UTS-CMU-D2DCRC Submission at TRECVID 2016 Video Localization

Linchao Zhu[§] Xuanyi Dong[§] Yi Yang[§] Alexander G. Hauptmann[†]
[§]FEIT, University of Technology Sydney
[†]SCS, Carnegie Mellon University

Abstract

In this report, we summarize our solution to TRECVID 2016 Video Localization task. We mainly use Faster R-CNN to localize objects in the spatial domain which is combined with frame-level and shot-level detectors to localize concepts in the temporal domain. We collected images with annotated bounding box from external sources, e.g.*, ImageNet Detection dataset and manually annotate bounding boxes for categories without any annotations. We trained frame-level detectors using ResNet-200 features pre-trained on ImageNet and for classes of "Running", "Sitting_Down" and "Dancing", we also use improved Dense Trajectories features. Finally, we fuse bounding box score, frame score and shot score to get the final score for each bounding box.*

1. Data collection

In the TRECVID Video 2016 Localization task [2], there are ten classes to be localized, which are "Animal", "Bicycling", "Boy", "Dancing", "Explosion_fire", "Instrumental_Musician", "Running", "Sitting_Down", "Skier" and "Baby". Different from previous year's settings, bounding box annotations are not provided this year. We thus need to collect bounding box annotations to train models in a supervised way. We use data from different sources, for example, images from ImageNet [8], videos from MPII Human Pose Dataset [1], HMDB-51 [4], Hollywood2 [5] for training. We manually select relevant categories in data sources and choose subsets from the original data. We construct the validation set using the provided development data. The detailed dataset construction can be found in Table 1. We annotate bounding boxes using the online tool[1].

2. Method

2.1. Bounding box score

We use Faster R-CNN [7] to detect objects in images. To train a Faster R-CNN model, we select about 500 images per category to train a model with ten classes. Note that "Animal" and "Instrumental_Musician" have subcategories and have more image annotations than other categories, we collected another set with about 4,500 images and trained a separate model for them.

We tried six different network structures *i.e.*, VGG-16, VGG-19 [9], GoogLeNet [10], ResNet-50, ResNet-101, ResNet-152 [3]. To combine different models, we use the region proposals generated by the ResNet-152 model and fuse scores of six models to obtain the final score for each region. We report the average precision of VGG-16 and ResNet-50 models in our validation set on Figure 1.

2.2. Frame-level and shot-level score

We train the frame-level and shot-level detectors to take the frame and video context into account. To localize objects or actions in the temporal domain, we annotated videos by checking whether the concept exists in the frame. We use the provided development videos as training and validation data.

To train a frame-level detector, We use the pre-trained ResNet-200 model[2] and extract the features from the layer before the final classification. We crop and resize the image to 320×320 and obtain the feature with dimension 2,048. We perform L_2 normalization and trained a linear SVM classifier. Empirically, we set $C = 1$.

For classes of "Running", "Sitting_Down" and "Dancing", we trained a shot detector using improved Dense Trajectories [11] features encoded by Fisher Vector [6].

For each region, we fused three scores to obtain the final score. The weight of scores and the thresholds are tuned on the validation set. We show the frame-level performance on Figure 2.

2.3. Submitted runs

We submitted four runs on the Localization task. Based on the testing results, we found that our runs are bad at localizing categories of "Boy" and "Sitting_Down". It may because that the distribution of our development data is different from the testing data.

[1] https://github.com/tzutalin/labelImg

[2] https://github.com/facebook/fb.resnet.torch

Concept Name	Data Source	Category ID	Number of examples (train,val)
Animal	ImageNet	n01443537,n01503061,n01639765,n01662784, n01674464,n01726692,n01770393,n01784675, n01882714,n01910747,n01944390,n01990800, n02062744,n02076196,n02084071,n02118333, n02129165,n02129604,n02131653,n02165456, n02206856,n02219486,n02268443,n02274259, n02317335,n02324045,n02342885,n02346627, n02355227,n02374451,n02391049,n02395003, n02398521,n02402425,n02411705,n02419796, n02437136,n02444819,n02445715,n02454379, n02484322,n02503517,n02509815,n02510455	880,264 (Two categories setting: 4400,484)
Instrumental_Musician	ImageNet	n02672831,n02787622,n02803934,n02804123, n03249569,n03372029,n03467517,n03800933, n03838899,n03928116,n04141076,n04536866	600,250 (Two categories setting: 4500,347)
Bicycling	ImageNet	n02834778,n02835271,n03792782,n04126066	744,101
Baby	ImageNet	n10353016	420,130
Boy	ImageNet	n09871229,n09871867,n10078719	350,101
Explosion_Fire	ImageNet	n03343560,n03346135,n10091450,n10091564, n10091861,n14891255	491,162
Skier	ImageNet	n04228054	703,101
Running	Hollywood2 HMDB-51 MPII Human Pose	Run Run running	351,101
Dancing	MPII Human Pose	dancing	596,101
Sitting_Down	Hollywood2 HMDB-51 MPII Human Pose	Sit Sit sitting	316,51

Table 1. We list image and video sources used in training. We also showed the number of training and validation examples. For "Animal" and "Instrumental_Musician", we additionally use a larger set to train a separate model.

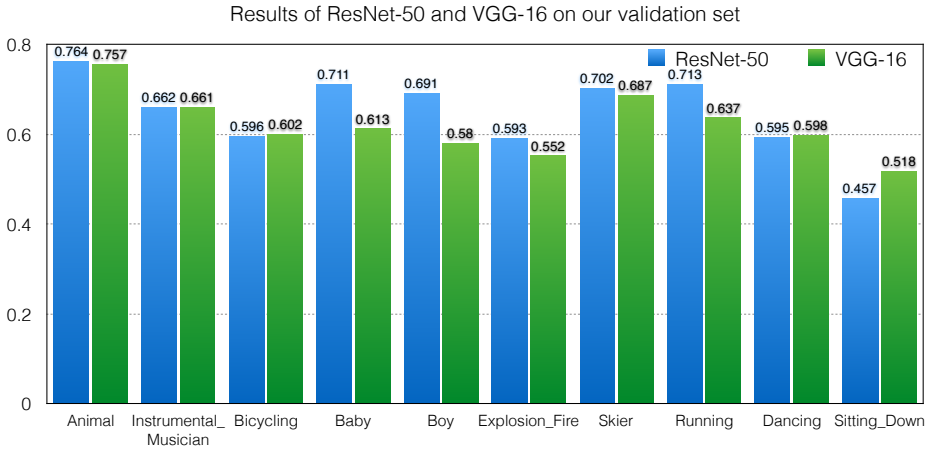

Figure 1. Results of spatial localization on our validation. We show the results of VGG-16 and ResNet-50.

In run *final_threshold_0_resnet50_10_cats_no_shot*, we use our spatial-only model where only bounding box score is used and we only use the ResNet-50 model. It achieves mean F1-score of 0.2780 on frame-level and mean F1-score of 0.1243 on pixel-level.

In run *final_threshold_0_merge_no_shot*, we use spatial-only model but fuse six Faster R-CNN models. It achieves mean F1-score of 0.2614 on frame-level and mean F1-score of 0.1157 on pixel-level which is comparable to the ResNet-50 only model. We tune the model to achieve high recall. Note that in this model, we train two separate models, one for "Animal" and "Instrumental_Musician" as these two categories have more training examples, and one for the rest categories.

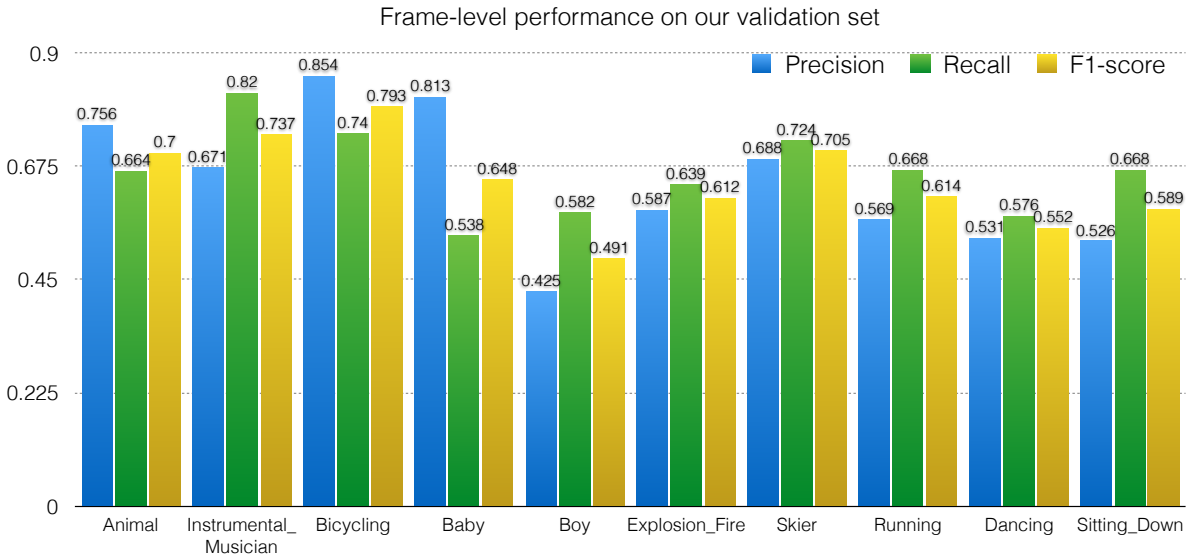

Figure 2. Frame-level performance on the validation set.

Run *final_threshold_0_resnet152* is the model which fuses bounding box score, frame-level score and shot-level score. It uses ResNet-152 model to spatially localize objects. It achieves mean F1-score of 0.4499 on frame-level and 0.2581 on pixel-level. Note that the fused model almost doubles the performance compared with the single model. Run *final_threshold_0_resnet50_10_cats* is similar to *final_threshold_0_resnet152*, but the network used is ResNet-50.

Acknowledgements

We thank TRECVID coordinators for providing detailed answers for our queries. This work is partially supported by the Data to Decisions Cooperative Research Centre www.d2dcrc.com.au. This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-1638429. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quenot, M. Eskevich, R. Aly, and R. Ordelman. TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

[5] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[6] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[7] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[8] 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. *IJCV*, 115(3):211–252, 2015.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[11] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.