Dense Optical Flow Prediction from a Static Image

Jacob Walker, Abhinav Gupta, Martial Hebert

1 The Robotics Institute, Carnegie Mellon University. 2 Pittsburgh, PA.

Consider the images shown in Figure 1. Given the man running left, we humans can easily predict that his entire body will move to the left. The surf wave will collapse to the bottom left, and the man performing the pushup will move his body downward. Humans have an amazing ability to not only recognize what is present in the image but also predict what is going to happen next. Prediction is an important component of visual understanding and cognition. In order for computers to react to their environment, simple activity detection is not always sufficient. For successful interactions, robots need to predict the future and plan accordingly.

While prediction is still a relatively new problem, there has been some recent work that has focused on this task. The most common approach to this prediction problem is to use a planning-based agent-centric approach: an object [2] or a patch [8] is modeled as an agent that performs actions based on its current state and the goal state. Each action is decided based on compatibility with the environment and how this actions helps the agent move closer to the goal state. The priors on actions are modeled via transition matrices. Such an approach has been shown to produce impressive results: predicting trajectories of humans in parking lots [2] or hallucinating car movements on streets [8]. There are two main problems with this approach. First, the predictions are still sparse, and the motion is still modeled as a trajectory. Second, and more importantly, these approaches have always been shown to perform in restrictive domains such as parking lots or streets.

In this paper, we take the next step towards generalized prediction — a framework that can be learned from tens of thousands of realistic videos. The framework can work in indoor and outdoor environments if the agent is an animal, a human, or even a car. It can account for one or multiple agents. Specifically, we look at the task of motion prediction — given a static image we predict the dense expected optical flow as if this image was part of a video. This optical flow represents how and where each and every pixel in the image is going to move in the future. Of course, we can see that motion prediction is a highly-context dependent problem. The future motion not only depends on what is active in the scene but also its context. For example, someone’s entire body may move up or down if they are jump-roping, but most of the body will be stationary if they are playing the flute. Instead of modeling agents and their context separately under restrictive assumptions, we use a learning based approach for motion prediction. Specifically, we train a deep network that can incorporate all of this contextual information to make accurate predictions of future motion in a wide variety of scenes.

We train our model from thousands of realistic video datasets, namely the UCF-101 [7] and the HMDB-51 [4].

Our paper makes threefold contributions. First, we present a CNN model for motion prediction. Given a static image, our CNN model predicts expected motion in terms of optical flow. Our CNN-based model is agent-free and makes almost no assumptions about the underlying scene. Therefore, we show experimental results on diverse set of scenes. Second, our CNN model gives state of the art performance on prediction compared to contemporary approaches. Finally, we also present a proof of concept extension of CNN model which makes long-range prediction about future motion. Our preliminary results indicate the this new CNN model might indeed be promising even on the task of long-range prediction.

Intuitively, motion estimation can be posed as a regression problem since the space is continuous. Indeed, this is exactly the approach used in [6], where the authors used structured random forests to regress the magnitude and direction of the optical flow. However, such an approach has one drawback: such an output space tends to smoothen results as the ambiguity is handled by averaging out the flow. Interestingly, in a related problem of surface normal prediction, researchers have proposed reformulating structured regression as a classification problem [5, 10]. Specifically, they quantize the surface normal vectors into a codebook of clusters and then the output space becomes predicting the cluster membership. In our work, we take a similar approach. We quantize optical flow vectors into 40 clusters by k-means. We can then treat the problem in a manner similar to semantic segmentation, where we classify each region as the image as a particular cluster of optical flow. We use a soft-max loss layer at the output for computing gradients.

However, at test time, we create a soft output by considering the underlying distribution of all the clusters, taking a weighted-probability sum over all the classes in a given pixel for the final output. Transforming the problem into classification also leads directly to a discrete probability distribution over vector directions and magnitudes. As the problem of motion prediction can be ambiguous depending on the image, we can utilize this probability distribution over directions to measure how informative our predictions are. Our network can rank upward and downward facing clusters much higher than other directions. Even if the ground truth is upward, and
Our network is similar to the standard 7-layer architecture [3] used for many recognition tasks. We take a 200x200 image as input. However, we use a spatial softmax as the final output. For every pixel in the image we predict a distribution of various motions with various directions and magnitudes. We can combine a weighted average of these vectors to produce the final output for each pixel. For computational reasons, we predict a coarse 20x20 output.

the highest ranked cluster is downward, it may be that the second-highest cluster is also upward. A discrete probability distribution, through classification, allows an easier understanding of how well our network may be performing.

Our model is similar to the standard seven-layer architecture from [3]. To simplify the description, we denote the convolutional layer as C(k,s), which indicates the there are k kernels, each having the size of s x s. During convolution, we set all the strides to 1 except for the first layer, which is 4. We also denote the local response normalization layer as LRN, and the max-pooling layer as MP. The stride for pooling is 2 and we set the pooling size as 20x20.

Our network architecture can be described as: C(96,11) → LRN → P → C(256,5) → LRN → P → C(384,3) → C(384,3) → C(256,3) → P → F(4096) → F(4096).

Our network architecture can be described as:

$$L(I, Y) = \sum_{i=1}^{M} \sum_{r=1}^{N} (1(y_i = r) \log F_{i,r}(I))$$

$$F_{i,r}(I)$$ represents the probability that the is pixel will move according to cluster r. 1(y_i = r) is an indicator function.

We automatically label our training dataset with an optical flow algorithm. With a publicly available implementation, we chose DeepFlownet [11] to compute optical flow. The UCF-101 and the HMDB-51 dataset use realistic, sometimes low-quality videos from a wide variety of sources. They often suffer from compression artifacts. Thus, we aim to make our labels somewhat less noisy by taking the average optical flow of five future frames for each image. The videos in these datasets are also unstabilized. [9] showed that action recognition can be greatly improved with camera stabilization. In order to further denoise our labels, we wish to focus on the motion of objects inside the image, not the camera motion. We thus use the stabilization portion of the implementation of [9] to automatically stabilize videos using an estimated homography.

Figure 1 shows some of our qualitative results. For single frame prediction, our network is able to predict motion in many different contexts. We find that while [6] is able to make reasonable predictions on the KTH, qualitative performance collapses once the complexity and size of the dataset increases. Although most of our datasets consist of human actions, our model can generalize beyond simply detecting general motion on humans. Our method is able to successfully predict the falling of the ocean wave in the second row, and it predicts the motion of the entire horse in the first row. Furthermore, our network can specify motion depending on the action being performed. For the man playing guitar and the man writing on the wall, the arm is the most salient part to be moved. For the man doing a pushup and the man riding the skyjet, the entire body will move according to the action.

In this paper we have presented an approach to generalized event prediction in static scenes. Namely, our framework focuses on motion prediction as a non-semantic form of action prediction. By using an optical flow algorithm to label the data, we can train this model on a large number of unlabeled videos. Furthermore, our framework utilizes the success of deep networks to outperform contemporary approaches to motion prediction. We find that our network successfully predicts motion based on the context of the scene and the stage of the action taking place. Possible work includes incorporating this motion model to predict semantic action labels in images and video. Another possible direction is to utilize the predicted optical flow to predict in raw pixel space, synthesizing a video from a single image.


