What Actions are Needed for Understanding Human Actions in Videos?

Liangjie Cao

Aug. 3, 2018

Abstract

What is the right way to reason about human activities? What directions forward are most promising? In this work, the authors analyze the current state of human activity understanding in videos. The goal of this paper is to examine datasets, evaluation metrics, algorithms, and potential future directions. They look at the qualitative attributes that define activities such as pose variability, brevity, and density. The experiments consider multiple state-of-the-art algorithms and multiple datasets. The results demonstrate that while there is inherent ambiguity in the temporal extent of activities, current datasets still permit effective benchmarking. They discover that fine-grained understanding of objects and pose when combined with temporal reasoning is likely to yield substantial improvements in algorithmic accuracy. They present the many kinds of information that will be needed to achieve substantial gains in activity understanding: objects, verbs, intent, and sequential reasoning. The software and additional information will be made available to provide other researchers detailed diagnostics to understand their own algorithms.

1. Introduction

Over the last few years, there has been significant advances in the field of static image understanding. There is absolutely no doubt that they are now closer to solving tasks such as image classification, object detection, and even semantic segmentation. On the other hand, when it comes to video understanding we are still struggling to figure out basic questions such as: What is an activity and how should they represent it? Do activities have well-defined spatial and temporal extent? What role do goals and intentions play in defining and understanding activities?

A significant problem in the past has been the absence of good datasets for activity detection and recognition. Most of the major advances in the field of object recognition have come with the creation of generic datasets such as PASCAL [2], ImageNet [1] and COCO [3]. These datasets helped define the problem scope and the evaluation metrics, as well as revealed the shortcomings of existing approaches.

But before we move forward and define the benchmarks, they believe it is worth pausing and thoroughly analyzing this novel domain. What does the data show about the right categories for recognition in case of activities? Do existing approaches scale with increasing complexity of activities categories, video data, or temporal relationships between activities? Are the hypothesized new avenues of studying context, objects, or intentions worthwhile: Do these really help in understanding videos?

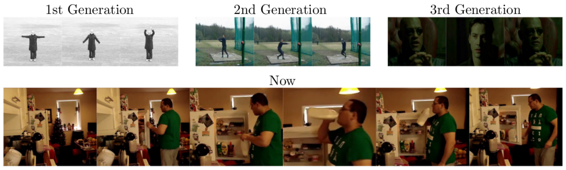

Figure 1. Now that the field of activity recognition has moved on from simple motions

2. What are the right questions to ask?

To start our discussion about activities, let us establish what they want to learn. When they talk about activities, they are referring to anything a person is doing, regardless of whether the person is intentionally and actively altering the environment, or simply sitting still. In this section, they will first look at how to define activity categories, and then investigate the temporal extents of activities.

3. What are the right activity categories?

Should we focus our analysis on general categories such as "drinking", or more specific, such as "drinking from cup in the living room"? Verbs such as "drinking" and "running" are unique on their own, but verbs such as "take" and "put" are ambiguous unless nouns and even prepositions are included: "take medication", "take shoes", "take off shoes". That is, nouns and verbs form atomic units of actions.

4. What are existing approaches learning?

The category plots are generated from (x, y) pairs where x is a an attribute for a category and y is the classification performance for the category. Finally, the pairs are clustered by the x coordinate and the average of the y coordinates visualized. Error bars for category plots represent one standard deviation around Two-Stream based on the y values in each cluster. In this section they report Pearson's p correlation between x and y. The video plots are generated similarly by clustering the videos based on the attributes. Finally the mAP is calculated in that group of videos. Error bars for video plots represent the 95% confidence interval around Two-Stream obtained via bootstapping.

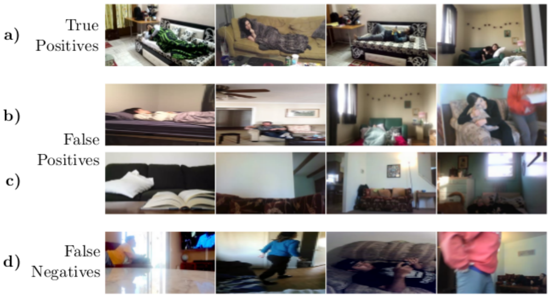

Figure 2. Example results from a Two-stream network

To understand what current methods are learning and motivate the rest of this section, they start by highlighting some of the errors made by current methods. First, they look at visual examples of the errors that a Two-Stream Network makes on Charades. In Fig 2 they see correct classifications, as well as three types of errors. The figure suggests that 1) models need to learn how to reason about similar categories; 2) methods have to develop temporal understanding that can suppress temporally simi- lar but semantically different information; and, 3) models need to learn about humans and not assume if couch is detected, then "Lying on a couch" is present.

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 *CVPR*, 2009. 1

[2] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 1

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1