Long Term Boundary Extrapolation for Deterministic Motion

Apratim Bhattacharyya, Mateusz Malinowski, Mario Fritz
Max Planck Institute for Informatics
Saarbrücken, Germany
{abhattac, mmalinow, mfritz}@mpi-inf.mpg.de

Abstract

We propose models for long-term prediction of boundaries that are learned from observations without making any strong model assumptions on objects or scene. We evaluate our approach in a billiard scenario that is only governed by the Newtonian laws of physics and is therefore fully deterministic. We argue that any model that succeeds in this task must have derived some notion of "intuitive physics". Our Recursive Convolutional Multi-Scale architecture turns out to be most effective.

1 Introduction

The ability of artificial neural networks to develop an intuitive understanding of physics from raw visual input has only been recently explored in [1, 2, 3]. Fragkiadaki et al. [1] has developed a model which predicts futures states of balls moving on a billiard table. Whereas, Lerer et al. [2] and Li et al. [3] have developed a model which predicts the stability and future states of towers made out of blocks. These models typically are parametric of some sort or only predict a qualitative outcome of the scene. Moreover, both Fragkiadaki et al. [1] and Lerer et al. [2] have an "object notion", meaning that the model knows a priori the location or type of the objects it is supposed to model.

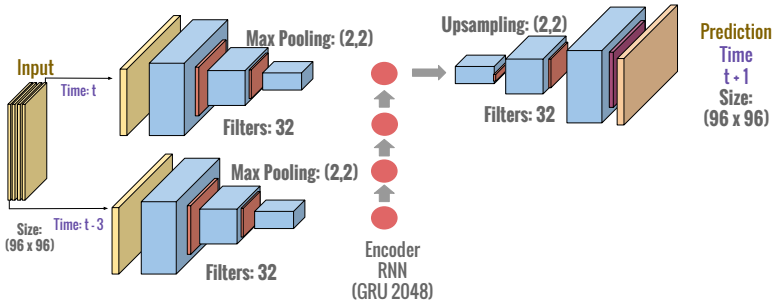
Concurrently, full future frame predication has been studied that is agnostic to the underlying cause of the change depicted in the sequence [4]. In contrast to the physical models, only very short range predictions have been shown and and blurry artifacts occur in the predicted future frames. Recently a lot of progress has been made in the field of video segmentation. Segmented videos not only discard many details of natural videos like color, texture etc which are hard for a model to extrapolate into the future, but also capture the important objects as boundaries. The boundaries between segments gives rise to boundary images.

Our main contributions are the first models to extrapolate image boundaries and to explore the performance of these models under deterministic motion. We show that our models develop an intuitive understanding of physics from raw visual input without any strong parametric model of the motion or "object notion" and is capable of making long term predictions in such deterministic settings. For an extended exposition of our methods and experiments, please refer to [5].

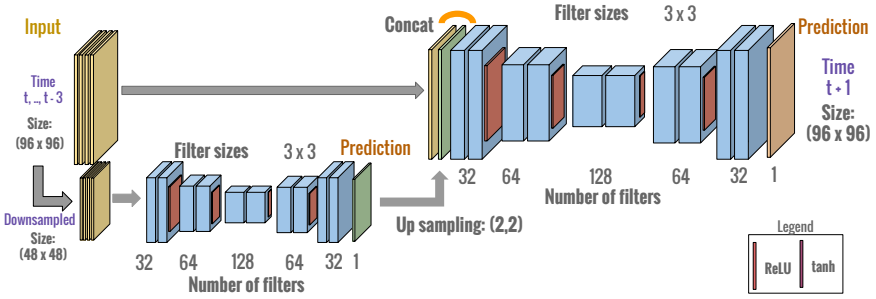
2 Models

We base our models on the recent success of Deep Learning. We approach long-term extrapolation by recursive schemes. However, this means that errors are potentially propagated and accumulated over time. In order to mitigate such effects, we need very accurate models that consolidate information. By analyzing prior work on frame prediction, we identify several key properties models have to fulfill,

Large Spatio-Temporal Context. The models should have a wide receptive field to preserve long range spatial and temporal dependencies and learn about interaction with other boundaries.



(a) Convolutional RNN - encoder decoder architecture



(b) Recursive Convolutional Multi-scale architecture (only 2 out 4 scales illustrated)

Figure 1: Our model architectures.

Preserve Resolution and Prevent Blurred Output. The models must maintain resolution in order to derive a high fidelity output boundary map. Excessive pooling or tight bottlenecks with fully connected layers have shown to induce image degradations for image synthesis tasks [4].

Globally Consistent Predictions from Local Models. In order to generalize across diverse input sequences while maintaining a tractable number of parameters, the models should observe and extrapolate on patches rather than the complete input image. Global consistency can still be achieved by facilitating "communication" between neighboring patches. We achieve this through the inclusion of a context i.e. the neighboring eight patches. This constitutes a read-write architecture, where the past frames serve as a kind of shared memory.

2.1 Model architectures

We propose the following model architectures for boundary extrapolation based on the key properties described above.

Convolutional-RNN (C-RNN) Encoder-Decoder Architecture. This model architecture is based on Srivastava et al. [6]. This architecture consists of an convolutional-encoder GRU unit which reads in the input frame sequence one time step at a time and produces a single vector as output. This vector is read by a convolutional-decoder unit to produce the final output frame (see Figure 1a). This creates a wide receptive field and fulfills the first key property. However, as the encoder produces a single vector as output, this acts a bottleneck layer.

Convolutional Multi-Scale (CMS) Architecture. Multi-scale architectures akin to a Laplacian pyramid has been used successfully for generating natural images [7] and predicting future natural frames [8]. We achieve long-term extrapolation by proposing a recursive scheme that samples from the output of a one frame CNN extrapolator, appending the frame and applying the extrapolator again. This results in a recursive convolutional extrapolator for arbitrarily long sequences. In detail, we use four scales, each downsampled by a factor of two from the one before. Each each level captures image structure present at a particular scale. This creates a wide receptive field for the output layer neurons. We use a fully convolutional network at each scale with moderate pooling (no tight bottleneck layer).

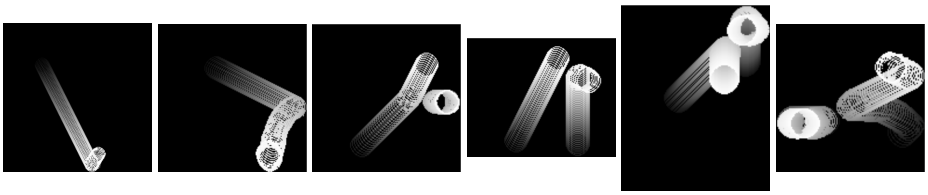


Figure 2: Trails produced by super-imposing extrapolated boundaries.

The output at a certain scale is upsampled and used as input at the next larger scale as a candidate future frame (see Figure 1b). This architecture fulfills both key properties.

Details Common to Both Architectures. As input we consider nine patches (eight neighbours as context) each of size 32x32 pixels. The models observe the patches at the current and previous three time-steps. The network are trained to locally predict the 3x3 patch neighborhood at the next time step, but at test time we only use the middle patch for prediction. We use a *tanh* non-linearity in order to produce binarized output in the range [0,1] and optimize for the L_2 loss during training.

3 Experiments

We use boundary precision recall (BPR) [9] as the evaluation metric. This metric can be defined for a set P of predicted boundary images and G of corresponding ground truth boundary images as:

$$P = \frac{\sum_{B_p \in P, B_g \in G} \mid B_p \cap B_g \mid}{\sum_{B_p \in P} \mid B_p \mid} \quad R = \frac{\sum_{B_p \in P, B_g \in G} \mid B_p \cap B_g \mid}{\sum_{B_g \in G} \mid B_g \mid} \quad F = \frac{2PR}{P+R}$$

where P is boundary precision, R is boundary recall and F is the combined F-measure.

3.1 Datasets

We sample synthetic sequences from worlds which consists of balls moving on a frictionless surface with a boundary, akin to a billiard table. We used the pygame module of python to create such worlds and sample binary boundary images from them. The **table sizes** were randomly sampled from {96,128,160,192,256} pixels.The **ball velocities** were randomly sampled from [{-3,..,3},{-3,..,3}] pixels. The **ball sizes** were constant, with a radius of 13 pixels.The **initial positions** were uniformly distributed over the table surface. For evaluation, we report the best F-measure obtained by varying the output threshold parameter.

Single ball worlds. We train the models by sampling 500 sequences from worlds having one ball. To keep our training set as diverse as possible we sample short sequences with collisions. We then test the models on 30 independent test sequences and report the results in Table 1. We include the last input image (image at the current time-step t) as a baseline. We include a Convolutional Multi-scale model without a context (CMS-WC). This model can only observe the central 32x32 patch as input. We also include a "blind" Convolutional Multi-scale model with context (CMC-BL), which cannot see the table borders. To beat this setup, our models need to learn the physics of ball-wall collisions.

Table 1: Evaluation of models on single ball billiard table worlds

Step	Last Input	C-RNN	CMC-BL	CMS-WC	CMS
$t+1$	0.141	0.013	0.957	0.282	**0.987**
$t+5$	0.038	0.006	0.841	0.035	**0.900**
$t+20$	0.002	0.005	0.347	0.009	**0.632**

The Convolutional Multi-Scale architecture (with a context) performs the best with accurate predictions 20 time-steps into the future. The Conv-RNN architecture (with its bottleneck layer) is unable to

learn the physics of the world and produces very blurred output. Without a context, the Convolutional Multi-scale architecture suffers heavily especially at larger time-steps.

Two and three ball worlds. Worlds with more than one ball also involve ball-ball collisions, which make the physics of such worlds much more complex. To test the models on such worlds we sample 100 and 50 training sequences with two and three balls respectively with a maximum length of 200 frames. We use a curriculum learning approach, that is, we initialize the models with the weights learned on single and two ball worlds respectively. We test the models on 30 independent sequences containing two and three balls respectively. We report the results in Table 2. We also include Convolutional Multi-Scale models (with a context) trained on single ball worlds (RCMS-1B) and two ball worlds (RCMS-2B) in the two and three ball world case respectively. To beat these models learning the physics of ball-ball collisions is necessary. Again, we see accurate extrapolation by the Convolutional Multi-Scale model (with a context) even at 20 time-steps in the future. The Conv-RNN architecture performs just as badly as in one ball worlds.

Table 2: Evaluation of models on complex billiard table worlds

	Evaluation on two ball worlds				Evaluation on three ball worlds			
Step	**Last Input**	**C-RNN**	**CMS-1B**	**CMS**	**Last Input**	**C-RNN**	**CMS-2B**	**CMS**
$t+1$	0.246	0.013	0.966	**0.969**	0.246	0.023	0.967	**0.968**
$t+5$	0.114	0.008	0.848	**0.896**	0.118	0.012	0.890	**0.892**
$t+20$	0.101	0.007	0.612	**0.681**	0.090	0.011	0.664	**0.700**

Extrapolation over very long time scales. Although we evaluate only 20 time-steps into the future, the Convolutional Multi-Scale model (with a context) is stable over longer time-horizons. In Figure 2 we extrapolate 100 frames into the future. We superimpose the frames to produce the trails in Figure 2. However, we noticed that sometimes the balls reverse direction mid table and the ball(s) get deformed or disappear altogether.

4 Conclusion

We demonstrate long-term boundary extrapolation that yields accurate and sharp results. Our proposed Convolutional Multi-Scale architecture performs best in the quantitative evaluation. Moreover, the accurate results on varied scenarios involving billiard balls show that this models can develop an intuitive notion of physics. The key ingredients that made long term prediction possible are, i) Accurate prediction at each time-step with a fully convolutional setup without any bottleneck layers and wide receptive field. ii) Overlapping receptive fields which allow for the sharing of information thus leading to global consistency.

References

[1] Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. *arXiv:1511.07404*, 2015.

[2] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *ICML*, 2016.

[3] Wenbin Li, Ales Leonardis, and Mario Fritz. Visual stability prediction and its application to manipulation. *arXiv:1609.04861*, 2016.

[4] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv:1412.6604*, 2014.

[5] A. Bhattacharyya, M. Malinowski, B. Schiele, and M. Fritz. Long-Term Image Boundary Extrapolation. *ArXiv e-prints*, November 2016.

[6] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015.

[7] Emily L Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.

[8] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv:1511.05440*, 2015.

[9] Fabio Galasso, Naveen Nagaraja, Tatiana Cardenas, Thomas Brox, and Bernt Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *CVPR*, 2013.