# OOPS! Predicting Unintentional Action in Video

Dave Epstein Boyuan Chen Carl Vondrick
Columbia University
Full paper at oops.cs.columbia.edu

## Abstract

*From just a short glance at a video, we can often tell whether a person's action is intentional or not. Can we train a model to recognize this? We introduce a dataset of in-the-wild videos of unintentional action, as well as a suite of tasks for recognizing, localizing, and anticipating its onset. We train a supervised neural network as a baseline and analyze its performance compared to human consistency on the tasks. We also investigate self-supervised representations that leverage natural signals in our dataset, and show the effectiveness of an approach that uses the intrinsic speed of video to perform competitively with highly-supervised pretraining. However, a significant gap between machine and human performance remains.*

## 1. Introduction

From just a glance at a video, we can often tell whether a person's action is intentional or not. For example, Figure 1 shows a person attempting to jump off a raft, but unintentionally tripping into the sea. In a classic series of papers, developmental psychologist Amanda Woodward demonstrated that this ability to recognize the intentionality of action is learned by children during their first year [3, 4, 1]. However, predicting the intention behind action has remained elusive for machine vision. Recent advances in action recognition have largely focused on predicting the physical motions and atomic actions in video, which captures the means of action but not the intent of action.

We believe a key limitation for perceiving visual intentionality has been the lack of realistic data with natural variation of intention. Although there are now extensive video datasets for action recognition, people are usually competent, which causes datasets to be biased towards successful outcomes. However, this bias for success makes discriminating and localizing visual intentionality difficult for both learning and quantitative evaluation.

We introduce a new annotated video dataset that is abundant with unintentional action, which we have collected by crawling publicly available "fail" videos from the web. Figure 2 shows some examples, which cover in-the-wild situations for both intentional and unintentional action. Our video dataset, which we will publicly release, is both large (over 50 hours of video) and diverse (covering hundreds of scenes and activities). We annotate videos with the temporal location at which the video transitions from intentional to unintentional action. We define three tasks on this dataset:

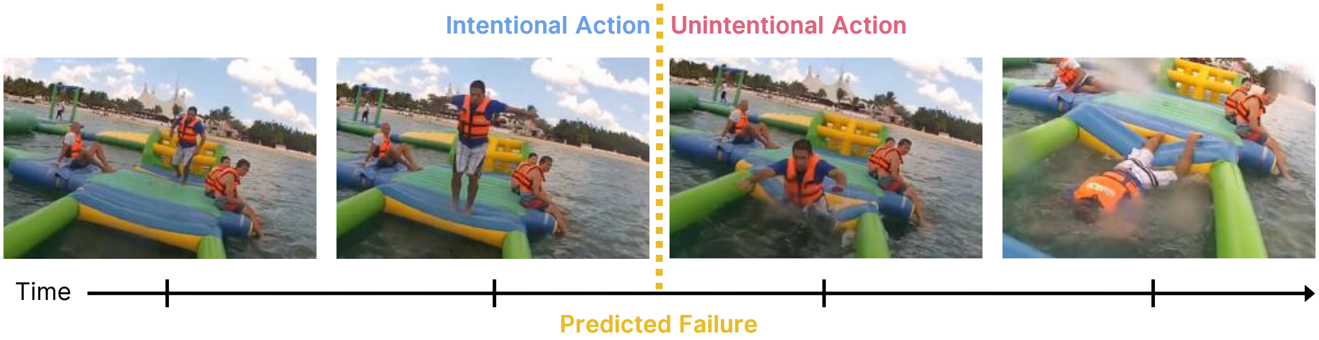

Figure 1: **Intentional versus Unintentional:** Did this person intend for this action to happen, or was it an accident? In this paper, we introduce a large in-the-wild video dataset of unintentional action. Our dataset, which we have collected by downloading "fail" videos from the web, contains over twenty thousand clips, and they span a diverse number of activities and scenes. Using this dataset, we study a variety of visual clues for learning to predict intentionality in video.

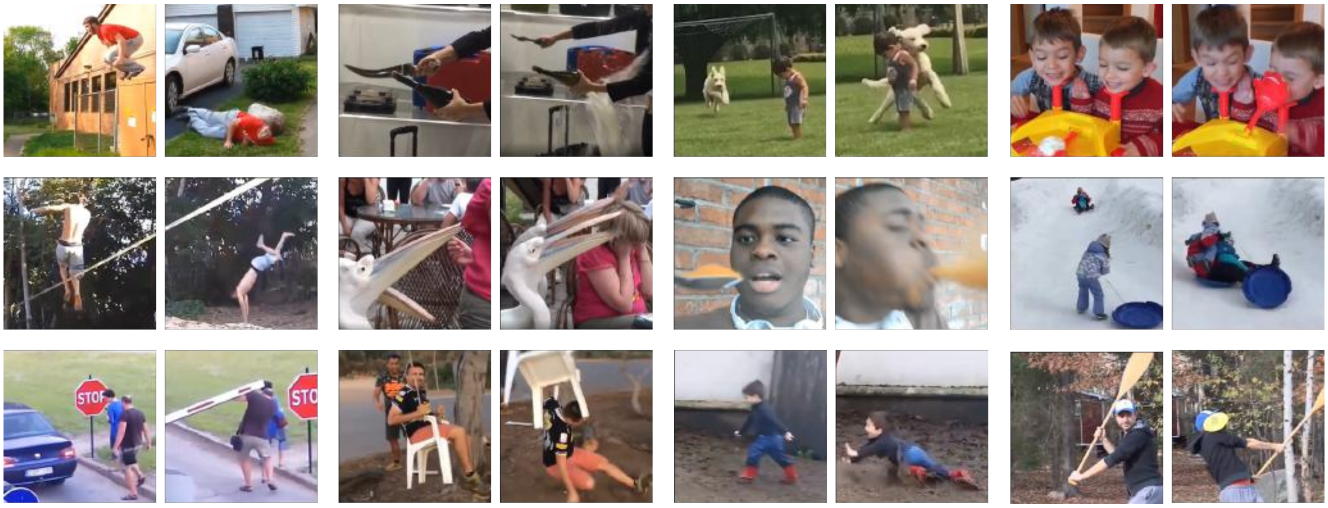

Figure 2: **The OOPS! Dataset:** Each pair of frames shows an example of intentional and unintentional action in our dataset. By crawling publicly available "fail" videos from the web, we can create a diverse and in-the-wild dataset of unintentional action. For example, the bottom-left corner shows a man failing to see a gate arm, and the top-right shows two children playing a competitive game where it is inevitable one person will fail to accomplish their goal.

classifying the intentionality of action, localizing the transition from intentional to unintentional, and forecasting the onset of unintentional action shortly into the future.

To tackle these problems, we propose a novel self-supervised task to learn to predict the speed of video, which is incidental supervision available in all unlabeled video, for learning an action representation. We also explore other visual clues that have been successful for representation learning.

Our results show that learning to predict the speed of video is a strong self-supervised task for recognizing intentionality. By ablating model and design choices, our analysis also suggests that models do not rely solely on low-level motion clues to solve unintentional action prediction. Moreover, although human consistency on our dataset is high, there is still a large gap in performance between our models and human agreement, underscoring that analyzing human goals from videos remains a fundamental challenge in computer vision. We hope this dataset of unintentional and unconstrained action can provide a pragmatic benchmark of progress.

This paper makes two primary contributions. Firstly, we introduce a new dataset of unconstrained videos containing a substantial variation of intention and a set of tasks on this dataset. Secondly, we present models that leverage a variety of incidental clues in unlabeled video to recognize intentionality. The remainder of this paper will describe these contributions in detail. Section 2 introduces our dataset and summarizes its statistics. Section 3 presents several self-supervised learning approaches to learn visual representations of intentionality. In Section 4, we present quantitative and qualitative experiments to analyze our model. We release all data, software, and models on the website. Please see the website also for the full version of this paper.

## 2. The OOPS! Dataset

We present the OOPS! dataset for studying unintentional human action. The dataset consists of 20,338 videos from YouTube "fail" compilation videos, adding up to over 50 hours of data. These clips, filmed by amateur videographers in the real world, are diverse in action, environment, and intention. Our dataset includes many causes for failure and unintentional action, including physical and social errors, errors in planning and execution, limited agent skill, knowledge, or perceptual ability, and environmental factors. We have released the dataset, along with pre-computed optical flow, pose, and annotations. We define three tasks on this dataset (shown in Figure 3) and later evaluate various approaches on these tasks. We believe that this dataset will facilitate the development and evaluation of models that analyze human intentionality.

### 2.1. Data Collection and Processing

We build our dataset from online channels that collate "fail" videos uploaded by many different users, since the videos they share display unconstrained and diverse situations. Figure 2 shows several example frames. We preprocess the videos to remove editorial visual artifacts.

### 2.2. Annotation

We labeled the temporal locations of failure in the entire test set and some of the training set using Amazon Mechanical Turk. We ask workers, whom we restrict to a ≥99% ap-

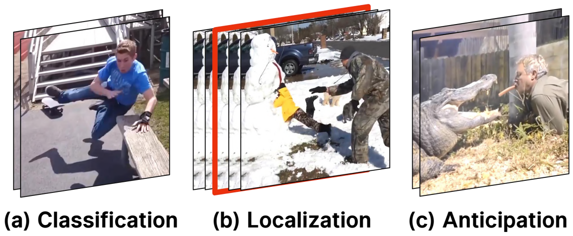
(a) Classification (b) Localization (c) Anticipation

Figure 3: **Tasks:** Our dataset has three tasks: classification of action as intentional or not, temporal localization of unintentional action, and forecasting unintentional action.

proval rating with at least 10,000 approvals, to mark videos at the moment when failure starts to happen (*i.e*. when actions start to become unintentional). We found that humans are very consistent across each other at labeling the time of failure. The median standard deviation across workers is about half a second, or 6.6% of the video duration.

## 3. Intentionality from Perceptual Cues

We investigate a variety of perceptual cues for learning to predict intentional action with minimal supervision. We can cast this as a self-supervised learning problem. Given incidental supervision from unlabeled video, we aim to learn a representation that can efficiently transfer to different intentionality recognition tasks. Principally, we introduce a new self-supervised representation learning task that requires predicting video speed. We also implement contrastive [2] and sort-order [5] pretext tasks, but omit their descriptions for brevity. Please follow the citations for more detail.

### 3.1. Predicting Video Speed

The speed of video provides a natural visual clue to learn a video representation. We propose a self-supervised task where we synthetically alter the speed of a video, and train a convolutional neural network to predict the true frame-rate. Since speed is intrinsic to every unlabeled video, this is a self-supervised pretext task for video representation learning.

Let $x_{i,r} \in \mathbb{R}^{T \times W \times H \times 3}$ be a video clip that consists of $T$ frames and has a frame rate of $r$ frames-per-second. We use a discrete set of frame rates $r \in \{4, 8, 16, 30\}$ and $T = 16$. Consequently, all videos have the same number of frames, but some videos will span longer time periods than others. We train a model on a large amount of unlabeled video using cross-entropy loss.

Our hypothesis, supported by our experiments, that speed is a useful self-supervisory signal for representation learning. For example, a person leisurely sitting down appears intentional, but a person suddenly falling into a seat appears accidental. Recently, fake news campaigns have manipulated the speed of videos to convincingly forge and alter perception of intent.

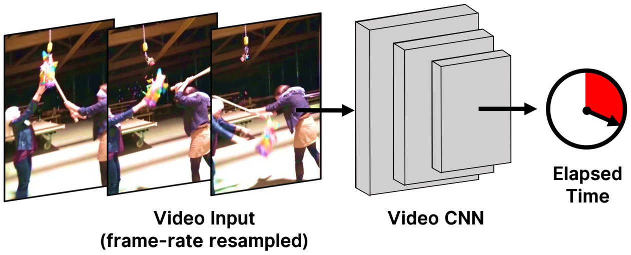

Figure 4: **Video Speed as Incidental Supervision:** We propose a new self-supervised task to predict the speed of video, which is naturally available in all unlabeled video.

### 3.2. Fitting the Classifier

We use these self-supervised clues to fit a classifier to discriminate action as intentional, unintentional, or transitional. We train the self-supervised models with unlabeled video, and fit a linear classifier with minimal annotation, allowing us to directly compare the quality of the learned representations for recognizing intentionality.

## 4. Experiments

The goal of our experiments is to analyze mid-level perceptual clues for recognizing intentionality in realistic video. To do this, we quantitatively evaluate the self-supervised methods on three tasks on our dataset (classification, localization, and anticipation). We also show quantitative ablations and qualitative visualizations to analyze limitations. Here, we show results on the challenging localization task. Please see the full paper for more extensive results.

### 4.1. Localization

We evaluate temporal localization, which is challenging because it requires the model to detect the temporal boundary between intentional and unintentional action. We use our classifier in a sliding window fashion over the temporal axis, and evaluate whether the model can detect the point in time that the action switches from intentional to unintentional. The predicted boundary is the one with the most confident score of transition across all sliding windows.

Table 1 reports accuracy at localizing the transition point. For both thresholds, the best performing self-supervised method is video speed, outperforming other self-supervised methods by over 10%, which suggests that our video speed task learns more fine-grained video features. Figure 5 shows a few qualitative results of localization as well as high-scoring false positives.

### 4.2. Analysis

Our results so far have suggested that there are perceptual clues in unlabeled video that we can leverage to learn to recognize intentionality. In this subsection, we break down

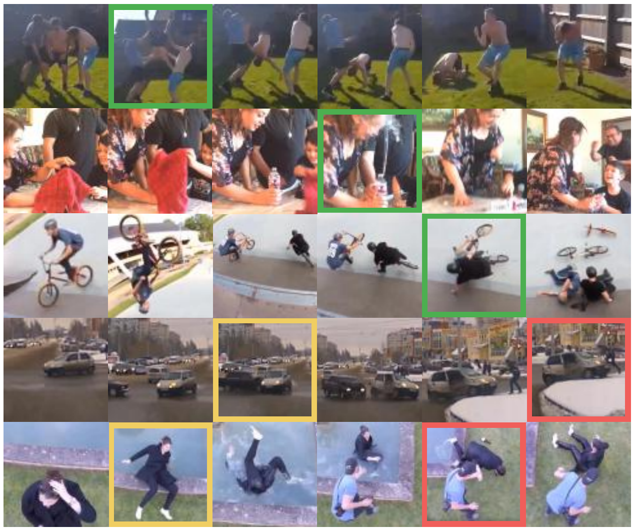

Figure 5: **Example Localizations:** We show example predictions for localizing the transition to unintentional action. Green indicates a correct prediction (within 0.25 sec). Red indicates an incorrect, yet reasonable, prediction. Yellow indicates a missed detection.

| **Method** | **Accuracy within**<br>**1 sec** | <br>**0.25 sec** |
|---|---|---|
| Human Consistency | 88.0 | 62.1 |
| Kinetics Supervision (Fine-tune) | 75.9 | 46.7 |
| Kinetics Supervision (Linear) | 69.2 | 37.8 |
| Video Speed (ours) | **65.3** | **36.6** |
| Video Context [2] | 52.0 | 25.3 |
| Video Sorting [5] | 43.3 | 18.3 |
| Scratch | 47.8 | 21.6 |
| Middle Prior | 53.1 | 21.0 |
| Chance | 25.9 | 6.8 |

Table 1: **Temporal Localization:** We evaluate the model at localizing the onset of unintentional action for two different temporal thresholds of correctness. Although there is high human agreement on this task, there is still a large gap for both supervised and self-supervised models.

performance to analyze strengths and limitations. Please see the full paper for more analysis.

**Visualization of Learned Features:** To qualitatively analyze the learned feature space, Figure 6 visualizes nearest neighbors between videos using the representation learned by predicting the video speed, which is the best performing self-supervised model. We use one video clip as a query, compute features from the last convolutional layer, and calculate the nearest neighbors using cosine similarity over a large set of videos not seen during training. Although this feature space is learned without ground-truth labels, the nearest neighbors are often similar activities and objects, suggesting that learning to predict the video speed is promising incidental supervision for learning features of activity.

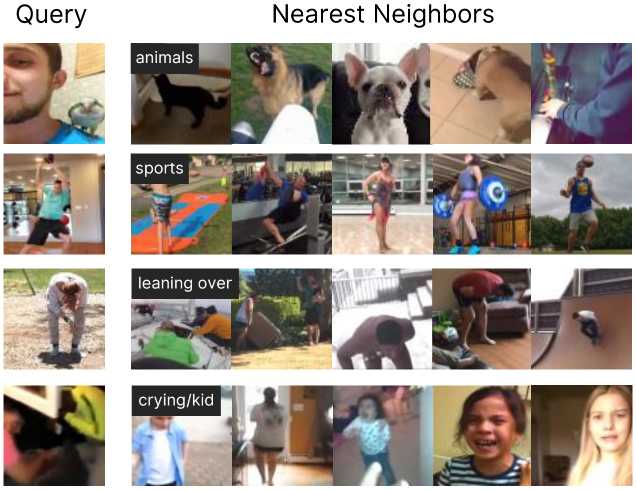

Figure 6: **Nearest Neighbors on Self-supervised Features:** We visualize some of the nearest neighbors from the feature spaced learned by predicting video frame rate. The nearest neighbors tend to be similar activities despite significant variation in appearance.

## 5. Discussion

This paper investigates mid-level perceptual clues to recognize unintentional action in video. We present an "in-the-wild" video dataset of intentional and unintentional action, and we also leverage the speed of video for representation learning with minimal annotation, which is a natural signal available in every unlabeled video. However, since a significant gain remains to match human agreement, learning human intentions in video remains a fundamental challenge.

## References

[1] Amanda C Brandone and Henry M Wellman. You can't always get what you want: Infants understand failed goal-directed actions. *Psychological science*, 20(1):85–91, 2009. 1

[2] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[3] Amanda L Woodward. Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, 22(2):145–160, 1999. 1

[4] Amanda L Woodward. Infants' grasp of others' intentions. *Current directions in psychological science*, 18(1):53–57, 2009. 1

[5] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 3, 4