Automated Feature Labelling: Head Detection and Localization Proposal

Anirudh Satish
Harvey Mudd College
asatish@hmc.edu

Abstract

This project seeks to perform head detection and localization of all heads in the frames of short video clips. If permissible, and depending on the complexity of the former task, I intend to extend this project to perform semantic segmentation on the localized head, and also classify the identified head with different meta-tags, such as visibility, sentiment, and the relationship between different heads. This project intends to use a three pronged Deep Learning approach that was employed by another group of researchers, and try to get similar results/optimize performance for this specific project (8).

1 Motivation

Object Detection has been a field of great a interest to the Computer Vision community for decades, with the earliest Neural Network approaches beginning all the way back to the 90s (5). Despite this early start with deep learning approaches, advances in this field were relatively slow until CNN and R-CNN approaches were developed along with support vector machine (SVM) architecture (2). While Face detection has come along way, with may of us unlocking our devices just by looking at a blank screen, the same cannot be said about the head detection subspace of object detection. The prevalence of state-of-the-art face detection applications can be seen all throughout the technology space, and even the very simplest of models can perform to staggering levels (4). For head detection and localization, state-of-the-art object detectors achieve on an upper bound only around a 65% Average Precision for people by the Pascal VOC (1) benchmark (8) (5).

It is this challenge in this field of object detection that is one of the driving forces of motivation behind this project, the goal of which, is to localize every head visible in short movie clips. There are a plethora of reasons for these challenges, some being that the human pose can take a lot of different forms, and various other features in the frame/image cause clutter and block key information about the head (8). Eliminating these obstacles, understanding how these features play a role in localization, and getting past them are huge motivators for me, and I hope to be able to tackle such issues, and answer these questions through this project. While it may seem that these obstacles are huge hurdles, there has been a lot of research in this specific field of object detection, using a variety of different deep learning methods that will help me in making this project successful.

Depending on the difficulty in achieving this goal, there is scope to extend the project to perform semantic segmentation of the localized head, and further classify the head with meta-tags such as whether the face is visible or not, and further trying to understand the relationship between multiple heads in any given frame or sequence of frames. While improving on existing state-of-the-art models in the project is highly unlikely given our level or experience and expertise in this field, I strongly believe that the success of this project would go a long way in understanding the semantics of movie/video clips, and could be the foundation of data analysis on theses clips, or a building block to a lot of other projects.

2 Related Works

I believe that this project is quite significant and applicable, especially since we are working with video clips from movies. A short test performed by Ivan Laptev, in his paper ”Modeling and visual recognition of human actions and interactions” suggests that about 35% of screen pixels in movies, TV programs and YouTube videos belong to people (3). Given the significant proportion of pixels that humans occupy in these films, and the fact that context can play a huge role in object detection (7), there is something to explore here.

3 Method of Approach

The data set that I have for this project is 15 short movie clips, of around 1-4 minutes long. This data set was annotated by rigorously recording the gaze behaviour for 21 participants from ages 22-73. At this point, it seems like additional annotation is required on the data set, to mark the heads of those appearing on screen.

For this project, the method of approach I intend to take is to re-purpose the models and pipeline used by Tuan-Hung Vu *et. al* in their paper "Context-aware CNNs for person head detection" (8), (that albeit was trained and used on a much larger data set) for the data set of movie clips that was mentioned earlier.

The methodology used by Tuan-Hung Vu *et. al* was to use information from surrounding context in the movie clips, rather than just treat the objects identified by bounding boxes independently (8) (7). The method that the authors used, and the one that I will try to follow is a three fold Deep learning pipeline. The models that comprise this pipeline are *Local Model, Global Model* and a *Pairwise Model* (8). Below is a very high level short summary of these three main models, to gain some big picture understanding of how these all tie together and give us the desired behaviour.

The *Local* Model follows the R-CNN architecture from (2) to detect objects in the scene, using selective search proposals to narrow down the number of objects (6). Following this, each bounding box is expanded outward linearly to include some supplementary environment information, and then the square like bounding boxes are selected as potential head targets (8).

The *Global* Model is based on CNN architecture, and is used to predict coarse locations and scales of objects when given as input the full low-resolution frame (8). It outputs this information by outputting a score for each cell of a multi-scale heatmap (**?**). The reason for not simply using this model for localization is that this model does not provide accurate localization bounding boxes (8). Rather, the scores from this *Global* Model are used to re-score the candidates of the *Local* and *Pairwise* Models, to then give the final bounding boxes (8).

Finally, the *Pairwise* Model is used to reason about pairwise relations between objects, and this information combined with the feature vectors that the *Local* and *Global* models output is used to determine the bounding boxes for human heads in video frames.

While most of this information might seem hard to digest, it becomes easier to understand when following the authors' paper and code instructions. The plan for this project is to try and get the existing models running on our data set of video clips. This would involve to meet the expectations of the network, annotating the data further if needed, formatting our data correctly, and potentially trying to optimize the performance of the model depending on its performance. This could be altering the code, to try and extract better performance, or formatting our data in a different way to best mimic the behaviour that was achieved with the original model.

4 Evaluation Metrics

Here, we use the same evaluation metrics that were used by the authors of the model that is being repurposed here. This evaluation metric would be a Precision-Recall metric. The reason for this is two fold. The first being that this is quite a standard metric to use, and a lot of Object detection methods are typically evaluated in terms of precision-recall (8). Secondly, since we are using the same metric that the authors' of the model used, the evaluation scores that we would get on running the model on our data set could be directly compared to those scores from the authors themselves, and used a measure of success and performance of the project.

5 Limitations

While the code and models are already available for public use by the authors of the paper, I believe that there could and will be hurdles in just getting the model running on our data set. This includes but is not limited to figuring out the pipeline, and making sure that the format of the data matches what the network expects. Furthermore, the requirements for getting these models running, according to the authors are MatConvNet, CUDA, and a reasonable GPU. While MatConvNet is a readily available toolbox on MATLAB, and we should have access to powerful computing via XSEDE, it remains to be seen if this process will be as seamless as possible. Lastly, and probably most importantly, I feel like I need to really dive deep into the paper, reading it repeatedly to understand the under the hood behaviour of the model. While this may not be absolutely necessary for getting the model running, it is imperative to know for any attempts at optimizing/fixing problems if the yielded performance is below par.

References

[1] Eslami S.M.A. Van Gool Everingham, M., Williams Christopher, K, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *Computer Vis*, 111(98–136).

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation tech report (v5). pages 1–3.

[3] Ivan Laptev. 2014. Modeling and visual recognition of human actions and interactions. page 1.

[4] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. 2014. Face detection without bells and whistles. In *Computer Vision – ECCV 2014*, pages 720–735, Cham. Springer International Publishing.

[5] C. Monrocq R. Vaillant and Y. LeCun. 2016. International Conference on Artificial Neural Networks (ICANN).

[6] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. 2011. Segmentation as selective search for object recognition. In *IEEE International Conference on Computer Vision*.

[7] Antonio Torralba. 2003. Contextual priming for object detection. volume 53.

[8] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev. 2015. Context-aware cnns for person head detection.