Kobe University and Kindai University at TRECVID 2018 AVS Task

Kimiaki Shirahama†, He Zhenying* and Kuniaki Uehara*
* Graduate School of System Informatics, Kobe University
jennyhe@ai.cs.kobe-u.ac.jp, uehara@kobe-u.ac.jp
† Department of Informatics, Kindai University
shirahama@info.kindai.ac.jp

***Abstract*—This paper presents our system developed for Ad-hoc Video Search (AVS) task (manually-assisted) in TRECVID 2018. Our system adopts a concept-based approach using the five sources of training data, 1) collaborative annotations for Semantic INdexing (SIN) in TRECVID 2013 [1], 2) ImageNet [2], 3) Places365 dataset [3], 4) Sports1M dataset [4] and 5) MS COCO dataset [5]. The following four runs were submitted:**

1) *M_D_kobe_kindai.18_1:* **Concepts for a topic are manually selected, and detection results for them are organised into a cascade. Shots are retrieved by gradually filtering out irrelevant shots at each stage of the cascade.**
2) *M_D_kobe_kindai.18_2:* **This run is an improved version of M_D_kobe_kindai.18_1. Some topics specify the number of objects or their spatial relation. This run examines object regions obtained by an object detector in order to filter out shots where detected regions do not satisfy the specified number or relation restriction.**
3) *M_D_kobe_kindai.18_3:* **Compared to M_D_kobe_kindai .18_1, this runs uses slightly different sets of manually selected concepts for some topics. The purpose of this run is to check the suitability of concepts used in M_D_kobe_kindai.18_1.**
4) *M_D_kobe_kindai.18_4:* **For all the topics, this runs uses the same sets of concepts to M_D_kobe_kindai.18_1. But, shots are retrieved by just summing up detection scores for the selected concepts. That is, this runs aims to examine the effectiveness of cascades in M_D_kobe_kindai.18_1.**

Unexpectedly, the evaluation results show that M_D_kobe_ kindai.18_4 outperforms all the other runs. This indicates that the cascade approach in M_D_kobe_kindai.18_1 does not work well. Apart from this, the comparison between M_D_kobe_kindai.18_1 and M_D_kobe_kindai.18_2 clearly shows the effectiveness of object detection to refine retrieval results. Finally, our team (kobe_kindai) is ranked at the third place among the six teams in AVS task (manually-assisted) and the performance of M_D_kobe_kindai.18_4 is ranked at the seventh place among all the 16 runs. Also, our runs lead to the best average precisions for six topics in the manually-assisted category.

I. Introduction

This paper introduces the video retrieval system that we (kobe_kindai team) have developed for AVS task in TRECVID 2018 [6], [7], [8]. This year we addressed the following two points: The first is how to fuse concept detection scores for accurate retrieval. Our systems in the past two years produce a retrieval result by examining shots just with the sum of detection scores for concepts related to a topic [9], [10]. However, we observed that the retrieval result includes clearly irrelevant shots, for which detection scores for some concepts are very high but those for the other concepts are not. It is impossible to exclude such shots by the simple summation of detection scores. Thus, a more sophisticated score fusion approach is necessary, and especially this year, we choose a cascade-based approach that uses a sequence of stages, at each of which irrelevant shots are gradually filtered out [11].

The second point is the adoption of object detection in order to accurately deal with the meaning of a topic, for which the number of objects or their relation is important. Until this year, we only used detection scores that only indicate probabilities of a concept's appearance without inspecting its position in a frame. Considering the recent advances in object detection [12], [13], [14], [15], we decide to incorporate object detection into our retrieval system.

II. Method

Similar to our past systems [9], [10], we adopt a concept-based approach where concepts related to a given topic are firstly selected, and then shots are retrieved by analysing detection scores for those concepts. But, different from [9], [10], this year's system takes advantage of a significantly large vocabulary of concepts, cascade-based approach, and object detection, which will be described below.

A. Concept Detection

Our system utilises the following five concept categories which include in total 11635 concepts:

1. 345 SIN concepts: Detection scores provided by the Centre for Research and Technology Hellas (ITI-CERTH) team [16] are used. They fine-tuned two pre-trained networks using the dataset collected by the collaborative annotation effort for TRECVID 2013 SIN task [1]. Each of these networks is used as a feature extractor where outputs of the last FC layer are utilised as a feature to train an SVM. For each concept, prediction scores of SVMs for the two fine-tuned networks are averaged as the final detection score.

2. 1000 ImageNet concepts: The ResNet152 [17] implementation in YOLO [18] is used to obtain detection scores for 1000 concepts defined in ImageNet [2].

3. 9418 ImageNet concepts: In YOLO [18], 9418 concepts defined in ImageNet [2] and MS COCO [5] are organised into a hierarchical tree, and a CNN (darknet9000) is trained for detecting them. One big advantage is that detection scores

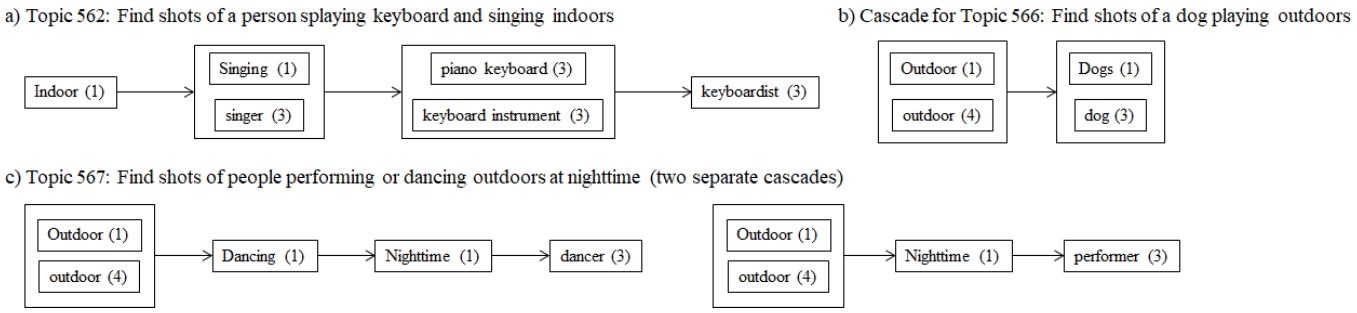

Fig. 1. Examples of concepts and the cascades build on them for Topics 562, 566 and 567. The number behind each concept name indicates the concept category ID (1: 345 SIN concepts, 2: 1000 ImageNet concepts, 3: 9418 ImageNet concepts, 4: 385 Places concepts, 5: Sports1M concepts). All the cascades used in our runs can be found on our website (https://www.info.kindai.ac.jp/ shirahama/tv18/cascade_list.html).

are conditional probabilities following the tree structure. So, the detection score of a concept (e.g., "hunting dog") is always smaller than that of its parent concept (e.g., "dog"). In addition, the sum of detection scores for concepts under the same parent concept is one.

4. 385 Places concepts: We use a ResNet fine-tuned for 365 scene concepts defined in Places365 [3]. Since these 365 concepts are leaves in the hierarchical tree, we also exploit 20 intermediate concepts such as "outdoor", "outdoor, natural", "outdoor, man-made". A detection score for each intermediate concept is defined as the maximum of detection scores for its child concepts (i.e., max-pooling).

5. 487 Sports1M concepts: A C3D which performs three-dimensional convolution to capture movements [19] is used to detect 487 concepts defined in Sports1M dataset [4].

For the detection of image-based concepts in the second, third and fourth categories, at most 10 frames are equidistantly sampled from a shot. Then, max-pooling is used to obtain the shot-level score.

B. Concept Selection and Cascade Construction

Figure 1 illustrates concepts and the cascades build on them for three topics. Here, concepts relevant to a topic are manually selected and organised into a cascade. Below, we describe how to select concepts and how to build a cascade of them.

Our concept selection is based on the following two policies:

1) **Generality:** A large part of our concept vocabulary is comprised of ImageNet concepts whose detection scores satisfy the generalisation/specialisation relations among them, as described above. Thus, it is only needed to use the most general concept for a term in a topic. For example, for "Topic 566: Find shots of a dog playing outdoors", we only have to use the concept "dog" and do not have to use its child concepts like "hunting dog", "working dog", "poodle" and so on. It was observed that using such specific concepts lead to a retrieval result which favours some of those concepts.
2) **Specificity:** Despite the generality policy, it is better to use a specific concept which is deduced from a phase in a topic. For example, for "Topic 563: Find shots of one or more people on a moving boat in the water", we use the concept "boatman" which indicates both "people" and "boat" in the topic. Instead, if the concept "Person" is used, it is likely that shots just including "people" are undesirably favoured.

In addition, based on our past experience, it is better to use "negative" concepts for a concept in order to improve the detection result of the latter. For example, shots showing the concept "indoor" should not display "outdoor". We specify such negative concepts for the following four concepts for which the opposite meaning is clear: 1. "outdoor" for "indoor", 2. "indoor" for "outdoor", 3. "Crowd" for "Two_People", and 4. "Daytime_Outdoor" for "Nighttime".

Next, selected concepts are organised into a cascade where each concept is associated with one stage. That is, this stage is used to filter out a fixed amount of shots whose detection scores for the associated concept are low. To construct such a cascade, we mainly address the following three points: The first is the order of stages. As a concept is more general, the corresponding stage is placed earlier. It is more difficult to detect specific concepts (e.g., "hunting dog") than general ones (e.g., "dog"). So, if the stage for the former is placed at the beginning, it is likely to falsely filter out many shots that should be retrieved. Hence, we adopt a conservative approach to firstly place stages for general concepts and gradually examine specific concepts at later stages.

The second point is based on the fact that there are multiple concepts representing the same (or very similar) meaning in our concept vocabulary. For example, there are two concepts named "Flags", one is from 345 SIN concepts and the other is from 9418 ImageNet concepts. The stages for such concepts are placed in parallel, which results in branches from the previous stage. In Figure 1, these concepts are put together in a rectangle, such as ("Outdoor from 345 SIN concepts (1)" and "outdoor from 385 Places concepts"), and ("Singing from 345 SIN concepts (1)" and "singer from 9418 ImageNet concepts (3)"). The last point is to use multiple cascades for one topic having two or more meanings due to "or". For example, "Topic 567: Find shots of people performing or dancing outdoors at nighttime" includes "people performing outdoors at nighttime" and "people dancing outdoors at nighttime". As shown in Figure 1 (c), separate cascades are created for each of these

meanings.

C. Cascade-based Retrieval

Based on the approach introduced in [11], our system performs retrieval by utilising a cascade of concepts in the following way: First, at each stage, detection scores for the associated concept are loaded, and power normalisation [20] is applied to those scores in order to reduce the influences of abnormally high scores (the hyper-parameter α is set to 0.15 for all the runs). Then, min-max normalisation is conducted so as to make the maximum and minimum detection scores 1 and 0, respectively. This is needed to make a fair evaluation of detection scores for the concept, compared to the concepts associated with the other stages. In addition, if some negative concepts are specified for the concept, detection scores for them are also normalised in the above-mentioned way. Subsequently, for each shot, the average score over negative concepts is multiplied with a pre-specified weight (0.5 for all the runs), and is subtracted from the detection score for the (non-negative) concept. The resulting score is used for shot filtering at the stage, where a half of shots with low scores are filtered out.

After sequentially performing this filtering at all the stages, the remaining shots are candidates to form a retrieval result. They are ranked based on their "final scores" which are computed as the sum of scores at all the stages, and the top-ranked 1000 shots constitute the retrieval result.

In the case where a cascade has branches for concepts placed in parallel (like the ones surrounded in rectangles in Figure 1), we disentangle it into separate cascades, and compute final scores of shots for each of them (zero is assigned to shots that are filtered out in the middle of the cascade). Finally, shots are ranked by average-pooling of final scores over disentangled cascades. Also, the same average-pooling approach is used for cascades which are constructed for a topic including multiple meanings.

D. Refinement by Object Detection

Compared with topics over past years, this year's topics are more complex. Most of the previous years' topics did not have clear requirements on the number and spatial relationship of objects. Therefore, in previous years, we usually extracted keywords from the topic as concepts, and then checked whether the keyframes of each shot contain all the concepts as the basis for judgement. However, most of this year's topics have requirements of number of objects or spatial relationship between them. Therefore, if we keep using last year's method, it is very likely that we will not be able to obtain good results.

For these complex topics like "Topic 561: Find shots of exactly two men at a conference or meeting table talking in a room" and "Topic 584: Find shots of a person lying on a bed", the correct number of objects and spatial relationships are essential to detect the correct shots. The simple combination loses the quantity and spatial relationship information of the topics. That is why we decide to take into account the presence or absence of concepts and also the number and spatial relationship of objects. The detection of the presence or absence of an object requires image classification, while the detection of the number and spatial information of objects requires object detection.

In order to detect objects in keyframes of a shot, we use R-CNN that combines CNNs with region proposals, which predict potential areas where objects may exist [12]. Since the proposal of R-CNN, the performance of object detection has been greatly improved. Thereafter, Fast R-CNN [13], Faster R-CNN [14] and Mask R-CNN [15] have also been proposed. Considering the accuracy and speed of detection, we use Mask R-CNN to detect the number of objects and the spatial relationship between them.

Mask R-CNN is an extension of Faster R-CNN by adding a branch that predicts an object mask. The branch is a small Fully Convolutional Network that predicts a segmentation mask in a pixel-to-pixel manner. Therefore, not only can it solve the classical computer vision tasks of object detection, which consists of classifying individual objects and localising each one of them, it can also solve semantic segmentation which aims to classify each pixel into a fixed set of categories without differentiating object instances. For each keyframe, Mask R-CNN outputs the label, probability and binary mask of all instances in the frame. Figure 2 shows an example of input and output of Mask R-CNN.

Fig. 2. An example of input and output of Mask R-CNN

In our implementation, we apply Mask R-CNN trained on the MS COCO dataset [5], which is a large-scale dataset for object detection, segmentation, and captioning. However, MS COCO only contains 80 object categories, consequently, there are very few related objects in the topics covered by MS COCO. Because of the limitation of related concepts, we have to use the result of the cascade-based approach as basis to get classification information for concepts not included in MS COCO and only apply Mask R-CNN to Topics 561, 563, 572, 584 and 586, whose related concepts are included in MS COCO.

Specifically, we take the top 10000 shots retrieved by the cascade-based approach. For at most five keyframes from each shot, we apply Mask R-CNN. When a topic requires the number of objects, we select shots where the number of related object instances detected by the Mask R-CNN matches the requirements. When a topic has a requirement for the spatial relationship between objects, we judge by the position of the centre of gravity of each object.

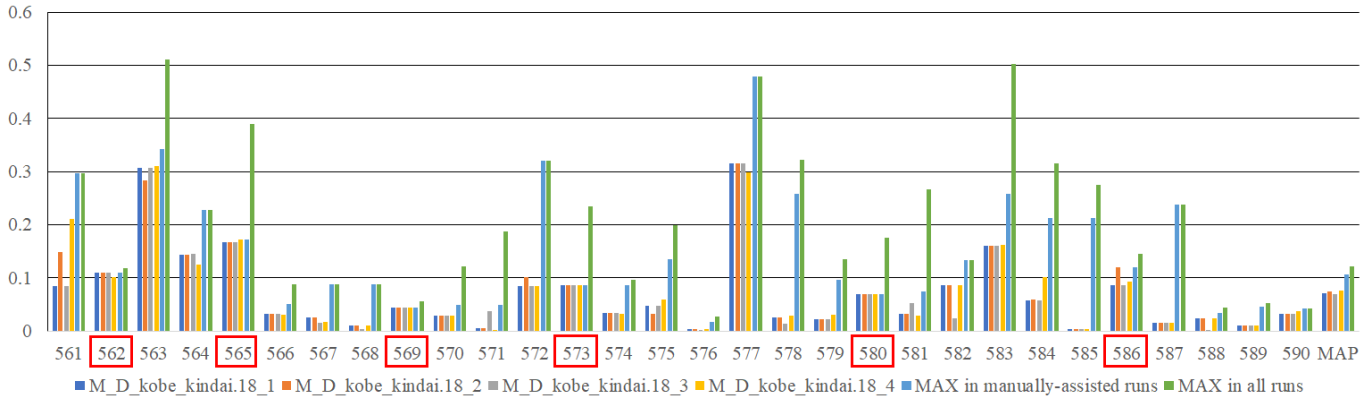

Fig. 3. An overview of results of our submitted four runs as well as the top scores in the manually-assisted category and the ones in the whole of AVS task.

To be precise, first of all, in order to calculate the number of objects, we use the number of instances with the same label in a keyframe as the number of objects. Second, to obtain the spatial relationship between objects, we consider that the average of the coordinates of all the positive pixels in the mask represents the position of the centre of gravity of each instance. By using the coordinates of the centre of gravity between objects, we can determine the spatial relationship between them easily. Furthermore, we filter out the shots whose quantity or spatial position relationship does not meet the requirements to obtain the final result.

As an example, for Topic 561, we define a correct shot as one where there are only two people in the keyframe. For Topic 584, a shot is considered correct when the person's position of centre of gravity is higher than the bed's. The top five results of these two topics are shown in the second row of Figure 4. As for Topic 563, we use a similar approach to Topic 584, a shot is considered correct when the person's position of center of gravity is higher than the boat's. Since the spatial relationship of Topic 586 is more difficult to define, we choose correct shots as the ones in which the overlap rate of the person's and truck's masks is higher than 50%.

III. Results

Figure 3 shows a bar graph that presents average precisions (APs) over 30 topics obtained by each of the submitted runs. For each topic, the four bars from the left represent APs by M_D_kobe_kindai. 18.1, M_D_kobe_kindai.18.2, M_D_kobe_kindai.18.3 and M_D_ kobe_kindai.18.4, which are defined as follows:

1) *M_D_kobe_kindai.18_1:* This is our baseline that uses the cascade-based approach without exploiting object detection.
2) *M_D_kobe_kindai.18_2:* This utilises object detection to refine shots retrieved by M_D_kobe_kindai.18_1.
3) *M_D_kobe_kindai.18_3:* This is a supplementary run where, compared to M_D_kobe_kindai.18_1, slightly different sets of concepts are used for some topics. This run just aims to check the retrieval performance using the concept sets, which could not be used in M_D_kobe_kindai.18_1.
4) *M_D_kobe_kindai.18_4:* To examine the effectiveness of the cascade approach, this runs uses the same sets of concepts to M_D_kobe_kindai.18_1. But, shots are retrieved by just summing up detection scores for the selected concepts.

In addition, for each topic in Figure 3, the second bar from the right and the rightmost bar represent the maximum AP among 16 runs in the manually-assisted category and the one among 51 runs in the whole of AVS task, respectively.

Unexpectedly, M_D_kobe_kindai.18_4 leads to the best performance among our submitted runs (the discuss about this will be provided later). In the manually-assisted category, its MAP 0.077 is ranked the seventh place among 16 runs and our team (kobe_kindai) is ranked at the third place among six teams. More closely, as shown in the six red rectangles in Figure 3, the best APs for Topics 562, 565, 569, 573, 580 and 586 in the manually-assisted category are obtained from our submitted runs. We think that one reason for these good performances is the adoption of the large concept vocabulary consisting of five different concept sets. Especially, the set of "9418 ImageNet concepts" includes key concepts for several topics. As seen from Figure 1 and our website[1], concepts from this set are used for almost all of topics, indicating the importance of this large concept set.

We cannot see any significant difference between using the cascade-based approach (M_D_kobe_kindai.18_1) and not-using it (M_D_kobe_kindai.18_4). Here, the latter's performance (MAP: 0.077) is higher than the former's one (0.072). But, M_D_kobe_kindai.18_1 is superior to M_D_kobe_kindai.18_4 for 11 of 30 topics, and is inferior for 10 topics. Thus, they can be actually considered comparable. In particular, for Topics 561 and 584 where M_D_kobe_kindai.18_1 is significantly outperformed by M_D_kobe_kindai.18_4, it is possible to improve the former's performance by changing the current average-pooling to max-

[1]https://www.info.kindai.ac.jp/~shirahama/tv18/cascade_list.html

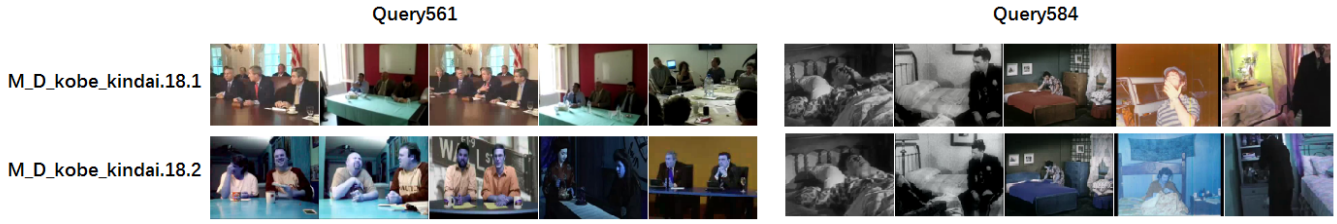

Fig. 4. Top five shots retrieved by M_D_kobe_kindai.18.1 and M_D_kobe_kindai.18.2 for Topics 561 and 584.

pooling[2]. This improves the APs of M_D_kobe_kindai.18_1 from 0.084 to 0.176 for Topic 561, and from 0.058 to 0.107 for Topic 584. In addition, by appropriately choosing either of average- or max-pooling for each topic, the MAP of M_D_kobe_kindai.18_1 can be improved to 0.0803. However, it is non-trivial and needs much heuristics to directly implement this kind of adaptive pooling selection. In the next section, we will describe our future work to avoid this implementation in the framework of neural network.

One clear advantage of the cascade-based approach is the reduction of search times. The average search time of M_D_kobe_kindai.18_1 over 30 topics is 4.0 seconds, and that of M_D_kobe_kindai.18_4 is 5.9 seconds. Especially, for topics involving many concepts like Topics 561, 584 and 586, M_D_kobe_kindai.18_1 is nearly two times faster than M_D_kobe_kindai.18_4.

Figure 4 shows the keyframes for the top five shots retrieved by M_D_kobe_kindai.18.1 (combining scores of related concepts using the cascade-based approach) and M_D_kobe_kindai.18.2 (adding object detection to M_D_kobe_kindai.18.1) for Topics 561 and 584.

From the first row of Figure 4, it is clear that when we only detect the presence of beds and people in keyframes, only a small part of retrieved shots meets the requirements of the topic and a large part of them is that people are sitting on the bed or standing by the bed, etc. For Topic 561, most of the results do not meet the "exactly two people" requirement.

In contrast to the results of the first row, the result of the second row to which we applied Mask R-CNN shows that all top five shots retrieved for Topic 561 meet the requirements of "exactly two people', while for the top five shots for Topic 584, although there are still incorrect shots, the accuracy has also increased slightly.

Figure 5 shows the performance comparison between M_D_kobe_kindai.18.1 and M_D_kobe_kindai.18.2. Topics 561 and 572 require the number of objects, while Topics 563, 584 and 586 require spatial relationships between objects. For queries to which we applied Mask R-CNN, the accuracy has increased except for Topic 563. Figure 5 clearly shows that combining Mask R-CNN allows us to further improve the accuracy of detection. This is especially true when the number of objects is required. As for queries with spatial relationship requirements, applying Mask R-CNN also slightly improved the results.

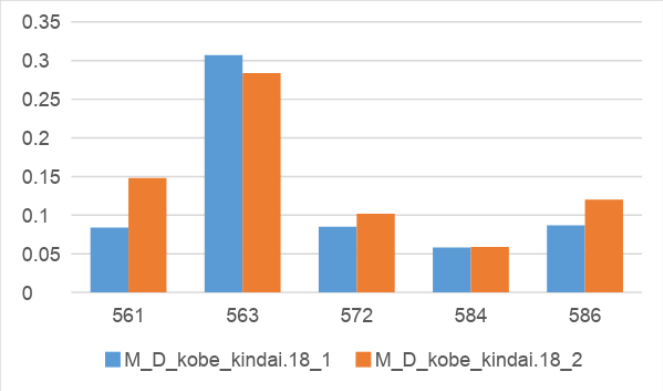

Fig. 5. Perfomance comparison between M_D_kobe_kindai.18.1 and M_D_kobe_kindai.18.2.

IV. Conclusion and Future Work

This paper introduced our video retrieval system developed for TRECVID 2018 AVS task. Our system uses a concept-based approach where concepts related to a topic are organised into cascades to gradually filter out irrelevant shots. In addition, retrieved shots are refined by examining the number or spatial relation of object regions detected by a state-of-the-art object detector (Mask R-CNN).

We strongly think to switch from the current concept-based approach to an "embedding-based" approach. This year, all the manually-assisted methods are actually outperformed by the fully automatic methods, which are expected to be using embedding methods like the one in [21]. Embedding projects both topics and shots into the same space, so they can be directly compared although they are originally from different media. Hence, embedding can avoid various unsolved issues in concept-based retrieval, such as how to select concepts and how to fuse detection scores. In addition, it is no need to suffer from what kind of pooling approach should be used, as discussed in the previous section. We plan to develop an embedding-based retrieval method where the examination of each shot is enhanced and accelerated using a cascade-like scheme based on gate units [22].

By observing the second column of Figure 4, we realize that although we used the spatial information of the objects to

[2]Overall average-pooling works much better than max-pooling. Actually, the MAP of M_D_kobe_kindai.18_1 using average-pooling is 0.072, but it is reduced to 0.065 when max-pooling is used. Thus, it is rare that max-pooling yields an improvement.

judge the relationship between them, the increase in detection accuracy does not reach our expectations. For example, in the shots of Topic 584, there are still many shots in which people are not lying on the bed. This means that the relationship between objects cannot be accurately determined by using only the position of the objects. This is because due to the angle of photography, the same spatial relationship may present different spatial distributions in an image. Therefore, in the future, we will consider using a model [23] that specifically predicts the relationship between objects in order to achieve better performance.

REFERENCES

[1] S. Ayache and G. Quénot, "Video corpus annotation using active learning," in *Proc. of ECIR 2008*, 2008, pp. 187–198.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. of CVPR 2014*, 2014, pp. 1725–1732.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. of ECCV 2014*, 2014, pp. 740–755.

[6] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Qunot, 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.

[7] J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad, "On influential trends in interactive video retrieval: Video browser showdown 2015-2017," *IEEE Transactions on Multimedia*, 2018.

[8] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proc. of MIR 2006*, 2006, pp. 321–330.

[9] Z. He, T. Shinozaki, K. Shirahama, M. Grzegorzekz, and K. Uehara, "Kobe university, nict and university of siegen at trecvid 2017 avs task," in *Proc. of TRECVID 2017*, 2017.

[10] Y. Matsumoto, T. Shinozaki, K. Shirahama, M. Grzegorzek, and K. Uehara, "Kobe university, nict and university of siegen at trecvid 2016 avs task," in *Proc. of TRECVID 2016*, 2016.

[11] L. Wang, J. Lin, and D. Metzler, "A cascade ranking model for efficient ranked retrieval," in *Proc. of SIGIR 2011*, 2011, pp. 105–114.

[12] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of CVPR 2014*, 2014, pp. 580–587.

[13] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," in *Proc. of CVPR 2017*, 2017, pp. 3039–3048.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. of NIPS 2015*, 2015, pp. 91–99.

[15] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," in *Proc. of ICCV 2017*, 2017, pp. 2980–2988.

[16] N. Pittaras, F. Markatopoulou, V. Mezaris, and I. Patras, "Comparison of fine-tuning and extension strategies for deep convolutional neural networks," in *Proc. of MMM 2017*, 2017, pp. 102–114.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR 2016*, 2016, pp. 770–778.

[18] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. of CVPR 2017*, 2017, pp. 6517–6525.

[19] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. of ICCV 2015*, 2015, pp. 4489–4497.

[20] H. Jgou, F. Perronnin, M. Douze, J. Snchez, P. Prez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[21] J. Dong, X. Li, and C. G. M. Snoek, "Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction," *ArXiv e-prints*, 2016.

[22] W. Pei, T. Baltrušaitis, D. M. J. Tax, and L. Morency, "Temporal attention-gated model for robust sequence classification," in *Proc. of CVPR 2017*, 2017, pp. 820–829.

[23] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. of CVPR 2017*, 2017, pp. 3298–3308.