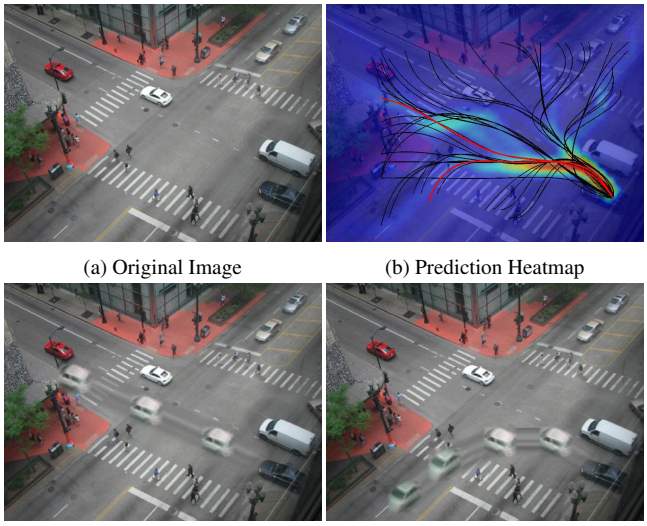
Patch to the Future: Unsupervised Visual Prediction (Extended Abstract)

Jacob Walker, Abhinav Gupta, and Martial Hebert
Robotics Institute, Carnegie Mellon University
{jcwalker, abhinavg, hebert}@cs.cmu.edu

1. Introduction

Consider the image shown in Figure 1. A reliable modern computer vision approach might at best recognize the objects and regions in the image and list the corresponding nouns — road, car, tree and grass. However, when we humans look at the same image, we can not only infer what is happening at that instant but also predict what can happen next. For example, in the same image, we can predict that the car on the bottom right is either going to go straight or turn left at the intersection.

In this work, we take a step toward this goal of generalized visual prediction — determining *what* is active in the scene as well as *how* the activity should unfold. However, this leaves us with major questions. What do we predict? What does the output space of visual prediction look like? Recent approaches have only focused on predicting the movements and transitions of agents treated as a point object [1] or optical flow of pixels [5]. In contrast, we humans can not only predict the motion but also how the appearances would change with that movement or transition. This allows us to create mental images of prediction. In a similar manner, we argue that the space of visual prediction should be richer and even include prediction of visual appearances. For example, we can guess how a car will look after it turns and how a book unfolds when opened. However, having a richer output space requires richer representation (elements of reasoning) and lots of data to learn the priors. Building upon the recent success of mid-level elements [3], we propose a new framework for visual prediction which uses these mid-level elements as building blocks of prediction. In our framework, we model not only the movement and transitions of these elements in the scene but also how the appearances of these elements can change. Our new framework has the following advantages over previous approaches: (a) Our approach makes no assumption about what can act as an agent in the scene. It uses a data-driven approach to identify the possible agents and their activities; (b) Using a patch-based representation allows us to learn the models of visual prediction in a completely unsupervised manner. We also demonstrate how a rich representation allows us to use a simple non-parametric approach to learn a state-of-the-art visual prediction model; (c) Finally, because our approach exploits mid-level elements instead of full scenes for creating associations, it allows for generalization and sharing across different instances. For more details and results, please refer to our CVPR 2014 paper: http://www.ri.cmu.edu/pub_files/2014/3/egpaper_final.pdf



(a) Original Image (b) Prediction Heatmap

(c) Predicted Path -1 (d) Predicted Path -2

Figure 1. Consider the scene shown in image (a). Our data-driven approach uses a large collection of videos to predict the likely future of an agent in the scene. The heatmap (b) shows the likely locations the car can visit in the future (along with a few possible trajectories.) (c) shows the hallucination of car moving straight and (d) shows the hallucination of the car turning left.

2. Our Approach

Given an input scene, our goal is to predict what is going to happen next — what parts of the image are going to remain the same, what parts of the image are likely to move, and how they move. The central idea is that scenes are represented as a collection of mid-level elements (detected using a sliding window) where agents can either move in space or change visual appearances. Each agent is predicted independently assuming a static scene. We model the distribution over the space of possible actions using a transition matrix which represents how mid-level elements can move

and transition into one another and with what probability. For example, an element that represents a frontal car can transition to a patch facing right if the car turns. Given the mid-level elements and their possible actions, we first determine which is the most likely agent and the most likely action given the scene. However, this notion of most likely action depends upon goals and the context/scene around the elements. For example, in Figure 1, the visual prediction of a car not only depends upon the goal but also on the other cars, pedestrians, and the sidewalk in the image. Therefore, as a next step, we need to model the interaction between the active element (agent) and its surrounding. We model this interaction using a reward function $\psi_i(x, y)$ which models how likely is it that an element of type i can move to location (x, y) in the image. For example, a car element will have high reward for road-like areas and low reward for grass-like areas — without modeling semantics explicitly. Given a goal, our approach then infers the most likely path using the transition matrix and computed reward. Finally, if the goal is unknown —which is the case here, we propose to sample several goals and select the most likely goal based on high expected reward.

3. Experiments

3.1. Car Chase Dataset

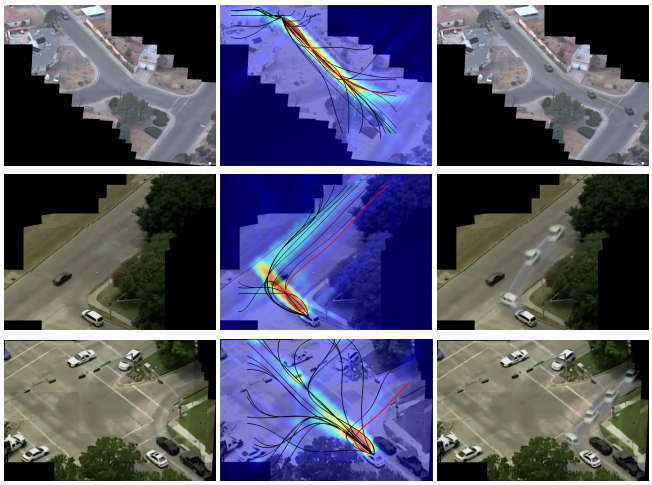
For our main experiments, we created a new dataset by downloading videos from Youtube of aerial car chase videos. In total we used 183 videos (from 48 different scenes) that lasted from 5 to 30 seconds. For extracting discriminative mid-level elements, we used random frames from the training set in addition to outdoor scenes from Flickr as the discovery dataset, and we use the MIT Indoor 67 [2] dataset for the negative dataset. We manually annotated the trajectories of the car in 44 test videos which were used as the ground-truth to evaluate the algorithm. Figure 2 shows some qualitative results.

3.2. VIRAT Dataset

For our second dataset, we chose a subset of the VIRAT dataset corresponding to a single scene A used in [1]. Since VIRAT data consists of only one scene, we used frames from the TUD-Brussels outdoor pedestrian dataset [4] to extract mid-level elements. Figure 3 shows some qualitative results.

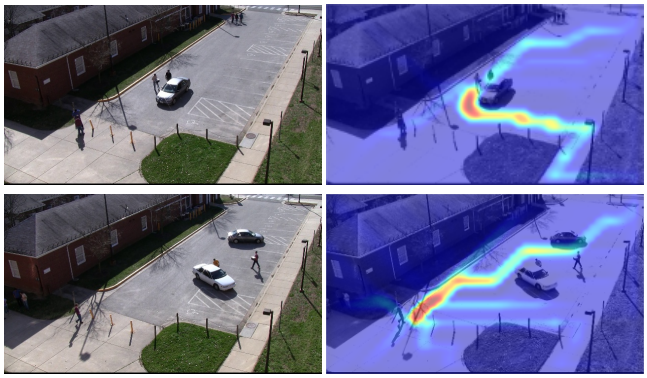
4. Conclusion

In this paper we have presented a simple and effective framework for visual prediction on a static scene. Our prediction framework builds upon representative and discriminative mid-level elements and combines this visual representation with a decision theoretic framework. This representation allows us to train our framework in a completely unsupervised manner from a large collection of videos. However, more importantly, we can also predict how visual appearances will change in time and create a hallucination of the possible future. It is important to note that this paper represents an initial step in the direction of general unsupervised prediction. Possible future work includes modeling the simultaneous behavior of multiple elements.



(a) Original Image (b) Prediction Heatmap (c) Predicted Path

Figure 2. Qualitative predictions for our approach. The far left shows the original image, the center shows a heatmap of possible paths, and the right shows a visualization of one of those paths.



(a) Original Image (b) Prediction Heatmap

Figure 3. Qualitative predictions for our approach on the VIRAT dataset. The left shows the original image; the right shows a heatmap of possible paths.

References

[1] K. Kitani, B. Ziebart, D. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012.

[2] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[3] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[4] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, 2009.

[5] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *ECCV*, 2010.