Video understanding using part based object detection models

Vignesh Ramanathan
Stanford University
Stanford, CA-94305
vigneshr@stanford.edu

1. Project proposal for CS 231A

Video understanding aims to identify spatial and temporal patterns in a video to recognize the events captured by it. Given a set of pre-defined events, the current project focuses on detecting the occurrence of an event in a video-clip. This is akin to the fundamental challenge of object recognition in images. The difficulty of the event detection task arises from the huge interclass variation in camera view points, appearance of objects/ persons involved in the event, resolution, illumination, video quality etc.

A part based model for object detection was proposed in [3] and shown to achieve state-of-the-art results on the PASCAL VOC benchmarks [2]. [3] represents object classes as multiscale models with deformable parts. I will use this part based model obtained from [4] to detect event related objects from training videos and iteratively train the model with the segmented objects. This would improve the performance of the model for videos belonging to the event set. The object detection algorithm will be further augmented by a temporal filter based on a tracking algorithm like Optical Flow [5] to remove spurious detections. The final improved detector can be used to extract object (pertaining to a certain event) sequences from a test video. If an object sequence is extracted with a high confidence from a video, it can be tagged with the corresponding event name.

The initial part based model for detecting event related objects in videos will be trained with images obtained from ImageNet [1]. TRECVID [6] event kits will be used for training and testing the proposed algorithm. TRECVID provides videos belonging to 15 event categories. Each event kit contains the definition and evidential description of the event. For a specific event, the event related objects are decided based on this evidential description. Some of the relevant literature providing background for the project is shown in References.

The performance of the final event detection algorithm will be evaluated by precision-recall (PR) curves. The PR curves will be plotted separately for each event class and compared with those generated by the original object detection based method without iterative training, to show the gain from proposed scheme. The Mean Average Precision (MAP) value will also be evaluated for each method.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 2010.

[4] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[5] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185 – 203, 1981.

[6] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.