VIREO-TNO @ TRECVID 2014:
Multimedia Event Detection and Recounting (MED and MER)

Chong-Wah Ngo[†], Yi-Jie Lu[†], Hao Zhang[†], Ting Yao[†], Chun-Chet Tan[†], Lei Pang[†], Maaike de Boer[‡∓], John Schavemaker[‡], Klamer Schutte[‡], Wessel Kraaij[‡∓]

[†] *Video Retrieval Group (VIREO), City University of Hong Kong*
http://vireo.cs.cityu.edu.hk

[‡] *Netherlands Organization for Applied Scientific Research (TNO), Netherlands*
[∓] *University of Nijmegen, Netherlands*

Abstract

This paper presents an overview and comparative analysis of our systems designed for TRECVID 2014 [1] multimedia event detection (MED) and recounting (MER) tasks, including all sub-tasks for *Pre-Specified* (PS) event detection, all sub-tasks except 100Ex for *Ad-Hoc* (AH) event detection, and 010Ex sub-task for both PS and AH event recounting.

Multimedia Event Detection (MED):

Our main focus for the MED task is on the study of a new zero-example system, which aims to solve the 000Ex and SQ problems. The system can run either fully automatically or semi-automatically. Specifically, we test the automatic run in 000Ex submission and the semi-automatic run in SQ submission. Our 7 runs are summarized below:

- MED14Full_PS_000Ex: Zero-example system with automatic semantic query generation and OCR matching.
- MED14Full_PS_SQ: Zero-example system with semi-automatic semantic query generation.
- MED14Full_PS_010Ex: Full system consists of visual system 010Ex, zero-example system and OCR system. Late fusion is used to combine classifier scores from multi-SVMs and zero-example system. The final score of a video is adjusted based on OCR matching.
- MED14Full_PS_100Ex: Full system consists of visual system 100Ex, audio system and OCR system. Late fusion is used to combine classifier scores from multi-SVMs. Same as MED14Full_PS_010Ex, video list is refined by OCR system.
- MED14Full_AH_000Ex: System design is the same as MED14Full_PS_000Ex.
- MED14Full_AH_SQ: System design is the same as MED14Full_PS_SQ.
- MED14Full_AH_010Ex: System design is the same as MED14Full_PS_010Ex.

Multimedia Event Recounting (MER):

For the 010Ex sub-tasks of both PS and AH event recounting, we design and implement a simple but effective system to optimize the concept-to-event relevance, evidence diversity and timing of evidential shots. We show that good performance can be achieved by only selecting the three shots with the highest confidence.

1 Multimedia Event Detection

1.1 System Overview

In TRECVID 2014, our MED system consists of 4 sub-systems which are built on different features. In the following, we will first describe all the used features, followed by presenting our 4 sub-systems: visual system, audio system, zero-example system and OCR system.

1.1.1 Features

In our MED system, we first decompose each video into two-granularity levels – keyframe level and shot level in a sense that different features require different granularity levels. For example, DCNN7 features are extracted from video frames, whereas improved dense trajectories are extracted from video shots. Another reason is that event evidences are required to be located at shot level for MER system. The keyframe sampling rate is set to be one frame per two seconds and the time duration of shot is set to be five seconds. For each video, we use different methods to generate feature vector. All the features used in our MED system are summarized in Table 1 and the detailed descriptions are given below:

	Visual Features	**Audio Feature** [2]
Low-level Features	DCNN7 [3][4][5] Improved Dense Trajectory [6]	MFCC, LSF, OBSI, LPC
High-level Features	Semantic Indexing Concepts (SIN14.346) [1] Research Collection Concepts (Research.Collection.497) ImageNet 1000 Categories [3] (ImageNet.ILSVRC12.1000)	
Text Features	Optical Character Recognition (OCR)	

Table 1: Features used for MED'14 system

- *DCNN7 and ImageNet.ILSVRC12.1000*
  Recently, deep convolutional neural networks (DCNN) have demonstrated their potential for learning image representation and classifiers simultaneously with a large number of training instances. Inspired by the success of DCNN, we use it to generate visual representations as visual features in our system. We use the same DCNN architecture proposed by G. Hinton in [4]. Specifically, the used DCNN architecture can be denoted as $Image - C48 - P - N - C128 - P - N - C192 - C192 - C128 - P - F4096 - F4096 - F1000$, which contains five convolutional layers (denoted by C following the number of filters) while the last three are fully-connected layers (denoted by F following the number of filters) ; the max-pooling layers (denoted by P) follow the first, second and fifth convolutional layers; local contrast normalization layers (denoted by N) follow the first and second max-pooling layers. The parameters of DCNN are learnt on ILSVRC-2012 [7], which

is a subset of ImageNet dataset with 1.26 million training images from 1,000 categories. For each keyframe, its representations are the neuronal responses of layer 7 (F4096), and layer 8 (F1000) by feeding the keyframe into the learnt DCNN. We use average pooling to fuse the two kinds of DCNN features for all the keyframes of one video to form the video-level feature vector respectively. For simplicity, the two features are named as DCNN7 and ImageNet.ILSVRC12.1000, respectively.

- *Improved Dense Trajectory* [6]
  We extract the state-of-the-art motion feature – improved dense trajectory – at shot level. Specifically, trajectory feature, histogram of oriented gradients (HOG), histogram of flow (HOF), and motion boundary histogram (MBH) are computed for each trajectory obtained by tracking points in video shots. We separately reduce the dimension of trajectory, HOG, HOF and MBH descriptors by a factor of two using Principal Component Analysis (PCA) and then concatenate them into one raw feature vector. After that, Fisher vector encoding is used to quantize the raw features and create a shot representation. For each video, the feature representation is obtained by average pooling the Fisher vectors of video shots. Finally, L2 normalization is applied to generate the video representation.

- *SIN14.346* [8]
  346 concept detectors trained on SIN'14 dataset are used to predict on all video keyframes. By concatenating 346 concept detectors' responses, each keyframe is represented by a 346 dimension feature vector. Then, a video level concept representation is obtained by average pooling the responses of keyframes. We name this feature as SIN14.346.

- *Research.Collection.497*
  Similar to [9], we select 497 concepts from the MED'14 Research Collection dataset [10], manually annotate at most 200 positive keyframes for each concept, and train 497 concept detectors using SVM. Similar, we concatenate the responses of 497 concept classifiers on each keyframe and further pool the keyframe level feature vectors to form a video feature representation. It is named Research.Collection.497.

- *Audio Features*
  For audio features, the following 12 features are used: line spectral frequency (LSF), octave band signal intensity (OBSI), line predictor coefficients (LPC), MFCC and their first and second derivatives. We extract the audio features from the audio signals and quantize them into BoWs. After that, a Fisher vector is used to encode the audio features without dimensionality reduction.

- *OCR*
  We use Tesseract OCR [11] to extract OCR (English) from video keyframes.

1.1.2 Visual System

We develop visual systems for 010Ex and 100Ex task separately.

- *Visual System 010Ex*
  SIN14.346, Research.Collection.497, ImageNet.ILSVRC12.1000 features and DCNN7 features are used in visual system 010Ex. Specifically, we concatenate SIN14.346, Research.Collection.497 and ImageNet.ILSVRC12.1000 features to one feature vector, and then train event classifiers using Chi-Square SVM. For DCNN7, event classifiers are trained using Chi-Square SVM as well. Average fusion is used to directly combine classifier scores of multiple SVMs described above.

- *Visual System 100Ex*

  Besides the features used in visual system 010Ex, the improved dense trajectory feature is also involved in visual system 100Ex. The setting of classifier training is the same as that used in visual system 010Ex on the common features. In addition, event classifiers are trained using linear SVM on the improved dense trajectory feature. Finally, fusion by standard averaging is used to combine prediction scores from multiple SVMs described above.

1.1.3 Audio System

For the the audio feature, event classifiers are trained using linear SVM.

1.1.4 Zero-Example System

Our zero-example system aims to pick up the event-relevant concepts in concept store by doing textual mapping between event kits and concept names. It also scores the selected concepts by their concept-to-event similarities. Therefore, given the event query and responses of concept detectors, the system can perform event search without the need of training examples.

In preprocessing, we generate a metadata store with 1843 concepts in total, which are collected from *SIN14.346*, *Research.Collection.497* and *ImageNet.ILSVRC12.1000* [3]. In addition, 5784 documents are collected from Wikipedia and indexed in the metadata store for measuring the *inverse document frequency* (IDF). These documents are collected by querying Wikipedia for all stemmed nouns and verbs and combination of nouns and verbs from the snippet texts in the MED'14 Research Collection dataset.

On the other hand, the keyframes are extracted uniformly from the video at the rate of one frame every two seconds. The detection responses of all 1843 concepts are predicted on all the keyframes of test videos. Then, max pooling is used to fuse the responses of keyframes from an identical video and generate a 1843 dimensional video representation. Each dimension denotes the confidence score of the corresponding concept detector for the video with respect to the dataset from which the detector is learned.

There are two main steps in the zero-example system: semantic query generation and event search.

A. Automatic Semantic Query Generation

In the query phase, our system takes the event description from each event kit as query. The queries are then parsed by Stanford CoreNLP parser [12] which analyzes the structure of sentences in terms of phrases and verbs. We excluded the event explications as complex sentences are too difficult to be well parsed. We also enforce lemmatization on both queries and concept names. The following steps are subsequently executed to generate the semantic query per event:

- Concepts are automatically selected by performing exact word-by-word matching between concept names and the query. These selected concepts are weighted by considering the term frequency in the query, term IDF, term relevance to the query, and term specificity. We refer to the weight as the importance of the concept. Specifically, given a concept name $\boldsymbol{c}_i$ and a phrase from the query $\boldsymbol{q}_j$, we denote the common words between them as $\boldsymbol{u}$. A similarity matrix $\boldsymbol{S}$ can be obtained by calculating the similarity for each $(\boldsymbol{c}_i, \boldsymbol{q}_j)$ pair, given

  $$sim(\boldsymbol{c}_i, \boldsymbol{q}_j) = \max(\boldsymbol{t}_{\boldsymbol{u}} \times \log \boldsymbol{s}_{\boldsymbol{u}}) \tag{1}$$

  where $\boldsymbol{t}_{\boldsymbol{u}}$ is the TF-IDF weights, in which TF (term frequency) represents the frequency appearance of words $\boldsymbol{u}$ in the text query, and IDF (inverse document frequency) is estimated based on a collection of Wikipedia pages downloaded from the Web. $\boldsymbol{s_u}$ is the word specificity vector defined by the minimum depths of words $\boldsymbol{u}$ in WordNet hierarchy. Then, for a phrase $\boldsymbol{q}_j$, we only retain the concepts with the maximum similarity to the phrase. This similarity is considered as the importance to the query.

- All the selected concepts are ranked by the importance. The top 8 concepts are picked up for each event. The concepts ranked after the 8th are picked as well if their importance is equal to the importance of the 8th concept. This would normally result in 8 - 10 concepts finally chosen for each event. These chosen concepts form the semantic query automatically generated by our system.

B. Semi-automatic Semantic Query Generation

In the query phase, human subjects can be involved to refine the automatically generated semantic query. Basically, the zero-example system will recommend more than 30 concepts automatically for a human subject to perform concept screening. Along this process, noisy concepts are expected to be removed, leaving only relevant concepts. This would usually result in around 10 concepts chosen for each event.

C. Event Search

In the search phase, for each event, we simply rank all the test videos based on the weighted sum of classifiers' confidence scores and the concept importance.

1.1.5 OCR System

The OCR system extracts text from video frames using the Tesseract OCR engine [11]. The engine is applied in a brute-force manner with post processing on the resulting extracted texts. From a video segment, every key video frame (roughly one per second) is decoded using the FFmpeg open-source library. All key frames are then fed to the Tesseract engine for OCR. The resulting texts are analysed to check whether meaningful text is extracted from the video frame. This post-processing checks the words in the extracted text per frame using the following rules:

1. every word has at least 3 characters
2. every word should have at least one vowel
3. every word should match with US-English dictionary (using Python Enchant spell-checking library)

Words that abide to all conditions are kept.

Keyword matching is performed between the keywords selected from event description and OCR extracted from videos. The matching score is calculated based on TF-IDF.

1.2 MED Results and Analysis

1.2.1 Feature Comparison

In this section, we present the individual feature performance on MED14-Test. As the experimental results shown in Table 2 and Table 3, we can easily observe that DCNN7 feature always gets the highest MAP in both 010Ex and 100Ex among all the low level features, indicating the effectiveness of DCNN architecture. Furthermore, we also observe that the DCNN7 feature outperforms concept-level feature when the number of concepts is small, e.g., 497 concepts of Research.Collection, 346 concepts of SIN14,

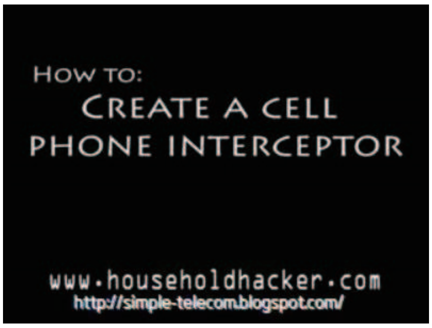


Figure 1: OCR text extracted: "HOWTO: CREATE A CELL PHONE INTERCEPTOR uwu householdhacker can "

Feature	MAP	Feature	MAP
Audio	0.042997	Research.Collection.497	0.163506
Improved Dense Trajectory	0.134402	SIN14.346	0.164958
DCNN7	**0.225104**	ImageNet.ILSVRC12.1000	0.19521
		All concept sets (Research.Collection.497 SIN14.346, ImageNet.ILSVRC12.1000)	**0.242445**

Table 2: MED PS_100Ex: Mean AP of single feature on MED14-Test

Feature	MAP	Feature	MAP
Audio	0.007861	Research.Collection.497	0.087166
Improved Dense Trajectory	0.03762	SIN14.346	0.086318
DCNN7	**0.114501**	ImageNet.ILSVRC12.1000	0.106746
		All concept sets (Research.Collection.497 SIN14.346, ImageNet.ILSVRC12.1000)	**0.129228**

Table 3: MED PS_010Ex: Mean AP of single feature on MED14-Test

1000 concepts of ImageNet.ILSVRC12. However, when we combine all these concepts and create a large vocabulary with 1843 concepts, concept-level feature (from all concept sets) can outperform DCNN7 feature, which basically indicates the advantage of exploiting a large number of concepts.

1.2.2 010Ex/100Ex Performance

Table 4 shows the performance comparison of each sub-system on Pre-Specified 100Ex and Pre-Specified 010Ex. An interesting observation is that visual system 100Ex by additionally utilizing improved dense trajectory on PS_100Ex sub-task leads to better performance. In contrast, the performance of visual system 100Ex decreased on PS_010Ex sub-task. Since audio system has a very low MAP on PS_010Ex, we don't involve audio system in full system for PS_010Ex.

For PS_100Ex, we evaluate our full system consisting visual system 100Ex, audio system and OCR system on MED14-Test and the MAP can achieve 0.3020. For PS_010Ex, our full system consists of visual system 010Ex, zero-example system and OCR system and the MAP can reach 0.155831 on MED14-Test. The observation is similar to [13], which also states that the performance of PS_010Ex is approximately

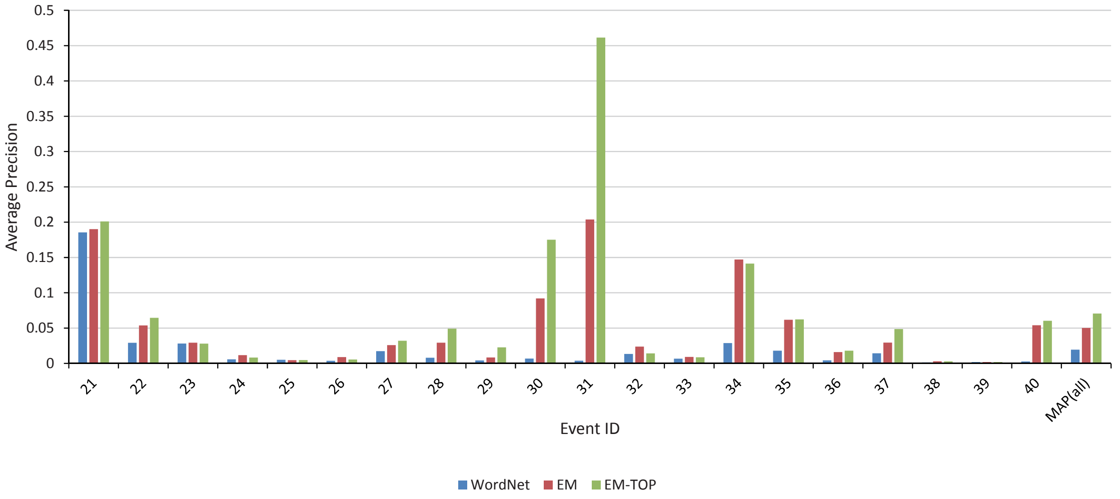


Figure 2: Performance comparison on MED14-Test among different methods for zero-example system

half of that of PS_100Ex.

	Visual System 010Ex	Visual System 100Ex	Audio System
PS-100Ex	0.2563	**0.2588**	0.0430
PS-010Ex	**0.1369**	0.1251	0.0079

Table 4: MAP of sub systems for PS-100Ex and PS-010Ex on MED14-Test

1.2.3 Zero-Example Performance

For zero-example system, we mainly focus on solving the main difficulty – how to accurately and reliably select concepts from the concept store and score their relevance?

In selecting the concepts, a simple and straightforward mapping method is the exact matching, which requires at least one term in a concept name matches with the event description. We also test the ontology-based mapping using WordNet and ConceptNet. This is based on the intuition that the ontology-based mapping would be helpful especially when there were only a small number of exactly matched concepts. However, in practice, WordNet/ConceptNet fails terribly in MED.

Figure 2 shows the performance comparison between exact maching (EM) and WordNet mapping (WordNet) on MED14-Test set. We use the WUP similarity as a typical method of WordNet mapping. For most events, EM significantly outperforms WordNet. Some are more than 10 times better. The mean average precision (MAP) over all events confirms this observation: EM achieves 5.0% of MAP versus 1.9% of MAP by WordNet. We don't include the performance of ConceptNet here because ConceptNet shows result similar to the WordNet. Compared to the more sophisticated mapping, the result demonstrates a clear advantage of the exact matching, which is by far much simpler.

We attribute this result to two causes: First, in our experiment on MED14-Test, on average there are 48 concepts, ranging from 20 to 110, that can be exactly matched to an event. We claim that the number is enough for describing an event. Second, the concepts discovered by WordNet/ConceptNet are mostly noisy or irrelevant to the event, the merely useful ones being the synonyms. Blindly matching words using WordNet or ConceptNet without the event context can easily divert to wrong concepts. For

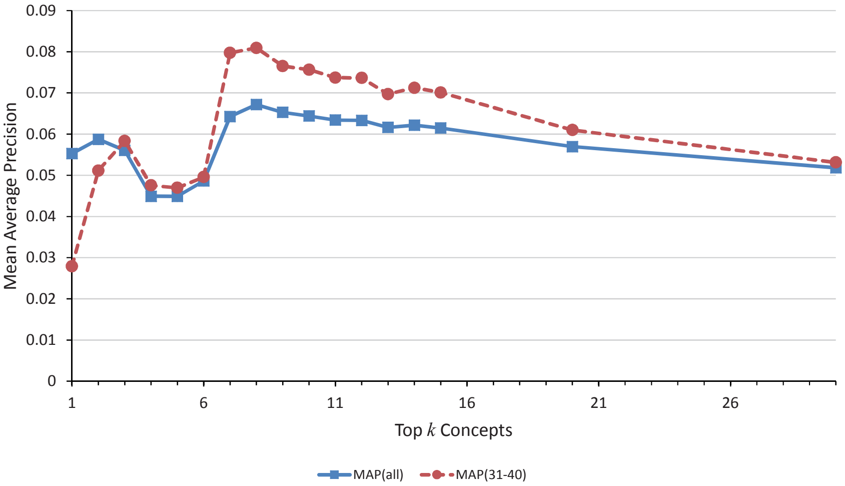


Figure 3: Top k concept selection for exact matching on MED14-Test

example, the concept "dog" in dog show is – according to WordNet – correlated to cat, horse and other animals which are not relevant to the dog show.

In addition, we also observe that the performance does not depend on the number of concepts being selected. Normally only the first few (around 8) can work very well. Selecting more concepts will only degrade the performance. We do experiment by choosing only the top few concepts using exact matching and then examining the MAP. As shown in Figure 3, both the overall MAP and MAP on last year's AH events comply with this conclusion.

Furthermore, we implemented query expansion methods using ConceptNet[14] and Wikipedia. The difference between this method and the previously mentioned ConceptNet method is that we apply this method on a different level in the process. Whereas the previous method is applied on the level of matching a word, found in the textual description, to the concept detector, this method used ConceptNet to find related concepts instead of the textual description. Within ConceptNet 5[14] all *IsA* relations to the event name are selected and used to expand our concept selection. The event name is also used as input for Wikipedia. Results show that both query expansion methods have no higher performance than the method using the textual description with EM. If the results from ConceptNet and Wikipedia are fused with the current Semantic Query Generation method, performance does not increase significantly. This finding supports our previous findings that concepts discovered from external web sources may be noisy or irrelevant because they do not include the context of the event.

Aside from these findings, we are experimenting with adding temporal relations between two concepts. The temporal relations between images or frames might be of great value, because it distinguishes video from a set of images. In procedural events such as *doing a bike trick* temporal relations might be very helpful in distinguishing between the event and a false positive. In this research the temporal relations *before*, *immediately before* and *while* are used, but the results have not yet shown a significant improvement in performance. Temporal relations are, therefore, not included in the system.

We implement the exact matching as a good baseline for the zero-example system. By only selecting the top 8 to 10 concepts for each event, combined altogether with the term frequency in the query, term IDF, term relevance to the query, and term specificity, we achieve the best zero-example MAP 7.0% on MED14-Test set compared to other combinations or without top concept selection. The EM-TOP in Figure 2 shows the performance of our final system.

1.2.4 Threshold Learning

For Pre-Specified evaluation, in order to calculate metric Minimum Acceptable Recall R0, a threshold score T_q is required for every event. For an event, positive and negative videos should theoretically have different distributions in view of their prediction scores. Similar to [13], we applied maximum entropy theory to set the threshold. Based on our experiments, maximum entropy theory is able to set appropriate threshold which helps to produce a large R_0. The threshold can be obtained by maximize the following formula.

$$H(T_q) = -\sum_{x_i=0}^{T_q} C(x_i) log \left(\frac{C(x_i)}{\sum_{x_i=0}^{T_q} C(x_i)} \right) - \sum_{x_i=T_q}^{1} C(x_i) log \left(\frac{C(x_i)}{\sum_{x_i=T_q}^{1} C(x_i)} \right)$$

x_i is the prediction score of a video, $C(x_i)$ is the count number of x_i. Table 5 shows our R0 results for MED14-Test.

	PS_100Ex	**PS_010Ex**
R_0	0.4615	0.2089

Table 5: R_0 results for PS_100Ex and PS_010Ex on MED14-Test

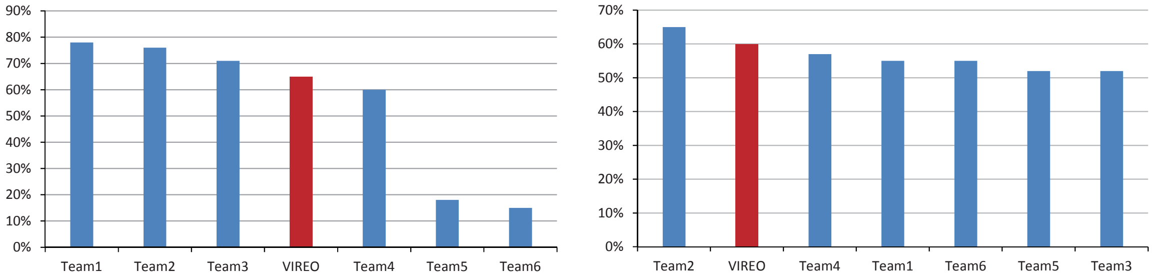
2 Multimedia Event Recounting

2.1 System Design

The VIREO's MER system mainly utilizes the semantic query generated for 010Ex task as input and takes advantage of 1843 concept indexing for evidence localization. The key criterions to be considered here are the importance of evidence and user experience. Specifically, how fast a user can declare a video is relevant by reading the presented evidences? General speaking, the general goal is to display the minimal amount of evidences in the shortest possible time. To optimize both criterions, we will consider three aspects of information: (1) concept-to-event relevance which prioritizes the importance of concepts to events, (2) evidence diversity which avoids redundancy by suggesting evidences of diverse content, and (3) the shorter the better which recommends only snippets that are just sufficient and necessary to evidence the presence of event.

The concept-to-event relevance is the most important concern in our system design, given that the key evidence should not only reflect the existence of chosen concepts, but also its relevance regarding the event. Therefore, the constraints are two-fold: On the one hand, certain concepts listed in the semantic query should appear in the evidential shot; on the other hand, the content of the shot should be event-relevant. Simply achieving only one of the two would be biased. For example, an evidential shot showing a boy playing a ball is insufficient to support the event "playing fetch" even though the important concept *ball* shows up in high confidence. In addition, this scenario is very much associated to the new MER workstation interface: Judges will be asked to score on both aspects, one being the concept-snippet correctness, the other being the snippet-event correctness.

To approach a high concept-to-event relevance, we first select candidate shots that are most relevant to an event. The selection is based on the weighted sum of concept-to-event similarities and concept detectors' responses for all concepts in the semantic query. Therefore, the concept-to-event relevance is



(a) Query conciseness

(b) Key evidence convincing

Figure 4: VIREO MER system performance (percentage of agree and strongly agree) compared to other teams.

constrained by votes from all event-relevant concepts rather than a single concept. Then, in the second step, we do concept-to-shot alignment which associates the most important concept to the candidate shot as the representative concept.

In optimizing the evidence diversity and viewing time, our recent study on MED 2012 training set shows that there are more than 50% of five-second shots, among the positive videos, can be considered as evidence. Randomly picking any three of the shots from a positive video can already achieve more than 60% of precision. As a result, basically we only select three most confident shots per positive video as key evidence, leaving the rest shots as non-key evidence. On the other hand, as we observe in MED 2012 training set that the evidential shots are highly redundant, we also enforce the criterion that the selected evidential concepts should be diverse. Specifically, in concept-to-shot alignment, we recount each shot with a unique concept different from other shots.

2.2 MER Results

We submitted both PS and AH event recountings of 010Ex task. Despite the simplicity of our strategy, the system works pretty well. The performance is showing in Figure 4. Basically, we achieve the second place among all teams concerning the quality of key evidence, and our query conciseness lies in good range as well. There are 28% (27%) of judges strongly agreeing that our submitted queries (key evidence) are concise (convincing), which ranks our system at the 1st position among all teams. Currently, we only use the visual concept detectors. The evidential shots are inferred from the candidate keyframes. But our system can be easily scaled to other types of detectors, e.g. motion-based concept detectors which indicate the window of time more precisely, and audio concept detectors which are the essential complements to the visual concept detectors.

3 Summary

For MED with training examples, we use the DCNN feature, concept feature, improved dense trajectory feature and audio feature to train event classifiers and draw two conclusions. The DCNN feature outperforms human-defined low level features, and the concept feature outperforms the DCNN feature when the vocabulary size of concepts is increased to 1843. Reviewing the sub-systems' performances for PS_010Ex and PS_100Ex on MED14-Test, we find that improved dense trajectory brings very little or no improvement (for PS_010Ex, involving improved dense trajectory even reduces MAP of MED14-Test), this may be due to a hidden technical bug.

For MED with no training example, we build a large concept store with 1843 concepts and focus on solving the difficulty of how to accurately and reliably select concepts and score their relevance.

Compared to the exact matching, we share the insights of why ontology-based mapping using WordNet and ConceptNet fails terribly in MED. In the end, we come up with a system which only chooses a few top concepts by exact matching. With several other improvements on scoring the concepts, we achieve a reasonable overall performance, which can be treated as a good baseline.

For MER, we implement a baseline system which makes use of 1843 concept indexing for evidence localization. Our strategy for system design, from the most important to least, is to optimize the concept-to-event relevance, evidence diversity and viewing of evidential shots, which aligns to what the new MER workstation interface wants to feature this year. Based on our recent study on MED 2012 training set, for each video, our system selects only three most confident shots as key evidence, and aligns the concepts diversely for each selected shot. This scheme, although simple, has demonstrated its effectiveness in event recounting.

Acknowledgment

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 120213).

References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot, "Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[2] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "YAAFE, an easy to use and efficient audio feature extraction software," *The International Society for Music Inforamtion Retrieval*, 2010.

[3] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," 2013. [Online]. Available: http://caffe.berkeleyvision.org/

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutionalneural networks," in *Advances in Neural Information Processing System*, 2012.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[6] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *International Conference on Computer Vision*, 2013.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierachical image database," in *Computer Vision and Pattern Recognition*, 2009.

[8] W. Zhang, H. Zhang, T. Yao, Y. Lu, J. Chen, and C.-W. Ngo, "VIREO @ TRECVID 2014: instance search and semantic indexing," in *NIST TRECVID workshop*, 2014.

[9] P. Natarajan, S. Wu, F. Luisier, X. Zhuang, and M. Tickoo, "BBN VISER TRECVID 2013 multimedia event detection and multimedia event recounting systems," in *NIST TRECVID workshop*, 2013.

[10] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel, "Creating HAVIC: Heterogeneous audio visual internet collection," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. DoAYan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[11] R. Smith, "An overview of the tesseract ocr engine," in *International Conference on Document Analysis and Recognition*, 2007.

[12] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, “The Stanford CoreNLP natural language processing toolkit,” in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[13] Z. Lan, L. Jiang, and A. Hauptmann, “CMU-Informedia @ TRECVID 2013 multimedia event detection,” in *NIST TRECVID workshop*, 2013.

[14] R. Speer and C. Havasi, “Representing general relational knowledge in conceptnet 5.” in *LREC*, 2012, pp. 3679–3686.