Newtonian Image Understanding: Unfolding the Dynamics of Objects in Static Images –Supplementary Material–

Roozbeh Mottaghi[1] Hessam Bagherinezhad[2] Mohammad Rastegari[1] Ali Farhadi[1,2]
[1]Allen Institute for Artificial Intelligence (AI2)
[2]University of Washington

1. Dataset collection details

Here we describe how we collected natural images/videos of the VIND dataset. For each Newtonian scenario, we queried on YouTube keywords that involve those scenarios. For example, for scenario (4), which represents *rolling* dynamics, we queried variations of *billiard*, *bowling*, *soccer pass*, and *golf rolling* and downloaded top 200 videos for each query. Then, we pruned out videos that were irrelevant. For each remaining video, we segmented out at most 8 clips that contained the Newtonian scenario of interest. This procedure resulted in more than 6000 video clips that contain more than 200K frames.

To collect our static images, we used a similar set of queries on Google Images. We removed low-quality and duplicate images and ended up with 4516 images for all Newtonian scenarios. We considered one third of images (randomly sampled) as the validation set and the remaining images as our test set.

For scenario (5), which represents *stability*, we augment our dataset with frames from 8 annotated sequences of the SUN3D dataset [1], which show stable objects in office/hotel environments. We use 4 sequences for training and the other 4 for testing. In Figure 2, we show some example images for each scenario.

We provide three types of annotations for each frame/image. First, we provide bounding box annotations for the objects that are described by at least one of our Newtonian scenarios. For video clips, we choose 5 frames randomly and annotate bounding boxes in those frames. Then, we find the location of the objects in other frames by interpolation. Second, we provide viewpoint annotations. We show the annotators the game engine videos that are rendered from different viewpoints of the same Newtonian scenario as that of the image/video clip and ask them which viewpoint better represents the scenario in the image/video clip (refer to Figure 3 of the paper). Finally, we provide *state* annotations for the objects. By *state*, we mean how far the object has moved on the expected scenario (*e.g.* is the object in the beginning of the projectile? or is it at the peak point?). For each video clip, we sample 10 equally spaced frames from its corresponding game engine video (the first frame is the first frame of the game engine video and the tenth frame is the last frame of the game engine video). For video clips, we have bounding box annotations across all frames (as mentioned above). We also know the 2D location of the object in the game engine video. For annotation, we need to solve an optimization problem that finds the correspondence between the projected 2D movement of the object in the game engine video and the frames in the natural videos. To solve this problem, we use Dynamic Time Warping (DTW). DTW provides the best assignment *i.e.* it specifies which video frame corresponds to which of the 10 frames of the game engine video. Figure 1 shows an example for this process. The annotation procedure is different for images since we do not know the movement of objects in images. We show the 10 frames of the game engine video

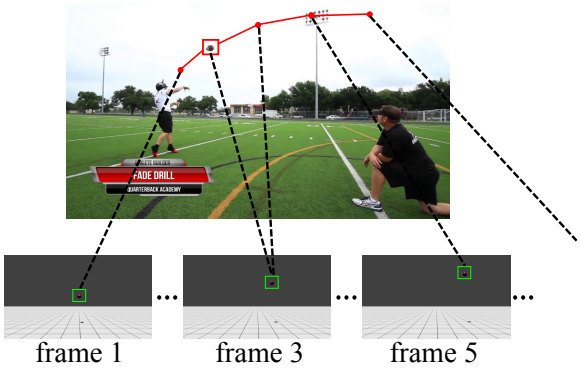
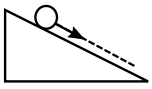
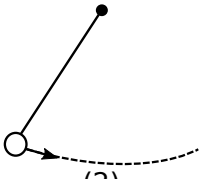


Figure 1: **State annotation**. We match the movement of the object in video (red curve) with the 2D projection of the movement of the object in the game engine video. Using Dynamic Time Warping, we infer which frame of the natural video corresponds to which frame of the game engine video.



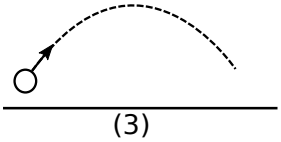
(1)



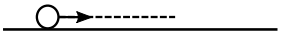
(2)



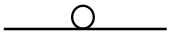
(3)



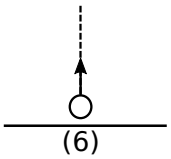
(4)



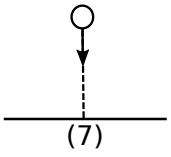
(5)



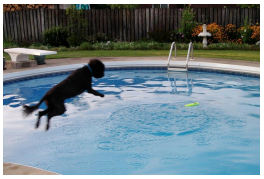
(6)

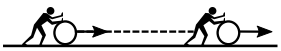


(7)





(8)




Figure 2: (left) The Newtonian scenarios. (right) Four example images that show the Newtonian scenario.

to annotators and ask them to specify which frame (out of 10 frames) best shows the state of the object. Note that we do not use this type of annotation for training N^3. It is just used for our ablation study and also to evaluate how well we can approximate the state of the object.

2. Unseen scene types (Section 6.2)

To test the generalization of our method, we removed one scene type per Newtonian scenario from our training set and evaluated our method on images that represented those scene types. The list of removed scene types for each Newtonian scenario is as follows:

- Scenario (1): Playground scene
- Scenario (2): Scenes showing swinging with a rope
- Scenario (3): Soccer scene
- Scenario (4): Bowling scene
- Scenario (6): Table tennis scene
- Scenario (7): Diving scene
- Scenario (8): Scenes including cars
- Scenario (9): Volleyball scene
- Scenario (10): Rugby scene
- Scenario (11): Tennis scene
- Scenario (12): Weightlifting scene

References

[1] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013. 1