Forecasting Hand and Object Locations in Future Frames

Chenyou Fan∗
fan6@indiana.edu
Jangwon Lee∗
leejang@indiana.edu
Michael S. Ryoo
mryoo@indiana.edu

School of Informatics and Computing
Indiana University
Bloomington, IN, USA

Abstract

This paper presents an approach to forecast future locations of human hands and objects. Given an image frame, the goal is to predict presence and location of hands and objects in the future frame (e.g., 5 seconds later), even when they are not visible in the current frame. The key idea is that (1) an intermediate representation of a convolutional object recognition model abstracts scene information in its frame and that (2) we can predict (i.e., regress) such representations corresponding to the future frames based on that of the current frame. We design a new two-stream convolutional neural network (CNN) architecture for videos by extending the state-of-the-art convolutional object detection network, and present a new fully convolutional regression network for predicting future scene representations. Our experiments confirm that combining the regressed future representation with our detection network allows reliable estimation of future hands and objects in videos.

1 Introduction

The ability to forecast future scene is very important for many real-time computer vision systems. Similar to humans predicting how the objects in front of them will move and what objects would newly appear, computer vision systems need to infer future locations of objects for their tasks. This is particularly necessary for interactive/collaborative systems, including surveillance systems, robots, and wearable devices. For instance, a robot working on a collaborative task with a human needs to predict what objects the human is expect to move and how they will move. This is also necessary for more natural human-robot interaction as well as better real-time surveillance, since the forecast will allow faster reaction of such systems in response to humans and objects.

In the past 2-3 years, there has been an increasing number of works on ‘forecasting’ in computer vision. Researchers studied forecasting trajectories [6, 17], convolutional neural network (CNN) representations [18], optical flows and human body parts [19], and video frames [6, 17]. However, none of these approaches were optimized for forecasting explicit
future locations of hands/objects appearing in videos. [17] only forecasts presence of objects without their locations, because of its design to regress lower dimensional (e.g., 4K) representations. [8] forecasts human pose only when the person is already in the scene (e.g., it is unable to predict whether a new person would appear) without any objects. Its learning was also not an end-to-end, since optical flows were used as its intermediate representation. [3, 7] were designed for forecasting direct image frames, not for object-level estimations. An approach to learn the entire forecast model in an end-to-end fashion so that they become optimized for the objective of future location estimation has been lacking.

This paper introduces a new approach to forecast hand/object locations in future frames (e.g., 5 seconds later). Given an image frame, the objective is to predict future bounding boxes of appearing objects even when they are not visible in the current frame. Our key idea is that (1) an intermediate CNN representation of an object recognition model abstracts scene information in its frame and that (2) we can model how such representation changes in the future frames based on the training data. For (1), we design a new two-stream CNN architecture with an auto-encoder by extending the state-of-the-art convolutional object detection network (SSD [6]). For (2), we present a new fully convolutional regression network that allows us to infer future CNN representations. These two networks are combined to directly predict future locations of human hands and objects, forming a deeper network that could be trained in an end-to-end fashion (Figure 1).

We evaluated our proposed approach with a couple of first-person video datasets with human hands and objects. Notably, in our experiments with the public ADL dataset [11], our accuracy was higher than the previous state-of-the-art method [15] by more than 0.25 mean average precision (AP).

2 Related work

Computer Vision researchers are increasingly focusing on ‘forecasting’ of future scene. Some earlier works include early recognition of ongoing human activities [4, 12] and more recent works include explicit forecasting of human trajectories/future locations [5, 9, 10, 16] and even future features or video frames themselves [3, 8, 15].

[15] proposed an approach to predict human trajectories in surveillance videos. There also were other works with a similar direction [9, 16]. [10] also tried to predict future location of the person, but using a different viewpoint: egocentric videos. However, most of these trajectory-based analysis are limited in the aspect that they assume the person to forecast is already present in the scene. This is insufficient particularly when dealing with hands and objects recorded in wearable/robot cameras, since a human hand often goes out of the scene (together with objects) and returns.

More recently, [8] showed that forecasting of fully connected layer results of convolutional neural networks (e.g., VGG [14]) is possible. The paper further demonstrated that such representation forecast can be used for forecasting the presence of objects in the scene (i.e., whether a particular object will appear in front of the camera 5 seconds later or not). However, due to the limited dimensionality of the representation this approach was forecasting (i.e., 4K-D), it was not directly applicable for forecasting ‘locations’ of objects in the scene. Similarly, [3] used a CNN regression to forecast optical flows, and used such optical flows to predict future human body pose. However, it requires the presence and the correct initial estimation of human pose. [8] predicted future video frames by learning dynamics from training videos, but did not attempt future localization of semantic entities such as objects.
Figure 1: Overview of our approach: It consists of two fully convolutional neural networks: The first network is the two-stream object detection network (the 1st row and the 3rd row of the figure). The 1st row and the 3rd row are the duplicates of the same model. The second network is the fully convolutional regression network to predict the future intermediate scene representation (the 2nd row). Only the colored layers are used in the actual testing stage.

The contribution of this paper is in (1) introducing the concept of future object forecast using fully convolutional regression of intermediate CNN representations, and (2) the design of the two-stream SSD model to consider both appearance and motion optimized for video-based future forecasting. We believe this is the first paper to present a method to explicitly forecast future hand and object locations using a fully convolutional network.

3 Approach

The objective of our approach is to predict future presence/location of human hands and objects in the scene given the current image frame. We propose a new two-stream convolutional neural network architecture with the fully convolutional future representation regression (Figure 1). The proposed model consists of two fully convolutional neural networks: (1) an extended two-stream video version of the Single Shot MultiBox Detector (SSD) [6] with a convolutional auto-encoder for human hand/object detection and (2) a future regression network to predict the intermediate scene representation corresponding to the future frame.

The key idea of our approach is that we can forecast scene configurations of the near future (e.g., 5 seconds later) by predicting (i.e., regressing) its intermediate CNN representation. Inside our fully convolutional hand/object detection network, we abstract scene information of the input frame as its intermediate representation (i.e., $\hat{F}_t$ in Figure 1) using
convolutional auto-encoder. Such intermediate representation gets further processed by the later layers of the network to finalize positions of hand/object bounding boxes. Our approach estimates the intermediate representation of future frame $\hat{F}_{t+\Delta}$ and combines it with the later layers of the network to forecast future bounding boxes of hands/objects.

### 3.1 Two-stream network for hand/object detection

This component predicts future hand/object locations given a video frame at time $t$. In this subsection, we newly introduce our two-stream network extending the previous fully convolutional object detection network SSD.

Our two-stream network is designed to combine evidence from both spatial- and motion-domain features to predict hand/object locations given the frame, as shown in Figure 1. The spatial stream receives one image frame, while the temporal stream receives the corresponding optical flow image. This design was inspired by the two-stream network of Simonyan and Zisserman [13], which was originally proposed for activity recognition. The intuition behind the use of the two-stream network is that it allows capturing of temporal motion patterns in activity videos as well as spatial information.

Here, we extend the SSD object detection network to construct the proposed two-stream network. We first insert a fully convolutional auto-encoder to our model, which has five convolutional layers followed by five transposed convolutional layers (also referred as de-convolutions) for the dimensionality reduction. We use $5 \times 5$ filters for each layers of the auto-encoder. The number of filters in the convolutional layers are: 512, 256, 128, 64, and 256. The transposed convolutional layers have the symmetric number of filters: 256, 64, 128, 256, and 512. We do not apply any pooling layer, but instead use stride 2 for the last convolutional layer. This design allows the abstraction of scene information in an image frame as a lower dimensional (256x25x25) intermediate representation.

We designed our object detection network to have both the spatial-stream and temporal-stream part. Instead of using late-fusion to combine spatial and temporal streams at the end of network as was done in [13], we design early-fusion in our two-stream network by combining two streams’ feature maps before the encoder-decoder component. Specifically, at conv5 layers, two $256 \times 25 \times 25$ feature blobs from both streams are combined to form a single $256 \times 25 \times 25$ blob with feature selection layer of one-by-one kernels. This selection layer is also learned during our training process, making it optimized for the hand/object detection. In addition, since our designed regression component can combine multi-frame information, we are able to reduce the amount of computations in our temporal stream by making it receive one single optical flow image instead of stacked optical flows. We will discuss this in detail in Subsection 3.2.

Let $f$ denote the proposed two-stream network to detect hands/objects given a video frame at time $t$. This function has two input variables $\hat{X}_I_t$ and $\hat{X}_O_t$, which represent a given current input frame and the corresponding optical flow image at time $t$ respectively. Then, this function can be decomposed as two sub functions, $f = g \circ h$:

$$\hat{Y}_t = f(\hat{X}_I_t, \hat{X}_O_t) = h(\hat{F}_t) = h(g(\hat{X}_I_t, \hat{X}_O_t)),$$

where a function $g : (\hat{X}_I, \hat{X}_O) \rightarrow \hat{F}$ denotes a convolutional encoder to extract compressed visual representation (feature map) $\hat{F}$ from $\hat{X}_I$ and $\hat{X}_O$, and $h : \hat{F} \rightarrow \hat{Y}$ indicates the remaining part of the proposed network that uses the compressed feature map as an input for predicting hands and object locations $\hat{Y}_t$ at time $t$. The upper part of Figure 1 shows the architecture.
3.2 Future regression network

The objective of this research is not about estimating hand/object locations in the ‘current’ frame \( \hat{Y}_t \), but to forecast the locations of them in the ‘future’ frame \( \hat{Y}_{t+\Delta} \). Thus, we design a new fully convolutional regression network.

We formulate the problem as a regression problem of forecasting future intermediate representation \( \hat{F}_{t+\Delta} \) of the proposed two-stream network based on its current intermediate representation \( \hat{F}_t \). The main idea is that the intermediate representation of our proposed network abstracts spatial and motion information of hands and objects, and we are able to take advantage of them to establish a mapping from the current frame to the future frame. Once the future intermediate representation \( \hat{F}_{t+\Delta} \) is regressed, we plug-in the predicted representation to the remaining part of the proposed network (i.e., \( h \)) to forecast future hand/object bounding boxes.

Let \( r \) denote our future regression network to predict the future intermediate scene representation \( \hat{F}_{t+\Delta} \) given a current scene representation \( \hat{F}_t \).

\[
\hat{F}_{t+\Delta} = r_w(\hat{F}_t),
\]  

(2)

The regression network consist of nine convolutional layers, each having 256 channels of \( 5 \times 5 \) filters except the last two layers. We use dilated convolution with \( 1024-D \) to cover a large receptive field of \( 13 \times 13 \) for the 8th layer, and \( 256-D \) with \( 1 \times 1 \) kernel is used for the last layer.

A desirable property of this formulation is that it allows training of the weights (\( w \)) of the regression network with unlabeled videos using the following loss function:

\[
w^* = \arg\min_w \sum_{i,t} \| r_w(g(\hat{X}_I^i_t, \hat{X}_O^i_t)) - g(\hat{X}_I^{i+\Delta}t, \hat{X}_O^{i+\Delta}t) \|^2_2
\]  

(3)

where \( \hat{X}_I^i_t \) indicates the frame at time \( t \) from video \( i \). Here, we use our compressed scene representation as \( \hat{F}_i \), taking advantage of its relatively low dimensionality, but we can use any intermediate scene representation from any layers of our two-stream network in principle. Once we get the future scene representation \( \hat{F}_{t+\Delta} \), it is fed into the two-stream network to forecast hand/object locations corresponding to the future frame:

\[
\hat{Y}_{t+\Delta} = h(\hat{F}_{t+\Delta}).
\]  

(4)

Figure 1 illustrates data flow of our proposed approach during testing phase. Given a video frame \( \hat{X}_I^i_t \) and its corresponding optical flow image \( \hat{X}_O^i_t \) at time \( t \), (1) we first extract the compressed intermediate representation using the feature extractor \( g \), then (2) feed it into the future regression network \( r \) to obtain future scene representation \( \hat{F}_{t+\Delta} \). Finally, (3) we can predict future location of hands/objects \( \hat{Y}_{t+\Delta} \) by feeding the predicted future scene representation into the remaining part of the proposed two-stream network \( h \) at time \( t \).

\[
\hat{Y}_{t+\Delta} = h(\hat{F}_{t+\Delta}) = h(r(\hat{F}_t)) = h(r(g(\hat{X}_I^i_t, \hat{X}_O^i_t))).
\]  

(5)

In addition to the above basic formulation, our proposed approach is extended to use previous \( K \) frames to obtain \( \hat{F}_{t+\Delta} \) as illustrated in Figure 1, instead of using just a single frame (i.e., the current frame) for the future regression:

\[
\hat{Y}_{t+\Delta} = h(r([g(\hat{X}_I^i_t, \hat{X}_O^i_t), \ldots, g(\hat{X}_I^{t-(K-1)}, \hat{X}_O^{t-(K-1)})])).
\]  

(6)
Our future representation regression network allows us to infer future hand/object locations while considering the implicit activity and object context in the scene. The intermediate representation $\hat{F}_t$ abstracts spatial and motion information in the current scene, and our fully convolutional future regressor can take advantage of it for the prediction.

4 Experiments

We conducted two sets of experiments to confirm the location forecast ability of our approach using the fully convolutional two-stream regression architecture. In the first experiment, we use a first-person video dataset to predict future human hand locations. In the second set of experiments, we use the public dataset with object annotations to train/test object bounding box forecasting.

4.1 Dataset

In our experiments, we used three datasets for the training and testing of our approaches.

**EgoHands [2]:** This dataset is a collection of 48 ego-centric videos that contains four types of human interactions (i.e., playing cards, playing chess, solving a puzzle, and playing Jenga). The original dataset has 15,053 ground-truth labels for hands in 4,800 frames, and we also newly annotated 466 frames with 1,267 hand bounding boxes. We used this dataset for the training our two-stream network to detect hands in a video frame.

**Unlabeled Human Interaction Videos:** This is our newly collected dataset that contains a total of 47 first-person videos of human-human collaboration scenarios, taken with a wearable camera. The dataset contains videos of two types of collaborative scenarios: (1) a person wearing the camera cleaning up objects on a table if another person approaches the table while holding a large box, making a room for him/her to put the box, and (2) the camera wearer pushes a trivet on a table to another person as he/she is approaching the table while holding a cooking pan. The duration of each video clip is between 4 and 10 seconds, and the videos do not have ground truth annotations of human hand bounding boxes.

**Activities of Daily Living (ADL) [11]:** This first-person video dataset contains 20 videos of 18 daily activities, such as making tea and doing laundry. This is a challenging dataset since frames display a significant amount of motion blur caused by the camera wearer’s movement. This dataset also suffers from noisy annotations. Object bounding boxes were provided as ground truth annotations. Although there are 43 types of objects in the dataset, we trained our model (and the baselines) for 15 most common categories, following the setting used in [15]. We split the ADL dataset into four sets, using three sets for the training and the remaining set for the testing.

4.2 Baselines

In order to confirm the benefits of our proposed approach quantitatively, we created multiple baselines based on our base network (SSD) [6].

(i) **SSD with future annotations** is the original SSD model taking the current image frame as an input, extended to forecast the future hands/objects. Instead of providing ground truth bound boxes of objects in the frame in the training step, we provided ‘future’ ground truth locations of hands and objects to make it learn to regress future boxes.
Figure 2: Examples of hand location prediction. The first row shows the input frames and the second row shows the optical flows. The third row shows our future hand prediction results and we overlaid our predictions on ‘future’ frames in the last row. Red boxes correspond to the predicted ‘my left hand’ locations, blue boxes correspond to ‘my right hand’, green boxes correspond to the opponent’s left hand, and the cyan boxes correspond to the opponent’s right hand.

Table 1: Future hand location forecasts measured with Human Interaction dataset.

| Method                      | Evaluation       |
|-----------------------------|------------------|
|                             | Precision | Recall  | F-measure |
| Hand-crafted features       | 0.30 ± 0.37      | 0.15 ± 0.19 | 0.20 ± 0.25 |
| Hands only                  | 4.78 ± 3.70      | 5.06 ± 4.06 | 4.87 ± 3.81 |
| SSD w/ future Annot.        | 27.53 ± 23.36    | 9.09 ± 8.96 | 13.23 ± 12.62 |
| Ours (one-stream): K=1      | 27.04 ± 16.50    | 21.71 ± 14.71 | 23.45 ± 14.99 |
| Ours (one-stream): K=5      | 29.97 ± 15.37    | 23.89 ± 16.45 | 25.40 ± 15.51 |
| Ours (one-stream): K=10     | 36.58 ± 16.91    | 28.78 ± 17.96 | 30.90 ± 17.02 |
| Ours (two-stream): K=1      | 37.21 ± 22.49    | 26.69 ± 14.28 | 30.21 ± 16.07 |
| Ours (two-stream): K=5      | 37.41 ± 22.97    | 26.19 ± 14.93 | 30.06 ± 17.16 |
| Ours (two-stream): K=10     | 42.89 ± 23.61    | 30.46 ± 13.08 | 34.18 ± 16.48 |

(ii) Hands only is the baseline only using estimated hand locations in the current frame to predict their future locations. The idea is to confirm whether the detection of the current hand locations is sufficient to infer their future locations.

(iii) Hand-crafted features uses a hand-crafted state representation based on explicit hand and object detection from the current frame. It encodes relative distances between all detected objects, and uses it to predict the future locations of them using neural network-based regression. More specifically, it detects objects using KAZE features [3] and hands using [4], then computes relative distances between all objects and hands for building the state representation which is a 20 dimensions of vector. It then performs a regression using a network of five fully connected layers.

In addition, we also implemented a simpler version of our approach, (iv) one-stream network, which uses the same CNN architecture as our proposed approach except that it only has the spatial stream (taking RGB input) without the temporal stream (taking optical flow input). We constructed this baseline to confirm how much the temporal-stream of our network helps predicting future hand/object locations.
### Evaluation

In this section, we evaluated the performance of our approach to forecast future locations of hands and objects using the two different datasets (our Unlabeled Human Interaction Videos and ADL dataset).

**Hand location forecast:** We first focused on evaluating the performance of our approach to predict future hand locations using the interaction dataset. This is a less noisier dataset than the ADL dataset. Here, we used hand detection results using the original SSD model trained on the EgoHands dataset [2] as the ground truth hand labels, since the interaction videos do not have any human annotations. We randomly split the dataset into the training set and the test set; we used 32 videos for the training and the remaining 15 videos for the testing. We used precision and recall values as our evaluation metric. Whether the forecasted bounding boxes are true positives or not was decided based on the “intersection over union” (IoU) ratio between areas of each predicted bounding box and the ground truth box (future). We only considered the prediction result as a true positive when the IoU ratio was greater than 0.5.

Table 1 shows quantitative results of 1-second future hand predictions on our Human Interaction dataset. Since our network may use previous $K$ frames as an input for the future regression, we reported the performances of our approach with $K=1$, 5 and 10 frames. We clearly observe that our approach significantly outperforms the original SSD trained with future hand locations. Even our one-stream model performed much superior to the SSD, suggesting the effectiveness of our concept of future regression. Also, our two-stream models performed better than our one-stream models. This indicates that the temporal stream was helpful to predict future locations of hands. Our proposed model with $K=10$ yielded the best performance in terms of all three metrics, at about 34.18 score in F-measure. Figure 2 shows example hand forecast results.

**Object location forecast:** We used the ADL dataset to measure the object location forecast accuracy. Similar to the hand forecast, (1 second or 5 second) future bounding box locations were estimated, and the accuracies were measured in terms of precision and recall. The IoU ratio of 0.5 was used. Table 2 shows average precision (AP) of each object category, and Figure 4 and 5 show PR-curves for predicting one second and five second future objects. The ADL dataset is particularly challenging in the sense that it is often the case that objects appear in 5-second-future frames are not visible in the current frame. We are able to observe that our approach significantly outperforms the SSD baseline. Figure 3 shows example results.

**Object presence forecast:** In this task, we use the ADL dataset to examine our approach to forecast ‘presence’ of objects in future frames. Specifically, we ignore the location of
the bounding boxes and decide that the object exists if confidence score of its bounding box is above the threshold. By iterating the confidence score threshold from 0 to 1, we obtained their PR-curves and calculated AP of each object category. Similar to the object location forecast task, we trained our model to predict presence of objects in 5-second-future frames. This experiment was particularly designed to directly compare our approach against the results of [15]’s AlexNet based architecture, following the same standard setting as theirs.

Table 3 compares our result with the baselines. Again, we are able to observe that our approach greatly benefits the overall forecast. Particularly, our approach significantly outperformed the results of [15] by more than 0.25 mean AP.

5 Conclusion

We presented a new approach to forecast human hand and object locations using a fully convolutional future representation regression network. The key idea was to forecast scene configurations of the near future by predicting (i.e., regressing) intermediate CNN representations of the future scene. We presented a new two-stream object detection model to abstract scene information of the given frame, and experimentally confirmed that we can map such
an intermediate scene representation of current frame to its future frame and thus forecast presence and location of future objects.

Acknowledgement

This work was supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-10-2-0016.

References

[1] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. Kaze features. In *European Conference on Computer Vision (ECCV)*, 2012.

[2] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[3] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 64–72, 2016.

[4] Minh Hoai and Fernando De la Torre. Max-margin early event detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[5] Kris M. Kitani, Brian D. Ziebart, J. Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision (ECCV)*, 2012.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

[7] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[8] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. *arXiv preprint arXiv:1701.01821*, 2017.

[9] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M. Kitani. A game-theoretic approach to multi-pedestrian activity forecasting. *arXiv preprint arXiv:1604.01431*, 2016.

[10] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854. IEEE, 2012.

[12] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[13] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014.
[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[15] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations with unlabeled video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3302–3309, 2014.