# Florida International University - University of Miami TRECVID 2019

Yudong Tao[1], Tianyi Wang[2], Diana Machado[2], Raul Garcia[2], Yuexuan Tu[1],
Maria Presa Reyes[2], Yeda Chen[1], Haiman Tian[2], Mei-Ling Shyu[1], and Shu-Ching Chen[2]

[1]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33146, USA
[2]School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
yxt128@miami.edu, wtian002@cs.fiu.edu, mpres029@cs.fiu.edu, dmach009@cs.fiu.edu, rgarc101@cs.fiu.edu,
yxt120@miami.edu, yxc806@miami.edu, htian005@cs.fiu.edu, shyu@miami.edu, chens@cs.fiu.edu

**Abstract**

This paper presents the framework and results from the team "Florida International University-University of Miami (FIU-UM)" in the TRECVID 2019 Ad-hoc Video Search (AVS) [1] task. We submitted 7 fully automatic runs as follows.

- run1: unweighted concept fusion + arithmetic mean + weighted W2VV score integration
- run2: weighted concept fusion + geometric mean + W2VV score with threshold integration
- run3: weighted concept fusion + geometric mean + weighted W2VV score integration
- run4: weighted concept fusion + geometric mean
- run5: unweighted concept fusion + geometric mean + W2VV score with threshold integration
- run6: unweighted concept fusion + geometric mean
- novelty run: weighted concept fusion + crawling concepts with description + geometric mean

Our framework includes the following processing steps: (1) automatically parsing the query and generating a concept tree, (2) generation of CNN features from keyframes, (3) generation of concept scores from multiple pre-trained models for image classification, object, scene, and action detection, (4) just-in-time concept learning for keywords not found in the concept bank, (5) word to visual vector (W2VV) image-text matching scores, and (6) integration of the scores based on the concept tree. The performance results show that our fourth run (run4), which includes our best-weighted combination scores and geometric mean, outperforms all the other runs. This year, the FIU-UM team achieved the highest novelty score among all the teams. The submission details are listed as follows.

- Class: F (fully automatic runs)
- Training type: E (used only training data collected automatically, using only the official query textual description)
- Team ID: FIU-UM (Florida International University - University of Miami)
- Year: 2019

## I. Introduction

The TREC Video Retrieval Evaluation (TRECVID) is a competition led by National Institute of Standards and Technology (NIST), which aims to accelerate the research and development in video-based content analysis and retrieval. Since 2010, TRECVID project [2] initiated the challenge of Semantic Indexing (SIN), which requests the participants to predict the semantic tags of the given video segments. In 2016, SIN has been elevated to a more comprehensive challenge, Ad-hoc Video Search (AVS) task, which aims to identify the video segments containing various contents based on a natural language description. Moreover, to track the advancement in this direction, provide a platform for evaluation of AVS tasks, and accommodate the trend in high-resolution video recording, a new V3C dataset [3] is released and used for AVS evaluation efforts.

In order retrieve and categorize the video contents, one main approach is to automatically compute a measure that evaluates the similarity between the contents in the video and the target contents [4–16]. However, many challenges, including data imbalance, scalability, and semantic gap problems [17–25], hamper the automatic content analysis being robust. Therefore, various approaches have been developed to mitigate the problem. Among those, some of the main approaches include: (1) learning and engineering robust video content features to detect concepts in the ad-hoc fashion; (2) integrating multi-modal data, such as image, video, and text, to extract more comprehensive information; and (3) data-driven learning for discriminating features for content analysis [20, 24, 26–35].

In the AVS task, 30 ad-hoc queries for 2019 evaluation are given along with the V3C1 dataset, a subset of the V3C dataset. The goal of the AVS task is to search the V3C1 dataset for the top-1000 video shots that fit the description of each ad-hoc queries. In each query, there could be a combination of any number of concepts described in various ways. The participants need to submit the shot IDs of the video shots in the V3C1 collection, whose video contents best fit the given queries based on their computed likelihood measures. The submission result is rated by using the mean inferred average precision (mean xinfAP) [36] based on the assessment of a 2-tiered random sampling (1-250@100% and 251-1000@11%).

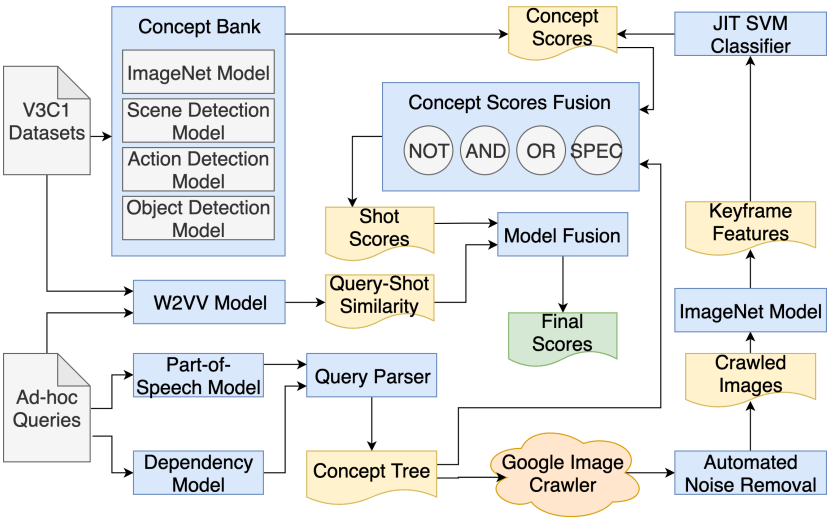


Fig. 1. The designed framework for the TRECVID 2019 AVS task

TABLE I
THE CONCEPT BANK DESCRIBING ALL THE DATASETS AND THE CORRESPONDING DEEP LEARNING MODELS USED IN THE PROPOSED FRAMEWORK

| Model Name | Database | # of concepts | Concept type(s) |
|---|---|---|---|
| InceptionResNetV2 | ImageNet | 1000 | Object |
| ResNet50 | Places | 365 | Scene |
| VGG16 | Hybrid (Places, ImageNet) | 1365 | Object, Scene |
| MaskR-CNN | COCO | 80 | Object |
| ResNet50 | Moments in Time | 339 | Action |
| TRN | Something-Something-v2 | 174 | Action |
| Kinetics-I3D | Kinetics | 400 | Action |

The remainder of this paper is structured as follows. Section 2 explains the proposed framework for the TRECVID 2019 AVS task and the details of different strategies used in each run. Section 3 evaluates the performance of each submission and demonstrates the submission results. Section 4 concludes the paper and suggests future directions for next year's submission.

## II. The Proposed Framework

As shown in Figure 1, the proposed framework incorporates several state-of-the-art pre-trained deep learning models. For image classification, InceptionResNetV2 pre-trained on the ImageNet dataset [37] is applied to detect the concepts in the ImageNet dataset and to extract the features from the keyframes. Moreover, additional advanced pre-trained models are utilized for object, scene, and action detection by generating the prediction scores of their specific concepts. For those concepts in the query not appearing in the concept bank, Google Image Crawler is used to downloads images from the Internet and a Just-In-Time (JIT) Support Vector Machine (SVM) model is trained to generate the prediction scores for them. To identify the interesting concepts (keywords) in the query, the given ad-hoc query is parsed into a concept tree based on its Part-Of-Speech (POS) tags and Dependency (DEP) tags. After all the concept scores are computed, they are fused based on the concept tree as well. Meanwhile, for some of the submitted runs, a pre-trained Word-to-Visual-Vector (W2VV) model is used to generate a query-shot similarity score, which will be integrated to produce the final scores.

### A. Concept Bank

Table I lists the datasets used to train different models for the proposed framework.

*1) ImageNet:* One of the basic datasets we apply is ImageNet [37] which includes large-scale images with various concepts from multiple domains such as animal, instrumentation, scene, and activity. These concepts are most commonly seen in the daily life and frequently appear in the TRECVID queries. There are 1.2 million images in 1000 categories included the ImageNet dataset. The classification accuracy of the models on this dataset has exceeded the human performance using the recent deep neural networks. As shown in Figure 1, an Inception-ResNet-v2 model is applied [38] to identify concepts listed in the queries and to extract the image features from the first dense layer for detecting unknown concepts in the queries.

*2) Places and Hybrid:* Since time and locations are important information required for accurate query answering, scene detection is included in the proposed framework as an essential part of improving the framework's performance. Among all the public scene detection datasets, PLACES365 which includes 365 scene categories is used [39]. A ResNet50 model is applied to detect the location and environment in the videos. In PLACES365 dataset, 1.8 million training images are provided and each class includes at most 5000 images. Furthermore, PLACES365 and ImageNet datasets are integrated to form the HYBRID1365 dataset and a VGG16 model is trained on all the images in both datasets to achieve a better classification accuracy.

*3) COCO:* Although ImageNet1000 provides a lot of object concepts, it has two shortcomings. First, it is specifically designed for image classification where we have a single and clear object that is the main focus in the picture. Therefore, the learning models based on this dataset do not produce a good performance on images with smaller objects compared to other object detection methods, such as Faster R-CNN [40] and Mask R-CNN [41]. Second, ImageNet models cannot detect all the instances of an object within a single image. Thus, an additional object detection dataset, COCO, is incorporated. COCO provides 80 object categories and over 200,000 images. A state-of-the-art object detection toolbox called MMDetection [42] is utilized, which contains a rich set of object detection and instance segmentation methods as well as related components and modules. The Cascade Mask R-CNN model (X-101-64x4d) is chosen as our backbone network which is pre-trained on the COCO dataset to generate the detection scores for each object instance. Cascade Mask R-CNN is a state-of-the-art object detection network that not only detects objects in an image but also provides pixel-level classification. It also solves the close false positive noise issue in Mask R-CNN due to the low IoU threshold and can generate more accurate confidence scores for each detected instance. For certain queries that require a specific number $N$ of an object $O$, the confidence score $P_{O,N}(I)$ of $N$ times of the object $O$ appearing in the image $I$ can be calculated using Equation (1).

$$P_{O,N}(I) = \begin{cases} 0 & n < N \\ \prod_{i=1}^{N} P_O^i(I) & n = N \\ \prod_{i=1}^{N} P_O^i(I) \cdot \prod_{i=N+1}^{n} (1 - P_O^i(I)) & n > N \end{cases} \tag{1}$$

where $n$ is the number of $O$ being detected by the model and $P_O^i(I)$ is the $i$-th highest confidence score among all the detected objects $O$ in image $I$. For example, for the query "608 Find shots of two people talking to each other inside a moving car", we want to obtain a score for "two people" (i.e., $O$ = "person" and $N$ = 2). Given a keyframe of the shot, assume that there are three detected "person" objects (i.e., $n$ = 3) in the image with the confidence scores of 0.99 ($P_O^1(I)$), 0.85 ($P_O^2(I)$), and 0.20 ($P_O^3(I)$). Therefore, the returning confidence score of "two people" in the image is $0.99 \times 0.85 \times (1 - 0.20) = 0.67$.

*4) Moments in Time:* To achieve the good performance in recognizing activity actions on various scenarios and objects, our framework needs to be able to model the spatial-temporal dynamics of the video clips. The Moments in Time dataset [43] contains over one million videos that cover 339 classes. Each video clip is 3 seconds long and includes action activities and scenes among people, animals, and objects. The Moments in Time model utilizes the ResNet50 model pre-trained on the ImageNet dataset as the foundation and fine-tuned on the Moments in Time dataset. Some of the classes show very good performance, such as smoking, sleeping, and running. The output of the model's last softmax layer is utilized as the score for the targeting video.

*5) Kinetics:* The Kinetics human action video dataset (Kinetics400) [44] is a large scale human action classification video dataset that contains 400 action classes performed by humans, in which each class is provided with more than 400 video clips. Our main rationale for choosing Kinetics400, the successor of HMDB51 [45] and UCF101 [46], is its advantage in data variation for each class. Unlike its predecessor, Kinetics400 includes all clips for each class from different video sources, which greatly enriches the data variation for each action by adopting a large variety in people, viewpoints, executions, etc. For this task, the Inception V1-based Inflated 3D ConvNet (I3d) [47] model pre-trained on ImageNet is used and it is subsequently fine-tuned on Kineticts400. Due to the limitation of the computational power, only the RGB stream is adopted to generate the score for the testing queries, which still provides an excellent performance. Each video shot was fed into the model and the results directly from the output layer are used as the scores.

*6) Something-Something:* This year, we have incorporated the concepts from the 20BN-SOMETHING-SOMETHING V2 dataset (Something-SomethingV2) [48] to improve the performance of action recognition. Something-SomethingV2 allows us to train models that can recognize basic interactions between humans and common objects in the physical world such as putting something on a surface and throwing something. In our framework, the Temporal Relation Network (TRN) [49] pre-trained on the Something-SomethingV2 dataset is used. TRN serves to identify temporal dependencies between video frames at multiple time scales. At the time, TRN-equipped networks demonstrated the best performance when predicting human-object interactions from the Something-SomethingV2 dataset with the ability to learn intuitive and interpretable visual common sense knowledge from videos. Exactly eight keyframes were extracted from each video clip and fed into the pre-trained TRN network to obtain the predicted Something-SomethingV2 concepts.

### B. Just-in-Time Concept Learning

Although the concept bank has collectively covered most of the commonly used concepts, there are many concepts unseen in the ad-hoc queries. More importantly, the concepts can be modified by adjectives, adverbs, etc., which creates a huge amount of more specific concepts in the query. Thus, the Just-in-Time Concept Learning method was proposed, which automatically crawls the related images in an image search engine, such as Google Image, as the training data, filters the outliers in the search engine results, and then trains the classifier to detect the concepts for the corresponding query. The phrases used in the query were automatically separated and identified as the searching keywords and fed into our proposed toolchain. For each new concept, around 10,000 images were crawled. After the reference images were downloaded, the features were extracted from the outputs of the first dense layers of the Inception ResNet-v2 model [38], followed by an SVM classifier to determine whether the video shots include the concepts or not.

### C. Concept Tree Formation and Score Fusion

*1) Concept Tree Formation:* To fully automate the video retrieval based on the ad-hoc queries, a natural language processing algorithm is necessary to extract semantic concepts from the given queries. In this year, our proposed model develop a query parser based on the SpaCy library[1]. The Part-Of-Speech (POS) tags and Dependency (DEP) relationships are extracted using the library, which are fed into a query parser along with the original query to build a grammar tree structure, called concept tree, where each leaf node is a semantic concept in the query and each non-leaf node represents the relationship among the concepts. The scores of each concept are obtained from the aforementioned deep learning model and fused based on the type of non-leaf nodes. The types of nodes in the concept tree are listed as follows.

- **Concept:** the basic leaf nodes. It represents a specific semantic concept in the query.
- **Numbered Concept:** an alternative leaf node. It represents that the concept is modified by a number and will be treated differently when using the object detection model.
- **Not Node:** a non-leaf node with only one child, which represents that the query includes a concept with complementary meaning of its child.
- **And Node:** a non-leaf node with two or more children, which represents that the query has its semantic meaning of all its children appearing concurrently.
- **Or Node:** a non-leaf node with two or more children. The query has its semantic meaning that any of its children exists in the video.
- **Spec Node:** a non-leaf node with exactly two children. One is the modifier and the other is the central concept. This relationship shows that the modifier adds specific constrains on the semantic of the central concept. For example, the phrase "moving car" includes a central concept "car" and its modifier "moving", which constrains the semantic meaning of "car" being moving around.
- **Sent Node:** an unique non-leaf node which is essentially an "And Node" while it has at most five children, namely subject, action, object, place, and time, respectively. They cover most of the semantic concepts in the sentences and all TRECVID queries can be summarized into such a structure. The root of a concept tree is a "Sent Node" while the clause and other complex structures in the sentences are parsed as "Sent Node" as well.

The concept tree is very similar to a parse tree using the English grammar; while the concept tree focuses more on the relationship between semantic concepts in the query and ignores the other details including article, preposition, etc. The prepositions in the query are handled during the parsing process, which is used to distinguish time and places.

*2) Score Fusion:* Once the concept tree is built, the likelihood of a video shot similar to a query can be achieved by computing the scores at the root node. Since all the leaf nodes represent a single concept, their scores are determined by the scores of the concepts. Meanwhile, the likelihood score of the non-leaf node is computed purely based on the scores of its children. Since there are different types of nodes in the concept tree, the ways to fuse the scores are different as well. The methods to compute the non-leaf node scores of the aforementioned relationships are defined as the following.

- **Concept:** Get the scores of the corresponding concept based on the concept bank or JIT concept learning.
- **Numbered Concept:** The same as "Concept" except when the concepts are available in object detection model, Equation 1 is used to compute the scores.
- **Not Node:** The score of this node is computed by $1 - s_{child}$, where $s_{child}$ is the score of its child.
- **And Node:** The score of this node is computed by the geometric mean of all the children of the node.
- **Or Node:** The score of this node is determined as the maximum of the scores among all its children.
- **Spec Node:** The score of this node is computed in one of the two ways: the weighted arithmetic or geometric mean of the central concept and the modifier, i.e., $w_c \times s_c + (1 - w_c) \times s_m$ or $s_c^{w_c} \times s_m^{(1-w_c)}$, where $w_c \in [0, 1]$ is the weight of central concept, $s_c$ is the score of its central concept, and $s_m$ is the score of its modifier.

[1]https://spacy.io/

- **Sent Node:** It is the same as the "And Node" except that for some runs, a weight is added to each type of the concepts in the sentence and a weighted geometric mean is applied.

### *D. Video-Text Matching Model*

Since it is hard to exhaustively train a model with all the concepts in the queries, the word embedding technique is utilized to integrate with the visual features and to map them to a common latent space for comparing the similarity between the text and visual data. Using deep neural networks, this structure is able to achieve an excellent performance in ad-hoc video retrieval. In our designed framework, a well-known text-video matching model, called Word2VisualVector (W2VV) [50], is applied to compute the similarity scores. The applied W2VV model is trained on the Flickr30k dataset [51].

### *E. Model Fusion*

Based on both the concept tree approach and W2VV approach, two different scores for each query-shot pair can be obtained. In some of our submitted runs, the scores from both models are utilized and integrated with the attempt to achieve better scores, though based on the evaluation results, the pure concept tree approach achieves the best performance among all our submitted results. This indicates that our W2VV model under-performs other team's video-text models and hampers the overall performance in the submitted runs.

- Weight-based Fusion: the scores from two approaches are weighted based on their distributions and aligned with the mean and variance. After that, these two scores are summed to produce a new score for each query-shot pair, which is used to generate the final top-1000 shots for each given query.
- Threshold-based Fusion: we assume that the more the concepts in the query appear in the training dataset, the higher performance the video-text matching model can achieve. Therefore, for each given query, the average frequency of those concepts appearing in the W2VV training dataset is calculated to decide whether the results from W2VV should be used or not based on an empirically learned threshold.

### *F. Submitted Runs*

The following seven runs are submitted to the TRECVID 2019 AVS task by our team. In all these runs, as mentioned earlier, various pre-trained models (e.g., ImageNet, COCO, YOLO, Moments, Places, etc.) are used to generate the likelihoods of each pre-defined concept in the concept bank, to extract the CNN features based on the Inception ResNet-v2 [38], and to train a linear SVM for all the unknown concepts based on the images crawled from the Internet using the automatically generated queries. The query is automatically parsed into a concept tree and the scores from various components are fused based on the tree. The differences between all the submitted runs are how the scores are fused and how the weight is assigned to each concept.

- **run1**: unweighted arithmetic mean for concept score fusion and integrating scores from the W2VV model using the weights learned by our empirical study;
- **run2**: weighted geometric mean for concept score fusion and integrating scores from the W2VV model using the thresholds learned by our empirical study;
- **run3**: weighted geometric mean for concept score fusion and integrating scores from the W2VV model using the weights learned by our empirical study;
- **run4**: weighted geometric mean for concept score fusion and without the W2VV model;
- **run5**: unweighted geometric mean for concept score fusion and integrating scores from the W2VV model using the thresholds learned by our empirical study;
- **run6**: unweighted geometric mean for concept score fusion and without the W2VV model;
- **novelty_run**: weighted geometric mean for concept score fusion and using only the most specific concepts.

## III. Results

### *A. Evaluation*

Our framework generates a list of at most 1000 video shot IDs based on the given 30 queries, the reference shots, and the V3C1 dataset [3]. This dataset contains 7475 Internet Archive videos with a total duration of around 1000 hours and 1.08 million video segments. The mean duration of each video is 8 minutes and 2 seconds. All the results are evaluated by the assessors at NIST as described in [52]. All the top-250 results and 11% of the remaining results of each query are evaluated, and the mean extended inferred Average Precision (mean xinfAP) metrics [36] are computed based on the performance of these evaluated results. Meanwhile, the detailed metrics such as inferred interpolated recall and precision at different depths are given by the *sample_eval* software provided by NIST.

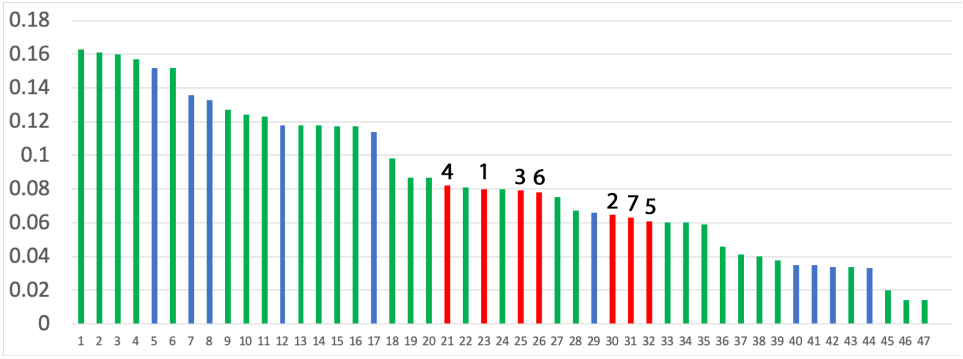


Fig. 2. Comparison of FIU_UM runs (red) with other runs for all the submitted fully automated (green) and manually-assisted (blue). The IDs of our runs are annotated on top of the bars.

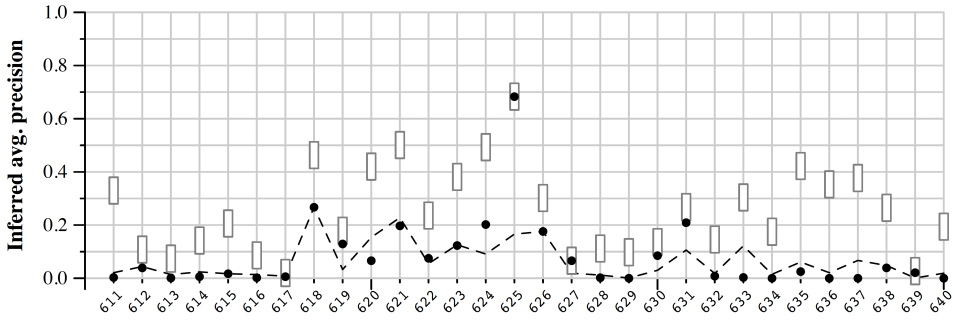


Fig. 3. Inferred precision (dot) of our best fully-automatic run (run4), the median (dashes), and the best (box) results for each query.

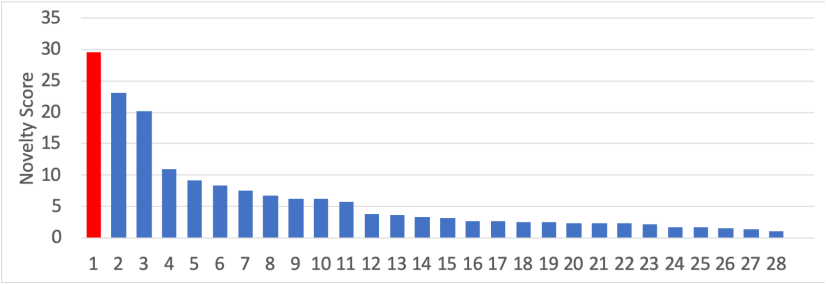


Fig. 4. Comparison of the novelty scores among FIU_UM novelty runs (red) with other runs

### *B. Performance*

The performance (xinfAP) of all the runs based on our proposed framework is shown in Figure 2. All our submitted runs are fully automatic runs and their xinfAP scores are 0.080, 0.065, 0.079, 0.082, 0.061, 0.078, and 0.063, which ranked 21st, 23rd, 25th, 26th, 30th, 31st, and 32nd among all the runs, respectively.

Figure 3 shows the inferred average precision of each query of our best run (run4). The x-axis of Figure 3 shows the query number; while the y-axis presents the xinfAP measures of our run (shown as a dot), median performance (shown as dashes), and the best result (shown as a box) for each query. These query-level metrics indicate that we perform the best in query 625 and query 627.

In addition to xinfAP, in this year, a novelty score is also evaluated, indicating how many unique shots are retrieved by the proposed model among all teams. As shown in Figure 4, our submitted novelty run ranks the first among all the runs. The novelty score is computed based on the average unique shot weights over all the queries.

$$S_{novelty} = \frac{1}{|\mathbb{Q}|} \sum_{q \in \mathbb{Q}} \sum_{s \in \mathbb{S}} \left[ 1 - \frac{N(q,s)}{N_{run}} \right] \tag{2}$$

where $\mathbb{Q}$ is the set of all queries, $|\mathbb{Q}|$ is the number of queries in $\mathbb{Q}$, $\mathbb{S}$ is the set of shots detected by the run being evaluated,

$N_{run}$ is the total number of submitted runs, which is 47 in 2019, and $N(q, s)$ is the number of times that the shot $s$ was retrieved by any run submission for query $q$. For each shot, if a shot is uniquely identified by one run, a score of 0.978 is assigned; while if a shot is retrieved by all runs, 0 novelty score is assigned. It is obvious that our framework is able to obtain the most number of shots that the other teams are hard to retrieve. This can be understood in two folds. On one hand, using the detailed description of a concept to form the queries can help obtain good results. On the other hand, it reveals one weakness of the deep-learning-based approaches that the performance of the model heavily depends on the training dataset, i.e., the retrieved results tend to converge to whatever the model has been seen in the training dataset but cannot generalize well to true concepts.

## IV. Conclusion and Future Work

In this notebook paper, the framework and results of the FIU-UM team in the TRECVID 2019 AVS task are presented. This year, in addition to the classic datasets such as ImageNet, Places, and UCF101, we leveraged several recently released datasets such as Moment339 for action recognition. Also, a new model "Mask R-CNN" is applied to improve the object recognition performance and also to estimate the number of objects for some queries (e.g., "exactly two men at conference"). Although we achieved a good performance this year, it can be seen that the overall score of the AVS task for all the teams is still very low. This problem is mainly due to the complicated queries (e.g., "a truck standing still while a person is walking beside or in front of it"), as well as the noisy and imbalanced nature of the TRECVID dataset which represents the real-world data. In the future, we will focus on utilizing more temporal information from video datasets and a better fusion model. In addition, we will try to generate a fully automatic video retrieval system.

## References


[1] J. Lokoc, W. Bailer, K. Schoeffmann, B. Münzer, and G. Awad, "On influential trends in interactive video retrieval: Video browser showdown 2015-2017," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3361–3376, 2018.

[2] G. Awad, C. G. Snoek, A. F. Smeaton, and G. Quénot, "Trecvid semantic indexing of video: A 6-year retrospective," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 187–208, 2016.

[3] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt, "V3c–a research video collection," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 349–360.

[4] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "Augmented transition networks as video browsing models for multimedia databases and multimedia information systems," in *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, 1999, pp. 175–182.

[5] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval," in *Proceedings of the 7th IEEE International Symposium on Multimedia*, 2005, pp. 37–44.

[6] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "Image retrieval by color, texture, and spatial information," in *Proceedings of the 8th International Conference on Distributed Multimedia System*, 2002, pp. 152–159.

[7] M.-L. Shyu, K. Sarinnapakorn, I. Kuruppu-Appuhamilage, S.-C. Chen, L. Chang, and T. Goldring, "Handling nominal features in anomaly intrusion detection problems," in *Proceedings of the 15th IEEE International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications*, 2005, pp. 55–62.

[8] S.-C. Chen, S. H. Rubin, M.-L. Shyu, and C. Zhang, "A dynamic user concept pattern learning framework for content-based image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 36, no. 6, pp. 772–783, 2006.

[9] T. Meng and M.-L. Shyu, "Leveraging concept association network for multimedia rare concept mining and retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, July 2012, pp. 860–865.

[10] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 715–734, 2001.

[11] X. Chen, C. Zhang, S.-C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 2, pp. 228–233, 2009.

[12] M.-L. Shyu, C. Haruechaiyasak, S.-C. Chen, and N. Zhao, "Collaborative filtering by mining association rules from user access sequences," in *Proceedings of the IEEE International Workshop on Challenges in Web Information Retrieval and Integration*, 2005, pp. 128–135.

[13] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, "Deep learning for imbalanced multimedia data classification," in *Proceedings of the IEEE International Symposium on Multimedia*, 2015, pp. 483–488.

[14] X. Huang, S. Chen, M. Shyu, and C. Zhang, "User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval," in *Proceedings of the Third International Workshop on Multimedia Data Mining*, 2002, pp. 100–108.

[15] M. Shyu, S. Chen, M. Chen, and C. Zhang, "A unified framework for image database clustering and content-based retrieval," in *Proceedings of the Second ACM International Workshop on Multimedia Databases*, 2004, pp. 19–27.

[16] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. E. P. Reyes, M. Shyu, S. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 92:1–92:36, 2019.

[17] C. Chen, Q. Zhu, L. Lin, and M.-L. Shyu, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, pp. 61:1–61:22, October 2013.

[18] L. Zheng, C. Shen, L. Tang, C. Zeng, T. Li, S. Luis, and S.-C. Chen, "Data mining meets the needs of disaster information management," *IEEE Transactions on Human-Machine Systems*, vol. 43, pp. 451–464, 2013.

[19] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Effective feature space reduction with imbalanced data for semantic concept detection," in *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 2008, pp. 262–269.

[20] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Effective supervised discretization for classification based on correlation maximization," in *Proceedings of the IEEE International Conference on Information Reuse and Integration*, 2011, pp. 390–395.

[21] ——, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *Proceedings of the 4th IEEE International Conference on Semantic Computing*, 2010, pp. 462–469.

[22] M.-L. Shyu, T. Quirino, Z. Xie, S.-C. Chen, and L. Chang, "Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 2, no. 3, p. 9, 2007.

[23] S.-C. Chen and R. L. Kashyap, "Temporal and spatial semantic models for multimedia presentations," in *Proceedings of the International Symposium on Multimedia Information Processing*, 1997, pp. 441–446.

[24] L. Lin, M.-L. Shyu, G. Ravitz, and S.-C. Chen, "Video semantic concept detection via associative classification," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2009, pp. 418–421.

[25] N. Rishe, J. Yuan, R. Athauda, S.-C. Chen, X. Lu, X. Ma, A. Vaschillo, A. Shaposhnikov, and D. Vasilevsky, "Semanticaccess: Semantic interface for querying databases," in *Proceedings of the VLDB conference*, September 2000, pp. 591–594.

[26] S.-C. Chen, M.-L. Shyu, and R. L. Kashyap, "Augmented transition network as a semantic model for video data," *International Journal of Networking and Information Systems*, vol. 3, no. 1, pp. 9–25, 2000.

[27] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems*, vol. 3, no. 3, pp. 319–337, 2001.

[28] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Category cluster discovery from distributed www directories," *Information Sciences*, vol. 155, no. 3, pp. 181–197, 2003.

[29] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, "Spatio-temporal analysis for human action detection and recognition in uncontrolled environments," *International Journal of Multimedia Data Engineering and Management*, vol. 6, no. 1, pp. 1–18, 2015.

[30] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative shot boundary detection for video indexing," in *Video Data Management and Information Retrieval*. IGI Global, 2005, pp. 217–236.

[31] L. Lin and M.-L. Shyu, "Weighted association rule mining for video semantic detection," *International Journal of Multimedia Data Engineering and Management*, vol. 1, no. 1, pp. 37–54, 2010.

[32] S.-C. Chen, M.-L. Shyu, and C. Zhang, "An intelligent framework for spatio-temporal vehicle tracking," in *Proceedings of the 4th IEEE International Conference on Intelligent Transportation Systems*, 2001, pp. 213–218.

[33] T. Meng, Y. Liu, M.-L. Shyu, Y. Yan, and C.-M. Shu, "Enhancing multimedia semantic concept mining and retrieval by incorporating negative correlations," in *Proceedings of the IEEE International Conference on Semantic Computing*, 2014,

pp. 28–35.

[34] S. Chen and R. L. Kashyap, "A spatio-temporal semantic model for multimedia database systems and multimedia information systems," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 4, pp. 607–622, 2001.

[35] S. Pouyanfar, Y. Yang, S. Chen, M. Shyu, and S. S. Iyengar, "Multimedia big data analytics: A survey," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 10:1–10:34, 2018.

[36] E. Yilmaz, E. Kanoulas, and J. A. Aslam, "A simple and efficient sampling method for estimating AP and NDCG," in *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 603–610.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *the 31th AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[39] 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, 2017.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[41] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[42] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Mmdetection: Open mmlab detection toolbox and benchmark," *CoRR*, vol. abs/1906.07155, 2019.

[43] M. Monfort, B. Zhou, S. A. Bargal, A. Andonian, T. Yan, K. Ramakrishnan, L. M. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva, "Moments in time dataset: one million videos for event understanding," *CoRR*, vol. abs/1801.03150, 2018.

[44] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[45] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2556–2563.

[46] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[47] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4724–4733.

[48] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 5843–5851.

[49] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, 2018, pp. 831–846.

[50] J. Dong, X. Li, and C. G. M. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3377–3388, 2018.

[51] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 74–93, 2017.

[52] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.