28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, 2020.

Object-centered Fourier Motion Estimation and Segment-Transformation Prediction

Moritz Wolter[1], Angela Yao[2], and Sven Behnke[1]

1- University of Bonn, Computer Science Institute,
Endenicher Allee 19A, 53115 Bonn, Germany

2- National University of Singapore, School of Computing,
13 Computing Drive, Singapore 117417

Abstract. The ability to anticipate the future is essential for action planning in autonomous systems. To this end, learning video prediction methods have been developed, but current systems often produce blurred predictions. We address this issue by introducing an object-centered movement estimation, frame prediction, and correction framework using frequency-domain approaches. We transform single objects based on estimated translation and rotation speeds which we correct using a learned encoding of the past. This results in clear predictions with few parameters. Experimental evaluation shows that our approach is accurate and efficient.

1 Introduction

In this paper, we investigate the problem of video frame prediction. Previous approaches have used recurrent neural networks [1] to directly synthesize the predicted image [2, 3]. These were trained using a mean squared error loss function in tandem with learned image synthesis. This approach introduces blur into the prediction as the network hedges its bets to minimize the error and smears predictions, to cover as many eventualities as possible.

To address this issue, we focus on single objects that we assume have been fully pre-segmented. We model their movement separately, decoupling transformation learning from image synthesis. Our main contributions are:

- Given a moving pre-segmented object we estimate its centroid, as well as translational and rotational velocity based on phase correlation.
- We model the estimated velocities and their changes, e.g. at image boundaries, using neural networks and
- demonstrate that the frequency-based three pass image transformation method is able to produce sharp predictions for multiple time steps by transforming the input according to the estimated parameters.

Source code for the parameter estimation framework, the correction network as well as the phase-shift based image transformation is available online. [1].

[1] at `https://github.com/v0lta/Fourier-Motion-Estimation-and-Segment-Transformation`

2 Related Work

Early works on future frame prediction [1] rely on general-purpose recurrent architectures such as gated recurrent units (GRU) to encode an input video sequence and use a decoder with an identical structure to predict future frames. The moving MNIST test-problem consisting of handwritten digits moving in a box was introduced in [1]. Later convolutional structures and flow models were integrated into recurrent cells [2, 3]. All of the aforementioned methods employ a mean squared error loss function and use learned weights to synthesize the prediction. Generative adversarial network (GAN) based formulations include [4], but these are computationally expensive and hard to train. As a step towards a more efficient approach, we leverage image registration methods to estimate motion and rotation parameters [5, 6, 7]. Work similar to ours couples CNNs and the log polar transform [8] or estimates image translation by phase correlation [9]. Compared to [9], we additionally estimate and predict *image rotation* using a log-polar transformation.

3 Methods for Motion Estimation

Due to its robustness, we estimate displacement and rotation of images by computing the normalized cross-correlation in the frequency domain [5]. Based on the two-dimensional discrete Fourier transformations of the current image $F_1 = \mathcal{F}(I_1)$ and that of its predecessor $F_2 = \mathcal{F}(I_2)$, we compute

$$C = \mathcal{F}^{-1}\Big(\frac{F_1 \odot \overline{F_2}}{\|F_1 \odot \overline{F_2}\|}\Big), \tag{1}$$

using the Hadamard product $\odot$. Afterwards, we find the displacement $\triangle\hat{x}$ and $\triangle\hat{y}$ by locating the correlation peaks in C. We employ the same strategy on log-polar transformed images to estimate the rotation velocity $\triangle\hat{\theta}$ [7]. We further apply high-pass filtering of the log-polar transformation [5] to increase the accuracy. As we assume to be working with a pre-segmented object, we compute the centroid c_x, c_y by multiplying a normalized image $I_n = I/\sum I$ with the coordinate grids X,Y:

$$c_x = \sum I_n \odot X \text{ and } c_y = \sum I_n \odot Y. \tag{2}$$

4 Neural Network Parameter Correction

We choose a machine learning approach to correct the motion estimates, with a residual formulation modelling velocity. This leads to a predictor-corrector setting, where the learned model produces a correction based on current and previous estimates. More formally, we evaluate

$$(\triangle x, \triangle y, \triangle\theta)^T = (\triangle\hat{x}, \triangle\hat{y}, \triangle\hat{\theta})^T + \text{net}(\hat{c}_x, \hat{c}_y, \triangle\hat{x}, \triangle\hat{y}, \triangle\hat{\theta}, \mathbf{s})^T, \tag{3}$$

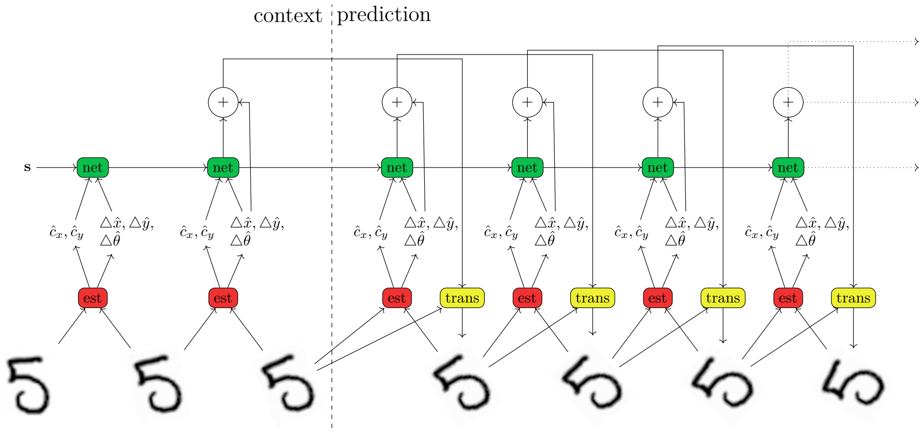

Fig. 1: Overview of our estimation, correction and transformation framework. The estimator (est) finds transformation parameters between the last and current frame based on phase correlation and computes the object centroid. The parameters are corrected by the network (net) based on its encoding of the history s, by computing a residual which is added to the current estimate. Finally, the transformer (trans) transforms the last image using the phase-shift property of the Fourier transform to create the prediction.

where the network computes the correction applied to the parameter estimation based on the object centroid, velocity estimations and its own internal state $\mathbf{s}$. Hats indicate estimates. The sum of network corrections and estimates are the transformation parameters which we use to transform the current image into the prediction. The object's centroid serves as prior knowledge to guide the motion estimation. Figure 1 illustrates our proposed approach. We pick a GRU structure in Equation 3 and update $\mathbf{s}$ accordingly.

5 Fourier Domain Image Transformation

We employ the three pass frequency domain method [10][11] to transform the current image based on the estimated parameters. It relies on the Fourier shift theorem for both translation and rotation. Given the desired translation $\triangle x$, we compute one-dimensional Fourier transforms and shift the phase using

$$I_t = \mathcal{F}^{-1}(\mathcal{F}(I)\exp(-i2\pi\triangle x f)) \tag{4}$$

to translate the image in x-direction. f denotes the frequencies and i the imaginary unit. For a translation of $\triangle y$ in y, we apply the same formulation but use the transposed image I^t instead of I. For a rotation by angle $\theta \in [-0.25\pi, 0.25\pi]$, we compute the shear parameters $a = \tan(\theta/2)$ and $b = -\sin(\theta)$ [11]. We modify

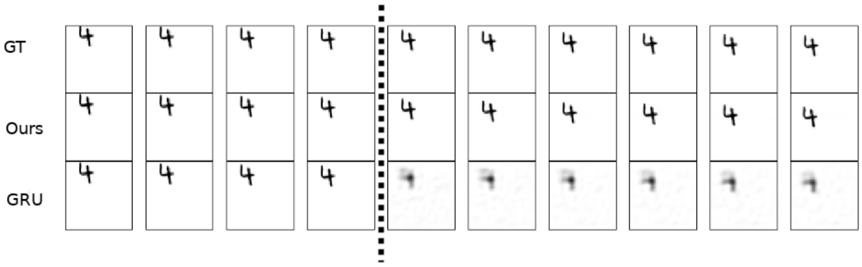

Fig. 2: Moving MNIST translation prediction. Ground truth (top), our estimation correction cell prediction (middle) and a standard GRU of size 512 (bottom) are shown. Predictions are made using 4 context frames. We observe that our approach produces predictions which remain sharp, while the much larger GRU cell's predictions are blurry.

the phase in order to obtain a shear effect in x-direction using,

$$I_{sa} = \mathcal{F}^{-1}(\mathcal{F}(I)\exp{(-i2\pi afy)}), \tag{5}$$

with the same notation as in Equation 4 with y being the y-coordinate. In the second pass we shear in y-direction by transposing I and using b instead of a,

$$I_{sb} = \mathcal{F}^{-1}(\mathcal{F}(I^T)\exp{(-i2\pi bfy)})^T, \tag{6}$$

where we replace the shear parameters a and b. The transpose shifts x and y coordinates, therefore y appears. Finally, the third pass is a repetition of the first so that we obtain a full rotation.

6 Video Frame Prediction

We evaluate our approach using the popular moving MNIST data set [1] and its rotating cousin [3]. We normalize inputs to be within $[0, 1]$ and choose a GRU for the correction-net in Equation 3. The state size is set to 50, the learning rate to 0.0005 and the batch size to 550. We stop training after 5000 iterations. In addition, we use the same parameters to train an off-the-shelf gated recurrent unit with a state size of 512 as a baseline.

6.1 Translation

In the classic MNIST translation setting [1], digits move with a random velocity on a 64 by 64 pixel canvas and bounce off the walls. An important limitation of the velocity estimation described in Equation 1 is it's restriction to pixel-level accuracy. When used in recurrent operation, this restriction can lead to instabilities: Systematically underestimating movement may cause it to halt eventually. Systematic overestimation can lead to acceleration over time. The

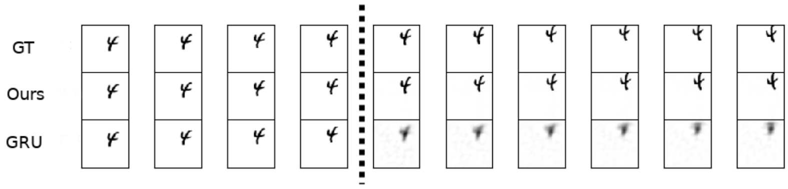

Fig. 3: Rotating moving MNIST prediction, ground truth (top), our estimation correction transformation cell output (middle) and standard GRU (bottom).

Table 1: Evaluation using 550 prediction and ground truth sequences. The mean and standard deviation of our predictions are very close to the ground truth. Our approach performs slightly worse in terms of mean squared error with significantly fewer parameters and remains sharp.

	mean	std	gt-mse	#weights
Ground truth	0.025	0.137	-	-
Ours	**0.026**	**0.135**	0.015	**32k**
GRU	0.036	0.103	**0.009**	9182k

high level GRU corrects these errors and handles wall bounces. The state size of this cell can be comparatively small, because it integrates velocity and centroid information over time instead of entire images. We compare our approach to a conventional cell without estimation and image transformation capabilities. This recurrent cell directly synthesizes its prediction of the upcoming frame based on its internal state. While this low-level approach is more flexible, it suffers from blurred prediction as well as miss-classification. Results are shown in Figure 2. We observe that our results are very hard to distinguish from the ground truth. Even though the vanilla GRU state is ten times as large and its architecture is more flexible, the result is blurry and can suffer from miss-classification. Our higher-level approach prevents both.

6.2 Rotation and Translation

In Equations 5 and 6, we introduced our Fourier shift approach to rotation. We have already shown in Figure 2 that gradients can be back-propagated through our frequency-domain image translation operation. Figure 3 demonstrates that this also works for our multi-step rotation by shearing procedure. We can stack multiple phase modification transforms within recurrent cells. The Fourier transform is a unitary operation and our phase modification matrices do not modify the magnitude and therefore do not change the scaling. This enables us to run a stable process with multiple transformations per time step in a recurrent manner. In Table 1, we compare mean standard deviation and mean squared error of the ground truth as well as our and the GRU-baseline predictions. We observe that our approach does not significantly alter the mean and standard deviation. Using the standard GRU leads to slightly better performance in terms of mean squared error, but its predictions are not naturally distributed, which significantly lifts the mean and reduces the standard deviation.

7 Conclusion

In this paper, we studied a new approach to avoid the smearing effect which arises when learned image synthesis and mean squared error functions are combined. By modelling object movement at a higher level, we prevent our system from spreading out its predictions. Our approach could be an effective potential replacement for expensive GAN-based approaches for transformation scenarios, which are used to achieve the same goal. Object segmentation methods could be combined with our approach in the future.

Acknowledgement: We thank DFG project YA 447/2-1 (Research Unit FOR 2535) and the Singapore Ministry of Education Academic Research Fund Tier 1 for funding. Moritz would like to thank Hafez Farazi for coding and Michael Weinmann for writing support.

References

[1] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning (ICML)*, 2015.

[2] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[3] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[4] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[5] B Srinivasa Reddy and Biswanath N Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996.

[6] Hongjie Xie, Nigel Hicks, G Randy Keller, Haitao Huang, and Vladik Kreinovich. An IDL/ENVI implementation of the FFT-based algorithm for automatic image registration. *Computers & Geosciences*, 29(8):1045–1055, 2003.

[7] Rittavee Matungka, Yuan F Zheng, and Robert L Ewing. Image registration using adaptive polar transform. *IEEE Transactions on Image Processing*, 18(10):2340–2354, 2009.

[8] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. In *International Conference on Learning Representations*, 2018.

[9] Hafez Farazi and Sven Behnke. Frequency domain transformer networks for video prediction. In *27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2019.

[10] Erwin Pang. *Parameter estimation and efficient implementation of affine transforms for digital images.* PhD thesis, National Library of Canada, 1999.

[11] Kieran G Larkin, Michael A Oldfield, and Hanno Klemm. Fast Fourier method for the accurate rotation of sampled images. *Optics Communications*, 139(1-3):99–106, 1997.