

oops! Predicting Unintentional Action in Video

Dave Epstein Boyuan Chen Carl Vondrick
 Columbia University
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 pre-training. 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 [70, 71, 6]. 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 [28, 18, 40], 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 [28, 18, 40], 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 investigate several visual clues for learning with minimal labels to recognize intentionality. First, 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. Second, we explore the predictability of the temporal context as a clue to learn features, as unintentional action often deviates from expectation. Third, we study the order of events as a clue to recognize intentionality, since intentional action usually precedes unintentional action.

Experiments and visualizations suggest that unlabeled video has intrinsic perceptual clues to recognize intentionality. Our results show that, while each self-supervised task is useful, learning to predict the speed of video helps the most. 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 first reviews related work in action recognition. Then, Section 3 introduces our dataset and summarizes its statistics. Section 4 presents several self-supervised learning approaches to learn visual representations of intentionality. In Section 5, we present quantitative and qualitative experiments to analyze our model. We release all data, software, and models on the website.

2. Related Work

Video datasets: Computer vision has made significant progress in recognizing human actions through video analysis. Critical to this success are datasets of diverse videos released to facilitate this research [50, 5, 32, 53, 64, 27, 7, 1, 41, 51, 28, 17, 16, 11, 18, 40]. Most modern datasets are intended for discriminating between human activities to perform action classification and localization [46, 69, 2, 10, 26, 26, 4, 76]. In our paper, we instead focus on analyzing goal-directed human action [59], and propose a dataset that allows for learning about failed goals and the transition from intentional to unintentional action. Our dataset includes both human errors caused by imperfect action execution (*e.g.* physical interference, limited visibility, or limited knowledge) and human errors due to mistakes in action planning (*e.g.* flawed goals or inadequate reasoning).

Action recognition and prediction: Our work builds on a large body of literature in action classification and prediction. Earlier research in action classification [34, 30, 63, 49, 45] focuses on designing features or descriptors for given input video frames. Recent progress has focused on using deep convolutional networks to solve these tasks, and many methods have been proposed to learn useful feature represen-



Figure 3: Dataset Statistics: We summarize our dataset with the **(a)** distribution of clip lengths, **(b)** the distribution of temporal locations where failure starts, and **(c)** the standard deviation between human annotators. The median and mean clip lengths are 7.6 and 9.4 seconds respectively. Median standard deviation of the labels given across three workers is 6.6% of the video duration, about half a second, suggesting high agreement. We also show the distribution of **(d)** action categories and **(e)** scene categories, which naturally has a long tail. For legibility, we only display the top and bottom 5 most common classes for each. Figure best viewed on a computer screen with zoom.

tations, such as visual information fusion [65, 8], two-stream CNNs [52], 3D convolutional networks that take in a chunk of video [56], and temporal reasoning for feature extraction [78, 13]. In this paper, we base our methods on 3D CNNs.

Previous work which studies future action prediction in video is also relevant to predicting unintentionality [48, 44, 72, 22, 11]. Many methods rely on action label supervision, along with other auxiliary information, to predict future actions [75, 77]. Other approaches [60, 57, 73, 82] focus on leveraging large unlabeled datasets to learn visual representations useful for action anticipation.

Self-supervised learning: Our work uses unlabeled video to learn useful representations without manual supervision. In recent years, self-supervision, which predicts information naturally present in data by manipulating or withholding part of the input, has become a popular paradigm for unsupervised learning. Various types of self-supervised signals have been used to learn strong visual representations, such as spatial arrangement [42], contextual information [12, 66], color [35, 62], the arrow of time [68, 73, 21, 36, 39, 15], future prediction [38, 61, 74, 43], consistency in motion [3, 24], view synthesis [81, 80], spatio-temporal coherence [58, 67, 14, 37, 33], and predictive coding [43, 55]. Learned representations are then used for other downstream tasks such as image classification, object detection, video clip retrieval, and action recognition. We introduce a new self-supervised pretext task to estimate video speed, which is effective for learning video representations.

3. 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 plan to release the dataset, along with pre-computed optical flow, pose, and annotations, in the near future. We believe that this dataset will facilitate the development and evaluation of models that analyze human intentionality.

3.1. 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. For example, after downloading the long compilation videos from these channels, we must delineate scene boundaries to separate between unrelated clips. We experiment with various such methods and found that `scikit-video` gives good results.¹ We discard all scenes under 3 seconds

¹We use the scenedet function with method='edges' and parameter1=0.7 from <https://github.com/scikit-video/scikit-video>



Figure 4: **Tasks:** Our dataset has three tasks: classification of action as intentional or not, temporal localization of unintentional action, and forecasting unintentional action.

long, since they are unlikely to contain a complete scene, as well as all scenes over 30 seconds, since they are likely to contain multiple scenes (due to false negatives in scene detection). Some videos were filmed in portrait orientation but collated in landscape, resulting in a “letterbox” effect. We run a Hough line transform to detect these borders, and crop out the border artifacts.

3.2. Annotation

We labeled the temporal locations of failure in the entire test set and some of the training set using Amazon Mechanical Turk [54]. We ask workers, whom we restrict to a $\geq 99\%$ approval 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).

Quality Control: We also use a variety of techniques to ensure high-quality annotation. We repeat annotation three times to verify label quality. We also ask workers to annotate whether the video contained unintentional action or not. We remove all videos where most workers indicate there is no failure or where the failure occurs at the very beginning or end of a video clip (indicating an error in scene detection). The majority of videos we label pass these checks. To control quality, we also manually label ground truth on a small set of videos, which we use to detect and remove poor annotations.

Human Agreement: We annotated the test set a fourth time, which we use to analyze human agreement on this task. 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.

3.3. Dataset Statistics

Figure 3a shows the distribution of video clip lengths and Figure 3b shows the distribution of failure time labels in the dataset. Figure 3c plots standard deviation of the three labels from different workers, which is around half a second on average. Figure 3d and Figure 3e show the action and scene class distributions, as predicted by models pre-trained on the Kinetics and Places [79] datasets. The dataset covers intentions for a variety of scenes and activities.

3.4. Benchmark

We use our dataset as a benchmark for recognizing intentional action. We split the dataset into three sets: an unlabeled set of videos for pre-training, a labeled training set, and a labeled test set. The entire dataset contains 20,338 videos, and the labeled training set contains 7,368 videos, which is kept relatively small because the goal of the benchmark is to evaluate self-supervised learning. The test set contains 6,739 videos, which are labeled only for quantitative evaluation. In our benchmark, models are allowed to train on any number of *unlabeled* videos and only a small number of labeled videos. Figure 4 shows the tasks for the benchmark.

4. Intentionality from Perceptual Clues

We investigate a variety of perceptual clues 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.

4.1. 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:

$$\min_f \sum_i \mathcal{L}(f(x_{i,r}), r) \quad (1)$$

where \mathcal{L} is the cross-entropy loss function. Figure 5 illustrates this task.

We hypothesize, supported by our experiments, that speed is a useful self-supervisory signal for representation learning. Firstly, estimating the speed requires the model to learn motion features because a single frame is insufficient to distinguish between frame rates. Secondly, this task will require the model to learn features that are correlated to the expected duration of events. For example, the model could detect that a video of a person walking is synthetically sped up or slowed down by comparing it to the average human walking speed. Finally, human judgement of intentionality is substantially affected by video speed [9]. For example, a

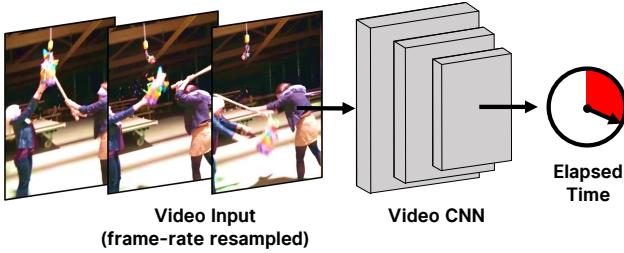


Figure 5: Video Speed as Incidental Supervision: We propose a new self-supervised task to predict the speed of video, which is naturally available in all unlabeled video.

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.²

4.2. Predicting Video Context

Since unintentional action is often a deviation from expectation, we explore the predictability of video as another visual clue for intentions. We train a predictive visual model on our unlabeled set of videos and use the representation as a feature space. Let x_t be a video clip centered at time t , and both x_{t-k} and x_{t+k} be contextual clips at times $t - k$ and $t + k$ respectively. We learn a predictive model that interpolates the middle representation $\phi_t = f_\theta(x_t)$ from the surrounding contextual frames x_{t-1} and x_{t+1} :

$$\max_{f,g} \sum_i \log \left(\frac{e^{z_t}}{e^{z_t} + \sum_{n \in N} e^{z_n}} \right) \quad \text{for } z_j = \frac{\phi_j^T \hat{\phi}_t}{\sqrt{d}} \quad (2)$$

where $\hat{\phi}_t = g_\theta(\{\phi_{t-k}, \phi_{t+k}\})$ such that f_θ and g_θ are convolutional networks. d is the dimension of the representation for normalization, and N is the negative set.

Maximizing this objective corresponds to pulling the features of the target frame, ϕ_t , closer to the contextual embedding, $\hat{\phi}_t$, while pushing it further away from all other negatives in the mini-batch. This objective is an instance of noise-contrastive estimation [25] and contrastive predictive coding [43, 55, 19], which obtains strong results on other self-supervised learning tasks. We use this as a baseline.

We compute the loss over mini-batches, so the negative set for a given middle clip includes all other clip representations in the mini-batch except itself. We set g_θ as a two-layer fully-connected network, with hidden dimension 1024, ReLU as activation, and output dimension $d = 512$ (same dimension as output of the video encoder f_θ).

²<https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/>

4.3. Predicting Event Order

We also investigate the order of events as a perceptual clue for recognizing unintentional action. Since unintentional action often manifests as chaotic or irreversible motion, we implement a convolutional model that is tasked with predicting the permutation applied to shuffled input video clips as in [73, 68], which we use as a strong baseline.

We sample 3 clips with a gap of 0.5sec between subsequent clips, so there are $3! = 6$ possible sort orders. We run all clips through a neural network f_θ , which yields a feature vector, then concatenate feature vectors for all pairs of videos and run them through another neural network g_θ , to represent pairwise clip relations. Finally, we concatenate these pairwise representations and input into a third network h_θ to predict the sort order. The networks g_θ and f_θ are both linear layers with a ReLU activation. The output dimensions of f_θ , g_θ , and h_θ are 512, 256, and 6.

4.4. 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.

Network Architecture: We use the same convolutional network architecture throughout all approaches. Since this is a video task, we need to choose a network architecture that can robustly capture motion features. We use the ResNet3D-18 [20] as the video backbone for all networks, which obtains competitive performance on the Kinetics action recognition dataset [28]. We input 16 frames into the model. Except for the video speed model, we sample the videos at 16 fps, so that the model gets one second of temporal context. We train each network for 20 epochs.

Classifier: After learning on our unlabeled set of videos, the self-supervised models will produce a representation that we will use for our intentionality prediction tasks. We input a video into the self-supervised model, extract features at the last convolutional layer, and fit a linear classifier. While there are a variety of ways to transfer self-supervised representations to subsequent tasks, we chose to use linear classifiers because our goal is to evaluate the self-supervised features, following recommended practice in self-supervised learning [31]. We train a regularized multi-class logistic regression using a small amount of labels on the labeled portion of our training set. We formulate the task as a three-way classification task, where the three categories are: a) intentional action, b) unintentional action, and c) transitioning from intentional to unintentional. We define an action as transitioning if the video clip overlaps with the point the worker labeled.

Method	Linear Classifier		Fine-tune
	All Labels	10% Labels	All Labels
Kinetics Supervision	53.6	52.0	64.0
Video Speed (ours)	53.4	49.9	61.6
Video Context [43]	50.0	47.2	60.3
Video Sorting [73]	49.8	46.5	60.2
Scratch	48.2	46.2	59.4
Motion Magnitude	44.0	-	44.0
Chance	33.3	33.3	33.3

Table 1: **Classification Accuracy:** We evaluate performance of each self-supervised model versus baselines. We also compare against a model trained with Kinetics supervision to understand the gap between supervision and self-supervision. This results suggests learning to predict video speed is a promising form of video self-supervision.

5. 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.

5.1. Baselines

Besides the self-supervised methods above, we additionally compare against several other baselines.

Motion Magnitude: We use simple motion detection as a baseline. To form this baseline, we compute optical flow [23] over the videos, and quantify the motion magnitude into a histogram. We experimented with several different bin sizes, and found that 100 bins performed the best. We then fit a multi-layer perceptron on the histogram, which is trained on the labeled portion of our training set to predict the three categories.

Kinetics Supervision: We compare against a model that is trained on the full, annotated Kinetics action recognition dataset, which is either fine-tuned with our labeled training set, or used as a feature extractor. Since the model is trained on a large, labeled dataset of over 600,000 videos, we do not expect our self-supervised models to outperform it. In-



Figure 6: **Multi-modal Evaluation:** Unintentional actions cascade. For example, in this video, the person “fails” twice. To handle this in evaluation, we consider a prediction correct if it is sufficiently near any ground-truth label.

stead, we use this baseline to understand the gap between supervised and self-supervised methods.

Linear versus Fine-tune: Unless otherwise noted, the classifier is a linear classifier on the features from the last convolutional layer of the network. However, we also evaluated some models by fine-tuning, which we do to understand the best performance that one could obtain at this task. To fine-tune, we simply use the method as the network initialization, change the last layer to be the three-way classification task, and train the network end-to-end with stochastic gradient descent on our labeled set of videos.

Fixed Priors: We also compare against naive priors. We calculate the mode on the training set, and use this mode as the prediction. Additionally, we use chance.

Human Agreement: To establish an upper expectation of performance on this dataset, we use a fourth, held-out worker’s labels to measure human performance.

5.2. Classification

We first evaluate each model on a classification task. Given a short video clip, the task is to categorize it into one of the three categories (intentional, unintentional, or transitional). We extract one-second temporal windows in increments of 0.25 seconds from the testing set.

Table 1 reports classification accuracy for each method. All of the self-supervised methods outperform baselines, suggesting there are perceptual clues in unlabeled video for intentionality. The model trained with full Kinetics supervision obtains the best performance overall, indicating there is still no substitute for labeled data. However, the gap between the self-supervised models and supervised models is relatively small. For example, the best performing perceptual clue (video speed) is tied with Kinetics when large amounts of labels are available for training a linear layer. We also experimented with the reducing the number of examples in our labeled training set. While accuracy is positively correlated with number of labels, reducing the number of labels by an order of magnitude only causes a minor drop in performance.

5.3. Localization

We next 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. Since videos can contain multiple transitional points, we consider the prediction correct if it sufficiently overlaps any of the ground truth positions in the dataset (Figure 6). We use two different thresholds of sufficient overlap: within one second, and within one quarter second.



Figure 7: **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.

Table 2 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. Human consistency at this task is high (88% agreement), however there is still a large gap to both supervised and self-supervised approaches, underscoring the challenge of learning human intent in video. Figure 7 shows a few qualitative results of localization as well as high-scoring false positives. The incorrect predictions our model makes (bottom two rows of Figure 7) are often reasonable, such as a car hitting a pedestrian on the sidewalk (ground truth: car first hits another car) and person falling when exiting fountain (ground truth: person first falling into fountain).

5.4. Anticipation

We also evaluate the representation at anticipating the onset of unintentional action. To do this, we train the models with self-supervision as before, but then fine-tune them for a three-way classification task to predict the labels 1.5 seconds into the future. Table 3 reports classification accuracy for the prediction. Features from the the video speed prediction model obtain the best self-supervised performance. However, the model with full Kinetics supervision obtains about 3% higher performance, suggesting there is still room for self-supervised learning to improve on this task.

Method	Accuracy within	
	1 sec	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 [43]	52.0	25.3
Video Sorting [73]	43.3	18.3
Scratch	47.8	21.6
Motion Magnitude	50.7	23.1
Middle Prior	53.1	21.0
Chance	25.9	6.8

Table 2: **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.

Method	Accuracy
Kinetics Supervision	59.7
Video Speed (ours)	56.7
Video Context [43]	51.2
Video Sorting [73]	51.0
Scratch	50.8
Chance	50.0

Table 3: **Anticipation:** We evaluate performance at predicting the onset of failure before it happens (1.5 seconds into future) by fine-tuning our models. The best performing self-supervised visual clue we considered is video speed.

5.5. 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 performance to analyze strengths and limitations.

Frequent Confusions: Figure 8 compares the confusion matrices for both the video speed representation and Kinetics supervised representation. In both cases, the most challenging point to predict is the boundary between intentional and unintentional action, which we label the “Failure” point. Moreover, a key difference between models is that the self-supervised model more often confuses intentional action with the start of failure. Since the supervised model performs better here, this suggests there is still substantial room to improve self-supervised models at fine-grained localization of intentionality.

Visualization of Learned Features: To qualitatively analyze the learned feature space, Figure 9 visualizes nearest neighbors between videos using the representation learned by predicting the video speed, which is the best performing

		Predicted Label		
		Intent.	Failure	Unintent.
True Label	Intention	62.2	10.2	27.6
	Failure	22.9	43.9	33.2
	Unintentional	23.3	9.3	67.4

(a) Kinetics + Fine-tune

		Intent.	Failure	Unintent.
True Label	Intentional	43.0	19.5	37.4
	Failure	24.8	43.5	31.5
	Unintentional	21.5	16.0	62.5

(b) Video Speed + Linear

Figure 8: **Classification Confusion Matrices:** We compare the confusion matrices for (a) Kinetics supervision and (b) self-supervision. One key difference is the self-supervised model often confuses intentional action with the start of failure, suggesting there is substantial room for improving fine-grained localization of intentionality.

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.

Performance Breakdown: We manually labeled a diagnostic set of 270 videos into nine types that characterize the cause of unintentional action. Figure 10 reports performance of models broken down by video type. The most challenging cases for our models are when the unintentional action is caused by environmental factors (such as slipping on ice) or unexpected interventions (such as a bird swooping in suddenly). Moreover, performance is comparatively low when the person in the video has limited visibility, such as due to occlusions, which motivates further work in gaze estimation [47, 29], especially in video. Another challenge is due to limited knowledge, such as understanding that fire is hot. In contrast, the model has better performance at recognizing unintentional action in multi-agent scenes, likely because multi-agent interaction is more visually apparent.

6. 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.

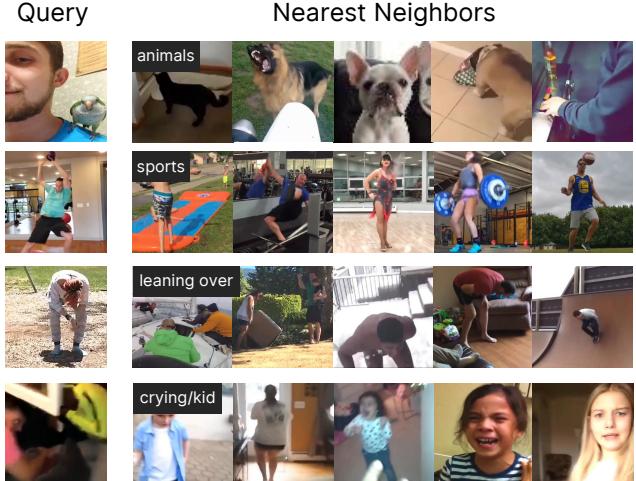


Figure 9: **Nearest Neighbors on Self-supervised Features:** We visualize some of the nearest neighbors from the feature space learned by predicting video frame rate. The nearest neighbors tend to be similar activities despite significant variation in appearance.



Figure 10: **Performance Breakdown:** We annotate a diagnostic set of videos into different categories of unintentional action in order to break down model performance and limitations. See text for discussion.

Acknowledgements: We thank Dídac Surís, Parita Pooj, Hod Lipson, and Andrew McCallum for helpful discussion. Funding was provided by DARPA MCS, NSF NRI 1925157, and an Amazon Research Gift. We thank NVidia for donating GPUs.

References

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sud-

- heendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2
- [2] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011. 2
- [3] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015. 3
- [4] Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. A survey on deep learning based approaches for action and gesture recognition in image sequences. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 476–483. IEEE, 2017. 2
- [5] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1395–1402. IEEE, 2005. 2
- [6] 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
- [7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2
- [8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3
- [9] Eugene M Caruso, Zachary C Burns, and Benjamin A Converse. Slow motion increases perceived intent. *Proceedings of the National Academy of Sciences*, 113(33):9250–9255, 2016. 4
- [10] Guangchun Cheng, Yiwen Wan, Abdullah N Saudagar, Kamesh Namuduri, and Bill P Buckles. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*, 2015. 2
- [11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2, 3
- [12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 3
- [13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 3
- [14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 3
- [15] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 3
- [16] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4991–5000, 2018. 2
- [17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The " something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 3, 2017. 2
- [18] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1, 2
- [19] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5
- [20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 5
- [21] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3
- [22] Minh Hoai and Fernando De la Torre. Max-margin early event detectors. *International Journal of Computer Vision*, 107(2):191–202, 2014. 3
- [23] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 6
- [24] Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1413–1421, 2015. 3
- [25] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. 5
- [26] Soo Min Kang and Richard P Wildes. Review of action recognition and detection methods. *arXiv preprint arXiv:1610.06906*, 2016. 2

- [27] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2
- [28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 5
- [29] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6912–6921, 2019. 8
- [30] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008. 2
- [31] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *arXiv preprint arXiv:1901.09005*, 2019. 5
- [32] Hildegarde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 2
- [33] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. *arXiv preprint arXiv:1905.00875*, 2019. 3
- [34] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005. 2
- [35] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017. 3
- [36] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 3
- [37] Xuetong Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *arXiv preprint arXiv:1909.11895*, 2019. 3
- [38] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 3
- [39] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3
- [40] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2
- [41] Phuc Xuan Nguyen, Gregory Rogez, Charless Fowlkes, and Deva Ramanan. The open world of micro-videos. *arXiv preprint arXiv:1603.09439*, 2016. 2
- [42] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 3
- [43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 5, 6, 7
- [44] Mingtao Pei, Yunde Jia, and Song-Chun Zhu. Parsing video events with goal inference and intent prediction. In *2011 International Conference on Computer Vision*, pages 487–494. IEEE, 2011. 3
- [45] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 612–619, 2014. 2
- [46] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010. 2
- [47] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015. 8
- [48] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, pages 1036–1043. IEEE, 2011. 3
- [49] Sreemananand Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1234–1241. IEEE, 2012. 2
- [50] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004. 2
- [51] Gunnar A Sigurdsson, Gülcin Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2
- [52] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 3
- [53] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2
- [54] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008. 4
- [55] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019. 3, 5
- [56] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

- [57] Vinh Tran, Yang Wang, and Minh Hoai. Back to the future: Knowledge distillation for human action anticipation. *arXiv preprint arXiv:1904.04868*, 2019. 3
- [58] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017. 3
- [59] J David Velleman. Intention, plans, and practical reason. *The Philosophical Review*, 100(2):277–284, 1991. 2
- [60] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016. 3
- [61] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. 3
- [62] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 391–408, 2018. 3
- [63] Heng Wang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin. Action recognition by dense trajectories. 2011. 2
- [64] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1(2):2, 2014. 2
- [65] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3
- [66] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015. 3
- [67] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 3
- [68] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 3, 5
- [69] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011. 2
- [70] Amanda L Woodward. Infants’ ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, 22(2):145–160, 1999. 1
- [71] Amanda L Woodward. Infants’ grasp of others’ intentions. *Current directions in psychological science*, 18(1):53–57, 2009. 1
- [72] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. Inferring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2224–2231, 2013. 3
- [73] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueteng 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, 5, 6, 7
- [74] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in neural information processing systems*, pages 91–99, 2016. 3
- [75] Gang Yu, Junsong Yuan, and Zicheng Liu. Predicting human activities using spatio-temporal structure of interest points. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1049–1052. ACM, 2012. 3
- [76] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5):1005, 2019. 2
- [77] Xin Zhao, Xue Li, Chaoyi Pang, Xiaofeng Zhu, and Quan Z Sheng. Online human gesture recognition from motion data streams. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 23–32. ACM, 2013. 3
- [78] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 3
- [79] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 4
- [80] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 3
- [81] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016. 3
- [82] Mohammadreza Zolfaghari, Özgün Çiçek, Syed Mohsin Ali, Farzaneh Mahdisoltani, Can Zhang, and Thomas Brox. Learning representations for predicting future activities. *arXiv preprint arXiv:1905.03578*, 2019. 3