

Multi-object Tracking with Neural Gating Using Bilinear LSTM

Chanho Kim¹, Fuxin Li², and James M. Rehg¹

¹ Center for Behavioral Imaging
Georgia Institute of Technology, Atlanta GA, USA
{chkim, rehg}@gatech.edu

² Oregon State University, Corvallis OR, USA
lif@oregonstate.edu

Abstract. In recent deep online and near-online multi-object tracking approaches, a difficulty has been to incorporate long-term appearance models to efficiently score object tracks under severe occlusion and multiple missing detections. In this paper, we propose a novel recurrent network model, the Bilinear LSTM, in order to improve the learning of long-term appearance models via a recurrent network. Based on intuitions drawn from recursive least squares, Bilinear LSTM stores building blocks of a linear predictor in its memory, which is then coupled with the input in a multiplicative manner, instead of the additive coupling in conventional LSTM approaches. Such coupling resembles an online learned classifier/regressor at each time step, which we have found to improve performances in using LSTM for appearance modeling. We also propose novel data augmentation approaches to efficiently train recurrent models that score object tracks on both appearance and motion. We train an LSTM that can score object tracks based on both appearance and motion and utilize it in a multiple hypothesis tracking framework. In experiments, we show that with our novel LSTM model, we achieved state-of-the-art performance on near-online multiple object tracking on the MOT 2016 and MOT 2017 benchmarks.

1 Introduction

With the improvement in deep learning based detectors [16, 35] and the stimulation of the MOT challenges [32], tracking-by-detection approaches for multi-object tracking have improved significantly in the past few years. Multi-object tracking approaches can be classified into three types depending on the number of lookahead frames: online methods that generate tracking results immediately after processing an input frame [33, 1, 22], near-online methods that look ahead a fixed number of frames before consolidating the decisions [24, 7], and batch methods that consider the entire sequence before generating the decisions [39, 38]. For tracking multiple people, a recent state-of-the-art batch approach [38] relies upon person re-identification techniques which leverage a deep CNN network that can recognize a person that has left the scene and re-entered. Such

