# Florida International University - University of Miami TRECVID 2018

Samira Pouyanfar[1], Yudong Tao[2], Haiman Tian[1], Maria Presa Reyes[1], Yuexuan Tu[2], Yilin Yan[2],
Tianyi Wang[1], Hector Cen[1], Yingxin Li[1], Saad Sadiq[2], Mei-Ling Shyu[2],
Shu-Ching Chen[1], Winnie Chen[3], Tiffany Chen[3], and Jonathan Chen[4]

[1]School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
[2]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33146, USA
[3]School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907, USA
[4]Miami Palmetto Senior High School
Miami, FL 33156, USA

spouy001@cs.fiu.edu, yxt128@miami.edu, htian005@cs.fiu.edu, mpres029@cs.fiu.edu, yxt120@miami.edu,
y.yan4@umiami.edu, wtian002@cs.fiu.edu, hcen001@cs.fiu.edu, yli120@cs.fiu.edu, saadsadiq@miami.edu,
shyu@miami.edu, chens@cs.fiu.edu, chen1219@purdue.edu, chen1791@purdue.edu, jc201923@gmail.com

**Abstract**

This paper presents the framework and results from the team "Florida International University-University of Miami (FIU-UM)" in the TRECVID 2018 Ad-hoc Video Search (AVS) [1] task. We submitted four manually-assisted runs as follows.

- M_D_FIU_UM.18_1 & M_D_FIU_UM.18_3: Convolutional Neural Network (CNN) features + linear Support Vector Machine (SVM), scores from other sources, two different sets of concepts and weighted combinations ("and", "or", & "mix" operations)
- M_D_FIU_UM.18_2: CNN features + linear SVM, scores from other sources, weighted combination ("and", "or", & "mix" operations) + rectified linear score normalization
- M_D_FIU_UM.18_4: CNN features + linear SVM, scores from other sources, weighted combination ("and", "or", & "mix" operations) + fuse different score sets ("merge" operation)

Our framework includes the following processing steps: (1) manual extraction of the most important keywords based on a given query, (2) generation of CNN features from keyframes, (3) generation of scores for each concept using the linear SVM classifier, (4) generation of additional scores from multiple pre-trained models for image classification, object, scene, and action detection, (5) just-in-time concept learning for keywords not found in the concept bank, and (6) integration of the scores using the "and", "or", "mix", and "merge" operators. The performance results show that our first run (M_D_FIU_UM.18_1), which includes our best-weighted combination scores, outperforms the other three runs. This year, the FIU-UM team achieved the second highest score in the manually-assisted run and ranked third among all the submitted runs (combining manually-assisted and automatic runs). The submission details are listed as follows.

- Class: M (Manually-assisted runs)
- Training type: D (IACC & non-IACC non-TRECVID data)
- Team ID: FIU-UM (Florida International University - University of Miami)
- Year: 2018

## I. Introduction

The core purpose of the TREC Video Retrieval Evaluation (TRECVID) is to stimulate progress in the domain of content-based analysis and content retrieval from digital video data. From the years 2010 to 2015, TRECVID project [2] addressed the challenge of Semantic Indexing (SIN), which aims to identify the semantic tags that a given video segment contains. This task was elevated, in 2016, to a more comprehensive Ad-hoc Video Search (AVS) task that looks for not only the video segments containing persons, objects, activities, locations, etc. but also segments with their combinations. The AVS task remains the same for this year as well, i.e., to model the end user search use-cases for concepts in video segments and their combinations.

The automatic metric-based evaluation of video segments is a fundamental process for retrieval and categorization of video content [3–10, 10–12]. However, there are several challenges that impede the automatic annotation of semantic concepts such as data imbalance, scalability, and semantic gap problems [13–21]. Some of the main research directions for semantic concept retrieval include: (1) developing robust learning approaches that adapt to the increasing size and the diversity of video content; (2) fusing information from other sources such as audio and text; and (3) detecting low-level and mid-level features that have high discriminating capabilities [16, 20, 22–29].

In the AVS task, there are 346 high-level semantic concepts provided by IACC, where each concept contains a list of ground truth labels provided for training. Given the master shot reference test collection (IACC.3) and 30 Ad-hoc queries, the goal is to return at most 1000 shot IDs for each query, where each query can be a combination of the 346 concepts and/or some

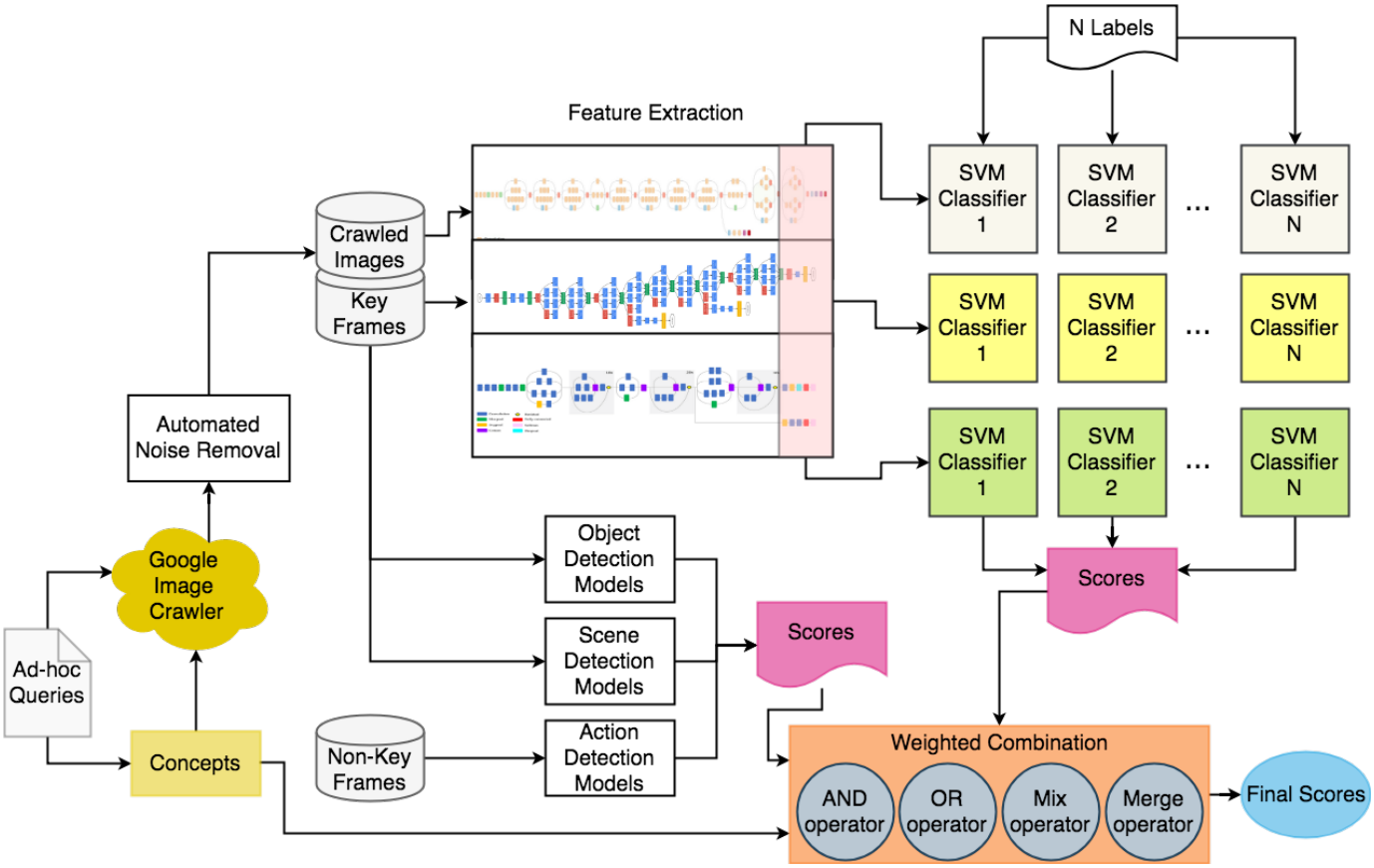


Fig. 1. The designed framework for the TRECVID 2018 AVS task

other concepts not included in the training set. The shot IDs from the test collection are ranked according to their likelihood of containing the target query. The submission result is rated by using the mean inferred average precision (mean xinfAP) [30] based on the assessment of a 2-tiered random sampling (1-150@100% and 151-10002.5%).

The remainder of this paper is structured as follows. Section 2 explains the framework proposed by our team, alongside with the details of the methods used at each run. Section 3 evaluates the performance of each submission and demonstrates the submission results. Section 4 concludes the paper with discussions and key takeaways, 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 the concepts provided by TRECVID, we used three deep learning models including InceptionV3, InceptionV4, and InceptionResNetV2 pre-trained on the ImageNet dataset [31] to extract the features from the keyframes. For each concept, a binary SVM classifier is trained to generate the final classification scores. Moreover, additional advanced pre-trained models are utilized for object, scene, and action detection by generating the prediction scores of their specific concepts. Interesting concepts (keywords) are selected based on their semantic similarity to the respective Ad-hoc query. The prediction scores from the identified concepts contribute to the calculation of the final scores.

### A. Concept Bank

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

*1) TRECVID:* We used three pre-trained deep learning models to extract features from both the training and testing keyframes using the 346 TRECVID concepts. These pre-trained models are described as follows.

- InceptionV3: a widely recognized image recognition model achieving up to 78.1% accuracy on the ImageNet dataset [32];
- InceptionV4: introducing more uniform modules to enable a boost in performance [33];
- InceptionResNetV2: a variation of the InceptionV3 model that borrows some ideas from Microsoft's ResNet [34].

To perform transfer learning, we extracted features from the next-to-last layer of the networks. In the case of our pre-trained models, we used the Average Pooling layer, resulting in a dimensional vector of size 2048 in the case of InceptionV3, and 1536

TABLE I
THE CONCEPT BANK DESCRIBING ALL THE DATASETS AND THE CORRESPONDING DEEP LEARNING MODELS WE USED IN OUR SYSTEM

| Model Name | Database | # of concepts | Concept type(s) |
|---|---|---|---|
| InceptionV3 | TRECVID | 346 | Object, Scene, Action |
| InceptionV4 | TRECVID | 346 | Object, Scene, Action |
| InceptionResNetV2 | TRECVID | 346 | Object, Scene, Action |
| ResNet50 | ImageNet | 1000 | Object |
| VGG16 | Places | 365 | Scene |
| VGG16 | Hybrid (Places, ImageNet) | 1365 | Object, Scene |
| MaskR-CNN | COCO | 80 | Object |
| YOLO | YOLO9000 | 9000 | Object |
| ResNet50 | Moments in Time | 339 | Action |
| Kinetics-I3D | Kinetics | 400 | Action |

for both InceptionV4 and InceptionResNetV2. For each model, we trained a binary SVM classifier on each TRECVID concept to generate the scores. Three training datasets from the 2010-2015 SIN task, namely the IACC.1.tv10.training, IACC.1.A-C, and IACC.2.A-C, were integrated.

*2) ImageNet:* ImageNet [31] is a well-known large-scale image dataset that includes concepts from multiple domains such as animal, instrumentation, scene, and activity, all of which appear in some of the queries. In total, ImageNet contains 1.2 million images belonging to 1000 classes. This dataset includes a lot of common objects in the real world, and the classification accuracy of models on this dataset has exceeded the human performance using the recent deep neural networks. In our framework, we used a ResNet [34] pre-trained model to generate the prediction scores for concepts in each keyframe from the last dense layer.

*3) Places and Hybrid:* Because some of the queries specify not only the objects but also the surroundings, scene detection is an essential part of improving the framework's performance. PLACES365 introduces 365 scene categories, which is very useful in the detection of location and environment [35]. It provides 1.8 million training images, where each class includes at most 5000 images. HYBRID1365 incorporates 365 scene categories from PLACES365 and 1000 object categories from ImageNet. Both of the datasets mentioned above are deployed on the VGG16 pre-trained CNN model to generate the prediction scores of all concepts for each keyframe, extracted from the last fully connected layer.

*4) COCO and YOLO:* 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; and 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 [36] and Mask R-CNN [37]. Second, ImageNet models cannot detect all the instances of an object within a single image. Thus, we incorporated two additional object detection datasets: COCO and YOLO9000. COCO provides 80 object categories and over 200,000 images. We used the Mask R-CNN model which is pre-trained on the COCO dataset to generate the detection scores for each object instance. 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. In addition, we incorporated another pre-trained object detection dataset called YOLO9k, which contains about 9,000 classes. Both models can generate 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 "561 Find shots of exactly two men at a conference or meeting table talking in a room", we want to obtain a score for "exactly two men" (i.e., $O$ = "men" 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 "exactly two men" in the image is $0.99 \times 0.85 \times (1 - 0.20) = 0.67$.

*5) Moments in Time:* Semantics in most of the queries are related to one or several agents (i.e., person, animal, etc.) performing some actions. The "Moments in Time" dataset [38] provides models trained on very large, comprehensive, and labeled datasets of three-second videos that capture people, animals, and objects in diverse and dynamic action scenes. This model was trained using a dataset of approximately one million 3-second videos and outputs the prediction scores over 339 classes. Examples of the classes are *bowling*, *surfing*, *hiking*, *sailing*, etc. The weights for training the Moments in Time model

are taken from a 50 layer ResNet network initialized on the ImageNet dataset. For our task, the scores for the shot keyframes were extracted from the model's softmax layer.

*6) Kinetics:* The concepts from the Kinetics human action video dataset (Kinetics400) [39] were incorporated to improve the performance of action recognition. Kinetics400 contains 400 human action classes with at least 400 video clips for each action and a total of 306,245 clips. Kinetics400's main advantage over its predecessors, such as HMDB51 [40] and UCF101 [41], is the large variation for each action. For HMDB51 and UCF101, multiple clips of the same action may be originated from the same video, which makes these clips less variant in terms of viewpoint, lighting, etc. In comparison, each clip in Kinetics400 was taken from different videos. The model of our choice is the InceptionV1-based Inflated 3D ConvNet (I3D) [42], which has one of the best action recognition result so far on the UCF101 dataset. The network model was pre-trained on Kinetics400, and the concept scores were generated directly from the output layer for all the non-keyframes extracted from the TRECVID dataset. We extracted the scores for all the concepts using the pre-trained network on all ten non-keyframes for each video-shot.

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

There are a few queries where all the models mentioned above cannot find any relevant concepts. 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 key concepts in the queries were manually 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 InceptionV3 model, followed by an SVM classifier to determine whether the video shots include the concepts or not.

### *C. Query Formulation and Score Combination*

*1) Query Formulation:* In our submission this year, all the four runs are manually-assisted runs. Given an Ad-hoc query phrase, our team members manually formulate it into a combination of concepts based on its topic and query interface without the knowledge of the collection or the search results. In comparison to last year, a large concept bank was built as described in Section II.A which covers most of the concepts in the queries. Additional models were trained if certain concepts were not included in the concept bank. For instance, given the Ad-hoc query "561 Find shots of exactly two men at a conference or meeting table talking in a room," the concepts "Adult", "Indoor", and "Male Person" from TRECVID, and the concept "conference room" from PLACES365 were extracted, while an additional model especially for "meeting table" was trained using the just-in-time concept training toolchain.

*2) Score Combination:* Four different methods, "and", "or", "mix", and "merge" operations were utilized to combine the scores from different concepts, based on the relationships among the concepts and the query.

- "and" Combination: Since the performance of the weighted geometric mean of the scores is often better than the weighted arithmetic mean, which also has been reported by several groups in past years, the scores of the selected concepts were fused by calculating their weighted geometric mean. To combine the score ($S_i$) related to each concept $c_i$, different weights ($w_i$) were used based on our empirical studies in all the runs as follows.

$$\text{Score}_{\text{query}}^{\text{and}} = \prod_{i=1}^{\mathcal{N}} S_i^{w_i} \tag{2}$$

  where $\mathcal{N}$ is the total number of concepts integrated.
- "or" Combination: We also leveraged "or" combination in which only the maximum score of all "or" concepts was used. Take the same query 561 as an example. There is no need for a video shot to include both "conference room" and "meeting table". In other words, it is sufficient if only one of these concepts is included in the video, and only the larger score of these two concepts is used in the integration. Furthermore, if a concept can be represented by several subclasses or different related classes, it is more reasonable to use the "or" operation to combine the scores rather than the "and" operation.

$$\text{Score}_{\text{query}}^{\text{or}} = \max_{i=1,\ldots,\mathcal{N}} S_i \tag{3}$$

- "mix" Combination: For some queries, we need to "mix" the "and" and "or" operations. For instance, for query 561, if two or more "or" operations are included in a query, they may be assigned with different weights based on our empirical studies as follows.

$$\text{Score}_{\text{query}}^{\text{mix}} = \prod_{i=1}^{\mathcal{N}_0} S_i^{w_i} \times \prod_{j=1}^{\mathcal{M}} {S'_j}^{w'_j} \tag{4}$$

where $\mathcal{N}_0$ is the number of concepts integrated directly by the "and" operation, $\mathcal{M}$ is the number of groups of concepts integrated by the "or" operation, $S'_j$ is the score obtained from Equation (3), and $w'_j$ is the weight for the $j$-th group obtained from the empirical studies.

- "merge" Combination: This year, we also proposed a "merge" operation as a new way to fuse the scores from the previous three operations. Since there are different ways to interpret the query, we can use different combinations of concepts to generate the results. The proposed "merge" operation assigns a weight to each combination of concepts ($\text{comb}_k$) to merge the results from various interpretations of the query. The weight is determined based on the number of concepts integrated by the "and" operation. Given the scores of several combinations ($S_{\text{comb}_k}$) with different weights ($w_{\text{comb}_k}$), they are merged using Equation (5).

$$\text{Score}^{\text{merge}}_{\text{query}} = \max_k w_{\text{comb}_k} \times S_{\text{comb}_k} \tag{5}$$

### D. Submitted Runs

The following four runs are submitted to the TRECVID 2018 AVS task by our team. In all these runs, as mentioned earlier, we used the CNN features for TRECVID and trained a linear SVM for each concept in this dataset. We also utilized all other scores from the proposed concept bank (e.g., ImageNet, COCO, YOLO, Moments, Places, etc.). The differences between these runs are the way we fused the scores and how we assigned the weight to each concept.

- **Manual1 (M_D_FIU_UM.18_1)**: CNN features + linear SVM for the TRECVID dataset, scores from other sources in the concept bank, the best set of concepts and the weighted combinations ("and", "or", & "mix" operations) based on our empirical study;
- **Manual2 (M_D_FIU_UM.18_2)**: CNN features + linear SVM for the TRECVID dataset, scores from other sources in the concept bank, the best set of concepts and the weighted combinations ("and", "or", & "mix" operations) based on our empirical study+ rectified linear score normalization introduced in [43];
- **Manual3 (M_D_FIU_UM.18_3)**: CNN features + linear SVM for the TRECVID dataset, scores from other sources in the concept bank, the second best set of concepts and the weighted combinations ("and", "or", & "mix" operations) based on our empirical study;
- **Manual4 (M_D_FIU_UM.18_4)**: CNN features + linear SVM for the TRECVID dataset, scores from other sources in the concept bank, fusing different score sets ("merge" operation).

## 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 TRECVID 2017 test dataset IACC.3 [44]. This dataset contains 4593 Internet Archive videos with a total duration of 600 hours. The duration of each video is between 6.5 and 9.5 minutes. All the results are evaluated by the assessors at NIST as described in [45]. All the top-150 results and 2.5% of the remaining results of each query are evaluated, and the mean extended inferred Average Precision (mean xinfAP) metrics [30] are computed based on the performance of these evaluated results. Meanwhile, the detailed metrics such as inferred interpolated recall precision and inferred precision at different depths are given by the *sample_eval* software provided by NIST.

### B. Performance

The performance (xinfAP) of all the runs based on our proposed framework is shown in Figure 2. All our submitted runs (Manual1, Manual2, Manual3, Manual4) are manually-assisted runs and their xinfAP scores are 0.089, 0.079, 0.079, 0.080, which ranked 7th, 13th, 14th, 15th among all the 51 runs, respectively. Our framework's overall performance ranked third among all 13 teams who submitted at least one result in the AVS task.

Figure 3 shows the inferred average precision of each query of our best run (Manual1). The x-axis of Figure 3 shows the query number; while the y-axis presents the infAP 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 queries 563, 568, 587, and 589. These good results are achieved by different parts of our proposed framework. The "boating", "hiking", and "raining" concepts in Moments339 are the key concepts to retrieve the videos for queries 563, 568, and 589, respectively. The just-in-time concept training helps to train the model to identify the concepts, including "look through the window" in query 587, "people queue" in query 569, etc. Our framework also performs well in query 572 in which the object detection models (COCO/YOLO) count the object "cat" which could be an important reason to obtain the current performance. Moreover, the performance of query 578 benefits from the "or" operation of the score fusion. Since the scenes of the concepts "in front of a garage" and "inside a garage" are significantly different, it is more reasonable to use the "or" operation to merge them. Finally, by combining this operation and our just-in-time concept training, we can identify various scenes suitable for "a dog playing outdoors" and train different concepts for each of them, which helps us to improve the retrieval results of query 566.

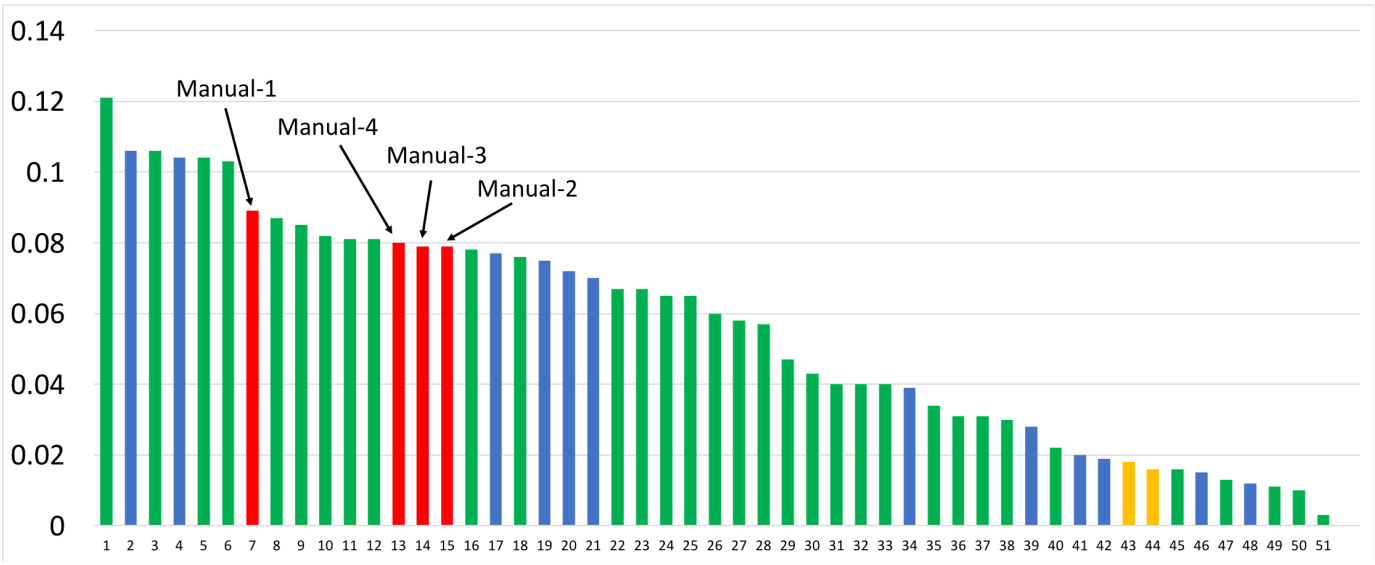


Fig. 2. Comparison of FIU_UM runs (red) with other runs for all the submitted fully automated (green), manually-assisted (blue), and relevance-feedback (orange) results.

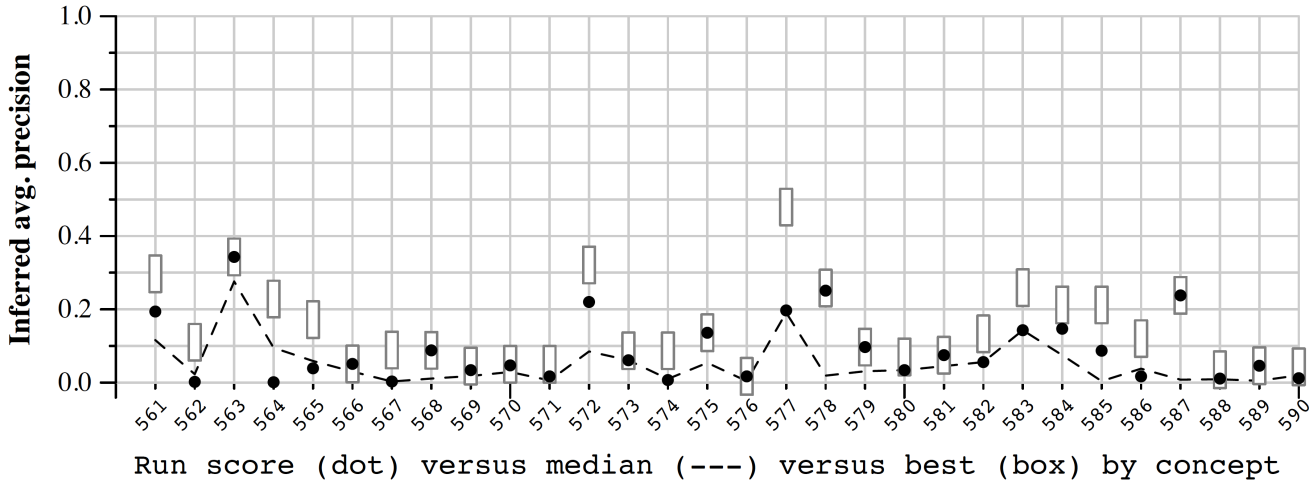


Fig. 3. Inferred precision (dot) of our best manually-assisted run (Manual1), the median (dashes), and the best (box) results for each query.

## IV. Conclusion and Future Work

In this notebook paper, the framework and results of the FIU-UM team in the TRECVID 2018 AVS task are presented. This year, in addition to the classic datasets such as ImageNet, Places, and UCF101, we leveraged 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. Muenzer, and G. Awad, "On influential trends in interactive video retrieval: Video browser showdown 2015-2017," *IEEE Transactions on Multimedia*, 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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. 04, pp. 715–734, 2001.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] ——, "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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[33] 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.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] Y. Yan, S. Pouyanfar, Y. Tao, H. Tian, M. P. Reyes, M.-L. Shyu, S.-C. Chen, W. Chen, T. Chen, and J. Chen, "Florida International University-University of Miami TRECVID 2017," *TRECVID*, 2017.

[44] G. Awad, A. Butt, K. Curtis, J. Fiscus, A. Godil, 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 *Proceedings of TRECVID 2018*. NIST, USA, 2018.

[45] 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.