NATURAL VIDEO UNDERSTANDING

Andrew Floren

University of Texas at Austin
Electrical and Computer Engineering
Austin, Texas

ABSTRACT

Abstract

Index Terms— One, two, three, four, five

1. INTRODUCTION

Want to understand the content of natural videos. Many applications such as ... Like natural language understanding, first need to parse the 'words' and 'sentences' that make up video. The scenes in a video correspond to these sentences while the objects in that scene make up the words.

We have focused first on parsing the sentences or scenes of a video. Image recognition has come a long way in the last few years and differentiating between thousands of natural objects in an image with high accuracy is now possible [refs]. By applying these techniques to video, we can begin detect scenes based on image content rather than intensity or motion based heuristics [refs].

We have developed an efficient method to detect, store, and search through scenes in a video based on image content. The algorithm is specifically designed to be used in conjunction with H.264 encoded video, however it is applicable to any video compression technique that utilizes I-frames. A deep convolutional network is used to classify image content of each I-frame in a video. The output of the neural network is a 1,000-dimensional classification-vector that represents how likely each of the 1,000 content classes is present in the image. Agglomerative clustering is used to create a hierarchical scene representation from the classification-vectors. A strong adjacency constraint is enforced during clustering both for algorithmic efficiency and to produce temporally consistent scenes. Finally, we introduce a method for constructing search-vectors based on textual labels and searching this hierarchical scene representation using these search-vectors.

The output of the image recognition algorithm also tells us what types of objects are likely in the scene. Eventually, we want to detect and track these objects within and across scenes to gain a broader understanding of the video such as ... [refs]. However, the state of the art object detection algorithms are not yet as robust as those in image recognition.

2. RELATED WORK

Lookup

3. IMAGE RECOGNITION

neural networks
    backpropagation training
    convolutional layers
    googlenet
    inception layer
    auxiliary outputs

4. AGGLOMERATIVE CLUSTERING

clustering
    agglomerative clustering
    ward clustering
    neighborhood constraints

5. VIDEO RECOGNITION AND SCENE DETECTION

Trained network on image-net to classify image contents

Classify I-frames of new H.264 encoded video using trained network. I-frames allow for parallel access patterns and tend to be both regularly spaced throughout the video and aligned along scene changes (because they create large residual errors due to bad motion vector reconstruction).

Implemented in parallel to speed up processing of video files. The trained classifier is slow to classify so we used a Hadoop cluster with Apache Spark to efficiently classify them in parallel.

Perform agglomerative clustering on I-frame classification vectors using frame adjacency constraints for efficiency to find scenes in video. Scenes based on similar content rather than similar images (e.g., jump cut between two horses at different times of day could still be considered a single scene).

To search a video, construct a search vector and compare (inner product) it with all cluster centroids. If similarity score exceeds a threshold report a successful match for that scene. Search vector constructed based on text search of synsets.

6. EVALUATION

Compare speed of evaluation versus non-parallel and naive approaches.

Don't have ground truth so hard to evaluate, just need to come up with something.

7. CONCLUSION

It works

Where it fails

How to fix

Future work and applications

Search vectors could be constructed from template images. Allows to search for similiar images in a video rather than text.

Instead of searching within a video, could be used to search across databases of videos based on content, e.g., for YouTube or Vine. Probably just want to use a single or few clusters per video in this case.

Could be used in conjunction with quality assessment algorithms to better account for content-based opinion score biases.

8. REFERENCES