Structural Inpainting of Road Patches for Anomaly Detection

Asim Munawar, Clement Creusot
IBM Research - Tokyo, Japan
{asim,clement}@jp.ibm.com

Abstract

Obstacle detection on the road is a key function for self-driving vehicles. A lot of research has focused on detecting large obstacles such as cars and pedestrians. Small obstacles can also be the source of serious accidents, especially at high speed. We present an approach for detecting anomalies on the road using a higher-order Boltzmann machine. As opposed to conventional anomaly detectors the proposed system learns to inpaint the road patches with commonly occurring road features such as lane markings and expansion dividers, depending on the context. The system does not consider these frequent road artifacts as anomalies and significantly reduces the number of obstacle candidates. We show initial empirical results for anomaly detection with this new approach.

1 Introduction

The automotive industry is facing increasing pressure from customers and governments to improve the safety of cars. Although fully autonomous cars might be the ultimate goal, partially autonomous or assisted driving is more urgent and more likely to be commercialized in the near future. The main objective of these driving assistance systems will be to save lives by avoiding accidents. We can already see many emerging technologies to make the cars safer, such as pedestrian detection [2] or collision warning system[8]. Current technologies either use passive sensors (cameras) or active sensors (millimeter-wave radar, LIDAR). Active sensors can provide accurate measurements of large obstacles at shorter distances. In spite of the expensive price tags, the precision of such sensors decreases quickly with the distance from the obstacle. On the other hand, human perception only uses passive sensing with cognition capability to detect small obstacles at larger distances with high accuracy.

In this paper, we argue that a machine learning based solution that mimics human perception should outperform the current technologies in detecting anomalies on the road. While obstacles are 3D objects an anomaly can also be a 2D marking on the road. When looking at a scene our brain constantly creates top-down predictions of what the next image should be. Anomalies are detected by comparing these predictions with the actual observations (see Egner et al. [3]). We consider a similar approach by inpainting parts of the road and by comparing the predictions with the actual observations to detect the anomalies.

Roads come in a variety of patterns and colors. Learning a pattern on one kind of road might not work for an other kind. Similarly, the road color changes widely in different patches of the road and depending on the light and weather conditions. In some countries different color road surfaces are used to mark bus lanes, car pooling lanes, or dangerous turns. Similarly the road markings might also vary between countries. Never the less, the local features of the road, including the lane markings, are basically consistent. This local information can be used to inpaint missing parts of the next frames. Some assumptions that are spatially and temporally true for a video of the road taken by a dashboard camera are:

- The car's speed and direction do not change significantly between two consecutive video frames.
- The road color, texture, and marking patterns do not change abruptly between two consecutive frames.
- An anomaly is easiest to detect when it is observed for the first time.

The main contribution of this research is to exploit these above assumptions to detect anomalies on highway video stream. The proposed approach exploits both the local temporal and spatial information to inpaint the regions of the next observed frame. To the best of our knowledge, no one has tried this approach for anomaly detection on roads.

The rest of the paper is organized as follows: Section 2 clarifies the problems that our work seeks to solve. We describe our proposed approach in Section 3. Quantitative analysis is presented in Section 4. Comparisons with existing techniques are given in Section 5. Finally we conclude the paper in Section 6.

2 Background and Motivation

In addition to detecting pedestrians and other vehicles on the road, smart cars will need to detect small obstacles at a considerable distance for safe stop or other corrective actions such as slowing down or changing lanes.

At normal highway driving speed, and even with a deceleration of 0.8 G (the upper limit for safe deceleration of a car) it may take over 100 m to come to a complete stop on a dry road. This does not include the latency caused by the cognitive and decision making processes. Even though the objective is to detect a 3D solid obstacle, the current paper only focuses on anomaly detections. It is intended as the first step toward full obstacle detectors.

From computer vision perspective, a road is a sequence of locally repeating patterns. To explain the motivation for the technique used in this research, assume a hypothetical world where the road is always empty, and where we have a system with an infinite memory and computational power. Such a system could remember all of the road scenes it has ever seen. In such a world the problem of detecting anomalies on the road is simple. The system could segment the road out of the image and find the closest match (nearest neighbor) in the previously observed images. The re-

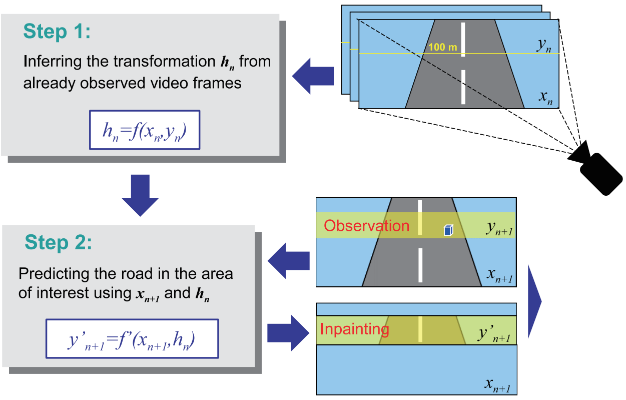

Step 1:
Inferring the transformation h_n from already observed video frames
$h_n=f(x_n,y_n)$
100 m
y_n
x_n
Step 2:
Predicting the road in the area of interest using x_{n+1} and h_n
$y'_{n+1}=f'(x_{n+1},h_n)$
Observation
y_{n+1}
x_{n+1}
Inpainting
y'_{n+1}
x_{n+1}

Figure 1. Image transformation h_n is a function of $f(x_n, y_n)$. The same transformation h_n can be used to inpaint the unknown region $y'_{n+1} = (x_{n+1}, h_n)$. The anomaly can be detected by comparing the inpainted region y'_{n+1} and the actual observation y_{n+1}.

trieved image would then be matched with the current observation to find any anomalies.

In reality, we don't have an infinite memory and computational capability. However, with a few realistic assumptions about the road and vehicle motion useful information can be inferred from a limited number of previously seen frames. Since the patterns of lane markings are usually consistent in two consecutive video frames, they can be learned by using the current frame and applied to the next frame to inpaint the road. This inpainting can be compared to the actual observation to find any anomalies on the road. Compared with a simple anomaly detection technique, the presented approach learns the commonly occurring structures on the road and does not consider them as anomalies. This significantly reduces the number of false positives. In addition, since the approach is based only on local information, it is robust to different weather and lighting conditions. Figure 1 is an abstract representation of the process.

3 Structural Scene Inpainting

The core idea behind this research is to learn the inpainting of the frequent road patterns. Since the patterns on the road keep repeating the transformations required to inpaint a part of the pattern can either be acquired from the current image frame or from the previously observed frames. The inpainted predictions can then be compared with the actual observations to find potential anomalies on the road.

The over simplified concept shown in Figure 1 is not feasible due to the limited learning resources. In this research we use patches of the shape shown in Figure 2-a. The region y is treated as the unknown region while the surrounding area x is used as the known context. There is a high probability that the last few observed frames have shown similar structure (e.g. discontinuous lane markings in the middle or a continuous marking on the side of the road). It is safe to assume that the spacing between the lane markings will stay constant for many frames and will not change abruptly. We argue that instead of defining global rules for the lane markings and other markings that commonly occur on the road, it is better to learn them from short-term memory.

Although various techniques can be applied to inpaint the missing part of an image we employ a higher-order Boltzmann machine (also known as gated Boltzmann machine [5]) to learn the transformation between x and y. The abstract structure of the network is shown in Figure 2-b.

3.1 Offline Learning

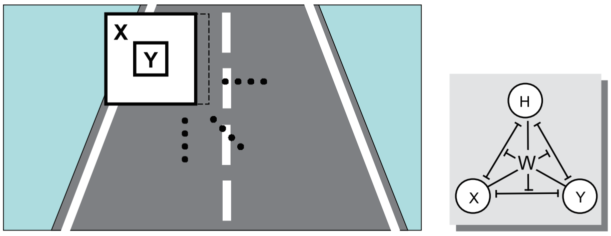
To capture the correlations between input x, output y, and the hidden variables h, Memisevic et al. [5] suggests using the following three-way energy function to define the conditional distribution:

$$-E(y, h; x) = \sum_{ijk} w_{ijk} x_i y_j h_k \tag{1}$$

where i, j, and k index the input, output, and hidden units. The biases are not shown in this equation. W_{ijk} is a three-dimensional tensor whose dimensions increase cubically with the increase in the size of the input, output, or hidden units. Learning such a large number of weights make the learning slow and inefficient. In another paper, Memisevic et al. [6] proposed solving this problem by factoring the interaction tensor. This technique is known as Factored Gated Boltzmann Machine (FGBM). In FGBM the three-way energy of a joint configuration of the visible and hidden units can be defined as:

$$-E(y, h; x) = \sum_{f=1}^{F} \sum_{ijk} x_i y_j h_k w^x_{if} w^y_{jf} w^h_{kf} \tag{2}$$

where f indexes the factors. In other words, the $I \times J \times K$ parameter tensor is replaced by three matrices with sizes $I \times F$, $J \times F$, and $K \times F$. If the number of factors is comparable to the number of units in the visible and hidden groups, the factorization reduces the number of parameters from $O(N^3)$ to $O(N^2)$. Using this factorization, the weight w_{ijk} in Equation 1 is implemented as $\sum_f w^x_{if} w^y_{jf} w^h_{kf}$. The bias terms remain unchanged. This factored higher-order Boltzmann machine allows efficient learning of the transformations between larger image patches.



(a) (b)

Figure 2. (a) Shape of the mask for x (known area) and y (unknown area) used for offline training and online testing. (b) Represents the structure of a higher-order Boltzmann machine. The x, y, and h are fully connected to each other by using a 3 dimensional weight matrix w.

Unsupervised learning is used to learn the commonly occurring shapes on the road. The patches for the training data are extracted by sliding a mask on the sample images (Figure 2-a).

3.2 Online Inpainting

After the offline learning of the weight tensors $w^x_{if}, w^y_{jf}, and w^h_{kf}$, the system can be used to inpaint missing parts of a road image. To infer the missing part y of an image, x and h must be known. While x is simply the surrounding region around y, there are several possible sources to compute h. Generally speaking h can either be obtained using the current image or previous images from memory.

When the current image is used to obtain the transformation $p(h_n|y_n;x_n)$ (n represents the video frame number) and then reconstruct the output $p(y_n|h_n;x_n)$, the system behaves like an autoencoder that tries to generate an image by reconstructing the patch using the learned features of the image. However, in this case if the anomaly covers a significant part of y_n the system will encode the wrong transformation and will ultimately inpaint an incorrect result.

Another more practical scenario is to keep a short term memory of M previously observed frames. In the n^{th} frame we can run the mask over the region of interest. The x_n of the n^{th} image frame is matched with the memory image database to find the most similar image x^s and its respective y^s. In the next step the transformation is inferred $p(h^s|y^s;x^s)$. The estimated transformation h_s can then be used to inpaint the desired region of the n^{th} frame $(p(y_n|h^s;x_n))$.

As mentioned in the introduction, the anomaly is easiest to detect when it is observed for the first time. Therefore the patch that contains an anomaly with a certain probability P_o should not be stored in the short-term image memory.

4 Results

For the training of the system we used 100,000 unique samples (with each one being a x,y pair) extracted from 2,000 VGA images. All 2,000 images represented empty roads with no vehicles. A mask was applied to the original images to remove the non-road regions. The size of x is 50 × 50, while the size of y is 15 × 15. The number of factors in our FGBM was 200 and the size of hidden layer (transformation layer) was also 200. The offline training was performed until the weights converged and did not change for 20 consecutive epochs. With this termination criteria the training stopped after 243 epochs.

In the first experiment, we compare the quality of inpainting by comparing it with the nearest-neighbor approach. As discussed in the introduction, a system with infinite memory and computational capabilities can reconstruct a patch by just recalling the nearest neighbor in the memory. However, the proposed system should be able to reconstruct the frequently occurring road patterns by using a very small amount of memory. The comparison of the proposed approach with nearest-neighbor approach is given in Figure 3. The upper-left image in the figure shows the original road image (black represents the road while white represents a lane marking or expansion divider). We can see a comparison of the proposed technique with the nearest-neighbor approach. The Mean Square Error (MSE) for reconstructing the patterns is 0.915 for the proposed approach when only the previous image is kept in memory while with nearest-neighbors even using last 20 images gives MSE of 4.524. The plot in the lower-right corner shows MSE comparisons for the two approaches. Even with a memory of 300 previous images the nearest-neighbor approach's performance does not come close to the presented approach.

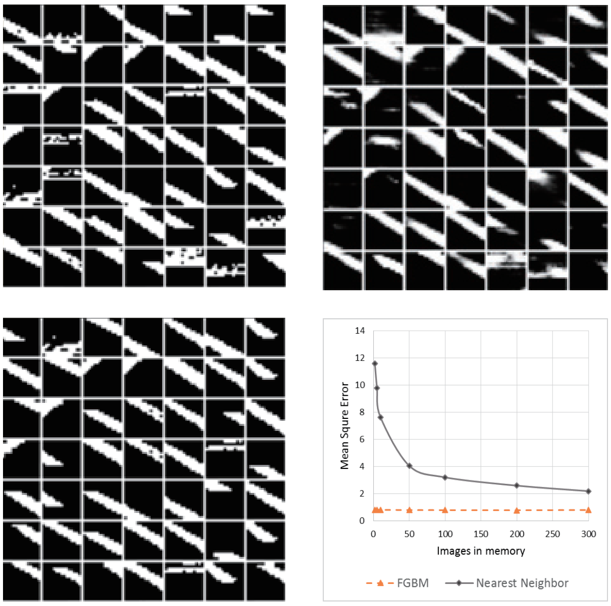
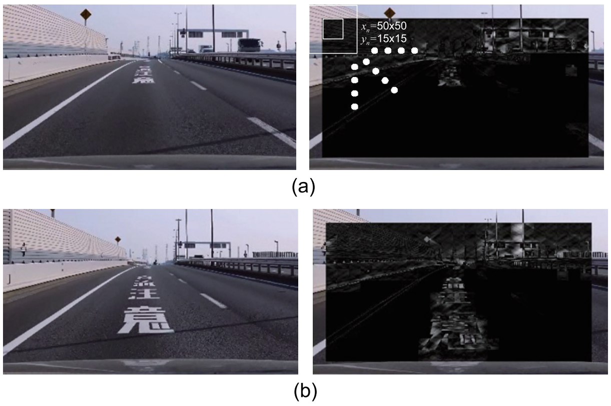


Figure 3. (upper-left) A sample of original output patches Y. (upper-right) Reconstructed output patches by using FGBM with a memory of only 1 previous image - MSE=0.915. (lower-left) Nearest-neighbor in the last 20 images - MSE=4.524. (lower-right) Reconstruction error for FRGBM and nearest neighbor approach. Even with a memory of 300 previous images the nearest neighbor approach cannot compete with reconstruction using only 2 previous frames by FGBM.



(a)

(b)

Figure 4. Left column shows the original image of a road scene. Right column images show the difference between the original and the reconstructed (inpainted) image.

Some of the results obtained by the technique are shown in Figure 4. In this case the proposed system is used as an autoencoder where the same frame is used to determine the transformation h_n between x_n and

y_n. The encoded transformation h_n is used to inpaint y'_n. The absolute difference $|y_n - y'_n|$ is then computed to highlight the anomalies. The right column side images of Figure 4 show in black the regions detected as road. Note that the lane markings, which are a standard feature of the road, are reconstructed and therefore appears black in the difference image. In contrast, the text written on the road which is relatively rare, was not reconstructed properly and appears white in the difference image.

Figure 5 shows another snapshot from a video with a small dummy obstacle added to the image. For this experiment we used the short-term memory of size one. This means the system only stores the last observed image to find the similarity transformation h^s. As shown, the system is able to detect the obstacle on the road as an anomaly. Note that the common road features like lane markings are not detected as anomalies. Some other noise that is produced by the patch reconstruction is removed by simple erosion and dilation operation.

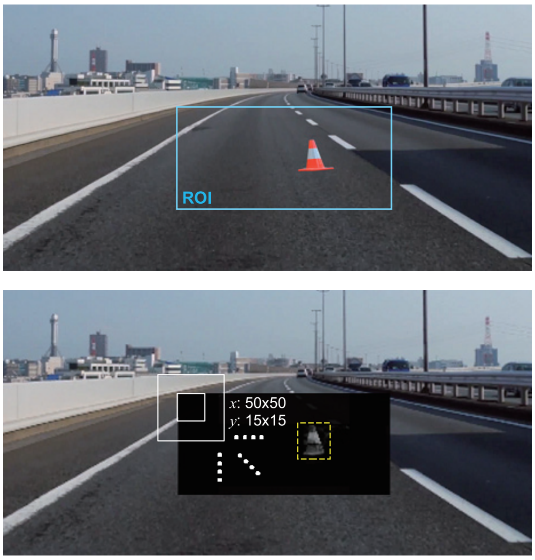


Figure 5. An Region of Interest (ROI) is defined for anomaly detection. The anomaly is detected and is shown by a dotted bounding box in the lower image.

5 Related Work

Significant amounts of work have been done in the field of obstacle detection on roads. Although active sensors are widely used for obstacle detection, such devices are often expensive and have very low resolutions at long distances, making them impractical for the problem discussed in this paper. In the passive sensors the obstacle detection problem have been researched for over a decade. Several techniques propose using stereo or Structure From Motion (SFM) to find a kind of homography transform between two images and then finding the anomalies by warping one image and comparing it with the other. T. William et al. [7] uses a multi-baseline stereo technique and claims to detect 14 cm obstacles at a distance of over 100 m. Similarly, H. Kyutoku et al. [4] compares the previous frame and the present frame to find any anomalies on the road. Although, such techniques should theoretically detect any obstacle on the road, in practice, these techniques require a very clean road environment with an accurate point matching for image warping and disparity computations. This is not practical for point matching since the real world images can be very noisy.

Many other systems commonly used in the Advanced Drivers Assistance System (ADAS) use a monocular camera [8] or stereo-based vision [1] to detect large obstacles very robustly, but fail to detect small obstacles at medium or long distance.

6 Conclusions

In this paper we present a technique that can inpaint lane markings and other commonly occurring features of the road. When compared with the actual observations, this technique can be used to detect anomalies on the road. During the offline learning stage the system learns the shapes of common structures on the road. The system cannot inpaint the anomalies that may occur on the road, hence considerably reducing the candidates for the anomalies.

One very obvious limitation of the system is that it will fail to detect an anomaly that looks like a frequently occurring road feature. In such a case, contextual information about the road must be used to identify anomalies. Although the proposed technique would work on most roads, highway environments are better maintained with fewer obstacles, hence making them a better choice for initial testing.

References

[1] Subaru eyesight: Driver assist technology - *http://www.subaru.com/engineering/eyesight.html*.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):743–761, April 2012.

[3] Tobias Egner, Jim M. Monti, and Christopher Summerfield. Expectation and surprise determine neural population responses in the ventral visual stream. *The Journal of Neuroscience*, 30(49), December 2010.

[4] Haruya Kyutoku, Daisuke Deguchi, Tomokazu Takahashi, Yoshito Mekada, Ichiro Ide, and Hiroshi Murase. On-road obstacle detection by comparing present and past in-vehicle camera images. In *Conference on Machine Vision Applications*, Nara, Japan, June 2011.

[5] Roland Memisevic and Geoffrey Hinton. Unsupervised learning of image transformations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[6] Roland Memisevic and Geoffrey E. Hinton. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Computation*, 22(6), June 2010.

[7] Todd Williamson and Charles Thorpe. Detection of small obstacles at long range using multibaseline stereo. In *IEEE International Conference on Intelligent Vehicles*, 1998.

[8] David B Yoffie. Mobileye: The future of driverless cars. Harvard Business School Case 715-421, October 2014.