Waseda_Meisei at TRECVID 2018: Ad-hoc Video Search

Kazuya Ueki[1,2], Yu Nakagome[1], Koji Hirakawa[1], Kotaro Kikuchi[1], Yoshihiko Hayashi[1], Tetsuji Ogawa[1], and Tetsunori Kobayashi[1]

[1] Faculty of Science and Engineering, Waseda University, Room 40-701, Waseda-machi 27, Shinjuku-ku, Tokyo, 162–0042, Japan
[2] Department of Information Science, Meisei University, Room 27-1809, Hodokubo 2-1-1, Hino, Tokyo, 191–8506, Japan
kazuya.ueki@meisei-u.ac.jp

Abstract. The Waseda_Meisei team participated in the TRECVID 2018 Ad-hoc Video Search (AVS) task [1]. For this year's AVS task, we submitted both manually assisted and fully automatic runs. Our approach focuses on the concept-based video retrieval, based on the same approach as last year. Specifically, it improves on the word-based keyword extraction method presented in last year's system, which could neither handle keywords related to motion nor appropriately capture the meaning of phrases or whole sentences in queries. To deal with these problems, we introduce two new measures: (i) calculating the similarity between the definition of a word and an entire query sentence, (ii) handling of multi-word phrases. Our best manually assisted run achieved a mean average precision (mAP) of 10.6%, which was ranked the highest among all submitted manually assisted runs. Our best fully automatic run achieved an mAP of 6.0%, which ranked sixth among all participants.

1 System Description

For ad-hoc video search, our approach is based on concept-based video retrieval. We first build a large concept bank comprising many types of concepts. It contains classifiers such as persons, objects, scenes, and actions. We then calculate the concept scores for every video sequence in advance. In the following section, we will introduce the concept bank used in our system and then demonstrate how to calculate concept scores for each video. Next, we will describe our three types of approaches for concept similarity scoring: Method 1: word-based, Method 2: sentence-based, and Method 3: phrase-based.

1.1 Concept bank

In the TRECVID AVS task in 2016 and 2017, we achieved the best performance by constructing a large concept bank using pre-trained concept classifiers and improving the coverage rate for keywords appearing in query sentences [2][3]. The teams in the competition achieving the highest performance were found to have a similar concept-based approach. In addition, this method is acknowledged as an effective solution for video search[4][5].

After analyzing the failure cases from 2016 and 2017, we found deficiencies wherein the system could not properly select corresponding concept classifiers for some query sentences. We also found that most concepts corresponded only to nouns. Action words or phrases composed of multiple words did not show any correspondence.

Table 1. Concept bank used in our systems.

Name	Database	# Concepts	Concept Type(s)	Models
TRECVID346	TRECVID SIN [6]	346	Person, Object, Scene, Action	GoogLeNet + SVM
FCVID239	FCVID [7]	239	Person, Object, Scene, Action	GoogLeNet + SVM
UCF101	UCF101 [8]	101	Action	GoogLeNet + SVM
PLACES205	Places [9]	205	Scene	AlexNet
PLACES365	Places	365	Scene	GoogLeNet
HYBRID1183	Places, ImageNet [10]	1,183	Person, Object, Scene	AlexNet
IMAGENET1000	ImageNet	1,000	Person, Object	GoogLeNet
IMAGENET4000	ImageNet	4,000	Person, Object	GoogLeNet
IMAGENET4437	ImageNet	4,437	Person, Object	GoogLeNet
IMAGENET8201	ImageNet	8,201	Person, Object	GoogLeNet
IMAGENET12988	ImageNet	12,988	Person, Object	GoogLeNet
IMAGENET21841	ImageNet	21,841	Person, Object	GoogLeNet
PASCAL20	Pascal VOC [11]	20	Person, Object	YOLOv2
ACTIVITYNET200	ActivityNet [12]	200	Action	GoogLeNet + SVM
KINETICS400	Kinetics [13]	400	Action	3D-ResNet
ATTRIBUTES300	Visual Genome [14]	300	Attributes of persons/objects	GoogLeNet + SVM
RELATIONSHIPS53	Visual Genome	53	Relationships b/w persons/objects	GoogLeNet + SVM

Therefore, in the 2018 system, the following two types of enhancements were implemented to expand the previous concept bank to the phrase level. The first involves an improvement that allows the system to handle verbs or verb phrases (verb + object) such as "swimming," "playing drums," "riding a horse," "wearing a scarf," among others. The second improvement enables the system to be able to deal with combinations of people/objects and their attributes (adjective + noun), also known as adjective noun phrases. Examples include "blue shirt," "brick building," "blonde female," among others.

Finally, we constructed a new concept bank shown in Table 1, which includes newly added concept classifiers that can deal with various forms of query sentences. Using this concept bank, we calculate all concept scores for all videos. In this paper, we will explain how to create concept classifiers for each database and pre-trained model.

1. TRECVID346, FCVID239, UCF101, and ACTIVITYNET200

   First, a maximum of ten frames from each shot were selected at regular intervals, and the corresponding images were input to the *GoogLeNet* model [15] pre-trained on the ImageNet database [10]. This allowed us to obtain 1,024-dimensional feature vectors from pool5 layers. These feature vectors (a total of ten at most) were then bound to one feature vector using element-wise max-pooling. We trained SVMs using data-given labels of concepts for each database as positive samples and randomly selected images from the TRECVID SIN dataset as negative samples. The shot score for each concept was calculated as the distance to the hyperplane in the SVM model.

2. PLACES205/365, HYBRID1183, and IMAGENET1000/4000/4437/8201/12988/21841

   We calculated the concept scores using a scene classification model, pre-trained with the Places database [9]. We also calculated concept scores using image classification models [16][17][18] pre-trained with the ImageNet database [10]. Since each unit of the convolution neural network (CNN) output layer identifies a concept in a scene/object, the values representing a concept of the CNN output layer (before softmax was applied) were used as concept scores. The maximum score of each concept for each video was obtained after inputting at most 10 images to the CNN.

3. PASCAL20

   To deal with two or more objects/persons in query sentences, such as "two or more men" and "at least two planes," we created concept classifiers that provide a higher score to videos with a large number of specific objects/persons. We accomplished this using 20 object detection models (*YOLOv2* [19]) trained on the Pascal VOC dataset [11]. Specifically, the score of a concept containing two or more objects/persons is the sum of the probabilities of the concepts occurring in all detected bounding boxes.

4. KINETICS400

   We used a *3D-ResNet* model [20][21] pre-trained with the Kinetics database [13]. Sixteen consecutive frames were input into 3D-ResNet from one to the next, and the maximum score of each concept obtained from the output layer was taken as the concept score for each video.

5. ATTRIBUTES300 and RELATIONSHIPS53

   Using annotations of attributes of persons/objects and relationships between persons/objects in the Visual Genome Database [14], 300 types of attributes and 53 types of relationship concepts were created. As for the attributes, we built "adjective + noun" concepts such as "blue_sky," "white_plate." As for the relationships, we selected the data whose subject is a person and that had specific verbs such as "wear" (34 types) and "hold/have" (19 types). Finally, we created concepts such as "wear_shirt", "have_ski_pole".

The score for each semantic concept was normalized over all training iterations using a min-max normalization, that is, the maximum and the minimum scores were 1.0 (most probable) and 0.0 (least probable), respectively.

1.2 Method 1: Word-based keyword selection

Video retrieval using word-based keyword selection was performed through the following pipeline:

1. Extract one or more keywords from a query sentence.
2. Select one or more concept classifiers related to a keyword. The corresponding concept may not exist in the concept bank.
3. For each video, a score is calculated for the query sentence by integrating the scores from multiple concept classifiers.

Given a query sentence, we selected some visually important keywords. For example, given the query sentence "one or more people driving snowmobiles in the snow," we picked out the keywords "people," "snowmobile," and "snow." We then matched the keywords with concepts using a concept classifier. Semantically similar concepts were also chosen using the word2vec algorithm [22] to select as many concept classifiers as possible.

1.3 Method 2: Similarity calculation between the word definition sentence and the whole query sentence

In our previous study, when associating the search keyword with the concept by Method 1, we observed that concept classifiers with the same classifier name but different

meaning were erroneously selected in some cases. Moreover, when we analyzed concept classifiers that were rated as having a high similarity by word2vec, we observed that word2vec failed to select the appropriate classifier in many cases. For example, "guitarist" could not be selected from the query sentence: "a person playing guitar." To tackle these problems, we have added auxiliary information corresponding to concept classifiers in order to select more appropriate classifiers.

Because the names of concept classifiers trained with ImageNet are linked to a lexical database of English, the WordNet dictionary, we used hypernyms, words with superordinate meanings, and these words' definitions. After restricting classification to the concepts whose name and query sentence have the same hypernyms, video retrieval was performed based on the similarity between the query sentence and the video definition sentence. We calculated the similarity by mapping both the definition sentence and the query sentence into a common space using Universal Sentence Encoder (USE) [23]. Finally we selected two concept classifiers with high similarity for each query sentence.

1.4 Method 3: Phrase-based concept selection

For the systems submitted in 2016 and 2017, concept classifiers and their selection of concepts corresponding to verb phrases appearing in the query sentence was insufficient. To deal with the verb phrases in the query sentence, we selected concept classifiers corresponding to phrases by chunking the query sentences as follows:

- Subject + intransitive verb,
- Subject + transitive verb + object,
- Subject + transitive verb + preposition + object.

As for the correspondence between phrase and concept classifier, USE was used in the same ways as Method 2. Specifically, we input both a phrase and a classifier name into USE and compared the similarities between feature vectors by mapping them into the common space.

2 Submissions

This year we submitted two manually assisted runs and four fully automatic runs to the TRECVID 2018 Ad-hoc Video Search (AVS) task.

2.1 Manually assisted runs

We submitted two manually assisted runs (Manual1 and Manual2). Our manually assisted runs used word-based keyword selection (Method 1). For this submission, there were no major changes between last year's system [3] except for the addition of training datasets which include the following four concepts: `ACTIVITYNET200`, `KINETICS400`, `ATTRIBUTES300` and `RELATIONSHIPS53`.

The difference between two runs was calculated through the score fusion methods used; Manual1 run used weighted product score, and Manual2 run used product score without weight.

2.2 Fully automatic runs

Our fully automatic runs were based on the combination of all three methods (Method 1, Method 2, and Method 3). We combined three methods by calculating the weighted sum. The weights for Methods 1, 2, and 3 used were

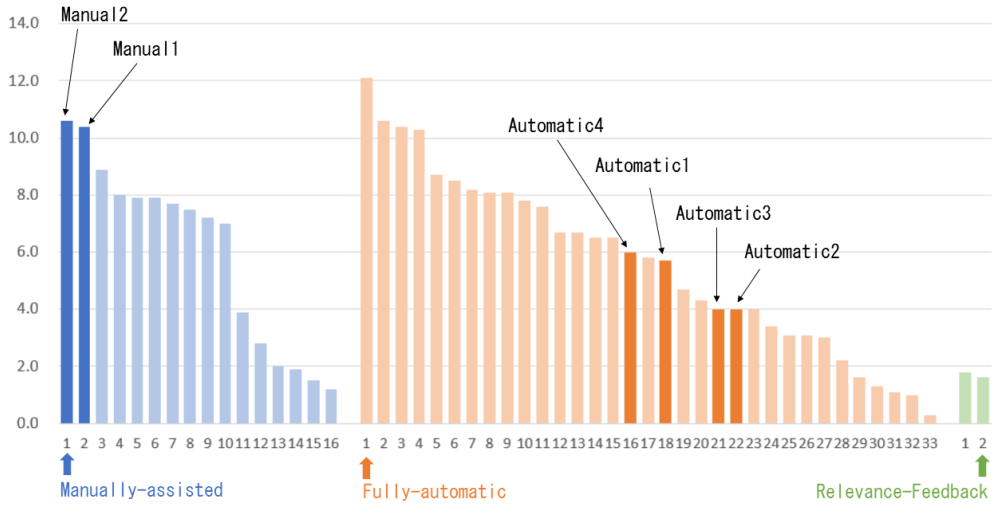

Fig. 1. Comparison of Waseda_Meisei runs with the runs of other teams for all the submitted runs including manually assisted (blue), fully automatic (red), and relevance feedback runs (green).

- Automatic1 1.0 : 1.0 : 5.0,
- Automatic2 1.0 : 1.0 : 1.0,
- Automatic3 1.0 : 0.0 : 2.0,
- Automatic4 1.0 : 1.0 : 99.0.

These weights were chosen based on the evaluation result using the 2016 and 2017 test data.

2.3 Results

Fig. 1 shows the results of all the submitted runs including manually assisted, fully automatic, and relevance feedback. The overall mAP was lower than that obtained over the previous year. Thus, we concluded that the 2018 query sentences made classification more difficult than that done in 2016 and 2017. The mAPs of our submitted manually assisted runs (Manual1 and Manual2) were 10.4% and 10.6%, which ranked 1st among all submitted manually assisted systems. However, the best automatic system achieved a higher mAP of 12.1%, more than that obtained by our manual system. The mAPs of our fully automatic runs (Automatic1, Automatic2, Automatic3, and Automatic4) were 5.7%, 4.0%, 4.0%, and 6.0%, respectively.

Fig. 2 shows the average precision for each query sentence. The average precision of our manual system was the best for many query sentences, whereas the average precision of the automatic system was lower than best and slightly better than the medians.

3 Conclusion

For this year's submissions, we solved the problem of ad-hoc video search using a combination of many semantic concepts in the same manner as that done for last year's submission. Our approach achieved a higher average precision when a simple query sentence was used. However, once a query sentence became complicated enough

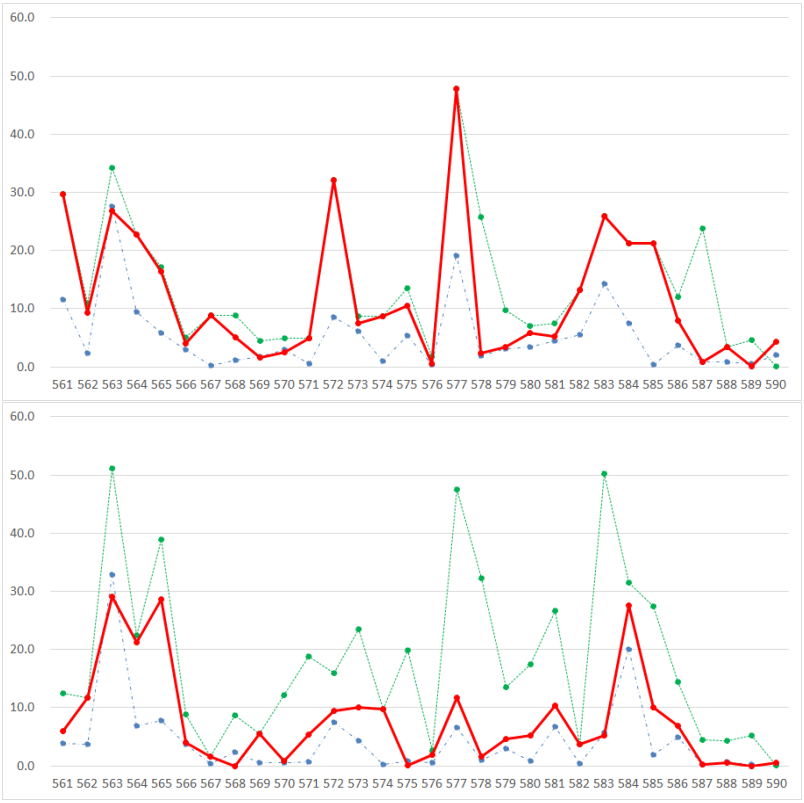

Fig. 2. The average precision for each query sentence. Manually assisted runs (top), and fully automatic runs (bottom). Our best run (red), the best run among all the submitted runs (green), and the median (blue).

to extract many keywords/concepts, the average precision worsened considerably. Even with a manual system that could select relatively ideal concepts, mAP was as low as 20% with concept-based methods. Therefore, our future work will focus on other approaches such as directly calculating the similarity between a video and a query sentence based on video captioning and multi-modal embedding techniques.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 15K00249, 17H01831, and 18K11362, Kayamori Foundation of Informational Science Advancement, and Telecommunications Advancement Foundation.

References

1. G. Awad, A. Butt, K. Curtis, J. Fiscus, A. Godil, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, J. Magalhaes, D. Semedo, and S. Blasi, "TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search," In Proc. of TRECVID 2018, 2018.
2. K. Ueki, K. Kikuchi, S. Saito, and T. Kobayashi, "Waseda at TRECVID 2016: Ad-hoc Video Search," In Proc. of TRECVID 2016, 2016.

3. K. Ueki, K. Hirakawa, K. Kikuchi, T. Ogawa, and T. Kobayashi, "Waseda Meisei at TRECVID 2017: Ad-hoc Video Search," In Proc. of TRECVID 2017, 2017.
4. C. G. M. Snoek, X. Li, C. Xu, and D. C. Koelma, "University of Amsterdam and Renmin University at TRECVID 2017: Searching Video, Detecting Events and Describing Video," In Proc. of TRECVID 2017, 2017.
5. P. A. Nguyen, Q. Li, Z. Cheng, Y. Lu, H. Zhang, and C. Ngo, "VIREO@TRECVID 2017: Video-to-Text, Ad-hoc Video Search, and Video hyperlinking," In Proc. of TRECVID 2017, 2017.
6. G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Quénot, "TRECVid Semantic Indexing of Video: A 6-Year Retrospective," ITE Trans. on MTA vol.4, no.3, pp.187–208, 2016.
7. Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, S.-F. Chang, "Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks," arXiv:1502.07209, 2015.
8. K. Soomro, A. R. Zamir, M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," arXiv:1212.0402, 2012.
9. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," Advances in Neural Information Processing Systems (NIPS), 2014.
10. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," In Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
11. M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," International Journal of Computer Vision, vol.111, no.1, pp.98–136, 2015.
12. F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles "ActivityNet: A large-scale video benchmark for human activity understanding," In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015), pp.961–970, 2015.
13. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T.r Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," arXiv: 1705.06950, 2017.
14. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalanditis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, and F.-F. Li, "Visual Genome : Connecting language and vision using crowdsourced dense image annotations," arXiv:1602.07332, 2016.
15. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions," In Proc. of Computer Vision and Pattern Recognition (CVPR), 2015.
16. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding", arXiv:1408.5093, 2014.
17. P. Mettes, D. C. Koelma, and C. G. Snoek, "The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection," In Proc. of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR), pp.175–182, 2016.
18. S. loffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv:1502.03167, 2015.
19. J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," arXiv:1612.08242, 2016.
20. K. Hara, H. Kataoka, and Yutaka Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018), pp. 6546-6555, 2018.
21. K. Hara, H. Kataoka, and Y. Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition," In Proc. of the ICCV Workshop on Action, Gesture, and Emotion Recognition, 2017.
22. T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space," arXiv:1301.3781, 2013.
23. D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, "Universal Sentence Encoder," arXiv:1803.11175, 2018.