The Open World of Micro-Videos

Phuc Xuan Nguyen[1], Gregory Rogez[2], Charless Fowlkes[1], Deva Ramanan[3]
[1]University of California, Irvine, [2]Inria Rhone-Alpes, [3]Carnegie Mellon University
{nguyenpx,fowlkes}@ics.uci.edu, gregory.rogez@inria.fr, deva@cs.cmu.edu

We examine an increasingly prevalent form of media known as *micro-videos*, time-constrained (typically 5-10 second) video clips commonly used on social networking sites such as Instagram and Vine. Micro-videos can be interpreted as the visual analog of a character-limited microblogs or "tweets". An estimated 12 million micro-videos are posted to Twitter each day. The number of micro-videos produced *surpasses the total inventory of YouTube every 3 months*. From an applied perspective, this flood of visual data is increasingly important and has unique characteristics that are not addressed by existing computer vision methodologies and benchmarks. We further argue that the microvideo format offers unique opportunities for basic research in building systems that address lifelong learning in *open-world* visual domains. In this work, we introduce and analyze a micro-video dataset with **260 thousand** micro-videos labeled with **58 thousand** tags. Example frames from the dataset are shown in Figure 1.

Ease of collection/processing: A particularly attractive aspect of micro-videos is the ease of large-scale collection, storage and processing. While large-scale datasets have revolutionized static-image processing, the counterpart for video-based recognition does not appear to exist. One reason is that it is notoriously challenging to collect, store, and process a large diverse video collection because of resource constraints. Indeed, existing video collections often contain multiple snippets cropped from a few longer videos.

Viewpoint: A unique technical aspect of micro-videos found on social networks is camera viewpoint. Current action datasets focus on *third-person* depictions of actions, where a person performing an action or activity is framed in the view. In contrast, a significant fraction of socially-driven micro-videos include *egocentric* viewpoints, where the photographer is participating in the action. Another camera configuration is a *self-facing* viewpoint, or "selfie", where a single user is both the photographer and subject. This is particularly interesting in the study of interactions photographers and their subjects. Such diverse camera viewpoints represent new modes of video acquisition that are not typically addressed in previous works. Example frames for different viewpoints are shown in Figure 1.

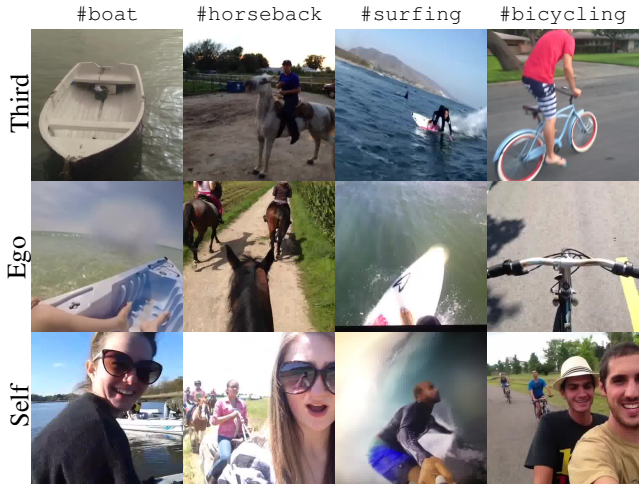

Figure 1: Example frames from six-second micro-videos with particular tags in each column, and different camera perspectives along each row.

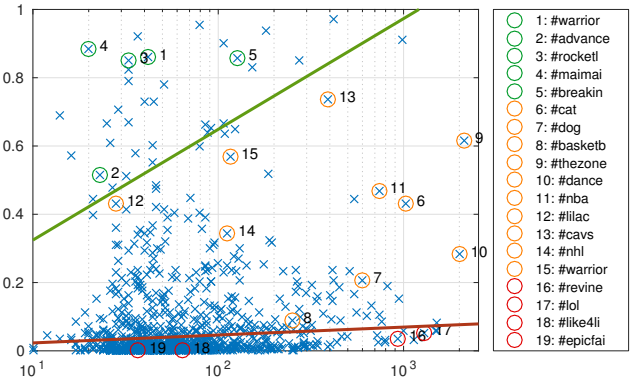

Figure 2: Open-world tag prediction. The scatter plot shows per-tag APs vs (log) number of training examples. We rank the "learnability" of a tag by the ratio of its AP to (log) number of training examples, and draw lines to loosely denote regions of easy, challenging, and unlearnable tags. Unlearnable tags appear to correspond to "stopwords" such as `#revine,lol` that do not capture video content. The per-tag mAP_T is 0.05.

Open-world dynamics: Micro-videos come labeled with hashtags that enable search and play an important

(a) Tag distribution

(b) Tag usage dynamics

Figure 3: Micro-videos lie in a regime between images and traditional videos, encoding semantically rich micro-narratives while remaining tractable to collect, store and process. Figure (a) shows the word cloud visualization of the tag distribution for micro-videos. Our dataset, **MV-58K**, includes common tags about actions and objects seen in other computer vision datasets such as `#cat` and `#dogs` as well as specific tags such as `#noseguitar` and `#puppetman` which are video-graphic styles unique to the micro-video sharing service Vine. Figure (b) shows the unique temporal structure in microvideos. We visualize changes in tag label priors over time by plotting their popularity rank as a heatmap (brighter denotes higher popularity). Note that many tags exhibit large fluctuations; `#gopdebate` is ranked first during the weeks of the associated events, but dramatically decreases otherwise. We can also observe interaction between the tags; `#cnn` rises in popularity as `#demdebate` and `#gopdebate` become popular. Some tags (`#music`) exhibit relatively stable popularity.

role in social communication. These tags provide a form of supervision, automatically labeling our diverse, multi-view, pre-trimmed dataset. In contrast to existing efforts to develop top-down ontologies for activities or events, tags form an open-world vocabulary whose usage and semantics changes dynamically over time. For example, the tag `#trump2016` did not exist until recently, while the visual meaning of the tag `#apple` expands whenever a new iPhone is released. Figure 3b includes more examples demonstrating the temporal dynamic of micro-video tagging. Micro-videos thus provide a unique opportunity to explore learning *in the open world*, where distributions of visual semantics follow long-tail statistics that change time over time. Taking a *bottom-up, data-driven* approach to video content semantics more accurately reflects naturally-occuring long-tail distributions that typically suppressed in hand-curated datasets. The dataset is thus large, has dynamic temporal variations and is continually growing is size. We analyze a snapshot as of Feb 2016 consisting of 264,327 videos with 58,243 tags.

Models for micro-video tag prediction: Recent work has shown that Convolutional Neural Networks (CNN) produce quite effective visual descriptors for recognition tasks including video analysis. We experimented with many open-source implementations of CNN architectures for video-feature extractors, and found a good tradeoff in speed, storage, and accuracy with the following simple pipeline: given a video, (1) run off-the-shelf VGG-16 models on 15 equally-spaced 8 frames from a video, (2) for each frame, extract multi-scale features across multiple layers, and (4) max-pool the resulting features across the 15 frames. We also include trajectory-based motion features in our analysis, focusing on Improved Dense Track (IDT) and extract oriented gradient histograms and optical flow. We train tag classifiers that aggregate features by combining multiple kernels.

Challenging un-curated tagging task: Current video benchmarks (HMDB and UCF101) have observed significant progress in recent years (91.5% UCF-101 and 65.9% on HMDB). We observed similar performance for a manually-curated subset of our dataset. However, the per-label and per-image mean Average Precision on the full un-curated dataset are 5% and 21% respectively (more details are shown in Figure 2). This suggests that while we are able to build excellent models for curated dataset, these models do not perform as well in the wild.

Conclusion: We introduce a challenging open-world dataset of micro-videos, which lie in a regime between single images and typical videos, allowing for easy capture, storage, and processing. They contain micro-narratives captured from viewpoints typically not studied in computer vision. Because they are naturally diverse, pre-trimmed, and user-annotated, they can be used a live testbed for open-world evaluation of video understanding systems. We conclude with an intriguing thought: rather than distributing a fixed benchmark dataset (which historically leads to eventual overfitting), we can instead distribute a benchmark script that evaluates models on live open-world micro-videos. We think the time is right to consider video recognition out in-the-open!