Automated risk assessment for scene understanding and domestic robots using RGB-D data and 2.5D CNNs at a patch level

Rob Dupre
Kingston University

Georgios Tzimiropoulos
Nottingham University

Vasileios Argyriou
Kingston University
Vasileios.Argyriou@kingston.ac.uk

Abstract

In this work the notion of automated risk assessment for 3D scenes is addressed. Using deep learning techniques smart enabled homes and domestic robots can be equipped with the functionality to detect, draw attention to, or mitigate hazards in a given scene. We extend an existing risk estimation framework that incorporates physics and shape descriptors by introducing a novel CNN architecture allowing risk detection at a patch level. Analysis is conducted on RGB-D data and is performed on a frame by frame basis, requiring no temporal information between frames.

1. Introduction

Scene analysis is a research topic covering a large range of topics, with applications in traffic analysis, domestic robotics, smart homes. The concept of scene analysis with regard to risk is a developing area of research. The ability to provide smart homes or domestic robotics with the capabilities to define and localise hazards in a scene is highly advantageous. This has applications in both the social and child care sectors as well as for assisted living.

Traditionally, the notion of risk detection focused on those that use the environment not the environment itself, for example gait analysis for indoor fall assessment for elderly adults [1]. Currently risk analysis research focuses on the definition of hazard detection mechanisms which allow the quantification of measurable elements of risk into some sort of coherent risk score. These detection mechanisms can be then combined in a framework to build a context-depended picture of risk for a scene made up of a number of elements [2]. These elements can include any form or measurable risk in a scene; in *et al*. [10, 5, 9, 12] the authors look at the concept of stability as a form of risk by modelling the amount of force required to dislodge an object. This is combined with an analysis of human interaction within a scene to give a risk score for each object.

Recently, much work in scene understanding has been conducted using Deep Learning techniques. In [11, 3] authors apply CNNs and learn deep features for identification of places with the addition of context. Socher *et al*. [8] introduce a convolutional-recursive deep learning model allowing the classification of RGB-D images without the need of additional input channels. Huang and Suya [4] address the task of semantic labeling in point cloud data by using voxel-based full 3D CNN's instead of 2D techniques which can lead to loss of structural information beneficial to the labeling process. Qi *et al*. [7] also look at the issue of semantic labeling and clustering in point cloud data proposing a novel deep net architecture for 3D shape classification, segmentation and semantic analysis.

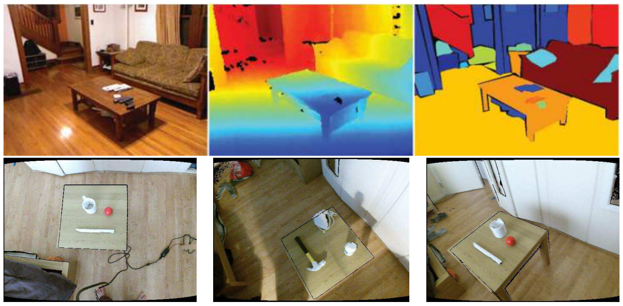

Figure 1: Example RGB-D frames from the NYU Depth Dataset V2, [6] and the 3DRS data set [2].

The following sections outline the proposed methodology for risk assessment of 3D scenes using multi-scale deep and then shallow CNNs networks. Here the goal is the identification of areas within an RGB-D frame that could correspond to a hazardous patch of an object. The classification to safe and risky areas is at a patch level with the term parch to represent a low-level semantic patch into the domain of a 3D scene and it is similar as the concept of superpixels.

2. Methodology

Within this section the proposed approach to risk based semantic labeling and scene understanding is reviewed. In this work an extension of the proposed framework in [2] is

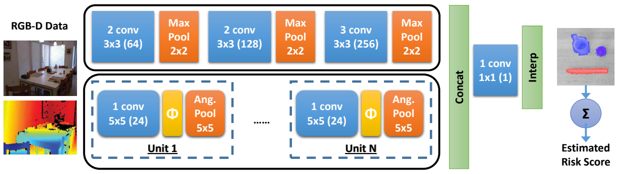

Figure 2: Overview of the proposed architecture for risk estimation.

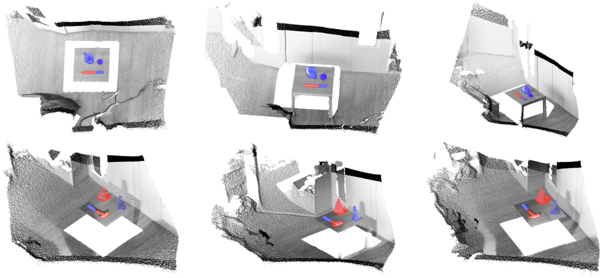

Figure 3: Example of patch based estimates with red colored areas to indicate high level of risk.

introduced considering a novel CNN architecture in order to produce patch based labeling of hazardous or safe areas. The Risk Estimation framework in [2] uses a component-based architecture, incorporating physics and shape based descriptors with supervised boosting techniques. In this work, the shape component and the boosting approach are replaced with an advanced CNN architecture.

We describe a novel approach which bypasses many of the difficulties encountered in scene understanding for risk evaluation by using a 2.5D representation of the 3D scene geometry, and an appropriate CNN architecture that is trained to detect high risk areas and object patches in a scene. An overview of our method is shown in Fig 2. Our architecture is based on two CNN networks, a deep one to distinguish shape-based risks on a fine grade and shallow one to detect hazardous patches in objects (e.g. blade of a knife). The multi-scale deep network that first estimates a global output based on the entire RGB-D image, and then refines it using a shallow CNN that consists of maximum two basic units placed in a sequential configuration. We call this unit-CNN and each one consists of a spatial convolutional layer, an activation function and a pooling layer.

3. Evaluation

In our evaluation process two datasets were used, the 3DRS (fig 1) and applicable scenes from the NYU Depth Dataset V2, (fig 1). From the obtained results the proposed approach improved the accuracy and the overall F1-score more than 5.2% with a 14.44% error drop. An example of the patch based estimates is shown in fig 3.

4. Conclusions

A novel CNN architecture integrated on an existing Risk Estimation framework has been created focusing on detecting high rich areas and patches in 3D scenes combining a deep and a shallow network to distinguish shape based risks on a global output and to detect hazardous patches, respectively. The overall results obtained from two datasets, demonstrate that proposed CNN architecture outperforms the classic shape descriptors and boosting.

References

[1] A. N. Belbachir, A. Nowakowska, S. Schraml, G. Wiesmann, and R. Sablatnig. Event-driven feature analysis in a 4D spatiotemporal representation for ambient assisted living. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1570–1577, nov 2011.

[2] R. Dupre, V. Argyriou, G. Tzimiropoulos, and D. Greenhill. Risk analysis for smart homes and domestic robots using robust shape and physics descriptors , and complex boosting techniques. *Information Sciences*, 372:359–379, 2016.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, Jan 2016.

[4] S. Y. J. Huang. Point Cloud Labeling using 3D Convolutional Neural Network. In *International Conference on Pattern Recognition*, pages 1–6, 2016.

[5] A. Lerer, S. Gross, and R. Fergus. Learning physical intuition of block towers by example. *CoRR*, 01312, 2016.

[6] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[7] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 1578–1584, 2016.

[8] R. Socher, B. Huval, B. Bhat, C. D. Manning, and A. Y. Ng. Convolutional-Recursive Deep Learning for 3D Object Classification. *NIPS*, 3(7):8, 2012.

[9] R. Zhang, J. Wu, C. Zhang, W. Freeman, and J. Tenenbaum. A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. *CoRR*, 01138, 2016.

[10] B. Zheng, Y. Zhao, J. Yu, K. Ikeuchi, and S. C. Zhu. Scene understanding by reasoning stability and safety. *Intern Journal of Computer Vision*, 112(2):221–238, 2015.

[11] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. *Learning Deep Features for Scene Recognition using Places Database*. Curran Associates, Inc., 2014.

[12] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S. C. Zhu. Inferring forces and learning human utilities from videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3823–3833, June 2016.