A 3D Scene Analysis Framework and Descriptors for Risk Evaluation

R. Dupre V. Argyriou D. Greenhill G. Tzimiropoulos

Abstract

In this paper, we evaluate the notion of scene analysis with regard to risk. We consider the problem of evaluating risk and potential hazards in an environment and providing a quantified risk score. A definition of risk is given incorporating two elements; Firstly scene stability, where Newtonian Physics are introduced into the scene analysis process, evaluating object stability within a scene. The effectiveness of which is demonstrated by conducting experiments on several scenes including a variety of stability levels. Secondly the analysis of the intrinsic risk related properties of an object such as sharpness. This being estimated using learning techniques and the introduction of the 3D Voxel HOG descriptor, analysed against the state-of-the-art descriptors. Finally a new dataset is provided that is designed for scene analysis focusing on risk evaluation.

Figure 1: Example Scenes of objects with a variety of (top) intrinsic properties (e.g. sharp, pointed) and (bottom) levels of stability.

1 Introduction

Scene analysis is a research area covering a large range of topics with applications in navigation systems [28], traffic analysis [4], domestic robotics [32], and smart homes [5] amongst many others. In this work we consider one such topic; scene analysis with regard to risk. With the aim of providing a quantified risk score for a given three dimensional scene. To achieve this a system is proposed that will, through a combination of novel feature selection mechanisms and Newtonian physics, analyse the potential risk in a given 3D scene.

The proposed system focuses on two concepts; firstly object and scene stability, derived from the resultant energy outputs due to an applied force, as well as its subsequent effect on other objects within a scene. This is analysed using a novel combination of vision based and physics simulation techniques. As an example consider the difference between the case of a glass bottle placed at the corner of a table against it being placed at the centre (see figure 1).

Secondly, determining the intrinsic properties of an object may add to the prospective hazard. To formulate this definition a novel voxel based three dimensional descriptor has been introduced that is based on the principles of Histogram of Oriented Gradients. The proposed 3D Voxel HOG (3D VHOG) descriptor tries to identify dangerous elements or characteristics of an object (e.g. 'hazard features'). When combined with a boosting technique (Ad-

aboost [11]), the resultant model aims to specify whether an object affects the potential risk in a given 3D scene (Figure 1). Importantly object recognition is not the goal allowing the proposed approach to be more general and operate at a lower level. In object recognition a 'feature' is defined in terms of a specific structure in the data. Here the term 'feature' relates to the actual physical properties of an object. We define 'hazard features' as any physical property of an object present in a given 3D scene that could increase risk (e.g. a knife's blade being sharp, pointed).

In this paper we evaluate the concept of risk estimation in static and dynamic scenes by combining the novel use of Newtonian physics mechanics and the introduction of a new feature descriptor. We define risk as a function of scene stability and the intrinsic properties of the present objects. Furthermore a novel dataset for 3D scene analysis was collected and will be available online.

The paper will continue as follows; in section 2 an analysis of the similar areas of research and related work. The proposed methodologies and processes used in the work will be presented in section 3. Section 4 will outline the experiment environments and analyse the results. Finally, in section 5 conclusions are drawn.

2 Related Work

2.1 Scene Analysis and Risk Assessment

To date very little research has been done in automated risk analysis systems. [31, 2] analyse indoor fall assessment for elderly adults; however in both proposed methods focus is given to analysing the person not the risk of the environment. [37] introduces the notion of analysing the fall potential of objects in a scene given the influence of environmental events such as human intervention or earthquakes. However the 'hazard features' of the objects are not analysed nor is the effect that the objects might have on each other.

Another emerging area of research within 3D scene analysis relates to Volumetric Reasoning. Here the application of logic based algorithms to existing object clusters is used to improve segmentation and clustering accuracy. [38] utilises the notion that clusters in a scene should be in a state of rest when simulation techniques are applied. Thus by using an iterative process clusters are grouped until such time as the scene is at equilibrium. [14] proposes a method that better fits bounding shapes to RGB-D clusters based on the premise that a good 3D representation of a scene is stable, fits the data well and is self-supporting. Battaglia [1] introduces the idea of a 'intuitive physics engine (IPE)' that tries to mimic a humans cognitive simulation process when analysing a scene.

It is worth mentioning the following papers that consider similar concepts and approaches for scene analysis [21, 10, 17, 19, 20, 8, 18]. Although the concept of risk evaluation is raised in some of this work, an automated form of risk evaluation for a given scene is not fully addressed.

2.2 Object Retrieval and Scene Descriptors

Object retrieval and feature selection are research subjects that have received a huge amount of work in recent years both in the 2D and 3D domains. The initial concept of HOG features [6] revolutionized the 2D object recognition world by creating a local descriptor that was resistant to geometric and photometric changes. Onishi [23] uses HOG on images from a monocular camera system to recognise human posture in a scene and estimate, using regression, the joint angles on a 3D human model. Buch in [3] implements a vehicle recognition framework, using a patch definition system on a 3D representation of the found vehicle, combined with a traditional two dimensional implementation of the HOG descriptor.

With the introduction of financially viable 3D depth camera hardware, such as the Microsoft Kinect [29], more research has been focusing on the 3D domain. Transfering Dalal and Triggs work into three dimensions, Scherer [26] performs gradient computation in 3D using a convoluted distance field. This provides an effective way of calculating the magnitudes of the gradients, scoring them highly when localised near a surface of a model (local maxima), however their method also scores highly those at local minima creating additional artifacts within the data. As such this particular implementation is unsuitable for local feature recognition. Another example that uses a variation of vectors within a histogram as a feature is [33]. Here the normal vectors are used as the feature to define an object. An alternative method in which HOG is extended into a third dimension is presented by Klaser

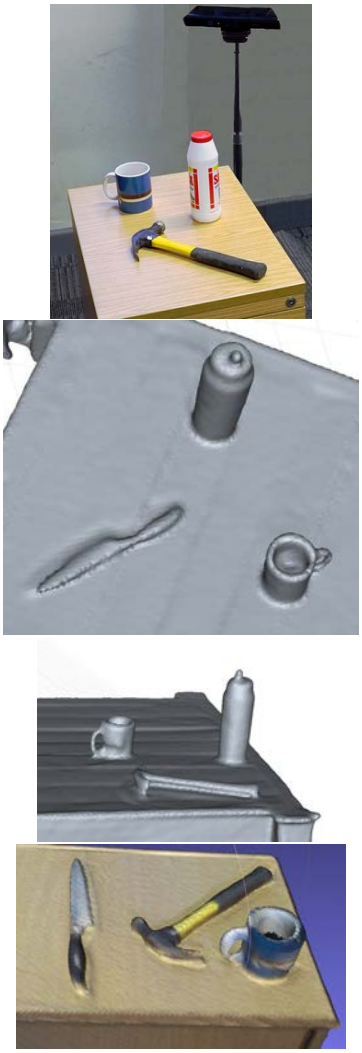


Figure 2: Example of the acquisition process using Kinect Fusion and some of the obtained 3D scenes.

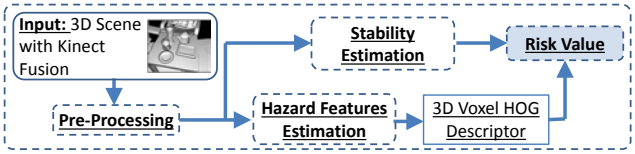


Figure 3: An overview of the proposed risk assessment framework.

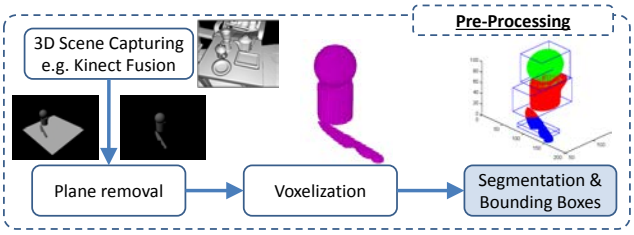


Figure 4: An overview of the pre processing stage.

[15, 24]. Here a method is proposed that tracks people and identifies their actions through a video sequence. They implement and then extend HOG through use of time as the third dimension. This approach is based on intensity gradients without taking into account concepts related to the density of an area and therefore is not an appropriate descriptor for local feature classification. Additionally, it is worth mentioning the following state of the art 3D descriptors [12, 27, 30, 9].

3 Proposed Methodology

The following section will discuss in detail the proposed methodologies used to define a weighted risk estimation model, and how the parameters in that model are established from a 3D scene. An overview of this framework is shown in Figure 3.

3.1 Pre Processing

Before the risk in a scene can be evaluated, pre processing steps are required to convert the input data, a mesh model of the scene, into a usable format (Figure 4). The scene data and 3D mesh model reconstruction of an environment is acquired using Kinect Fusion [13] (Figure 2). Alternatively other multi-camera acquisition systems [35] or sensors (e.g. thermal or acoustic cameras) could be used. The surface on which the objects are set requires removal, the work in [34] provides solutions for these problems.

Voxels are defined along the faces of the scene's 3D mesh, voxels enclosed within a mesh are also classified as part of the scene allowing us to consider features based on an objects' density. This resultant volume represents a binary classification of either object or not. With this classification in place, clustering of the voxel volume can be applied grouping voxels into object clusters. A num-

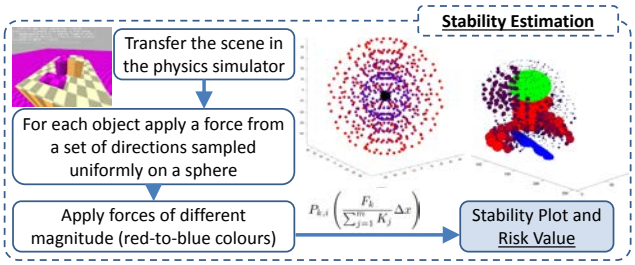


Figure 5: The proposed stability estimation mechanism.

ber of different clustering algorithms were tested, using modified versions of the work presented in [34, 7] while a range of other clustering algorithms were also considered. A bounding box for each object cluster is defined (Figure 4 right), with the number of voxels within that bounding box counted and used as a rudimentary measure of mass. For the purposes of physics simulation, bounding boxes must not intersect; as such a recursive reduction process is applied reducing bounding boxes until no overlap is detected.

3.2 The Risk Estimation Framework

Let us define the cumulative risk score R for a scene as the weighted sum of n risk elements E.

$$R = \sum_{i=1}^{n} w_i E_i \tag{1}$$

A risk element E is a measure of risk. In this work these include stability S and hazard features H. Other risk features obtained from related acquisition devices, such as thermal cameras, can also be utilised. This ensures the proposed framework is extendable as required.

For the purpose of this paper we define the cumulative risk score R as a function of the weighted elements, stability S and hazard features H.

$$R = f(w_S S, w_H H) \tag{2}$$

3.3 Proposed Stability Estimation Mechanism

The proposed novel methodology for scene stability estimation is based on Newtonian physics. To evaluate stability, forces are applied to objects within the scene and the resultant outputs measured (Figure 5). Statistical analysis on those outputs can be performed providing information about the properties of the scene. Consequently, allowing us to model the behavior of segmented objects and compute the energy output from the applied forces.

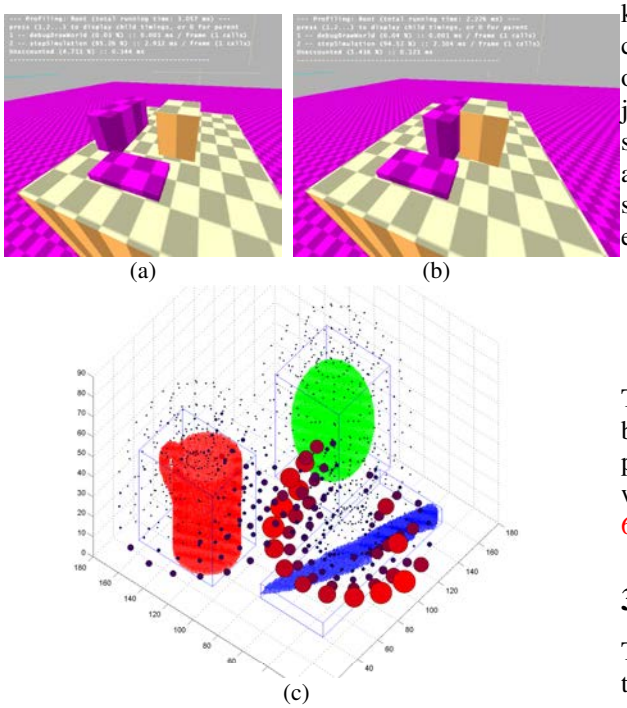
Using 'collision shapes' a 3D model can be redefined into a simplistic form, reducing the computational power needed to emulate its behaviour during simulation. Attached to these collision shapes are parameters such as position, size, mass, friction and angular dampening coefficients. These parameters define the inputs required for the simulation and therefore what information must be extracted from a scene. The bounding shape calculated from an object cluster serves as the guidelines for the collision shape. Global fixed parameters are defined for the friction and angular dampening coefficients based on existing models. The mass is estimated based on the number of voxels each object is made up of. Estimation of an objects material would provide a better approximation of mass but is beyond the scope of this work, however methods based on the estimation of an object's BRDF function could be utilised, [36, 16]. The proposed framework supports multiple acquisition devices but in this paper Kinect Fusion is utilised. The limitation of this acquisition device is that it cannot identify solid and non-solid objects, considering both as solid (e.g. tennis and golf balls). To overcome these limitations and provide further accuracy in the simulation process acoustic or thermal cameras could be utilised.

Stability s for a force k on a given object i is considered as a dimensionless quantity and is defined as the ratio of the applied force F_k over the summed kinetic energy K_j for all the objects m in the scene. This is scaled by the probability $P_{k,i}$ of the force being applied.

$$s_{k,i} = P_{k,i}\left(\frac{F_k}{\sum_{j=1}^{m} K_j}\Delta x\right) \tag{3}$$

where $K_j = \sum_{t=1}^{T} \frac{1}{2} M {V_t}^2$ represents the accumulated kinetic energy of the object j over time T from a simulation obtained using numerical integration. Here M represents mass and V the velocity of the object j at a given time t. Δx is the object's displacement, but since the kinetic energy is calculated numerically over fixed length intervals, this value is equal to one.

Probability $P_{k,i}$ represents the likelihood of a given



(a) (b) (c)

Figure 6: Stability evaluation process using Newtonian Physics (a) Initial layout in the physics simulation (b) Collision occurring during the simulation and (c) Stability Plot with the circles around the objects indicating the direction of instability and the radius corresponding to the severity.

force F_k being applied to object i. This is defined as whether the force could collide with the object without hitting first another entity within the scene. For example forces from below an object on a plane would collide with the surface first therefore would not be considered.

Forces F of different strengths are applied to the center of each bounding box (object) during the simulation, with directions sampled from a uniform distribution of angles over a sphere. The resultant overall kinetic energy K for each object j is calculated. By analysing the amount of kinetic energy produced by each object for each force we can ascertain if, during the course of that simulation, an object has been dislodged from the surface or if other objects within a scene have been affected. By varying the strength of force we build up a picture of how unstable an object is in its environment. The total stability S of a scene is given as the sum of the estimated stability s for each force k applied to each object j.

$$S = \sum_{k=1}^{r} \sum_{j=1}^{m} s_{k,j} \tag{4}$$

The outcome of this process will allow us to differentiate between the case of an object (e.g. glass bottle) being placed at the center of a table or at the edge, evaluating with enough precision the stability of each scene, (Figure 6).

3.4 Novel Hazard Feature Descriptor

The application of three dimensional descriptors to identify the properties of an object is a new concept. In the proposed framework rather than focusing on object recognition, the detection of hazard features is the core vision problem. This introduces the novel classification task of recognising sharp and pointed areas in a scene. The novel 3D Voxel HOG descriptor is introduced, which is specifically designed to be suitable for local feature recognition whilst also considering an objects' density. An overview of the proposed approach is shown in Figure 7.

The traditional HOG uses the normalized combination of gradient vectors from a given number of pixels to build up a histogram of binned angles that relate to the feature. We extend this process to the third dimension though the use of voxels and 2D histograms. The process begins by breaking the voxel volume up into set features spaces f

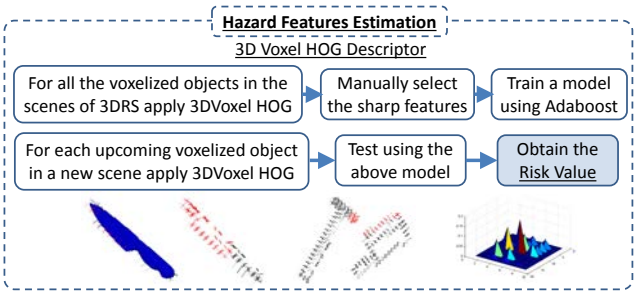


Figure 7: The proposed hazard classification system.

comprised of a number of cubic 3D cells c, which in turn is made up of voxels v. For each voxel within a cell the filter mask $[-1, 0, 1]$ is applied to its neighbouring voxels in all three dimensions giving us the gradient vector $\vec{g}$.

The magnitude $\|\vec{g}\|$ of the gradient vector is obtained and then its orientation is expressed using the azimuth θ and zenith ϕ angles.

$$(\theta, \phi) = (\cos^{-1}(\frac{g_z}{\sqrt{g_x^2 + g_y^2 + g_z^2}}), \tan^{-1}(g_y, g_x)) \quad (5)$$

Additionally a weight w is defined for each voxel, which is used to scale its contribution to its cell's 2D histogram. This is given by the mean value of the voxels within a given three dimensional kernel indicating the density over this area. By applying this weight, the proposed approach provides accurate estimates even in the presence of noise.

Once these values are established the voxels within each cell are binned into a 2D histogram h according to their θ and ϕ angles. The value added to the specified bin is given as the weighted magnitude of the vector $w\|\vec{g}\|$. Finally all the histograms for each cell within a feature h_f are normalised using the L_2 norm.

$$h_f \rightarrow \frac{h_f}{\sqrt{\|\vec{g}\|_2^2 + e^2}} \quad (6)$$

The obtained features are then vectorised and used by the learning mechanism to create a classification model.

$$\vec{x}^{3DVHOG} = \{h_{1,1}, ..., h_{1,\varphi}, ..., h_{\theta,\varphi}\} \quad (7)$$

The resultant 2D histograms can be visualised, and present a way of identifying different types of features within an object (Figure 8c). Another form of visualisation plots each possible gradient vector within local 3D

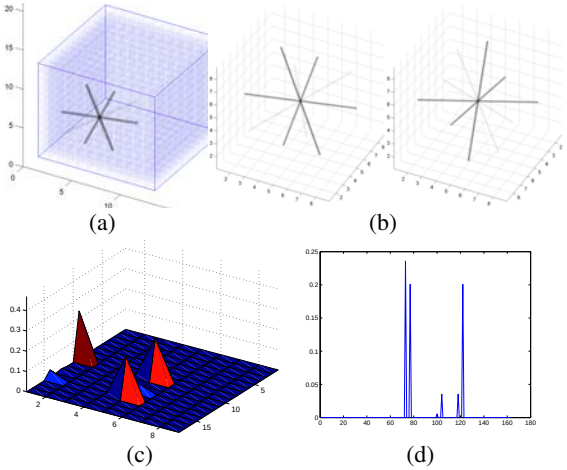


Figure 8: 3D Voxel HOG feature from a cube wall test case, (a) visualised on its object in 3D, (b) the same 3D representation in two different orientations, (c) as a 2D Histogram and (d) as a 162 dimension feature vector

histograms, showing the most common gradient vectors as stronger (see figure 8a,b).

Training is then carried out to create a model to classify safe and unsafe local features. The defined features of an unknown object can then be tested against this model and return a binary classification for each feature as either hazardous or not. This data then forms the hazard component of the Risk score.

One of the primary advantages of the proposed 3D Voxel HOG (3D VHOG) is the consideration of not only the faces of a mesh but also the area within as well. This ensures that no additional artifacts are created within the data that may lead to false classifications, additionally the density of an object is also taken into consideration (e.g. empty and full cup). This allows transference of the methodology to other areas such as medical imaging, for example the proposed method could detect defects such as osteoporosis in bone MRI scans which existing methods would not. A visual comparison of the 3D HOG features suggested in [3, 26] against 3D VHOG is shown in figure 9 indicating that the proposed method does not introduce erroneous information in the internal areas of an object. Importantly 3D VHOG returns one 2D histogram (visualised in 3D) per cell (Figure 8), as apposed to the other methods that provide multiple 1D histograms (visualised in 2D).

The use of voxel weighting smooths the edges of an ob-

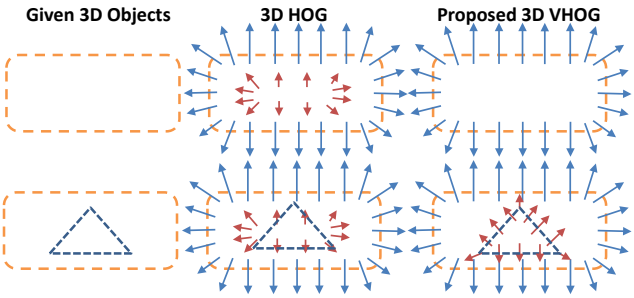


Figure 9: Example showing the differences of the proposed 3D Voxel HOG features with the 3D HOG [26] indicating that the objects' internal density affects the proposed 3D VHOG descriptor.

ject cluster ensuring robustness against noisy input data. Due to the local nature of the proposed feature, issues related to the normalization of a mesh are avoided, removing a potentially complex pre-processing step.

The pseudo code for the 3D Voxel HOG implementation is outlined below.

```
1. choose Size of Cell and Block
2. FOREACH Voxel v DO
3.   compute Weight w, Gradient (θ, φ),
       Vector Magnitude ||g||
4. FOREACH Cell c in Feature f DO
5.   create blockHist(theta_bins, phi_bins)
6.   FOREACH voxel v in c DO
7.     insert w||g|| into blockHist (θ, φ)
8. L2Normalize(blockHist in Feature)
```

These features are used to create a trained model that unknown features can be tested against. A binary classification is returned defining the object as either being hazardous or not. Adaboost is a learning technique that creates a non linear classifier to separate data into two groups. Weak classifiers are defined with a final strong classifier being a combination of these. At each iteration the weak classifiers with the lowest error margin are used to define the next in a 'greedy fashion'. Regarding the proposed features in both cases given N training examples $(\vec{x}_1, ..., \vec{x}_N)$, the corresponding labels $(y_1, ..., y_N)$ with $y_i \in \{-1, 1\}$, and an initial distribution of weights $W_1(i)$ a strong classification model $H(x)$ is obtained based on the weak classifiers h. The weak classifiers are trained over a number of iterations Q using the weights' distribution W_t. In each iteration the error ϵ_t is estimated based on the current weights W_t, that are updated before the next iteration.

$$W_{t+1}(i) = W_t(i) \exp(-a_t y_i h_t(x_i))/Z_t \quad (8)$$

where $a_t = -\frac{1}{2}\log(\epsilon_t/(1-\epsilon_t))$ and $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$ is a normalization factor. The strong classifier is defined as $H(x) = sign(f(x))$, where $f(x) = \frac{\vec{a}\cdot\vec{h}(x)}{\|\vec{a}\|_1}$.

Regarding the boosting approach, because of the way weak classifiers are selected a complicated feature problem can be broken down and classified using a sparse classification rule, based on only a few features. This makes computation much faster as only a subset of the features are used. This is essential if the methodology is to be implemented in a real time scenario. Another advantage of this approach is the explicit minimisation of error, whilst implicitly maximising the margin. This ensures the final strong classifier is general avoiding the problems of overfitting. Another similar boosting technique uses Support Vector Machines(SVM). This also provides a non linear, robust classifier, however tends to have higher computational requirements. This is due to their classifier taking into account all the features in a vector as apposed to a subset [22].

Finally, in order to define a second element E of the risk score R in equation (1) related to the 'hazard features' the obtained outcomes from the classification process above are utilised.

$$H^{3DVHOG} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{k=1}^{M} w_H G(j,k)}{\sum_{k=1}^{M} G(j,k)}\right)$$

$$H^{\omega} = \frac{1}{m}\sum_{j=1}^{m} w_D(j) \quad (9)$$

where $w_D = f(x)$ normalised and $G = \frac{1}{2}(sign(f(x)) + 1)$.

4 Results

4.1 Experiment Environment

In our experiments, scenes containing mainly household objects and toys placed on a surface were utilised. In order to obtain a ground truth for each scene and to ensure

(a) (b) (c)

Figure 10: An example scenario with each of its iterations. The level of complexity is increased from a simple layout(a) to a complex (c).

that the parameters of the tests are fully controllable and repeatable the objects were manually placed in specific locations on the given surface. To effectively test this problem and as no existing dataset fits the proposed work, a new dataset for risk estimation in 3D scenes (3DRS) was created comprising of 36 scenes captured using RGBD cameras. These 36 scenes are split into 12 different scenarios, each containing 3 objects. Each of these has 3 stability levels in terms of scene complexity, i.e. the objects are moved closer together on the plain (Figure 10). These include examples of 'hazardous' (scissors, stanley (open), screwdriver, plane, pencil, pen knife (open), knife, fountain pen, fork, cleaver, ballpoint pen, axe) and safe (vase, stanley (closed), spoon, rubix cube, pen knife (closed), mug, mouse, laptop, lamp, frame, bowl, bottle, ball) objects, with multiple instances for most of them. Additionally as an object's material is not defined in this dataset; it is given that the objects within a scene are made from the same material and have the same friction (1) and angular dampening coefficients (0.4). The 3D reconstructed models were filtered to remove some noise but also to close any gaps in the mesh since this is essential for the voxelization stage. Regarding the voxelization process of the 3DRS dataset, a set resolution of 256 cubic voxels was defined for the voxel volume. The resolution has a direct impact on computation time for each stage and as such this represents a reasonable trade off for processing time against object detail. For segmentation the the Mean Shift algorithm was found to be the most efficient at separating the objects across all the complexity levels.

4.2 Scene Stability

To demonstrate this concept, initially 3 experiments were conducted in which an example bounding shape is passed to the physics simulation and the resultant stability visualised, (Figure 11). The simulation software employed is

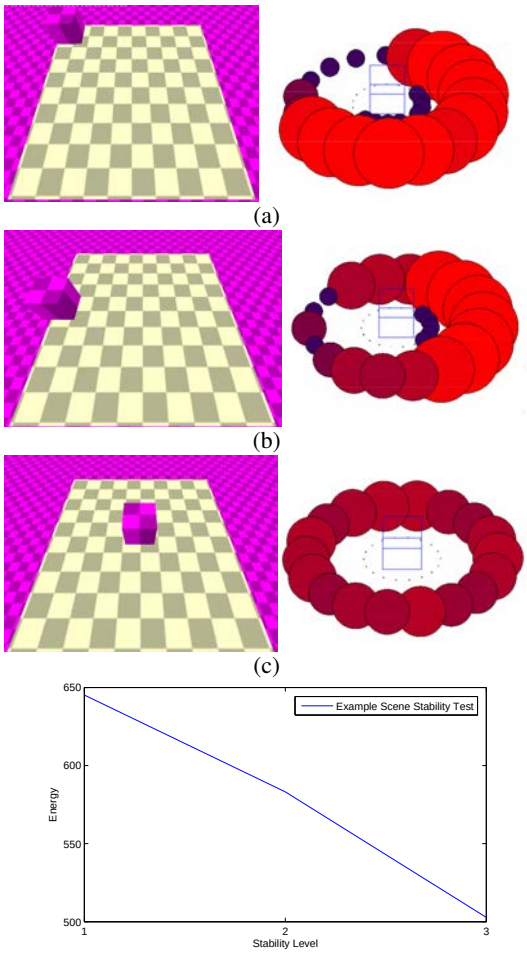


Figure 11: Results visualised using spheres placed around the object indicating the direction and the level of instability in case (a) Far Left Corner, (b) Left Side and (c) Centered and (d) Scene energy per stability level in graph form.

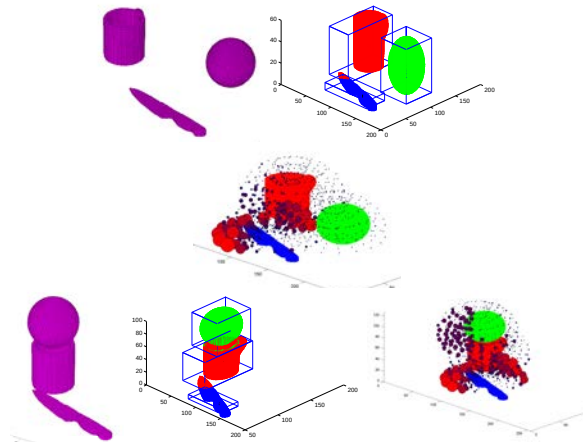


Figure 12: Scenes with the risk and stability levels visualised.

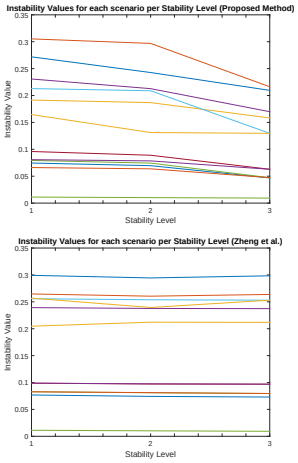


Figure 13: Stability Value for each scenario per Stability Level, (a) Proposed methods, (b) Work presented in [37]. The vertical axis indicates the stability value obtained using equation (4), and the horizontal axis indicates the four different stability levels shown in figure 10. Each of the lines corresponds to one of the scenes. Higher the stability value less stable is the scene.

based on the Bullet 3D Real-Time Multi-physics Library [25]. The output from which is the velocity and angular velocity information for each object at each time frame. In Figure 11 the stability for the object is visualised, position of the spheres represent the source (direction) of the force and their magnitude. Colour and size represent the resultant instability. In this example force was applied from points around the object on a single plane. It can be seen that as the object moves closer to the centre of the surface the energy output decreases, representing an increases in stability (Figure 11)

The stability of each scene within the 3DRS dataset was analysed. For these experiments, force was applied from uniform sampled points along a sphere and the scenes' overall stability quantified according to equations (3), (4). Some stability visualizations are presented in Figure 12. The resultant graph demonstrates that as the objects get closer together and positioned more centrally the scenes risk is reduced (Figure 13). Due to the nature of this type of evaluation, a ground truth is unnecessary as we are quantifying or estimating an attribute of a scene. This allows for good generalisation to other scene scenarios as the technique relies on the measurement of a physics simulation rather than a supervised classification technique.

4.3 3D Voxel HOG Experiments

The second component proposed to evaluate the risk level of a scene relates to an objects' properties or in this case it's 'hazard features'. Here we extract and analyse the novel 3D Voxel HOG descriptor over the 3DRS dataset, conducting comparisons with state of the art 3D descriptors. In the proposed 3D VHOG method a number of variables are defined. Based on experimental results the values were set for feature and cell size; 2 cubic cells and 16 cubic voxels respectively. The bins for the 2D histogram were set at 18 for θ and 9 for ϕ. Each of the preprocessed objects had their 3D VHOG features extracted. As is normal when evaluating a classification based task (e.g face detection/object recognition etc) the ground truth for the hazardous areas was manually labeled. Once established, the histogram data from each feature (8 cells) was arranged into a mean 162 dimension feature vector for training using Adaboost. Testing was carried out based on the 'leave-one out' protocol. The results for all the descriptors are summarised in Table 1. From the results obtained the overall sharp object recognition accuracy was highest for the suggested 3D VHOG method indicating a strong potential of the proposed descriptor for risk estimation over Harris, Sift and the original HOG. In each case the comparison method has been converted to work with a 3D environment. Due to the nature of the 3D Harris operator results across the dataset were inconsistent. The operator classified almost all features as hazardous, highlighting it as ineffective for use in the local space. In

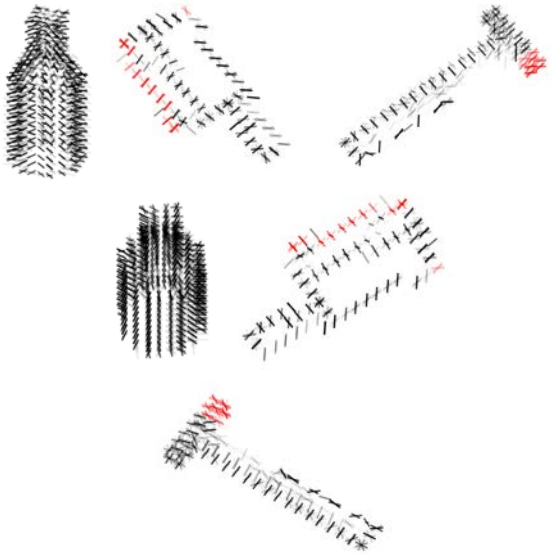


Figure 14: The 3D Voxel HOG visualisation of objects with the classified areas as sharp coloured red.

Feature	F1	Sensitivity	Accuracy
3D HOG [26]	0.533	0.500	0.363
3D Sift [27]	0.333	0.250	0.272
3D Harris [30]	0.200	0.110	0.272
3D VHOG	**0.625**	**0.625**	**0.454**

Table 1: Feature comparison against other existing 3D descriptors.

Figure 14 examples of the 3D VHOG features for some of the objects are shown.

Since the 'hazard features' of the testing objects have been estimated based on the proposed 3D VHOG descriptor and the classification mechanism, the level of risk based on the objects' characteristics is defined using the equation (9). The obtained results for some of the testing objects are shown in Table 2 indicating that the proposed method provides reasonable and accurate estimates.

4.4 Overall scene risk evaluation

An overall confidence (risk) score for each scene is finally estimated combining the previous partial results using equation (1). All the results are shown in table 3.

Object	B	C	F	H	K	M
Hazard Score	0	0.5	0.5	0.6	0.5	0
Object	P	Pl	S	Sd	Bt	
Hazard Score	0.7	0.6	0.3	0.6	0	

Table 2: Hazard scores for the testing objects with higher values indicating higher risk (e.g. presence of sharp features). Some of the objects are listed below: Ball, Cleaver, Fork, Hammer, Knife, Mug, Pencil, Plane, Scissors, Screwdriver, Bottle. Values obtained using Equation (9).

Scenario	1	2	3	4	5
Risk	0.8	0.5	0.6	0.5	0.6
Scenario	6	7	8	9	10
Risk	0.6	0.7	0	0.7	0.8

Table 3: The overall risk value for each one of the scenes. Values obtained using Equation (1) (4) (9).

5 Conclusions

In this work the concept of risk analysis is considered for 3D scenes. A novel approach to evaluating scene stability is given using Newtonian Physics. The 3D Voxel HOG descriptor is introduced and designed to represent the intrinsic properties of an object. When compared with other state of the art features, 3D VHOG provided the highest accuracy in risk detection. Additionally 3D VHOG has the advantages of considering an object's density as well avoiding issues relating to the normalization of a mesh. Furthermore, a new dataset was developed for 3D scene risk analysis and experiments were performed showing that the proposed framework has the potential to accurately measure risks in scenes.

References

[1] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America*, 110(45):18327–32, Nov. 2013. 2

[2] 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

[3] N. Buch, M. Cracknell, J. Orwell, and S. A. Velastin. Vehicle localisation and classification in urban CCTV streams. *Proc. 16th ITS WC*, pages 1–8, 2009. 2, 6

[4] N. Buch, S. A. Velastin, and J. Orwell. A Review of Computer Vision Techniques for the Analysis of Urban Traffic. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):920–939, Sept. 2011. 1

[5] L. Chen, C. D. Nugent, and H. Wang. A Knowledge-Driven Approach to Activity Recognition in Smart Homes. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 24(6):961–974, 2012. 1

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I, pages 886–893, 2005. 2

[7] C. Do and B. Javidi. 3D integral imaging reconstruction of occluded objects using independent component analysis-based K-means clustering. *Display Technology, Journal of*, 6(7):257–262, 2010. 4

[8] D. Engel and C. Curio. Towards robust scene analysis: A versatile mid-level feature framework. *kyb.tuebingen.mpg.de*, page 2009, 2009. 2

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. 3

[10] A. Fossati, H. Grabner, and L. Van Gool. Exploiting Physical Inconsistencies for 3D Scene Understanding. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 136–143. IEEE, 2012. 2

[11] Y. Freund and R. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. *Computational learning theory*, 1995. 2

[12] A. Godil and A. Wagan. Salient local 3D features for 3D shape retrieval. *IS&T/SPIE Electronic Imaging*, (Shrec), 2011. 3

[13] S. Izadi, A. Davison, A. Fitzgibbon, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, and D. Freeman. Kinect Fusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, page 559, 2011. 3

[14] Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 3D-Based Reasoning with Blocks, Support, and Stability. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2013. 2

[15] A. Kläser, M. Marszaek, C. Schmid, and A. Zisserman. Human focused action localization in video. In *Trends and Topics in Computer Vision*, volume 6553 LNCS, pages 219–233, 2012. 3

[16] Y. Kobayashi, T. Morimoto, I. Sato, Y. Mukaigawa, and K. Ikeuchi. BRDF Estimation of Structural Color Object by Using Hyper Spectral Image. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 915–922, Dec. 2013. 4

[17] R. Koch. Dynamic 3-D scene analysis through synthesis feedback control. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(6):556–568, 1993. 2

[18] S. J. Koppal and S. G. Narasimhan. Clustering appearance for scene analysis. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1323–1330. IEEE, 2006. 2

[19] S. J. Koppal and S. G. Narasimhan. Appearance derivatives for isonormal clustering of scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(8):1375–1385, 2009. 2

[20] D. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2143, June 2009. 2

[21] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3D Scene Analysis from a Moving Vehicle. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. 2

[22] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga, and P. M. Thompson. Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. *IEEE transactions on medical imaging*, 29(1):30–43, Jan. 2010. 7

[23] K. Onishi, T. Takiguchi, and Y. Ariki. 3D human posture estimation using the HOG features from monocular image. *2008 19th International Conference on Pattern Recognition*, pages 3–6, 2008. 2

[24] A. Prest, V. Ferrari, and C. Schmid. Explicit modeling of human-object interactions in realistic videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):835–48, Apr. 2013. 3

[25] Real-Time Physics Simulation. Bullet User Manual and API Documentation, 2012. 9

[26] M. Scherer, M. Walter, and T. Schreck. Histograms of oriented gradients for 3d object retrieval. *Proceedings of the WSCG*, 2010. 2, 6, 7, 10

[27] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. *Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07*, (c):357, 2007. 3, 10

[28] C. Sharp, O. Shakernia, and S. Sastry. A vision system for landing an unmanned aerial vehicle. *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*, 2:1720–1727, 2001. 1

[29] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *Cvpr 2011*, pages 1297–1304, June 2011. 2

[30] I. Sipiran and B. Bustos. Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes. *The Visual Computer*, 27(11):963–976, July 2011. 3, 10

[31] E. Stone and M. Skubic. Evaluation of an Inexpensive Depth Camera for Passive In-Home Fall Risk Assessment. *Proceedings of the 5th International ICST Conference on Pervasive Computing Technologies for Healthcare*, 2011. 2

[32] A. Swadzba, N. Beuter, S. Wachsmuth, and F. Kummert. Dynamic 3D scene analysis for acquiring articulated scene models. *2010 IEEE International Conference on Robotics and Automation*, pages 134–141, May 2010. 1

[33] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao. Histogram of oriented normal vectors for object recognition with a depth sensor. In *Computer VisionACCV*, volume 7725 LNCS, pages 525–538, 2013. 2

[34] A. Trevor, S. Gedikli, R. B. Rusu, and H. I. Christensen. Efficient organized point cloud segmentation with connected components. In *Proceedings of Semantic Perception Mapping and Exploration*, pages 1–6, 2013. 3, 4

[35] J. Wang, C. Zhang, W. Zhu, Z. Zhang, Z. Xiong, and P. A. Chou. 3D scene reconstruction by multiple structured-light based commodity depth cameras. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 5429–5432, 2012. 3

[36] O. Wang, P. Gunawardane, S. Scher, and J. Davis. Material classification using BRDF slices. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2805–2811, June 2009. 4

[37] B. Zheng, Y. Zhao, and J. Yu. Detecting Potential Falling Objects by Inferring Human Action and Natural Disturbance. *IEEE Int. Conf. on Robotics . . .* , 2014. 2, 9

[38] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond Point Clouds: Scene Understanding by Reasoning Geometry and Physics. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3134, June 2013. 2