER) – given an event kit, and a video clip that contains the event, the task is to produce a textual summary of the key evidence of the event. We conducted the MER task only on the six test video clips for each of the five MER events because of computation limits. Figure 5 illustrates the recounting process.

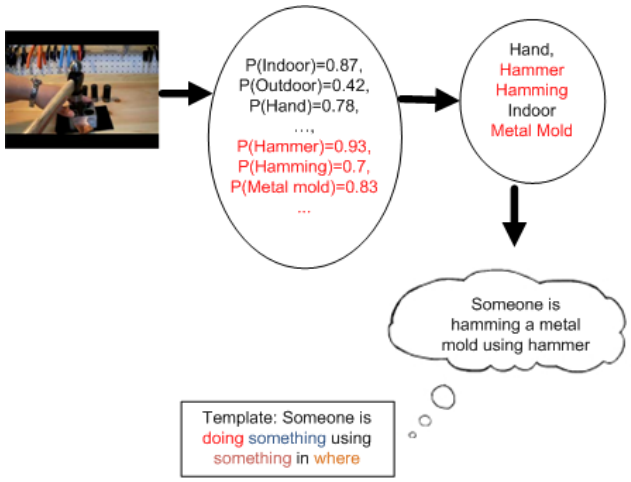

Fig. 5. Illustration of the MER process.

The sub-steps in our MER process include 1) semantic concept selection, 2) shot-level based concept and object detection, 3) semantic concept filtering and 4) template-based textual recounting. In the selection of the semantic concepts, we balanced the generic concepts and the event-specific concepts. The semantic concepts, including person, scene, and action concepts, are from the event kits provided by the organizera following the method outlined in [8]. Concept detection is similar to MED task. Furthermore, we also performed the object detection using the Object Bank [10].

In order to filtering the semantic concepts and objects, we consider the semantic relation hyponymy, meronym/holonymy using the co-occurrence information trained from the training set. Finally, textual recountings are constructed according to the predefined templates.

6 Conclusions

This year, our team participated in four tasks: INS, KIS and MED and MER. Since this was our first participation in the INS task, the high dimensional bag of visual words representation and vector space model approach was adopted. As the results indicated, limited performance has been achieved in this. In the Known-item search task we implemented two search systems based on a visualization of single vs multiple keyframe representations: We showed that the method based on a multiple keyframe approach performed better with our group of eight users. This year we chose not to use concepts in our evaluation, relying solely on the meta-index for our searches, though this left us with a median placing. In the pre-specific MED task, we followed the work of last year, but added an audio module. For the Ad Hoc MED task, we only consider the audio-visual features. Released results show that large false alarm errors are reported, but better recall figures are obtained. Summarizing, there are significant improvements that can be done in the framework. Results show that introducing semantic concepts for event detection results in the false alarm error rate reducing drastically, even if the performance of semantic concept detection is far from perfect. Finally, in the MER task, we used the selected semantic concepts and objects as the middle-level semantic "intermedia" and the template-based method to generate the textual description.

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. 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)

4. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision **65**(1/2) (2005) 43–72
5. 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
6. Lucene: The Lucene search engine (2005)
7. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: British Machine Vision Conference. (sep 2008) 995–1004
8. Guo, J., Scott, D., Hopfgartner, F., Gurrin, C.: Detecting complex events in user-generated video using concept classifiers. In: CBMI. (2012) 1–6
9. Guo, J., Gurrin, C.: Short user-generated videos classification using accompanied audio categories. In: Teh First ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis (AMVA). (2012)
10. Li, L.J., Su, H., Xing, E.P., Li, F.F.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS. (2010) 1378–1386