VIREO-TNO @ TRECVID 2015: Multimedia Event Detection

Hao Zhang[†], Yi-Jie Lu[†], Maaike de Boer [†‡∓], Frank ter Haar[‡], Zhaofan Qiu[†]
Klamer Schutte[‡], Wessel Kraaij[‡∓], Chong-Wah Ngo[†]

[†] *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 the TRECVID 2015 [1] multimedia event detection (MED) task. We submitted 17 runs, of which 5 each for the zero-example, 10-example and 100-example subtasks for the *Pre-Specified* (PS) event detection and 2 runs for the 10-example subtask for the *Ad-Hoc* (AH) event detection. We did not participate in the Interactive Run. This year we focus on three different parts of the MED task: 1) extending the size of our concept bank and combining it with improved dense trajectories; 2) exploring strategies for semantic query generation (SQG); and 3) combining our visual classifiers with audio and/or textual classifiers. Among our 17 submitted runs, the following runs achieved top performances:

- VIREO_MED15_MED15EvalFull_PS_0Ex_MED_p-manualfused_1: zero-example system with manual SQG, fused with textual (OCR) and speech (ASR) information.
- VIREO_MED15_MED15EvalFull_PS_10Ex_MED_p-ConceptBankIDTEK0OCR_1: 10-example system using our Concept-Bank feature fused with the improved dense trajectories and the 0Ex manual visual system and OCR.
- VIREO_MED15_MED15EvalFull_PS_100Ex_MED_c-ConceptBankIDTJointProb_1: 100-example system using Concept-Bank feature fused with the improved dense trajectories, using the joint probability to create the score.

1 Introduction

The main difference between TRECVID MED 2014 and 2015 is that we were allowed to provide five runs in the *Pre-Specified* (PS) case and two runs in the *Ad-Hoc* (AH) case. With these comparison runs, we may reasonably evaluate how each component of our system contributes to the result. Furthermore, speech recognition information was provided by LIMSI [2]. We used these changes to conduct research that focuses on 1) extending the size of our concept bank and combining it with improved dense trajectories; 2) exploring strategies for semantic query generation (SQG); and 3) combining our visual classifiers with audio and/or textual classifiers. The next section explains how we use different modalities in our system. Section 3 describes our systems and runs for the different amount of training examples, i.e., zero, ten or hundred.

2 Modalities

In our system we use three type of modalities: visual, textual and speech.

2.1 Visual

For our visual system, we first decompose each video into two-granularity levels – keyframe level and shot level. 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 generate 6 kinds of high-level concept feature and 1 kind of low-level motion feature. All the features used in our visual system are summarized as below:

- *ImageNet_1000*
  We use the same DCNN architecture proposed by G. Hinton in [3]. Specifically, the used DCNN architecture can be denoted as $Image - C48 - P - N - C128 - P - N - C192 - C192 - C128 - P - F4096 - F4096 - F1000$, in which C are the convolutional layers followed by the number of filters, F are the fully-connected layers, P are the max-pooling layers and N are the local contrast normalization layers. The parameters of DCNN are learnt on ILSVRC-2012 [4], which is a subset of ImageNet dataset with 1.26 million training images from 1,000 categories. The neural responses of the eight layer (F1000) for each keyframe are average pooled to form the video-level feature vector.

- *SIN_346* [5]
  A set of 346 concept detectors fine-tuned with a DCNN structure on the TRECVID SIN 2014 dataset is applied on all keyframes and average pooled to a video level representation.

- *RC_497*
  Similar to [6], we select 497 concepts from the MED'14 Research Collection dataset [7]. We manually annotate at most 200 positive keyframes for each concept and fine-tuned 497 concept detectors using DCNN architecture. As with the previous methods, we concatenate the responses of the concept classifiers on each keyframe and average pooled the resulting feature vectors to form a video feature representation.

- *Places_205*
  Multimedia events usually happen in notable sceneries, e.g., "Bike Trick" contains outdoor backgrounds such as "Park" or "Road" scenes, whereas "Cleaning an Appliance" contains indoor scenes such as "kitchen". To capture the scenery information for multimedia events, we fine-tune 205 scene categories on the MIT places dataset [8] using a DCNN architecture. The scenery concept responses are extracted on each keyframe and further pooled to form video-level representation.

- *FCVID_239*
  To capture the action/motion information of multimedia events, 239 concept detectors are trained with an SVM on the FCVID [9] dataset, which contains 91,223 web videos. Since concepts from this dataset are mainly action/motion concepts annotated at video-level, we extract DCNN layer 7 [3] responses from frames and pool it across the temporal domain to generate training features and train video-level concept detectors. We concatenate the responses of 205 concept detectors of an image.

- *Sports_487*
  487 concept detectors are trained with 3D CNN structure [10] on the Sport-1M dataset [11] containing 1 million videos. Similar to FCVID, Sports-1M are mainly action/motion concepts annotated at video level.

To capture semantical meaning of multimedia event, we collect all the above concept features and form a large ***Concept-Bank*** with 2,774 semantical concepts related to visual objects, background scenery and actions.

- *Improved Dense Trajectory* [12]
  We extract the state-of-the-art motion feature - improved dense trajectory (IDT) - at video level. The IDTs contain the following features: histograms of oriented gradients (HOG), histograms of oriented flow (HOF), and motion boundary histograms (MBH). For each of these three features, we reduce the descriptor size by a factor of two using principal component analysis (PCA), and encode a fisher vector (FV) with the use of a pre-learned Gaussian Mixture Model (k=256). Then the fisher vectors are concatenated and normalized using the power norm and L2 norm, and subsequently classification is performed with a linear SVM.

2.2 Textual

Tesseract OCR[13] is used to extract text from video frames. 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 analyzed 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.

The results from Tesseract per video are fed into Lucene to index and search in the text. A manually defined Boolean query based on the event description and Wikipedia is used in combination with the term frequency to retrieve the positive videos.

2.3 Speech

For the speech information, we use the data provided by LIMSI [2]. A manually defined Boolean query in combination with a PhraseQuery is used to search for relevant speech information. The standard Dirichlet Language Model in Lucene is used to calculate the similarity between this manually defined query and the audio file of a video and used to rank the videos.

3 Systems per Subtask

In this section, we will explain the systems used for each of the subtasks as well as which runs we submitted.

3.1 Zero Example

For the 20 PS events, five runs are submitted for 0-Ex subtask:

- VIREO_MED15_MED15EvalFull_PS_0Ex_MED_c-automaticfused_1: zero-example system with automatic SQG, fused with textual (OCR) and speech (ASR) information.

- VIREO_MED15_MED15EvalFull_PS_0Ex_MED_c-manualvisual_1: zero-example system with manual SQG using only our visual classifiers.

- VIREO_MED15_MED15EvalFull_PS_0Ex_MED_c-word2vecfused_1: zero-example system with SQG by word2vec, fused with textual (OCR) and speech (ASR) information.

- VIREO_MED15_MED15EvalFull_PS_0Ex_MED_c-word2vecvisual_1: zero-example system with SQG by word2vec using only our visual classifiers.

- VIREO_MED15_MED15EvalFull_PS_0Ex_MED_p-manualfused_1: zero-example system with manual SQG, fused with textual (OCR) and speech (ASR) information (primary run for the 0-Ex case).

The following subsections explain the automatic SQG, manual SQG, Word2Vec method and fusion strategy in more detail.

3.1.1 Automatic SQG

Zero-example system aims to pick up and score the event-relevant concepts in concept bank by doing textual mapping between event kits and concept names. This year we follow our last year's pipeline [14]. Based on the findings last year, improvements are focused on the following two folds: 1) a large concept bank which has 2,774 concepts that cover a broad range of topics with varying granularity; 2) a strategy of how to wisely pick up the right concepts for each event.

An important observation we recall from last year is that by choosing only a few top concepts from the concept bank we can achieve higher performance compared to choosing more concepts as adding more concepts might increase the noise. This, however, does not indicate that enlarging the concept bank is useless, because a larger concept bank increases the chance of an accurate event-to-concept match.

Compared to the 1,843 concepts used in last year, this year we additionally include *Sports_487*, *FCVID_239*, and *Places_205*. In *Places_205* are detectors for scenes which can compensate the 1,843 concepts that for the majority consists of objects. *FCVID_239* contains detectors for high-level topics and events. *Sports_487* is the dataset proven to have good contribution by CMU's report in MED '14 [15] and a latter summarization on zero-example case [16]. We also observe a similar trend in performance after necessary edits to the concept names, e.g. "*charreada*" is a horse riding sport. To this end, the whole concept bank has much broader coverage with varied granularity, thus better potential in representing an event.

While increasing the size of concept bank helps in increasing the chance of hit, it also raises the risk of picking up noisy concepts. Therefore a good strategy is required to ensure we only select the right concepts for event representation. From manual SQG, we find that when there already exists a concept detector for detecting the whole event, it would be generally wise to only include concepts that are distinctive to this event. Take the event *playing fetch* as an example, if a detector exactly named "*playing fetch*" is found, simply adding the concept "*yard*" would degrade the performance as yard may also appear in many other events, which inevitably brings noise. On the contrary, if no detector for the whole event is found, it is beneficial to use a few accurate concepts to represent the event as concluded in our last year's report [14]. However, it's difficult to require the concepts to be event-distinctive without human knowledge. Hence, we simply ignore other concepts if a detector for the whole event is found for automatic SQG.

3.1.2 Manual SQG

In the query phase, human subjects can be involved to refine the automatically generated semantic query. For convenience, our manual SQG is based on the results of automatic SQG. Typically, the automatic SQG system will loosely recommend more than 30 concepts for a human subject to perform concept screening. Along this process, noisy concepts are expected to be removed, leaving only relevant concepts.

This year, thanks to the involvement of FCVID concepts, many of which have similar semantics of MED events, we basically follow the steps below for manual concept screening:

- Remove false positives by looking at the names of concepts;
- Remove concepts for which training videos appear in very different context based on human's common sense;
- Only carefully include concepts that are distinctive to this event if a concept detector with the same name of the event can be found.

3.1.3 Word2Vec

For the semantic query generation with Word2Vec we use the Gensim code and the pre-trained GoogleNews model [17] to find the similarity between the event name and each of our concept classifier labels. This similarity is used as our weight. This weight is used in combination with the following strategy to determine which classifiers to use:

- 1. Direct match: the classifier label has a similarity of 1.0 with our event name. We use only the direct matches;
- 2. Indirect match: for each word in the event name a direct match with a classifier label can be found, for example "win" and "race" in "winning a race". Only these indirect matches are used;
- 3. Top 3: if no indirect or direct matches can be found, we use the three classifier labels with the highest similarity to the event name.

The score of a video is the linear combination of the weighted concept detector output scores.

3.1.4 Fusion

In the fusion we combine the scores of the SQG method with the ASR and OCR output. Because the score distributions of the ASR and OCR outputs from Lucene and the visual detectors are different, it is hard to normalize. We, therefore, choose to implement a high precision OCR and ASR re-ranking method, so we could use the fusion to boost the events that are almost certainly positive to the top of the ranked list. This is done by adding the OCR and/or ASR scores of the retrieved videos to the scores of the visual system.

3.2 Ten Example

In the ten example Pre-Specified subtask we submitted five runs and in the Ad Hoc subtask we submitted two runs:

- VIREO_MED15_MED15EvalFull_PS_10Ex_MED_c-ConceptBank_1: 10-example system based on our Concept-Bank feature.

- VIREO_MED15_MED15EvalFull_PS_10Ex_MED_c-ConceptBankIDT_1: 10-example system using Concept-Bank feature fused with the improved dense trajectories.

- VIREO_MED15_MED15EvalFull_PS_10Ex_MED_c-ConceptBankIDTEK0_1: 10-example system using our Concept-Bank feature fused with the improved dense trajectories and the 0Ex manual visual system.

- VIREO_MED15_MED15EvalFull_PS_10Ex_MED_c-ConceptBankIDTEK0OCRASR_1: 10-example system using our Concept-Bank feature fused with the improved dense trajectories, the 0Ex manual visual system, ASR and OCR.

- VIREO_MED15_MED15EvalFull_PS_10Ex_MED_p-ConceptBankIDTEK0OCR_1: 10-example system using our Concept-Bank feature fused with the improved dense trajectories, the 0Ex manual visual system and OCR.

- VIREO_MED15_MED15EvalFull_AH_10Ex_MED_c-visual_1 : 10-example ad hoc system using our Concept-Bank feature fused with the improved dense trajectories.

- VIREO_MED15_MED15EvalFull_AH_10Ex_MED_p-visualtextual_1: 10-example ad hoc system using our Concept-Bank feature fused with the improved dense trajectories, the 0Ex manual visual system and OCR.

The following subsections explain the training of the visual classifiers, the combination of the Concept-Bank feature with the improved dense trajectories (IDT) and the fusion of the 10-Ex system with the 0Ex system. The fusion method with ASR and OCR is the same as explained in the previous section.

3.2.1 Visual Classifiers

The Concept-Bank features, i.e., ImageNet_1000, SIN_346, RC_497, Places_205, FCVID_239 and Sports_487, are first concatenated to one feature vector and then used to train an event classifier using Chi-Square SVM.

For the event classifiers based on the improved dense trajectory feature, we follow the standard pipeline as in [12] and train the classifier with linear SVM.

3.2.2 Fusion

To properly fuse results of different system, we empirically design three stages of fusion strategy for 10-Ex system summarized as below:

- Concept-Bank + IDT
  Average fusion is used to directly combine scores of Concept-Bank based SVM and IDT SVM. This output is named as Visual-System.

- Visual-System + 0-Ex
  Average fusion is used to combine output scores of Visual System and 0-Ex System. This output is named as Visual-Zero System.

- Visual-Zero + OCR/ASR
  We use similar strategy presented in section 3.1.4 to fuse OCR/ASR system with Visual-Zero system.

3.3 Hundred Example

In the hundred example Pre-Specified subtask we submitted five runs:

- VIREO_MED15_MED15EvalFull_PS_100Ex_MED_c-ConceptBank_1: 100-example system using only our Concept-Bank feature.
- VIREO_MED15_MED15EvalFull_PS_100Ex_MED_c-ConceptBankIDT_1: 100-example system using our Concept-Bank feature fused with the improved dense trajectories.
- VIREO_MED15_MED15EvalFull_PS_100Ex_MED_c-ConceptBankIDTJointProb_1: 100-example system using Concept-Bank feature fused with the improved dense trajectories, using the joint probability to create the score.
- VIREO_MED15_MED15EvalFull_PS_100Ex_MED_c-ConceptBankIDTOCRASR_1: 100-example system using both our Concept-Bank feature fused with the improved dense trajectories and the OCR and ASR information.
- VIREO_MED15_MED15EvalFull_PS_100Ex_MED_p-ConceptBankIDTOCR_1: 100-example system using both our Concept-Bank feature fused with the improved dense trajectories and the OCR information (primary run for the 100Ex).

The following subsection explain the score generation with joint probability. The other methods are already explained in the section of the ten examples.

3.3.1 Joint Probability

In Visual-System, the standard fusion strategy is averaging prediction scores obtained by different features, i.e. Concept-Bank and IDT. This strategy could achieve reasonable results, but in our experiments we show that using a joint probability instead of the standard fusion strategy we could obtain better performance. The main reason is that ranking list of different features are quite different and with an average fusion strategy a low score of one type of classifier downgrades a possibly relevant video. By using the joint probability, only videos that receive a low score from both classifiers will be put at the bottom of the list. The formulas of average fusion and joint probability are separately shown below:

$$ave = \frac{P_{CB} + P_{IDT}}{2} \tag{1}$$

$$JointProb = 1 - (1 - P_{CB}) \times (1 - P_{IDT}) \tag{2}$$

Where, P_{CB}, P_{IDT} denote prediction scores obtained by Concept-Bank feature and IDT feature.

4 MED Results and Analysis

In this section we will explain the results per subtask, emphasizing our main research objectives.

4.1 Zero Example

Figure 1 shows comparison of manual/automatic SQG system with different similarity measurements on the evaluation set:

- *manualfused*: our primary run for manually picked concepts fused with OCR and ASR;

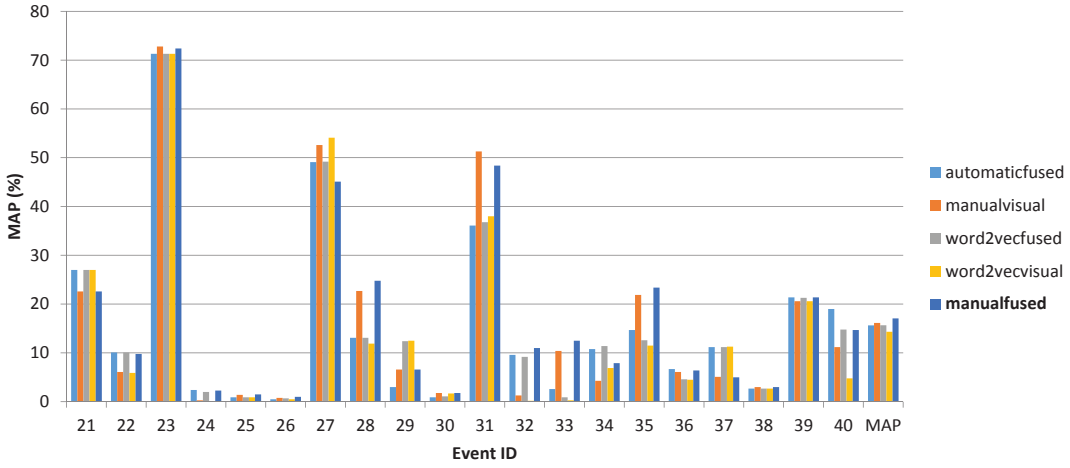

Figure 1: Comparison runs on the MED15-EvalFull set for PS_0-Ex task

- *manualvisual*: a comparison run for manually picked concepts only, without fusion of OCR/ASR;
- *word2vecfused*: concepts are weighted by word2vec similarity in automatic SQG and the results are fused with OCR and ASR;
- *word2vecvisual*: a comparison run for word2vec-based automatic SQG only, without fusion of OCR/ASR;
- *automaticfused*: concepts are weighted by our last year's automatic SQG system, but the concept selection strategy is the same as in the *word2vecfused*.

The *manualfused* run ranks at the top of the five, which is not a surprise. But interestingly we see our automatic SQG (*automaticfused*) does not drop the performance by much compared to the manual SQG (*manualfused*). For some individual events, e.g. E021, E027 and E037, it even performs better than the manual runs. In these particular events, the number of concepts that are picked up manually is significantly more than those picked up automatically. For example, in the event *bike trick*, automatic SQG simply picks two concepts with a same name "bike trick" from *FCVID_239* and *RC_497* respectively while manual SQG additionally includes a variety of bikes, e.g. *off-road mountain bike* and *motorcycle*. We conclude that these additional concepts only increase recall but decrease precision for the events. On the other hand, the runs using word2vec as similarity measurement basically fall in the same scale of performance compared to the similarity measurement we used in the last year.

4.2 Ten Examples

In this section, we first present performance of individual feature and fusion strategy for PS_10-Ex on MED14-Test dataset, then, we will report and analysis the performance of our systems for PS_10-Ex and AH_10-Ex on MED15-EvalFull/Sub data set.

4.2.1 Results on MED14-Test dataset

Table 1 shows individual feature performance on MED14-Test for 10-Ex task. As is shown, we can easily observe that **Concept-Bank** achieves the highest MAP in 10-Ex among small-scale concept features and IDT, indicating that our large-scale concepts pool works pretty well for multimedia event

Feature	**PS_10-Ex** MAP	**PS_100-Ex** MAP
ImageNet_1000	0.117892	0.200446
SIN_346	0.07411	0.152399
RC_497	0.097898	0.194238
Places_205	0.071228	0.119576
FCVID_239	0.121821	0.176341
Sports_487	0.150613	0.19892
Improved Dense Trajectory	0.122318	0.25525
Concept-Bank	**0.216718**	**0.310635**

Table 1: MED PS_10/100-Ex: Mean AP of single feature on MED14-Test

Fusion System	**MED14-Test** MAP	**MED14-EvalSub** MAP	**MED15-EvalFull** MAP
Baseline (Concept-Bank)	0.216718	0.229	0.163
Visual-System (Concept-Bank + IDT)	0.24243	0.231	0.168
Visual-Zero (Visual-System + 0-Ex)	0.258491	0.265	0.202
Full-System_V1 (Visual-Zero + OCR + ASR)	–	0.277	0.208
Full-System_V2 (Visual-Zero + OCR)	**0.310107**	**0.278**	**0.213**

Table 2: MED PS_10-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full

detection. Inspiringly, the vocabulary size of our Concept-Bank is only 2,774, which still has potential to be further expanded. Table 2 shows the effects of fusion strategy with three stages. Our observations are shown below:

- The fusion of Concept-Bank feature with IDT feature improves the MAP from 0.217 to 0.242, indicating the complementary of high-level Concept-Bank feature with low-level IDT feature.
- The fusion of Visual-System and 0-Ex system improves the MAP from 0.242 to 0.258, which indicates that textual information makes up for visual information especially when only 10 positive exemplars are provided.
- Since not all positive videos contain OCR/ASR information, our designed high-precision OCR/ASR system boosts relevant videos in the ranking list and improves the overall MAP from 0.258 to 0.310.

4.2.2 Results on MED15-EvalFull/Sub dataset

To further verify our fusion strategy with three stages, we submitted 5 PS_10-Ex runs for the MED15-EvalFull/Sub dataset. The returned results are shown in Table 2. Our three stages of fusion almost works well for large-scale test set MED15-EvalFull/Sub, except that adding ASR drops the overall performance

a bit. This is due to that the precision of our ASR system is not as high as our OCR system. The 0-Ex and OCR system does really improve the overall MAP with a relative of 20% on Visual-System.

For PS_10-Ex, our primary run (Full-System_V2) ranks 1st among 16 teams on MED14-EvalSub dataset and 2nd among 7 teams on MED15-EvalFull dataset, which outperforms last year's best results by 2.1% on MED14-EvalSub dataset and 1.9% on MED15-EvalFull dataset.

For AH_10-Ex, we submitted two runs: Visual-System and Full-System_V2, which share the same settings as our PS_10-Ex. Our primary run (Full-System_V2) ranks 4th among 7 teams on MED15-EvalFull. Compared with our PS_10-Ex results, the MAP of AH_10-Ex results are lower. The possible reason might be that our Concept-Bank is still not large enough, relevant concepts to AH Events are not sufficient which will have negative impact on our Visual-System and 0-Ex System.

4.3 One Hundred Examples

In this section, we first present performances of individual features and fusion strategies for PS_100-Ex on MED14-Test dataset, then, we will report and analyze the performance of our systems for PS_100-Ex on MED15-EvalFull/Sub dataset.

4.3.1 Results on MED14-Test dataset

Fusion System	**MED14-Test** MAP	**MED14-EvalSub** MAP	**MED15-EvalFull** MAP
Baseline (Concept-Bank)	0.310635	0.301	0.23
Visual-System (Concept-Bank + IDT)	0.352886	0.323	0.251
Visual-System with Joint Prob (Concept-Bank + IDT)	0.354844	0.338	0.267
Full-System_V1 (Visual-System + OCR + ASR)	–	**0.327**	0.254
Full-System_V2 (Visual-System + OCR)	**0.360628**	0.325	**0.256**

Table 3: MED PS_100-Ex: Mean AP of fusion strategies on MED14-Test/EvalSub/Full

Table 1 shows individual feature performance on MED14-Test for 100-Ex task. Same conclusion as 10-Ex can be drawn: Concept-Bank performs best.

Table 3 shows the effects of fusion strategies. The fusion strategy we designed for 100-Ex task have two differences from 10-Ex task. Firstly, we give up fusing 0-Ex results, due to 0-Ex results barely help 100-Ex results in our internal test. Secondly, we tested fusion strategy with joint probability instead of average fusion.

For PS_100-Ex on MED14-Test, joint probability is slightly better than average fusion in overall MAP, but when tested with larger-size dataset (MED15-EvalFull/Sub), the improvement is larger. Adding IDT, ASR/OCR feature, we observe expected improvement as PS_10-Ex on MED14-Test.

To further verify our fusion strategies, we submit 5 PS_100-Ex runs for MED15-EvalFull/Sub dataset. The returned results are shown in Table 3. Similar to 10-Ex, our ASR system drops overall performance a bit. The IDT, OCR and joint probability bring expected improvements to our baseline.

Our primary run (Full-System_V1) on PS_100-Ex ranks 2nd among 6 teams on MED14-EvalSub datset and 1st among 2 teams on MED15-EvalFull dataset.

5 Conclusion and Discussion

For the TRECVID Multimedia Event Detection task results of 2015, we can conclude that our bigger Concept-Bank and our efforts in combining different features and classifiers as well as a good strategy for the 0Ex pay out in a much higher performance compared to 2014 and even top ranked performance for 0Ex and 10Ex. Adding both IDT and OCR to the visual features improves performance on each task, whereas ASR surprisingly decreases performance on the evaluation. For the 10Ex task, the fusion of the system trained on the ten visual examples and the 0Ex system boosts performance with relatively 20%. For the 100Ex task, using the joint probability of the Concept-Bank features and IDT surprisingly improves performance on the evaluation set.

Acknowledgment

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 120213), and TNO acknowledges the financial support from ERP Big Data.

References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, G. Quenot, and R. Ordelman, "Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[2] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1, pp. 89–108, 2002.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.

[9] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," in *arXiv preprint arXiv:1502.07209*, 2015.

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *arXiv preprint arXiv:1412.0767*, 2014.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

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

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

[14] C.-W. Ngo, Y.-J. Lu, H. Zhang, T. Yao, C.-C. Tan, L. Pang, M. de Boer, J. Schavemaker, K. Schutte, and W. Kraaij, "VIREO-TNO @ TRECVID 2014: Multimedia event detection and recounting (MED and MER)," 2014.

[15] S.-I. Yu, L. Jiang, Z. Xu, Z. Lan, S. Xu, X. Chang, X. Li, Z. Mao, C. Gan, Y. Miao, X. Du, Y. Cai, L. Martin, N. Wolfe, A. Kumar, H. Li, M. Lin, Z. Ma, Y. Yang, D. Meng, S. Shan, P. D. Sahin, S. Burger, F. Metze, R. Singh, B. Raj, T. Mitamura, R. Stern, and A. Hauptmann, "Informedia@TRECVID 2014 MED and MER," 2014.

[16] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann, "Bridging the ultimate semantic gap: A semantic search engine for internet videos," in *International Conference on Multimedia Retrieval*, 2015.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

VIREO @ TRECVID 2015: Video Hyperlinking (LNK)

Lei Pang and Chong-Wah Ngo

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

Abstract

This paper presents an overview and comparative analysis of our system designed for TRECVID 2015 [1] video hyperlinking (LNK) task. The application scenario for video hyperlinking is to satisfy the needs of users to find further information on the content of interest contained within an anchor. The task here is given an anchor to generate a ranked list of video segments relevant to the name entities extracted from the anchor. Our four runs are summarized below:

- ***tf-idf***: The subtitles of video segments are indexed with Lucene [2] and more like this (MLT) query is adopted for tf-idf based retrieval.
- ***word2vec***: Each video segment is represented by uniformly summing the vector representations of words in subtitles using Word2Vec [3, 4] and the relevant segments are ranked based on cosine similarity.
- ***weighted_word2vec***: The words in subtitles are weighted by document frequency and a video segment is represented by summing the weighted vector representations.
- ***linear_tf-idf_weighted_word2vec***: The relevant segments are ranked based on average fusion of *tf-idf* and *weighted_word2vec*.

1 System Overview

The dataset is composed of 3,518 BBC videos. The videos are accompanied by archival metadata (e.g., subtitles, list of popular UK celebrities) and automatic annotations (e.g., speech transcripts, shot segmentation, face detection, multiple versions of concept detectors). Among these rich information, we only explore the usage of ***subtitles*** to investigate the effectiveness of name entities for hyperlinking. The system consists of three stages: scene extraction based on topic detection (1.1), name entity detection (1.2) and hyperlinking with word vectors (1.3).

1.1 Scene Cutting

Since the dataset only provides "shot" segmentations and we actually want to work on "scene" level, we adopt TextTiling [5] to split the subtitles of videos into multi-paragraph subtopics, where each subtopic corresponds to a scene segment. The discourse cues for identifying major subtopic shifts are patterns of lexical co-occurrence and distribution. The algorithm has been proven to be useful for many text analysis tasks, including information retrieval and summarization [5]. We also constrain the algorithm to avoid

Table 1: P@N and MAP results for video hyperlinking

	P@5	P@10	P@20	MAP
word2vec	0.366	0.321	0.207	0.105
weighted_word2vec	0.338	0.344	0.227	0.126
tf-idf	**0.462**	0.406	0.300	0.180
linear_tf-idf_weighted_word2vec	0.436	**0.423**	**0.313**	**0.190**

splitting the consecutive speeches from the same speaker. Finally, a total of 100,917 scene segments are extracted.

1.2 Name Entity Detection

To facilitate the detection of hyperlinking targets, we extract name entities from the subtitles of each scene segment. Here, we adopt the Stanford Name Entity Recognizer (NER) [6] and a total of 98,601 distinctive name entities are detected. These name entities are classified into four categories – person, organization, location and others. We filter the noisy name entities based on document frequencies. The name entities with document frequency less than 10 are removed as noises and totally 8,168 name entities are retained. During indexing and retrieval, each name entity is treated as one single word.

1.3 Hyperlinking with Word Vectors

Word vector representation [3, 4] has shown great performance in measuring syntactic and semantic word similarities. Hence, different from the previous works, which usually enrich the text information with synonyms or conceptually connected words [7] or visual concepts [8], we directly represent each scene segment with word vector representation. As mentioned in [3], each scene segment is represented by summing all the words in the subtitle. The vector representation for each word is finetuned on the GoogleNews model[1]. The model contains 300-dimensional vectors for 3 million words and phrases. When finetuning, each word is initialized based on GoogleNews model and the name entities are initialized by summing the words. Since name entities are not very frequent, we adopt the skip-gram architecture with hierarchical softmax. Sub-sampling of the frequent words is also used for improving accuracy and speed.

In addition to uniformly summing all the word vectors as representation, we also want to measure the importance of different words. Here, we use document frequency (df) as weight. To further investigate the effectiveness of the vector representation, we further compare with tf-idf based retrieval and linearly combine the scores of these two measures.

2 Evaluation Results

Table 1 presents the evaluation results of our four runs, where P@N are precision-oriented metrics at different cutoff points and MAP is mean average precision. From the table, we can easily observe that the fusion of *tf-idf* and *weighted_word2vec* achieves the best performance. But it is surprise that *tf-idf* achieves better performance than both *word2vec* and *weighted_word2vec*. By observing the results, we find that *word2vec* and *weighted_word2vec* perform better when the anchors contains name entities, such as person names "James Humbert Craig" and locations "Cartley Hole". This is mainly due to the

[1] https://code.google.com/p/word2vec/

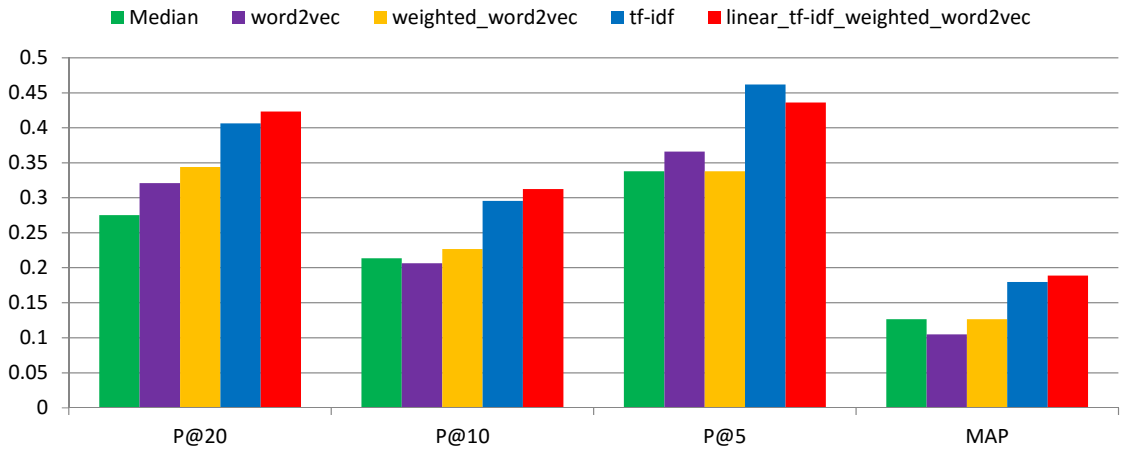

Figure 1: Results of four submitted runs in comparison with median perfromance

fact that word2vec can better represent the name entities by considering the context information. For example, the locations "Cartley Hole" is surrounded by words such as "house", "castle" and "garden" and the vector representation will be closely related to these words. However, *tf-idf* is attempting to retrieve the segments containing the exactly same entity, often resulting in lower recall. On the other hand, word vector representation also introduces semantic noises. For example, video segments about the traffic network is retrieved for the anchor showing tennis game the word "net" is mentioned. Since most of the anchors do not contain definite entities, the semantic noises degrade the performance of *word2vec* and *weighted_word2vec*. The linear combination achieving the best performance is in consistent with our observation, in which that the vector representation can be complementary to *tf-idf* by extending the keyword semantically.

We also compare our four runs with the median performance of other teams. As shown in Figure 1, *tf-idf* and *linear_tf-idf_weighted_word2vec* consistently performs better than median performance in all of the four measurements. Based on the previous observation, the performance of *word2vec* can be further improved through representing the segments with *interesting* words, such as nouns, verbs and name entities.

3 Summary

We submitted four runs mainly based on the word2vec representation. The word2vec indeed semantically extends the keywords for hyperlinking. However, some noises are introduced and as a result degrade the average performance. In the future, we are targeting to locate "interesting" words and representing the video segments based on the intensity of interest rather than document frequency. In addition, we will also consider visual entities such as faces and objects for hyperlinking.

4 Acknowledgement

The work described in this paper was supported by grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11210514).

References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quenot, and R. Ordelman, "Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[2] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Greenwich, CT, USA: Manning Publications Co., 2010.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 3111–3119.

[5] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, 1997.

[6] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 363–370.

[7] Z. Paroczi, B. Fodor, and G. Szücs, "Re-ranking the image search results for relevance and diversity in mediaeval 2014 challenge," in *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.*, 2014.

[8] B. Safadi, M. Sahuguet, and B. Huet, "When textual and visual information join forces for multimedia retrieval," in *International Conference on Multimedia Retrieval*, 2014, p. 265.