Visual Stability Prediction and Its Application to Manipulation

Wenbin Li
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
wenbinli@mpi-inf.mpg.de

Aleš Leonardis
School of Computer Science
University of Birmingham, United Kingdom
a.leonardis@cs.bham.ac.uk

Mario Fritz
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
mfritz@mpi-inf.mpg.de

Abstract

Understanding physical phenomena is a key compe-
tence that enables humans and animals to act and inter-
act under uncertain perception in previously unseen en-
vironments. Developmental psychology has shown that
such skills are acquired by infants from observations at
a very early stage. We contrast a more traditional ap-
proach of taking a model-based route with explicit 3D
representations and physical simulation by an *end-to-
end* approach that directly predicts stability from ap-
pearance. We explore how such a skill can directly be
acquired in a data-driven way—bypassing explicit sim-
ulations at run-time. We present a learning-based ap-
proach based on simulated data that predicts stability
of towers comprised of wooden blocks under different
conditions and quantities related to the potential fall of
the towers. We evaluate the approach on synthetic data
and compared the results to human judgments on the
same stimuli. Then we extend this approach to reason
about future states of such towers that in turn enables
successful stacking.

Introduction

Scene understanding requires, among others, understanding
of relations between and among the objects. Many of these
relations are governed by the Newtonian laws and thereby
rule out unlikely or even implausible configurations for the
observer. Although objects simply obey these elementary
laws, which can very well be captured in simulators, uncer-
tainty in perception makes exploiting these relations chal-
lenging in artificial systems. In contrast, humans understand
such physical relations naturally, which enables them to ma-
nipulate and interact with objects in unseen conditions with
ease. We build on a rich set of prior experiences that allow
us to employ a type of commonsense understanding that,
most likely, does not involve symbolic representations of 3D
geometry or physics simulation engines. We rather rely on
"naïve physics" or "intuitive physics" , serving as a good
enough proxy to make us operate successfully in real world.

It has not yet been shown how to equip machines with a
similar set of physics commonsense – and thereby bypassing
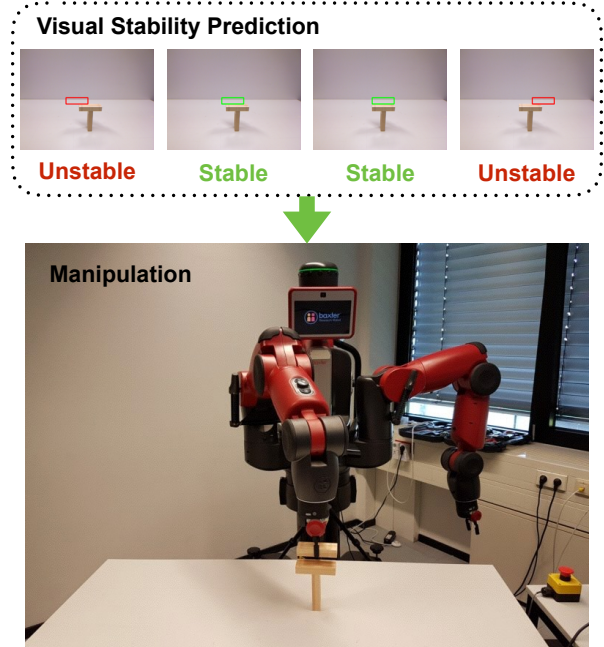a strong model representation and a physical simulation. In

Figure 1: Given a wood structure, our visual stability clas-
sifier predicts the stability for future placements, the robot
then stacks a block among the predicted stable placements.

fact, it has been argued that such an approach is unlikely due
to e.g., the complexity of the problem (Battaglia, Hamrick,
and Tenenbaum 2013). Only recently, several works have re-
vived this idea and reattempted a fully data drive approach
to capturing the essence of physical events via machine
learning methods (Mottaghi et al. 2015; Wu et al. 2015;
Fragkiadaki et al. 2015; Lerer, Gross, and Fergus 2016). In
contrast, studies in developmental psychology (Baillargeon
1994) have shown that infants acquire knowledge of physi-
cal events by observation at a very early age, for example:
support, how an object can stably hold another object; colli-
sion, how a moving object interact with another object.

In this work, we focus on support event and construct a
model for machines to predict object stability. We revisit the
setup in (Battaglia, Hamrick, and Tenenbaum 2013) and ex-
plore how to predict physical stability directly from appear-
ance cues. We approach this problem by synthesizing a large
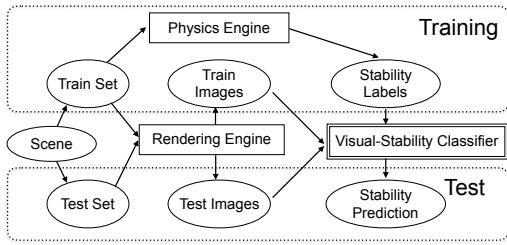
Figure 2: Overview of our approach for learning visual stability. Note physics engine is only used during training time to get the ground truth to train the deep neural network while at test time, only rendered scene images are given to the learned model to predict the physical stability of the scenes.

set of block towers under a range of conditions and then running them through a simulator (*only at training time!*) to generate stability labels. We show for the first time that aforementioned stability can be learned and predicted in a purely data driven way. Further, we successfully guide a robot to stack blocks based on the stability prediction shown in Figure 1. For more details, please refer to (Li, Fritz, and Leonardis 2016).

Visual Stability Prediction

Synthetic Data

Scene Parameters   As in Table 1, we include scenes with 4, 6, 10 and 14 blocks and vary the depth of the tower from a one layer setting as 2D to a multi-layer setting as 3D. The former only allows a single block along the image plane at all height levels while the latter does not enforce such constraint. We also include: *Uni*, the towers with uniform block size as in (Battaglia, Hamrick, and Tenenbaum 2013) and *NonUni*, towers with varying block sizes where two of the three dimensions are randomly scaled with respect to a Normal distribution. In total, there are 16 groups of 1000 scenes in each group.

Simulation   We deliberately decided against colored bricks as in (Battaglia, Hamrick, and Tenenbaum 2013) in order to challenge perception and make identifying brick outlines and configurations more challenging. The lighting is fixed across scenes and the camera is automatically adjusted so that the whole tower is centered in the rendered image. The final stability label is automatically decide by the displacement of blocks during simulation.

Stability Prediction from Still Images

Research in (Battaglia, Hamrick, and Tenenbaum 2013) suggests the combinations of the most salient features in the scenes are insufficient to capture people's judgments, yet, contemporary study reveals human's perception of visual information, especially some geometric features, like critical angle (Cholewiak, Fleming, and Singh 2013) play an important role in the process. Regardless of the actual inner mechanism for humans to parse the visual input, it seems clear that there is a mapping f involving visual input I to the stability prediction P: $f : I, * \rightarrow P$, where $*$ denotes other

possible information, i.e., the mapping can be inclusive, using it along with other aspects, like physical constraint or the mapping is exclusive, using visual cues only.

Here we directly predicts the physical stability from visual input. We use deep convolutional neural networks as they have shown great success on image classification and capable of adapting to various tasks through re-training/fine-tuning. We interchange the image classes labels with the stability labels so that the network can learn "stability salient" features by fine-tuning the pre-trained VGG Net(Simonyan and Zisserman 2014).

Intra-Group Experiment   We train and test on the scenes by varying only one type of scene parameter at a time: *Number of Blocks* , a consistent drop in performance can be observed with more blocks in the scene; *Block Size* , the performance generally decreases for varying block size over uniform block size; *Stacking Depth* , for simple scenes, prediction accuracy increases when moving from 2D stacking to 3D but vise versa for the complex scene.

Cross-Group Experiment   To see how the learned model transfers across scenes with different complexity, we divide the scene into: a *simple scene* group for scenes with 4 and 6 blocks and a *complex scene* for scenes with 10 and 14 blocks. We investigate: train on simple scenes and predict on complex scenes and vise versa. For the former direction, it gets 69.9%, which is significantly better than random guess at 50%, and for the latter we observe significant performance boost which can be explained by the richer feature learned from the complex scenes with better generalization.

Generalization Experiment   Similar to human's prediction in the task, we use training images from all different scene groups and test on any groups. While the performance exhibits similar trend to the intra-group, namely increasing recognition rate for simpler settings and decreasing rate for more complex settings, there is a consistent improvement over the intra-group for individual groups. Together with the result in the cross-group, it suggests a strong generalization capability of the model.

Human Subject Test

8 human subjects are recruited to predict stability for give scene images. Due to large number of test data, we sample images from different scene groups for human subject test. Each subject is presented with a set of captured images from the test split. Each set includes 96 images where images cover all 16 scene groups with 6 scene instances per group. For each scene image, subject is required to rate the stability from *Definitely Unstable* to *Definitely Stable*.

For simple scenes, human can reach close to perfect performance while for complex scenes, the performance drops significantly to around 60%. The image-based model outperforms human in most cases: while showing similar trends in performance with respect to different scene parameters, it is less affected by a more difficult scene parameter setting.
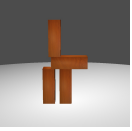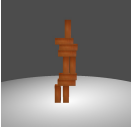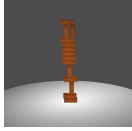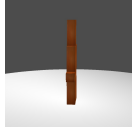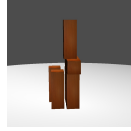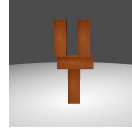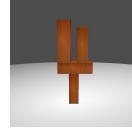
Block Numbers				Stacking Depth		Block Size	
(a) 4B	(b) 6B	(c) 10B	(d) 14B	(e) 2D	(f) 3D	(g) Uni	(h) N-Uni

Table 1: Scene parameters in our rendered scenes: number of blocks, stacking depth and block size.
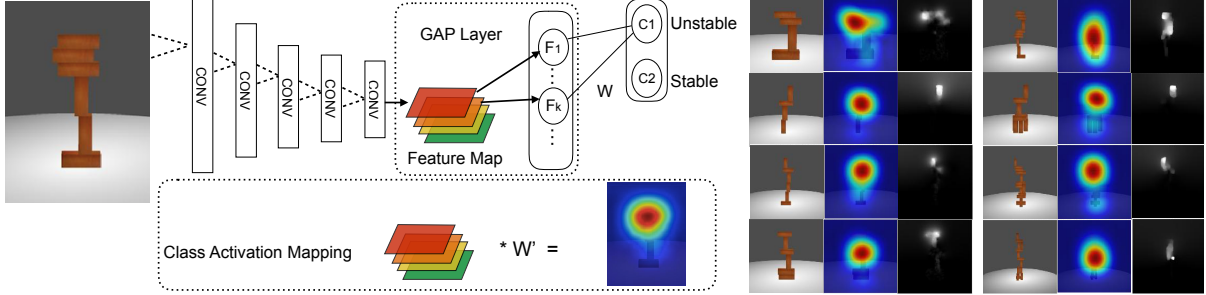

Figure 3: Model interpretation: (left) pipeline, (right) example visualizations.

Num.of Blks	Uni.		NonUni.
	2D	3D	2D
4B	93.0	99.2	93.2
6B	88.8	91.6	88.0
10B	76.4	68.4	69.8
14B	71.2	57.0	74.8

Table 2: Results for intra-group experiments.

Num.of Blks	Uni.		NonUni.	
	2D	3D	2D	3D
4B	93.2	99.0	95.4	99.8
6B	89.0	94.8	87.8	93.0
10B	83.4	76.0	77.2	74.8
14B	82.4	67.2	78.4	66.2

Table 3: Results for generalization experiments.

Model Interpretation

We apply the technique from (Zhou et al. 2016) to visualize the learned discriminative image regions from CNN for individual category. We investigate discriminative regions for unstable predictions to see if the model can spot the weakness in the structure. We compute the optical flow magnitude between the frame before the physics engine is enabled and afterwards as a coarse ground truth for the structural weakness, assuming the collapse motion starts from such weakness. Though not universal among the unstable cases, we do find significant positive cases showing high correlation between the activation regions in CAM for unstable output and the regions where the collapse motion begins. The pipeline and example visualizations are shown in Figure 3.

Num.of Blks	Uni.		NonUni.	
	2D	3D	2D	3D
4B	79.1/**91.7**	93.8/**100.0**	72.9/**93.8**	92.7/**100.0**
6B	78.1/**91.7**	83.3/**93.8**	71.9/**87.5**	89.6/**93.8**
10B	67.7/**87.5**	72.9/72.9	66.7/**72.9**	71.9/68.8
14B	71.9/**79.2**	68.8/66.7	71.9/**81.3**	59.3/**60.4**

Table 4: Results for human subject test: a/b denotes accuracies for human and image-based model respectively.

From Visual Stability Test to Manipulation

We set up a testbed where a Baxter robot stacks one Kapla block on a given block structure over 6 scenes without breaking their stability. We enforce some constraints for simplicity: (1) the structure is restricted to be single layer; (2) the block to be put is limited to {vertical, horizontal} and assumed to be held in robot's hand before placement; (3) the block has to be placed on the top-most horizontal surface (stacking surface) of the structure; (4) the depth of the structure is calibrated so only the horizontal and vertical displacements to the stacking surface are to be decided.

Prediction on Real World Data

Considering the significant difference between the synthesized and real world data, we directly apply the model trained on the RGB images to predict stability on the real data in a pilot study, but only got results close to random guessing. Hence we decided to train the visual-stability model on the binary-valued foreground mask on the synthesized data after observing comparable results and deal with the masks at test time for the real scenes. For each scene, an captured image is first converted to a foreground mask via background subtraction. The top-most horizontal boundary is detected as the stacking surface and then used to generate
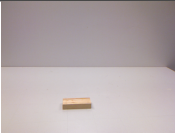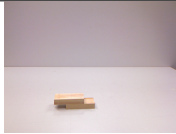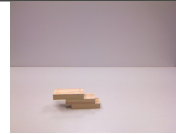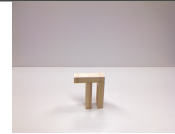
Id.	1		2		3		4		5		6	
Scene												
Pred.(%)	66.7	100.0	66.7	60.0	88.9	100.0	77.8	80.0	100.0	40.0	66.7	60.0
Mani.	4/5	5/5	2/3	3/3	2/3	1/1	2/2	2/3	3/3	1/4	0/3	0/1
Placement	H	V	H	V	H	V	H	V	H	V	H	V

Table 5: Results for real world test. "Pred." is the prediction accuracy. "Mani." is the manipulation success rate with counts for successful placements/all possible stable placements for each scene. "H/V" refer to horizontal/vertical placement.
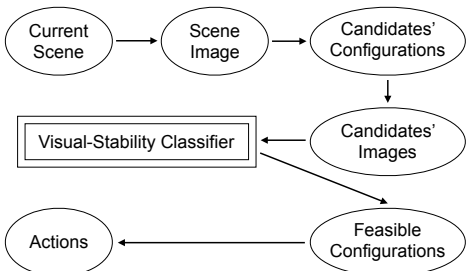

Figure 4: An overview of our manipulation system.

candidate placements: the surface is divided evenly into 9 horizontal candidates and 5 vertical candidates, resulting in overall 84 candidates. Afterwards, the candidates are put to the visual-stability model for stability prediction. Each candidate's actual stability is manually tested and recorded as ground truth. The model trained with synthetic data is able to predict with overall accuracy of 78.6% across different candidates in real world.

Manipulation Test

When the model predicts a given candidate placement as stable, the robot will execute a routine to place the block (Figure 1) with 3 attempts. We count the execution as a success if any of the attempt works. The manipulation success rate is defined as the ratio between the number of successful placements made by the robot and all ground truth stable placements. The manipulation performance (Table5) is generally good across most of the scenes except for the 6-th scene where the classifier predicts all candidates as unstable hence no attempts have been made by the robot.

Discussion

As a future work, our current system can be extended to integrate multimodal sensory input to handle more challenging real world scenarios. For example, when including different materials for objects, while the material information can be again inferred through visual input, a tactile sensor can be used to acquire relevant information from a second channel to further improve the prediction accuracy.

Summary

We propose a model to predict physical stability directly from visual input bypassing explicit 3D representations and physical simulation. The model is evaluated on towers with great variations. To further understand the results, we conduct a human subject study on a subset of our synthetic data, showing our model achieves comparable result to humans. We also investigate the discriminative image regions found by the model and spot correlation between such regions and initial collapse area in the structure. Finally, We apply our approach to a block stacking setting, showing it can guide a robot for placing new blocks by predicting the stability of future states. By integrating multimodal input, our system can potentially handle more complex real-world scenarios.

References

Baillargeon, R. 1994. How do infants learn about the physical world? *Current Directions in Psychological Science*.

Battaglia, P. W.; Hamrick, J. B.; and Tenenbaum, J. B. 2013. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*.

Cholewiak, S. A.; Fleming, R. W.; and Singh, M. 2013. Visual perception of the physical stability of asymmetric three-dimensional objects. *Journal of vision*.

Fragkiadaki, K.; Agrawal, P.; Levine, S.; and Malik, J. 2015. Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*.

Lerer, A.; Gross, S.; and Fergus, R. 2016. Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*.

Li, W.; Fritz, M.; and Leonardis, A. 2016. Visual stability prediction and its application to manipulation. *arXiv preprint arXiv:1609.04861*.

Mottaghi, R.; Bagherinezhad, H.; Rastegari, M.; and Farhadi, A. 2015. Newtonian image understanding: Unfolding the dynamics of objects in static images. *arXiv preprint arXiv:1511.04048*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wu, J.; Yildirim, I.; Lim, J. J.; Freeman, B.; and Tenenbaum, J. 2015. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *NIPS*.

Zhou, B.; Khosla, A.; A., L.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. *CVPR*.