Visual Attributes in Videos

Marielle Morris mariellem@rice.edu
Mahdi M. Kalayeh mahdi@eecs.ucf.edu
Mubarak Shah shah@crcv.ucf.edu

Center for Research in Computer Vision
University of Central Florida

Abstract

Attributes are descriptive characteristics with cross-category applicability. They can be used to provide additional information that is not found in simple object classification, giving the computer a complete picture of the item and the state in which it appears. While attribute prediction has been studied extensively in still images, this is not the case in videos. However, there exist visual clues in a video that are not present in a still frame. In this paper, we discover and annotate visual attributes for the Youtube-BoundingBoxes dataset. This is the largest video attribute dataset to date. Using the Youtube-Attributes dataset, a model can be trained for scalable multi-label classification. We have collected a dataset of 20,000 videos paired with over 85 attribute annotations describing 10 different classes of objects. We demonstrate that our dataset can be used to produce attribute classifiers that are both accurate and highly scalable.

1. Introduction

Efforts in computer vision have traditionally focused on categorical object recognition - classification, segmentation, localization, and so on. But a simple label does not tell the full story. When humans characterize objects, they create distinctions and consider visual clues that cannot be encapsulated in a single label. A standard categorical model will not recognize intra-category variations, such as the distinction between a "startled cat" and a "sleeping cat." This model limits the information that helps researchers understand the state of objects relative to others. Furthermore, the recent progress in image and video caption generation asks for robust attribute prediction models to enhance the quality of captions.

By definition, attributes are inherent characteristics that describe objects and the state in which they appear. They provide visual clues that are instantly detectable to a human observer. Additionally, they are typically general enough to apply to a very large number of objects: "startled" applies not only to cats, but to humans and many other animals.

While attribute prediction has been studied extensively in images, this is not the case in videos. However, there exists descriptive information in a video reel that is not detectable in a still frame. There are temporal attributes, such as descriptive actions, that will only be present in a video; additionally, a clip will provide more details about a particular object than will exist in one picture. Although video attributes are a pertinent area of interest in vision, there is no modern dataset that can be used for training deep learning networks.

With current advances in computational architectures, the research community demands large, diverse video datasets. In this paper, we provide a human-annotated video attribute dataset consisting of 20,000 short video clips from YouTube. These clips describe 10 distinct classes of objects: bear, bird, cat, cow, dog, elephant, giraffe, horse, person, and zebra. It features video-wise labels from a set of over 85 characterizing attributes from 3 categories, action attributes, descriptive attributes, and emotion attributes.

2. Related Work

Many video datasets with "attribute" annotations are available to the community. Below are some of the most relevant.

- UCF-101 [10], HMDB-51 [4], and the more recent Kinetics [3] provide annotations for a variety of human actions - however, this is not an attribute dataset because it does not consider actions as descriptions but rather as verbs.
- USAA Attributes [1] consists of 1500 "weakly labeled" videos from 8 distinct classes, focusing on social activity scenes. The 69 "attributes" which are annotated can be broken into categories of actions, objects, scenes, sounds, and camera movements.
- Other datasets, such as WWW Crowd [8] are highly specified, containing only video of crowds of people.

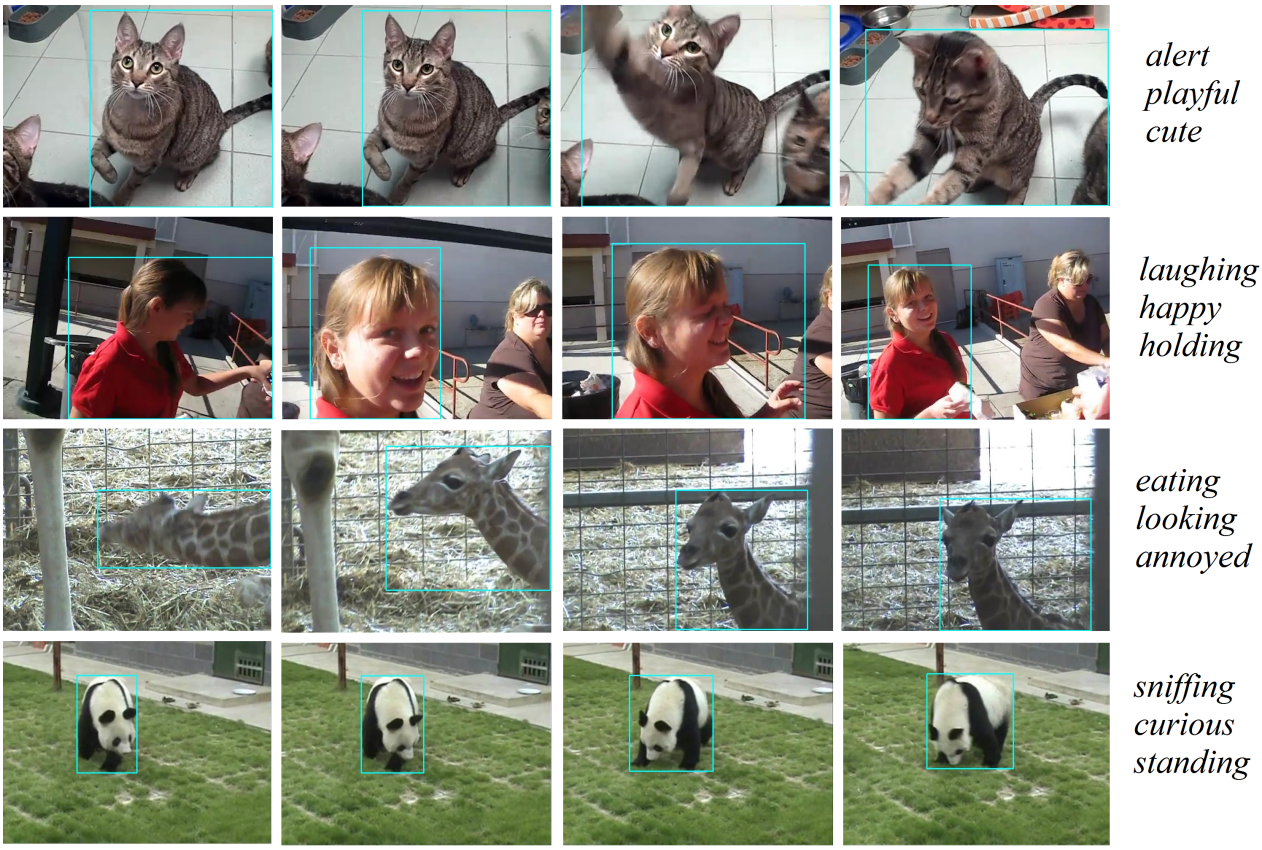


Figure 1. Sample annotated videos from the YouTube-Attributes Dataset. Each row shows frames from one video segment with attributes that describe the object in the video.

Still image attribute prediction datasets are much more abundant. Notably, the COCO Attributes [6] dataset, which includes exhaustive attribute annotation over the existing pixel-wise object segmentation from the original MS COCO dataset [5].

It should be considered that the majority of existing video attribute datasets maintain a distinct definition of the term ”attribute” than that in this paper and in COCO Attributes. In this paper, we consider multi-sourced attributes rather than event detection ”attributes”.

3. Methodology

3.1. Base Dataset

We chose to annotate attributes onto the Youtube-BoundingBoxes Dataset [7]. This dataset is a large-scale set of video URLs with single-object bounding box annotations. It consists of about 380,000 video clips of 15-20 seconds taken from a diverse set of publicly-accessible Youtube videos, featuring 23 distinct object categories. The objects are primarily in natural settings without editing or post-processing, with similar quality to a hand-held cell phone camera. The videos are all hand-annotated with classification labels and bounding boxes at one frame per second. A large fraction of this dataset is from the ”person” class, which is especially important since the research community has a vested interest in human-focused vision problems. However, one of our factors in choosing a base dataset was increasing video diversity. Annotating attributes over both animals and humans featured in YT-BB will provide a more general dataset which will have greater applications than a set focusing on a single class.

3.2. Attribute Selection

A list of 85 attributes was created by hand to apply to the 10 categories in the dataset. Of these, more than half apply to all of the object classes (bear, bird, cat, cow, dog, elephant, giraffe, horse, human, zebra).

These attributes were generated based on a large subset of YT-BB, and were written to be as general as possible. There are a few exceptions, however, particularly pertaining to the human and bird categories. Only birds can be described as ”perched” or ”flying,” and similarly the attribute ”laughing” specifically applies to human objects. However, the majority have more applications- general emotional descriptions like ”happy,” ”annoyed,” or ”excited,” could ap-

ply to all ten classes, and many additional animal classes that are not part of YT-BB. This stability makes our dataset very valuable for attribute prediction on nearly any video dataset. We determined that an average of 60 attributes were applicable to each animal or human class, and these were sorted accordingly.

The list of 85 attributes can be further separated into three categories: action attributes, emotion attributes , and descriptive attributes. Action attributes are primarily -ing adjectives, which describe something that the animal or human is actively doing, such as "trotting" or "playing." Emotion attributes are, as expected, characterizations of the object's emotional state, such as "angry." Finally, descriptive attributes are other adjectives that are not actions or emotions, but can be identified in the video clip - for example, describing a dog as "obedient" if it is walking calmly on a leash, or "strong" if it is shown pulling the leash taut.

Of the 60 attributes that pertain to each category, about half or 30 terms are action attributes, 20 are descriptive, and 10 are emotions. In our annotation platform (see Section 3.3), attributes were grouped in this manner for greater labeling simplicity.

It should be noted that action attributes are distinctly different than action or fine-grained action annotations that have previously appeared in other action/activity datasets (See Section 2). This is primarily because the object is not solely human, but rather one of ten distinct categories. Hence unlike previous action recognition datasets, a model which learns "playing" attribute from our dataset can generalize it to a wide range of human and animal videos, instead of simply humans.

3.3. Human Annotations

Like most existing large datasets, we used by-hand annotations to label the videos from YT-BB. We created an annotation platform based heavily on a tool created at the University of California at Berkeley for bounding-box annotations [9]. Each video, along with a subset of attributes, was provided to two different human annotators, who were instructed to watch the short clip while paying attention to the animal or human in the bounding box. Then, they would read through the list of 15-20 provided attributes, and check the ones that they believe apply to the object in the video. A user-input text box was also provided, for annotators to add their own attributes as desired. We required a minimum of five words for each video, either chosen from the provided list or added in manually.

We used *Amazon Mechanical Turk* to annotate attributes onto YouTube-BB. This allowed for quick results and feedback, but reduced quality of annotation that often is the drawback of crowd-sourcing. This issue was mitigated by video replication, and showing users a constant set of attributes rather than a new list to be read for each task.

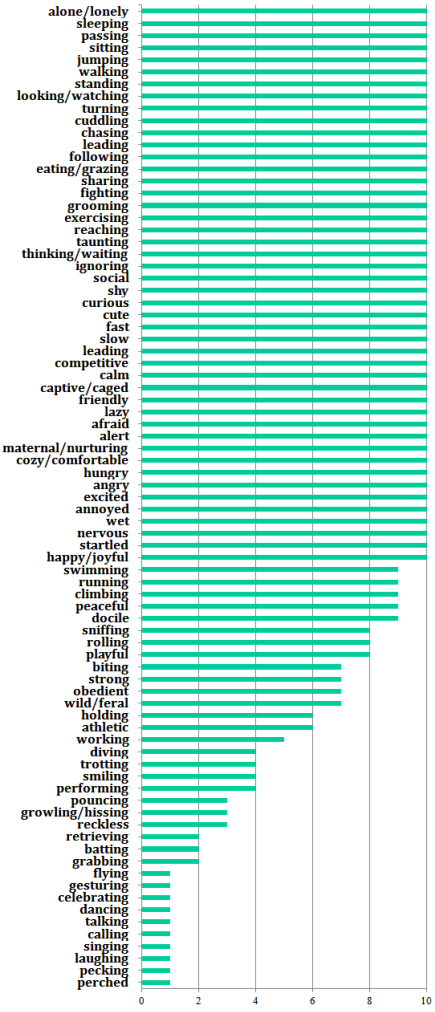


Figure 2. Attributes and number of applicable categories

An important aspect for generating accurate human annotations is the worker's user experience. We designed our annotation platform to be as simple as possible, with clear-cut bulleted instructions and a minimalist layout. Users were also given the option to write in feedback comments. We used these comments to frequently fine-tune the tasks, making both functional and aesthetic adjustments as needed.

3.3.1 User Interface

Figure 3 shows the user interface as seen by workers in Amazon Mechanical Turk. For ease of use, a subset of the 60 pertinent attributes were shown to each worker, and the attributes that were not shown are displayed to a different

worker in the second copy of the task. Turkers are asked to click the checkboxes of the attributes that they believe apply, as well as write any additional annotations if they feel a term is missing from the displayed attribute list. We determined that each task should take a worker around 30 seconds, before accounting for any sort of browser-based lag.

The annotation platform also included a landing page for monitoring active annotation tasks. Part of this is shown in figure 4. Here, we can view annotations from AMT in real time, and view information like annotation time, operating system used by workers, and any comments that may have been included. We used this page to quickly assess annotation quality and track overall progress. Based on feedback and on resulting annotations, we adjusted the platform to account for server traffic, user experience bugs, and more. If tasks are annotated with 5 or more attributes, they are approved, and if not they are rejected; approved annotations are stored with information about the video itself, therefore building YouTube-Attributes.

4. Discussion

In this paper, we introduced YouTube-Attributes, a new dataset with 20,000 video clips, hand-annotated with over 85 generalized attributes, describing objects in 10 categories. This represents the largest and most diverse video attribute dataset (See Section 2) available to the research community. It can be used for multi-task learning - object recognition, object localization, and video attribute prediction.

5. Future Work

Future work on the dataset could refine and improve YouTube-Attributes in a variety of ways. Additional videos from YT-BB can be annotated in a similar process, to extend the existing attribute dataset. Or, new videos containing different object classes could be added and annotated, generalizing the data even further. Alternatively, existing annotations could be improved with a review process.

When the YouTube-Attributes dataset is fully annotated, it will be used for training, validation, and testing using a deep attribute prediction network. As a baseline, we have implemented C3D [11], an established model consisting of a system 3-D convolution neural networks.

We can then supplement this architecture for improved performance in recognizing attributes, and tested on a variety of existing video datasets with class labels. Our proposed model for end-to-end attribute prediction in videos consists of C3D and a T-CNN [2], as shown in figure 5. This model aggregates results from C3D and from the T-CNN architecture, which employs region proposal networks as spatio-temporal tubes. We predict that this model will be a robust system for attribute prediction in videos of many classes.

Acknowledgments

This work was supported by NSF Grant number 1461121, as a part of University of Central Florida NSF Research Experience for Undergraduates in Computer Vision.

References

[1] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity, 2012. EECV.

[2] R. Hou, C. Chen, and M. Shah. Tube convolutional neural network (T-CNN) for action detection in videos. *CoRR*, abs/1703.10664, 2017.

[3] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset, 2017. arXiv preprint arXiv:1705.06950.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. 2011. *International Conference on Computer Vision (ICCV)*.

[5] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[6] G. Patterson and J. Hays. Coco attributes: Attributes for people, animals, and objects. *European Conference on Computer Vision*, 2016.

[7] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. 2017. *arXiv:1702.0082*.

[8] J. Shao, K. Kang, C. Change Loy, and X. Wang. Deeply learned attributes for crowded scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] A. Shen. Beaverdam: Video annotation tool for computer vision training labels. Master's thesis, EECS Department, University of California, Berkeley, 2016.

[10] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human action videos in the wild, 2012. *arXiv preprint arXiv:1212.0402*.

[11] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

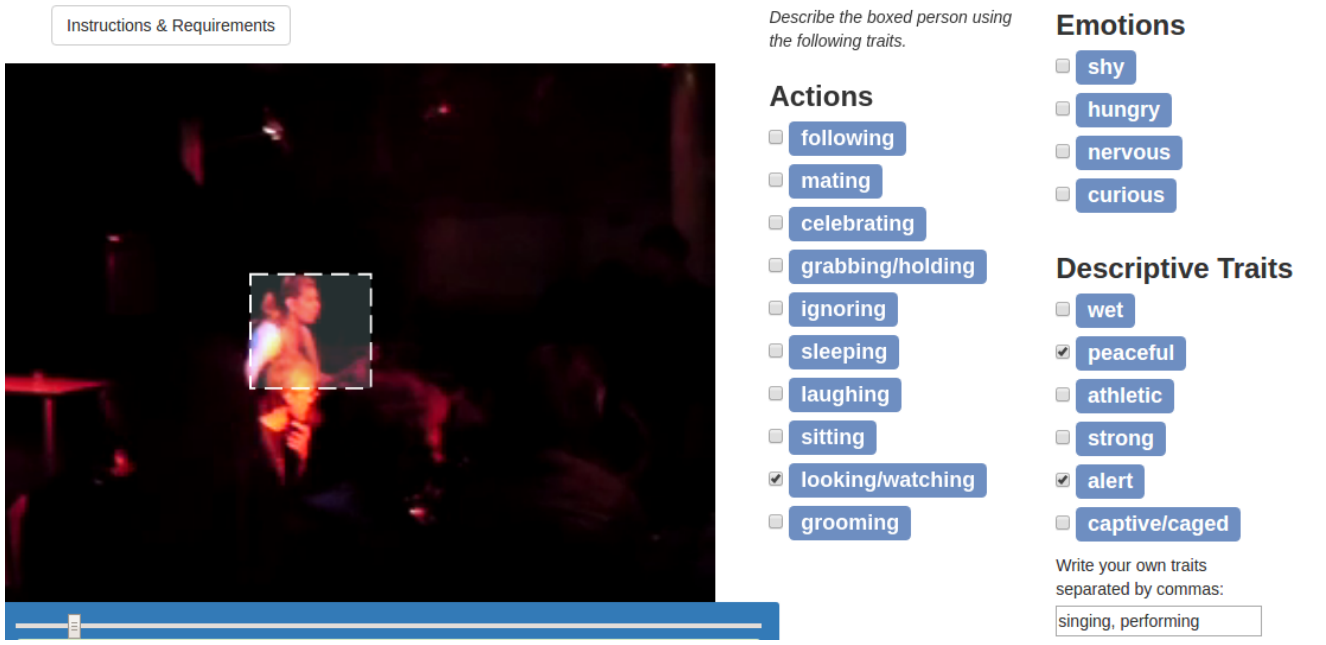


Figure 3. Screenshot of the human annotation tool used to gather video-wise attribute labels. For each short video, users click the attributes that they believe apply, and can add their own.

Class	Annotations?	Annotations	Verified	File	Link
person	yes	[{"color": "#bceef4", "keyframes": [{"x": 114.88, "y": 94.0000032, "w": 50.24000000000001, "h": 46.39999679999998, "continueInterpolation": true, "frame": 0}], "type": "looking/watching,peaceful,alert,*singing, performing"}]	no	AsEWLcMd1aE.mp4	view
bear	no		no	AA19zjGEPvg.mp4	view
cat	no		no	ZyDqefuyQfU.mp4	view
dog	no		no	AV9y4LnUV84.mp4	view
cat	no		no	AvdgweWTeeg.mp4	view
cat	no		no	AxMFQ834oGg.mp4	view
person	no		no	Ast7PFTEKdw.mp4	view
horse	no		no	AIQReB1+AF8-Jlg.mp4	view

Figure 4. Screenshot of requester platform, which can be used to manage tasks, track progress, and assess overall annotation quality. Tasks can be approved/rejected within this interface, or from the command line controls.

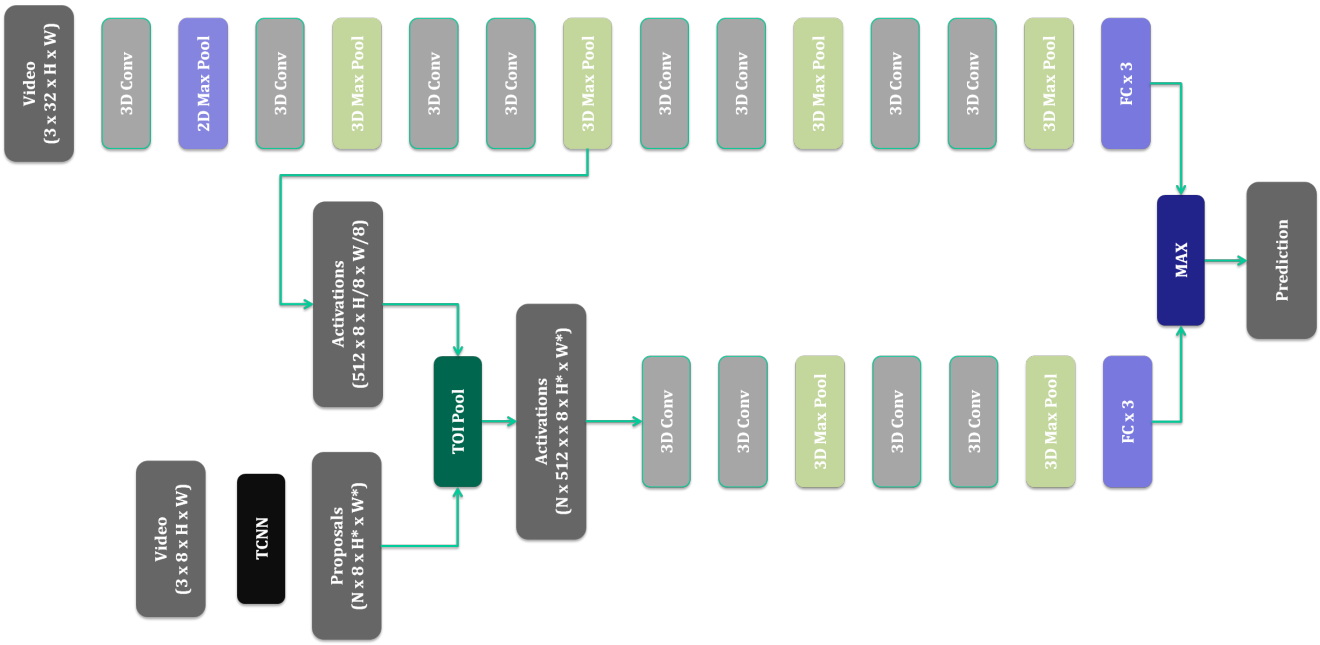


Figure 5. Proposed model for end-to-end attribute prediction, to be trained on the YouTube-Attributes dataset. The top component is C3D, running concurrently with a T-CNN (on bottom). The results of these two networks are aggregated to make video predictions. This model can be tested on a variety of video datasets.