Florida International University – University of Miami: TRECVID 2018

Ad-hoc Video Search (AVS) Task

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]Florida International University, Miami, FL, USA
[2]University of Mimai, Coral Gables, FL, USA
[3]Purdue University, West Lafayette, IN, USA
[4]Miami Palmetto Senior High School, Miami, FL, USA

Agenda

UNIVERSITY OF MIAMI COLLEGE of ENGINEERING

FIU Computing & Information Sciences

1. **Submission Details**
2. **Introduction**
3. **Proposed Framework**
   - Concept Bank
   - Incorporating Object Detection
   - Just-In-Time Concept Learning
   - Score Combination
4. **Experimental Results**
   - Evaluation
   - Performance
5. **Conclusion**

Submission Details

UNIVERSITY OF MIAMI
COLLEGE of ENGINEERING
FIU
Computing & Information Sciences

- **Class**: M (Manually-assisted runs)
- **Training Type**: D (Used any other training data with any annotation)
- **Team ID**: FIU-UM (Florida International University – University of Miami)
- **Year**: 2018

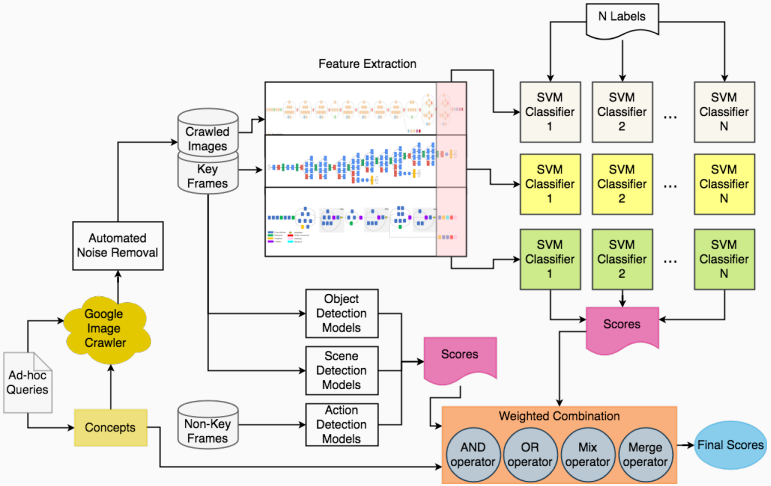
Introduction TRECVID 2018 AVS Task

UNIVERSITY OF MIAMI COLLEGE of ENGINEERING

FIU Computing & Information Sciences

- **Test Collection**: IACC.3 dataset with 4593 Internet Archive videos (144GB, 600 total hours)
- **Video Duration**: Between 6.5 and 9.5 minutes
- **Queries**: 30 new queries
  - Object (with specific description): 5 queries (570-572, 577, 585)
  - Scene: 1 query (580)
  - Object + Action: 12 queries (562, 568, 573-576, 581-584, 587, 588)
  - Object + Scene: 6 queries (561, 563, 578, 579, 589, 590)
  - Object + Action + Scene: 6 queries (564-567, 569, 586)
- **Results**: A maximum of 1000 possible shots from the test collection for each query

Proposed Framework



The designed framework for the TRECVID 2018 AVS task

Concept Bank

The concept bank contains 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

Image Classification Model

- To train image classification model on TRECVID dataset, 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;
- ImageNet contains 1.2 million images belonging to 1000 classes;
- PLACES365 introduces 365 scene categories, which is very useful in the detection of location and environment;
- HYBRID1365 incorporates both PLACES365 and ImageNet.

Places data for query 579 “Find shots of one or more people in a balcony”

ImageNet data for query 566 “Find shots of a dog playing outdoors”

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.

Action Detection Model

- The “Moments in Time” dataset includes approximately one million 3-second videos over 339 classes;
- The weights for training the “Moments in Time” model are taken from a 50 layer ResNet network initialized on the ImageNet dataset.

Query 563 “Find shots of one or more people on a moving boat in the water”

Query 568 “Find shots of one or more people hiking”

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.

Incorporating Object Detection

- Count the number of objects;
- Detect small objects;
- Query 572 “Find shots of two or more cats both visible simultaneously.”

Confidence Score of the Object Count

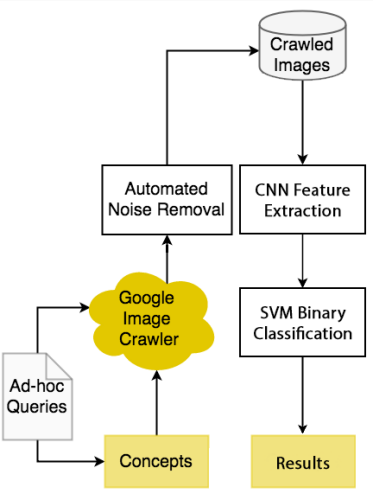
- $P_{O,N}(I)$: the confidence score object O appearing N times in the image I;
- n: the number of object O in the image I detected by the model;
- $P^i_O(I)$: the i-th highest confidence score among all the detected objects O in image I;

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

K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

Just-In-Time Concept Learning

- Automatically crawls the related images in an image search engine for the missing concepts;
- For each new concept, around 10,000 images are crawled;
- Filters the outliers in the search engine results with auto-encoder;
- Inception-V3 model is used to extract features;
- Trains the classifier to detect the concepts for the corresponding query.
- Query 587 “Find shots of a person looking out or through a window”.



Score Combination

- Four types of score combination operations: “AND”, “OR”, “Mix”, and “Merge”
- S_i: The score of the *i*-th concept;
- w_i: The weights of the *i*-th concept, determined by the concept rarity;
- $\mathcal{N}$: Number of the concepts;
- Query 578 “Find shots of a person in front of or inside a garage”
  Handle the heterogeneity of the garage from inside and outside views.

“AND” Operation

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

“OR” Operation

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

Score Combination (Cont.)

- S'_j: The score of "OR" operation of the j-th group of concepts;
- w'_j: The weights of the j-th group of concepts, determined by the concept rarity;
- $\mathcal{M}, \mathcal{N}_0$: Number of the groups, and remaining concepts;
  Query 578 "Find shots of a person in front of or inside a garage":
  $\mathcal{M} = 1$: The concept group "garage", combining "garage indoor" and "garage outdoor";
  $\mathcal{N}_0 = 1$: the concept "person";
- $S_{\text{comb}_k}, w_{\text{comb}_k}$: Scores from different combination of concepts and their weights.

"Mix" Operation

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

"Merge" Operation

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

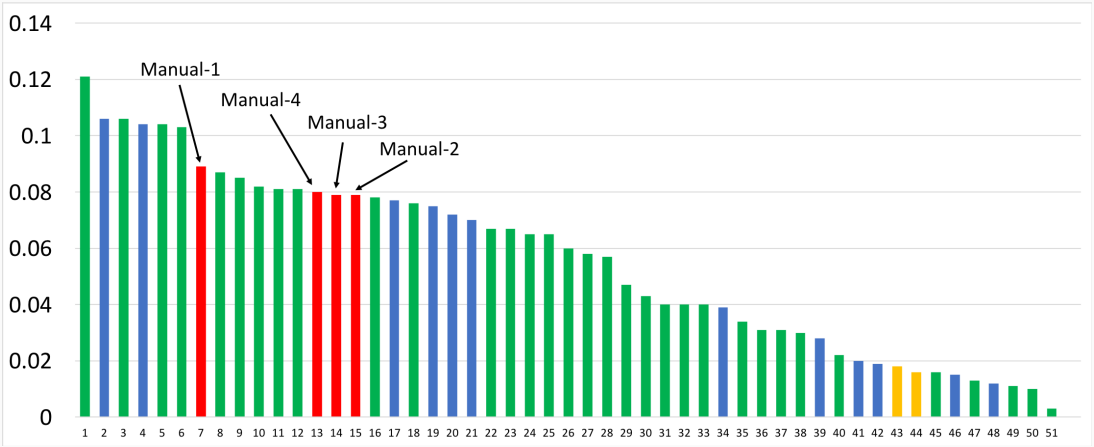
Evaluation

- **Metrics**: Mean extended inferred average precision (mean xinfAP);
- **Sampling**: All the top-150 results and 2.5% of the remaining results;
- As in the past years, the detailed measures are generated by the *sample_eval* software provided by NIST.

Submission Details

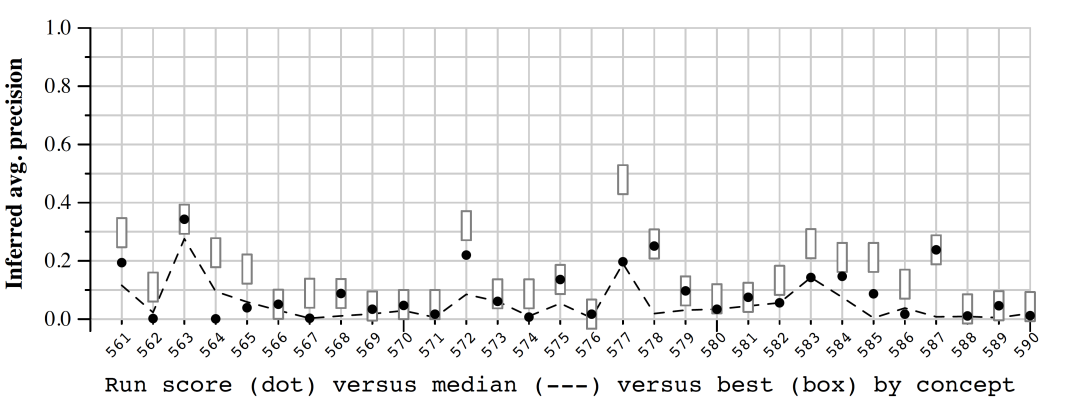
1. **Common Setting**: CNN features + linear SVM for the TRECVID dataset, scores from other sources in the concept bank;
2. **Manual-1**: use the best set of concepts and the weighted combinations (“and”, “or”, & “mix” operations);
3. **Manual-2**: use the best set of concepts and the weighted combinations (“and”, “or”, & “mix” operations) + rectified linear score normalization;
4. **Manual-3**: use the second best set of concepts and the weighted combinations (“and”, “or”, & “mix” operations)
5. **Manual-4**: fuse different score sets (“merge” operation)

Performance



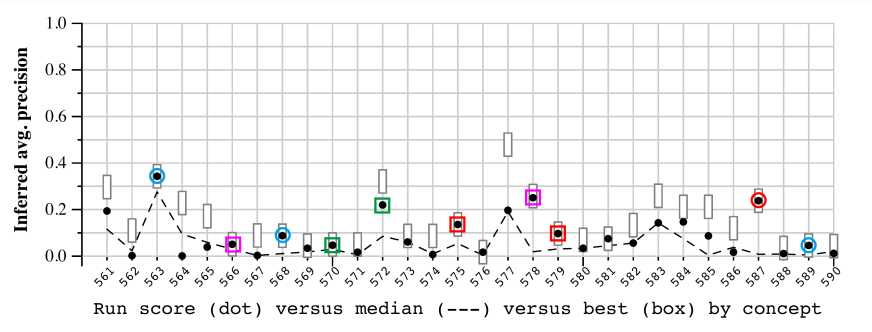
Comparison of FIU UM runs (red) with other runs for all the submitted fully automated (green), manually-assisted (blue), and relevance-feedback (orange) results.

Performance



Detailed scores of run Manual-1

Performance



Performs the best in queries 563, 568, 587, and 589 (circle) and achieves a good performance in queries 566, 570, 572, 575, 578, and 579 (square).
The good performance is benefited by Moments339 (blue), JIT concept learning (red), Object detection model (green), and the new score combination (purple).

Conclusion

- In addition to the classic datasets such as ImageNet, Places, and UCF101, we leverage recently released datasets, such as Moment339 for action recognition, and achieve improvements in several queries;
- “Mask R-CNN” and “YOLO” are applied to improve the object recognition performance and also to estimate the number of objects for some queries;
- We plan to utilize more temporal information from video datasets and a better fusion model;
- We plan to automate our video retrieval system.

Thanks!

Any questions?

Acknowledge to GAFAC at University of Miami to financial support Yudong Tao to attend the TRECVID 2018 workshop