Video understanding using part based object detection models

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

Abstract

We will explore the use of event specific object detectors (part based models) in multimedia event detection. The challenge is to choose a well-trained object detector specific to the event videos. The presence of dataset bias would make an object detection model trained on an unrelated image database less useful for the video dataset in hand. Given such a model,we propose an iterative method to build a more effective detector, trained only on frames from the training video dataset. The given model, in combination with an optical flow based filtering method is used to extract objects with high confidence from training videos, which are then used to train a new object detection model. This model is again used to extract objects, used for training in the next iteration. The process is repeated to finally obtain a model trained only on frames from training videos. The performance of the final model is evaluated on manually annotated frames form test videos. It is compared with the original object detector to show the gain of the proposed method.

1. Introduction

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, multimedia event detection identifies 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.

In this project, we will focus on the task of event detection using event specific object detectors. In general, such detectors are a part of a larger framework, where the motion of the object is also identified in successive video frames and compared with corresponding motion in training videos [6, 7]. However, in this project we will restrict the analysis to tagging videos based only on detection of event specific objects. In particular, we propose a method to build an object detector which would perform well for a given video dataset. An object detector trained on a generic image database like Imagenet [1] would not be effective on the video dataset, due to the presence of inherent dataset bias. For instance, in the case of detecting "skateboards" in skateboarding videos, it can be seen that the videos mostly contain frames showing people moving on skateboards. On the other hand, Imagenet skateboard images show skateboards from different views often occluded by other objects. The effect of such dataset bias has been explored in [9]. The paper has analyzed the performance of detectors trained on one dataset and tested on others. The performance was seen to degrade even for the two class classification problem. Hence, in order to achieve best results, we would like to train the object detector only on video frames from the training video dataset. However, it is impractical to manually annotate video frames, every time we are given a video dataset. Instead, we will use the object detector trained on a readily available annotated image dataset like [1] to build an object detector specific to the given video dataset. This object detector will be used to extract relevant object sequences from event videos and tag them according to the presence or absence of such object sequences.

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. we 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. While detecting objects from videos for training, an optical filter[5] based filtering is applied in the temporal domain to ensure that only objects which are detected consistently in successive frames and with motion along the path predicted by optical flow are retained. This minimizes the chance of spurious detections. The final improved detector can be used to extract object (pertaining to a certain event) sequences from a test video.

The initial part based model for detecting event related objects in videos is trained with images obtained from ImageNet [1]. TRECVID [8] event kits is used for training and testing the proposed algorithm. 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.

The performance of the final iteratively trained object detection model is evaluated on a set of manually annotated frames from test videos. In this report, preliminary results are shown for a single event class namely "attempting a board trick". The model obtained after three iterations trained exclusively on objects extracted from training video frames is compared with the initial model trained on images from Imagenet as well as a model trained both using images from Imagenet and training video frames.

2. Approach

Given an object detector trained on images from a generic database, we wish to gradually remove the dataset bias from the model and move towards a model more specific to the video dataset in hand. We first initialize the training procedure with a model trained on an image database. Secondly, we use this model in combination with an optical flow based filtering method to detect objects with high confidence and annotate corresponding frames from training videos. Thirdly, the newly annotated video frames along with the original database images are used to train a new object detection model. This procedure is repeated iteratively to train object detection models. The different steps are explained below.

2.1. Initialization

The event specific object is decided based on the evidential description of the event provided in the event-kit. In this project, we consider only one event related object for each event class. For instance, "skateboards" are chosen as the object relevant to the event class "attempting a board trick". The corresponding annotated images from the Imagenet database are used to train an initial part based model as described in [3].

2.2. Bounding box detection in video

This part based model is used to detect object bounding boxes from all training video frames. The top four bounding boxes B_i^j where, $j \in \{1, 2, 3, 4\}$ with highest detection scores in each frame is retained along with their corresponding scores S_i^j. These four bounding boxes are used to assign score values to pixels in the image to form a scoremap. Each pixel in an frame from a video is assigned the score of the best bounding box it falls into. If a pixel does not fall into any bounding box, it is assigned a very high negative score S_{min}. Let $S_i(x)$ denote the score assigned to the pixel at position x in the i^{th} frame of a video. The score values are now filtered using a optical flow based filter to ensure that only objects which are consistently detected in a sequence of frames are retained.

2.3. Optical flow based temporal filtering

Optical flow can be used to track a dense set of points across a sequence of frames. For every pixel in a frame, the corresponding position in another frame can be obtained. Let $u_{i,k}(x)$ denote the displacement of the pixel at position x from the i^{th} frame to k^{th} frame. We use this information to filter the scores S_i. This filtering is carried out across a window of frames. Let $(2N + 1)$ denote the window size and $R_i(x)$, the modified score at pixel x of i^{th} frame. Then,

$$R_i(x) = \sum_{k=-N}^{N} w_k S_i(x + u_{i,i-k}(x)) \tag{1}$$

The filter coefficients w_k are chosen according to a Gaussian kernel. This smoothens out any irregularities in object detection across successive frames. After obtaining R_i, the filtered scores R_i^j of bounding boxes B_i^j in images are computed as the average of all pixels belonging to the bounding box.

$$R_i^j = \frac{\sum_{x \in B_i^j} R_i(x)}{|B_i^j|} \tag{2}$$

Here, $|.|$ represents the size of the bounding box.

Having obtained the filtered score values, we still need to eliminate a large number of spurious detections and retain only "good" detections. Again, we impose the criteria that good detections will be consistent across a sequence of frames. Hence, we retain the detection only if the same region is detected in neighboring frames as well. We impose a set of hard conditions to achieve this. Let, B_i^{max} be the box with the highest score R_i^{max} in the i^{th} frame. We reject all other bounding boxes in the frame. Let B_i' denote the bounding box obtained by displacing B_i^{max} from i^{th} frame to $(i + 1)^{th}$ using optical flow. We reject the detection B_i^{max} if the overlap between the dispalced box B_i' and B_{i+1}^{max} is less than a threshold δ. Then, we move a window of size M in the temporal domain and retain only those detection sequences which have a length greater than M. A detection sequence in this context refers to a set of consecutive frames where an object has been detected (bounding box retained according to the previous conditions). Finally, the average velocity of the object in the sequence is computed using optical flow measurements. The detection sequence is rejected if this velocity is less than a threshold τ. This condition helps eliminate noisy detections particularly from background clutter. Moreover, objects which are static

through a sequence of frames will add less value to training. The conditions are enumerated below.

1. Only the bounding box B_i^{max} with naximum score in each frame is retained
2. B_i^{max} is rejected, if the overlap between the dispalced box B_i' and B_{i+1}^{max} is less than δ
3. Only detection sequences(consecutive frames, where a bounding box has been retained) with length greater than M are retained
4. A detection sequence is rejected if the average velocity of the object in the sequence is less than τ

These stringent conditions enforce the criteria, that only good detections are retained. This method is used to extract object sequences from all traiining videos belonging to the event class.

2.4. Iterative training

Let $P_{Imagenet}$ represent the object detection model obtained by training only with the Imagenet object images. The object detection results from Sec. 2.3 are used to obtain a set of image annotated with the bounding box information. These new images are now added to the pool of Imagenet images to train a new object detection model $P_{Imagenet+video}$. Alongside, another object detection model P_{video} is obtained by training only with the objects detected from Sec. 2.3. Both the models are cross validated on image frames from test videoset using $5 - fold$ validation. The model with the better average precision score is used in the next iteration. The steps discussed in Sec. 2.2 and 2.3 are repeated with this new object detection model. Finally an object detection model trianed only on frames from training video sequence is obtained which outperforms the original model $P_{Imagenet}$. It is to be noted that in our experiments the negative training examples remain consistent throughout all iteations. However, this need not be the case. An equal number of negative training samples can also be extracted in a similar fashion form training videos and used for training.

3. Preliminary experiments and results

In this section, we present the preliminary results for the event-class"attempting a board trick" from the TRECVID video dataset. The training was iteratively carried out, by extracting objects from a training set of 40 videos. The resultant models $P_{Imagenet+video}$ and P_{video} were evaluated on a set of 177 manually annotated images obtained from a test video set of 40 videos. The Average Precision (AP) values are shown for 3 iterations in Tab. 1. After the third iteration, the P_{video} outperforms the remaining models. In other words, we have gradually moved from an object detector trained on a generic image database to a detector specific the vide dataset of interest. The number of training objects detected at the end of each iteration also increases with the number of iterations (from 180 after initialization to 640 after 2^{nd} iteration). It was seen that, roughly 85% of the detected objects pertained to a skateboard or atleast a large part of the skateboard, while the remaining were spurious detections. The performance of $P_{Imagenet+video}$ and P_{video} is also seen to be vastly better than $P_{Imagenet}$.

Iteration	AP of $P_{Imagenet}$	AP of $P_{Imagenet+video}$	AP of P_{video}
0	0.124	-	-
1	0.124	0.301	0.251
2	0.124	0.348	0.328
3	0.124	0.368	0.372

Table 1. Comparision of performance (average precision AP) of different models at the end of each iteration.

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] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV 2007*, pages 1–8, 2007.

[7] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.

[8] 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.

[9] A. Torralba and A. Efros. Unbiased look at dataset bias. In *CVPR11*, 2011.