Predictive-Corrective Networks for Action Detection

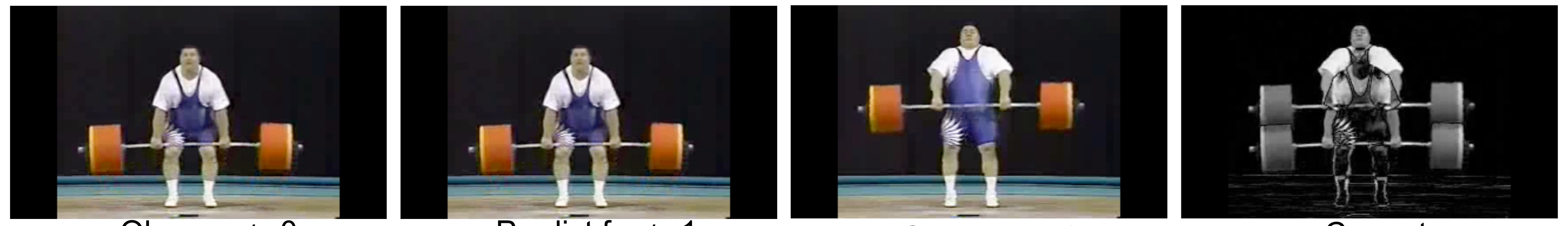
Achal Dave, Olga Russakovsky, Deva Ramanan

Carnegie Mellon University

CVPR 2017 July 21-26 HONOLULU

Introduction

- Deep architectures for video analysis largely based off of those for static images (e.g. two-stream networks)
- The **human vision system** relies on continuously **predicting** the future and then **correcting** for the unexpected
- Classic theory for **linear dynamical systems** provides a principled approach for incorporating this intuition



Observe t=0 | Predict for t=1 | Observe t=1 | Correct

Related Methods

- **Two-Stream** [Simonyan 2014] incorporates motion cues with optical flow. Our method models motion efficiently through "corrections"
- Our model is a **recurrent network** that ameliorates the issue of correlated data, and maintains a spatial memory
- **Clockwork RNN** [Koutnik 2014] maintains memory states that evolve at fixed rates; our model dynamically updates memory
- **ResNets** [He 2015] learn efficiently by focusing on "residuals" at each layer. Our model focuses on "residuals" at each time step

Contributions

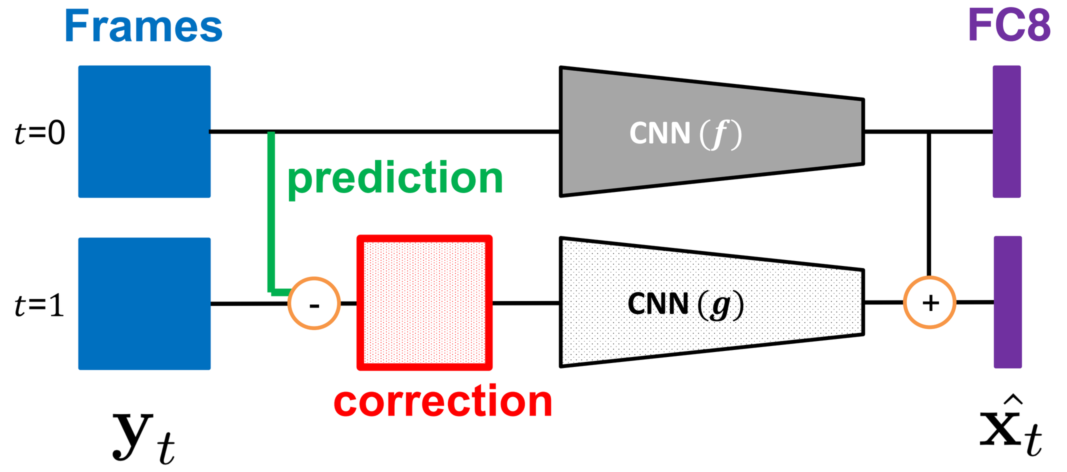
- Lightweight, interpretable model for incorporating temporal cues
- Competitive with two-stream [Simonyan 2014] networks without needing to compute optical flow

Method

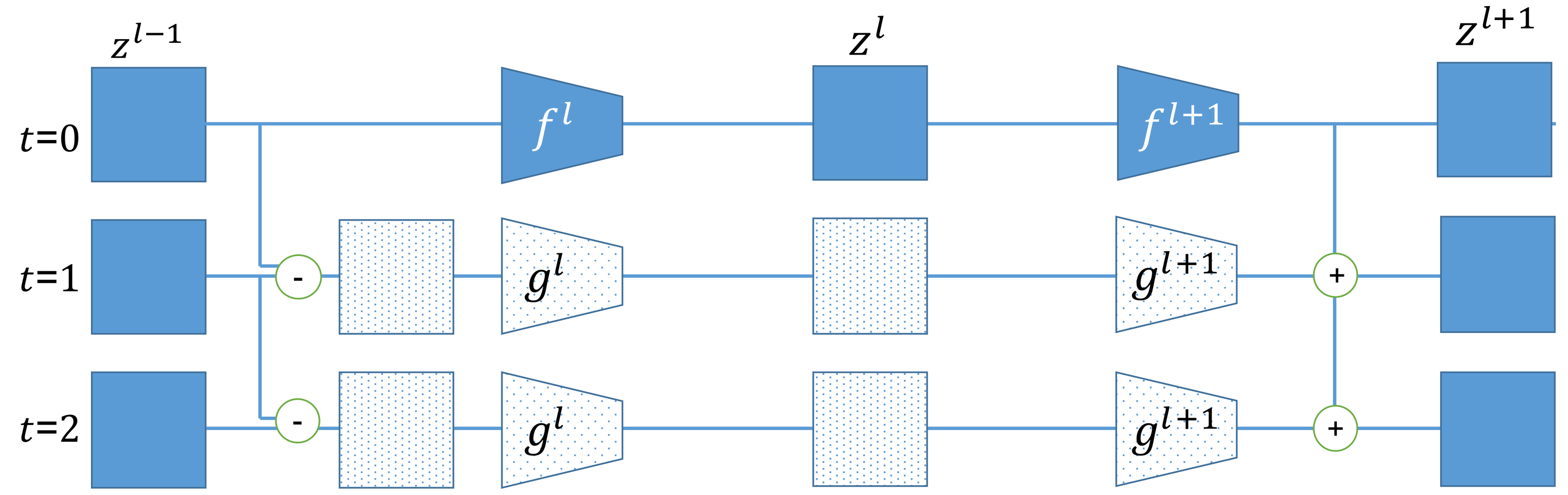
- Linear dynamical model inspired by Kalman Filters::
- Improve estimate using:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + noise \quad \text{Latent state}$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + noise \quad \text{Observation}$$
$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t-1} + \underbrace{g(\mathbf{y}_t - \underbrace{\hat{\mathbf{y}}_t}_{\text{Prediction}})}_{\text{Correction}}$$

- Predictive-corrective block applies this motivation to deep networks:



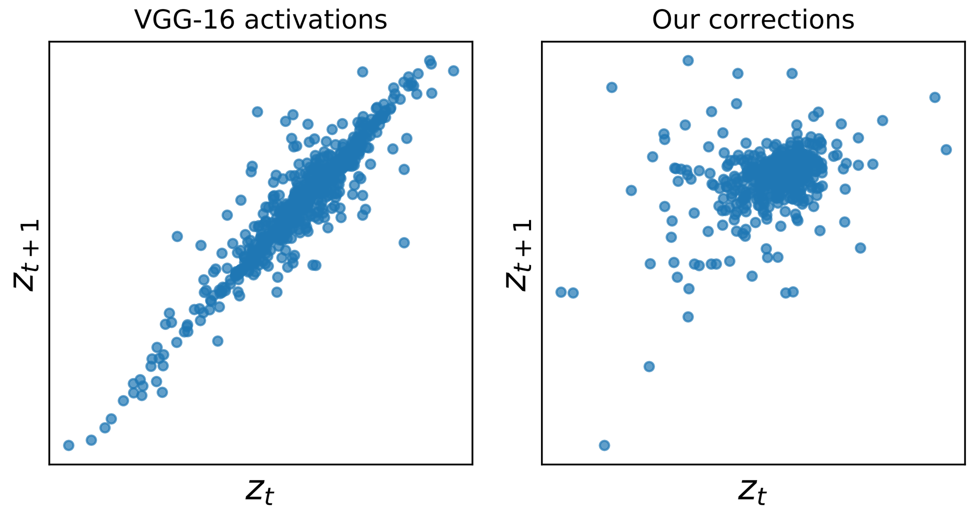
- Observations can lower layer activations (e.g. `conv3`), while latent states can be higher layer activations (e.g. `fc7`)
- Can efficiently be applied hierarchically



Properties

- Adaptively focus computation on "surprising" frames: ignore small corrections, re-initialize on large corrections
- Simplify learning by focusing on "residual-like" corrective terms
- Naturally de-correlate input stream in a hierarchical fashion

Experiments



Our model de-correlates inputs at each layer. While `conv4–3` activations (left) of consecutive frames are highly correlated, `conv4–3` *corrections* (right) are not.

	THUMOS	MultiTHUMOS	Charades
Single-frame	34.7	25.4	7.9
Two-stream	36.2*	27.6*	**8.9**
LSTM (RGB)	**39.3**	28.1	7.7
Predictive-Corrective	38.9	**29.7**	**8.9**

* reported from [Yeung 2017], using a single optical flow frame
Prior work achieves 29.6% on MultiTHUMOS [Yeung 2017] and 12.5% on Charades [Sigurdsson 2017]

Our update mechanism correctly recognizes the start of actions after initialization, and even corrects errors from initialization (last row).

References

- Sigurdsson, Gunnar A., et al. "Asynchronous Temporal Fields for Action Recognition." *CVPR* 2017.
- Koutnik, Jan, et al. "A clockwork rnn." *ICML* 2014.
- He, Kaiming, et al. "Deep residual learning for image recognition." *CVPR* 2016.
- Simonyan, Karen, et al.. "Two-stream convolutional networks for action recognition in videos." *NIPS* 2014.
- Yeung, Serena, et al. "Every moment counts: Dense detailed labeling of actions in complex videos." *IJCV* 2017.

Funding by NSF Grant 1618903 and 1208598, Intel Science and Technology Center for Visual Cloud Systems.