# Exploiting Physical Inconsistencies for 3D Scene Understanding

Andrea Fossati[1] Helmut Grabner[1] Luc van Gool[1,2*]
BIWI, ETH Zürich ESAT-PSI / IBBT, KU Leuven
{fossati,grabner}@vision.ee.ethz.ch vangool@esat.kuleuven.be

## Abstract

*Reliable 3D object tracking can provide strong cues for scene understanding. In this paper we exploit inconsistencies between measured 3D trajectories and their predictions using a physical model. In a set of proof-of-concept experiments we show how to retrieve the camera rotation and translation and how to detect surfaces that are hard to visually discern by simply tracking a rigid object. Furthermore we introduce the class distinction between active and passive objects. Prototype examples demonstrate the usability of the visual input for this type of classification. In all the presented experiments, additional information and a deeper understanding about the scene can be obtained, which would not be possible by analyzing solely the image measurements.*

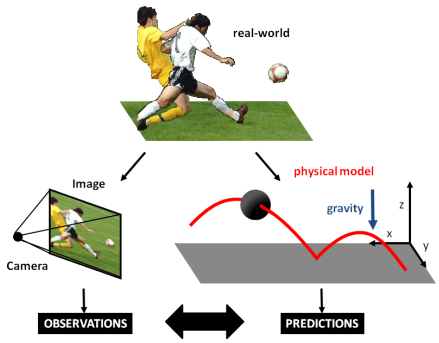
## 1. Introduction

One of the fundamental goals of computer vision research is to understand *what* an image is depicting and to reason about the scene, its objects and their behavior. A lot of research effort has been spent over the last decades in order to get closer to this ambitious objective. Many methods for detection and tracking of individual objects and analysis of their behavior have been developed as well as methods for global scene categorization. Most of these algorithms process only the pixel data from the image but do not take into account the real 3D scene structure.

Images can, in theory, depict scenes which are physically implausible. Famous examples are the drawings by M. C. Escher, one of which is shown in Fig. 1(a). By being aware of the 3D environment that is described by the image, we can enforce the objects to behave according to the laws of physics, which only makes sense when considering the real 3D world, as sketched in Fig. 1(b).

In this paper we propose to analyze trajectories of objects taking into account basic physical knowledge for scene understanding. Predicted, physically plausible trajectories are compared to actually observed ones. As we are focusing on real-world scenes, the laws of physics must hold and predictions according to this physical model should explain the observations. Inconsistencies provide additional information about the objects and the environment. For example, the knowledge of gravity and thus of the expected behavior of free falling objects already gives an idea about the rotation and motion of the camera with respect to the

(a) Escher: *Waterfall*, 1961




(b) Understanding the 3D world

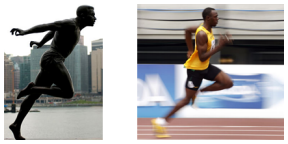
Figure 1. (a) M. C. Escher showed the importance of physics to judge if the depicted scene can exist in reality. (b) In our work we aim at reasoning in the 3D world and not on a 2D projected image of the scene. As we are focusing on real-world scenarios, we furthermore explore the inconsistencies between measurements and physically plausible predictions of objects movement for scene understanding.

*This research has been supported by funding from the EC project RADHAR and the SNF project NCCR IM2.

(a) Rotation of the camera with respect to the world

(b) Invisible structures and external forces

(c) Motion of the camera with respect to the world



(d) Passive vs. active objects

Figure 2. Reasoning about the scene and its objects based on inconsistencies between physical predictions and the observed world. (a) Just by analyzing the physical behavior of the wine flow one has a very clear idea of the camera rotation with respect to the world. (b) Inconsistencies might be explained by introducing hidden scene structures or external forces and hence allow for a more complete understanding of the scene. (c) Despite the main image target has a variable appearance, the water flow alone can give an idea of the relative motion of the camera between the two images. (d) Whereas the behavior of the runner is hard to predict, the behavior of the statue is predictable, despite their similar appearance. This gives evidence that the runner is an active object whereas the statue is passive, because its (non-)motion is only influenced by the laws of physics.

scene (Fig. 1(a,b)). Other inconsistencies can be explained by introducing external forces or hidden scene structures (Fig. 1(c)). Finally, the behavior of certain classes of objects might be only described by introducing internal forces. This leads to the classification of active objects, which distinguish them from passive objects that only undergo external forces like gravity (Fig. 1(d)). We show a series of experiments that distinguish such classes based on their observed motion. This classification, very difficult to solve with only appearance features is relevant for many fields, including object categorization and autonomous robot navigation.

### 1.1. Related Work

The fact that an image is a 2D projection of the 3D world has been widely used in computer vision for various applications, such as measuring distances in the world [10, 4]. Furthermore, the beneficial role of including physical knowledge was already noticed in early computer vision works [16]. In a similar context, in [2] the authors estimate some physical simulation parameters, in order to fit a model to the 2D measurements, but they need a very good manual initialization of the object position and velocity, which is not necessary in our case. More recently, physical constraints have been used in visual tracking to impose motion models [3, 22] or to restrict tracking to only allow for physically plausible configurations [12, 17].

Finally, it has been shown that using additional information from the world limits the number of free parameters, e.g. [5] uses a gravity sensor to reduce the number of point correspondences for relative pose estimation. In general, physical knowledge plays an important role in autonomous robotics, e.g. [20]. What we propose goes instead in a different direction: We use inconsistencies of the observed trajectories with respect to a current physical model to infer enhanced scene properties. In our case, tracking is a tool rather than an objective. In [3] contact dynamics are estimated in terms of forces, and there is a toy 1D experiment with a vertically bouncing ball. The forces are only qualitatively estimated because the mass of the ball is not known. In our case instead we use the full 3D measurements, but our purpose is to estimate surfaces, both in terms of 3D locations and normals. Most common approaches for scene interpretation do not take into account the 3D structure of the scene. Whereas it has been shown that context helps in object detection [21, 25], recognition [18], tracking [7] and scene understanding [15], these approaches are only working in the 2D image plane. Sometimes 3D structure is used as a constraint (e.g. the estimation of the ground plane) in order to improve detection/tracking results [14].

Hoiem et al. [11] and Saxena et al. [23] were among the first to exploit some geometric analysis to better *interpret the scene*. More recently, it has been shown that analyzing the 3D scene reduces ambiguities and is a new paradigm for scene understanding [8]. For example, it enables the use of functional features which improve object detection [6] and human centered scene interpretation [9]. The work by Gupta et al. [8] goes in a direction similar to ours. From a single image, they reason about the 3D structure and additionally make use of simple physical constraints (e.g., physical stability checks). However, they focus on static scenes; our approach is complementary to theirs and reasons about dynamics. By analyzing the object trajectories in the 3D space, we explore inconsistencies with a physical model of the world to infer important properties of the scene, which could not be estimated otherwise.

## 2. Physics as Universal Invariant

In the proposed framework, our purpose is to reason about the scene through the analysis of moving objects in the 3D environment. We assume to have as input some knowledge about:

**Environment E:** The environment defines all the parameters of the world model, i.e. it defines the world coordinate system, the underlying 3D structure, the shape of moving objects and their material properties.

**Trajectories T:** Moving objects in the 3D environment can be (at least partially) detected and tracked reliably over

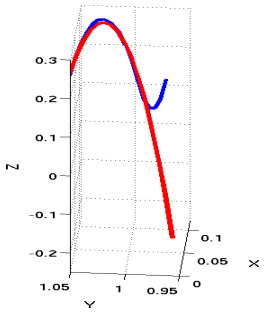



Figure 3. Inconsistency between the measured trajectory of a bouncing ball (blue) and its prediction (red) computed using the initial velocity. In this case it is due to a surface which was unknown to the system before the experiment.

time.

**Physical laws $\mathcal{L}$:** A set of universally valid physical laws which determine how objects move and interact in the environment.

Since we assume that we are observing a real-world scene, the laws of physics must hold. If our observations are not consistent with the expected motion, either some external or internal force must have acted, as shown in Fig. 1.

### 2.1. Inconsistencies between observations and predictions

Let us consider

$$\mathbf{T}_t := [x_t, y_t, z_t]^\mathsf{T} \tag{1}$$

as the 3D position of an object at time $t$ in the world coordinate system. Over time, $\mathbf{T}$ is the trajectory of the object, which in our case is measured through an RGB-depth sensor. Note that other 3D devices can be used as well.

Given the environment $\mathbf{E}$, the set of physical laws $\mathcal{L}$ would univocally determine the trajectory $\mathbf{T}$ of the analyzed object if no other forces were applied. Hence, based on the history $\mathbf{T}_{t'<t}$ of the object position[1] we predict its location $\hat{\mathbf{T}}_t$ for the current time step as

$$\hat{\mathbf{T}}_t = f(\mathbf{T}_{t'<t}, \mathbf{E}, \mathcal{L}). \tag{2}$$

More details about the concrete implementation are given in Sec. 3, when we illustrate some of the applications. An example of predicted and observed trajectories for the case of a bouncing ball is depicted in Fig. 3.

The predicted quantities are compared with the observations as

$$\text{Inconsistency}_t := \begin{cases} 1 & \text{if } ||\hat{\mathbf{T}}_t - \mathbf{T}_t||_2 > \theta \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In other words, a prediction is said to be *consistent* with the actual observations when the distance of the target's position in 3D space to its expected position is below a certain threshold $\theta$, which captures the inaccuracies due to (i) imperfect modeling (e.g., neglecting frictions) as well as to (ii) measurement uncertainties by the sensor. According to our experimental validation, we set the value of $\theta$ to $6mm$ in all our experiments, which lead to consistent and meaningful results.

[1]Please note that, when using the motion history, also higher order terms such as velocity and acceleration are assumed to be computed.

### 2.2. Exploiting inconsistencies

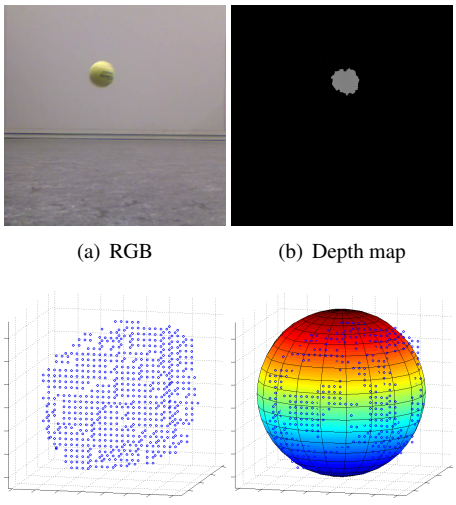
As stated above the detected inconsistencies must be explained. For this reason our knowledge of the objects and of the 3D environment can be revised and refined after such detections. For a better illustration of this important paradigm we present three proof-of-concept scenarios. In all three examples physical inconsistencies are explored in order to acquire a deeper understanding of the scene, which would not have been possible if considering only the image measurements alone.

The shown scenarios consider different assumptions and let us reason about unknown variables of the 3D environment or of the objects that move through it, as described in Fig. 1. First, we assume only passive objects in the environment without any forces exerted during the motion other than gravity. The object motions then strictly follow the laws of physics that are consistent with our scene interpretation. This allows us to estimate the camera pose (at least partially) with respect to a world reference. Secondly, the interactions of the objects with the scene are studied and invisible structures are retrieved. In the third scenario, we introduce the distinction between active and passive objects.

## 3. Experimental Results

**3D data.** Several sensors are nowadays available for acquiring 3D data. This can be done using multi-camera setups, depth cameras, or to some extent even from a single RGB image, e.g., [11]. If the measurements are accurate enough and taken at a sufficiently high sampling rate, e.g. through consumer depth cameras, the obtained data can give reliable estimates of 3D position, velocity and acceleration of captured objects. For our experiments we used the Microsoft Kinect, which can capture an image size of $640 \times 480$ pixels with the corresponding depth values at a frame rate of 30 Hz.

**Tracking.** In the first set of experiments, we track a tennis ball. The 3D tracking can be easily performed on RGB-depth data in real time, by performing a least squares fitting of a sphere of known radius to the measured 3D point cloud. This gives, at every time step $t$, the 3D location of its center $\mathbf{T_t}$. Since in this case the structure of the environment $\mathbf{E}$ is assumed to be known, measured 3D points which belong to the scene can be filtered out, see Fig. 4. The 3D location of the ball center, together with the time stamp pro-



(a) RGB (b) Depth map

(c) Point cloud (d) Fitted sphere

Figure 4. In order to measure the position $\mathbf{T_t}$ of a the tennis ball we perform a least square fitting of a sphere to the 3D point cloud obtained through the depth image.

vided by the camera, is the only input needed by our system in this experiment.

### 3.1. Camera Pose Estimation

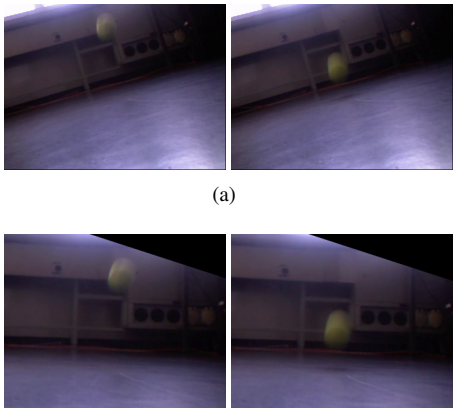
When observing a moving passive object, information about extrinsic camera parameters can be inferred by simply tracking the object and analyzing its trajectory. If there is no interaction with the environment, then gravity is the only force acting on the object (air friction is considered negligible in all our experiments). This can provide information about the camera orientation and translation with respect to the world reference system. In fact, as we assume the gravity to be perfectly vertical, i.e., $\mathbf{g} = [0, 0, -9.81]^{\mathsf{T}}$ m/s$^2$ a detected inconsistency can be explained by analytically computing the best solution, i.e. making the observations as consistent as possible with the physical model. In the context of this experiment we analyzed two different scenarios: (i) Rotation around the principal axis by the camera and (ii) pure translation by the camera.

*Assumptions:* The tracked object is passive and no external forces are present during motion, except gravity.

*Inconsistencies:* The measured motion relative to the camera is not consistent with the physical model.

*Solution:* Adapt the camera pose in order to render the absolute object motion to be physics-compliant.

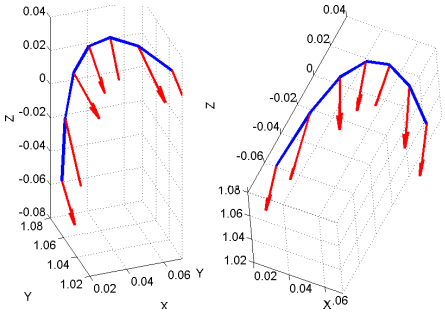
**(i) Camera rotation estimation.** In the first experiment we rotate the camera around its principal axis over a certain angle $\alpha$ to obtain a tilted view during capture. We then tracked the motion of a bouncing tennis ball, which is a passive object, during a time interval without interactions with the environment. As depicted by Fig. 6, we estimated its 3D



(a)

(b)

Figure 5. (a) Two successive sample frame from the video sequence used as input to estimate the camera orientation. (b) The same frames after rotation of $\bar{\alpha} = 18$ degrees and cropping.




(a) observed (b) rotated (gravity pointing downwards)

Figure 6. Estimating the camera orientation: Measured trajectory $\mathbf{T}$ of the ball and computed frame-wise acceleration $\hat{\mathbf{a}}$.

acceleration $\hat{\mathbf{a_t}}$ at each time step $t$. This can be computed quite accurately knowing the timestamps corresponding to each one of the 3D position measurements:

$$\hat{\mathbf{v}}_\mathbf{t} = \frac{\mathbf{T_t} - \mathbf{T_{t-1}}}{\Delta t}, \qquad \hat{\mathbf{a}}_\mathbf{t} = \frac{\hat{\mathbf{v}}_\mathbf{t} - \hat{\mathbf{v}}_\mathbf{t-1}}{\Delta t}, \tag{4}$$

where $\Delta t$ indicates the actual time difference between the measurements taken at time step $t$ and $t-1$. Then, by averaging such acceleration estimations at all the considered time steps, we obtain a reliable measurement of the direction of the gravity in the scene. In this experiment, the average magnitude of the measured acceleration vector was $||\bar{\mathbf{a}}||_2 = 9.72$ m/s$^2$. In the depicted case the recovered orientation $\hat{\alpha}$ of the camera with respect to the horizontal plane was of $\bar{\alpha} = 18$ degrees, which is consistent with the video sequence shown in Fig. 5.

**(ii) Camera translation estimation.** In a second experiment we instead assumed the camera orientation to be horizontally constant and translated the camera manually during the capture of the video. Also in this case we tracked

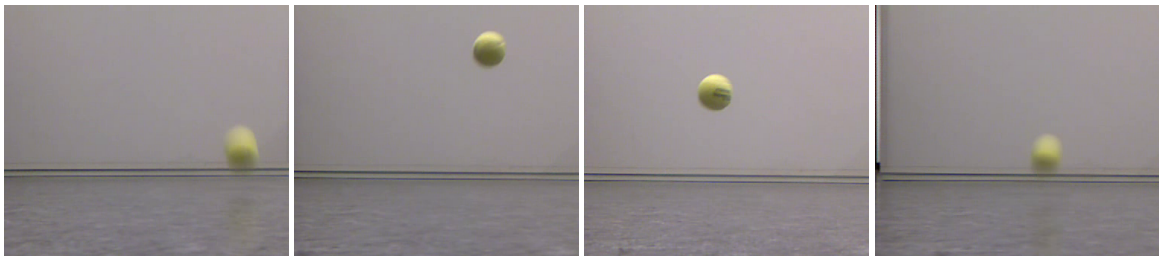
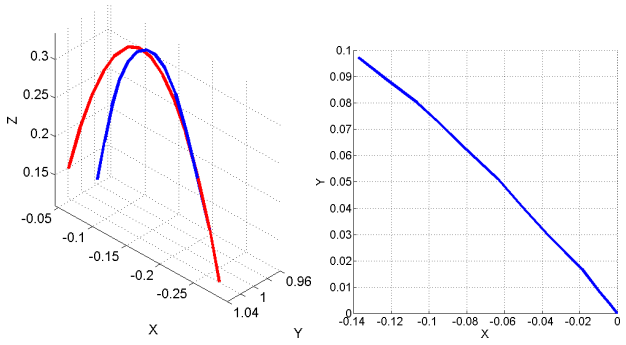


Figure 7. Four sample frames from the video sequence used as input to estimate the camera translation. Standard structure-from-motion approaches would face both the problem of missing texture and - if that could be resolved - of having an unknown relative scale (and trajectory) between the independently moving ball and the background.




(a) trajectories (b) estimated camera translation

Figure 8. Estimating the camera translation: (a) Predicted trajectory of the ball in a global reference (red) and measured trajectory in the local camera reference (blue). (b) Inferred camera movement to explain the inconsistencies shown in (a).

the motion of a bouncing tennis ball during a time interval when no interactions with the environment $\mathbf{E}$ occurred. By measuring, as described in Eq. (4), the 3D velocity of the ball at the beginning of the sequence, we could then compute a frame-by-frame prediction $\hat{\mathbf{T}}$ of its 3D position. This was performed considering that the only force, and thus acceleration, experienced by the ball was the gravity, and neglecting the effect of the air friction:

$$\hat{\mathbf{T}}_{\mathbf{t}} = \mathbf{T}_{\mathbf{t-1}} + \hat{\mathbf{v}}_{\mathbf{t-1}}\Delta t + \frac{1}{2}\mathbf{g}\Delta t^2. \qquad (5)$$

By then matching $\hat{\mathbf{T}}_{\mathbf{t}}$ with the measured trajectory $\mathbf{T}$, we could obtain the time evolution of the camera 3D position through a simple difference. In the experiment, the translation of the camera was performed manually approximately along a straight line. The distance from the start point to the end point was 15.5 cm, and our estimate yields to 16.5 cm. Please note that, since the only element present in the scene is the tracking target, as shown in Fig. 7, typical approaches for camera motion estimation would fail in this case. Fig. 8 shows the predicted motion and the measured trajectory for this experiment.

**Potential practical applications.** As briefly explained throughout this section, calibration of the extrinsic parameters of depth cameras can be improved by additional measurements of moving objects in the environment. This is particularly relevant for pre-existing sequences, for which calibration data is not available. The trajectory of passive moving objects, if present in the scene, can be used to obtain a rough estimate of the orientation of the camera. Furthermore, Structure-from-Motion (SfM) has a problem with dynamic scenes, even if all objects are rigidly moving. There exist unknown relative scales between all objects that move with respect to each other. There have already been studies to analyze to which of those scales different motion trajectories correspond [19]. Looking for those trajectories that are consistent with physical laws is an important cue for disambiguating the SfM process, which has not been studied in [19].

### 3.2. Invisible Surface Detection

In our second experiment, we explored an environment whose 3D shape is *a priori* unknown to the system, simply through the tracking of a tennis ball and the awareness of the forces controlling its motion. We assume that the camera is static and that the ball is a totally passive object. If there are no forces except gravity, then the motion can be described through the standard physical equations of a free falling body, see Eq. (5). However, if this is not the case, then some external forces, invisible with respect to our sensor, must have interacted with the ball.

*Assumptions:* The tracked object is passive and the camera is static.

*Inconsistencies:* The measured motion is not consistent with the physical model.

*Solution:* Adapt (infer) the scene structure.

**Setup.** We placed the Kinect on the ground, such that the principal axis was horizontal. This made the ground plane invisible to the camera, because the structured light pattern cannot be properly projected onto it and then recovered in this configuration. We also put inside the scene a transparent plastic box, whose shape could not be captured either by the depth camera. In this configuration, the depth

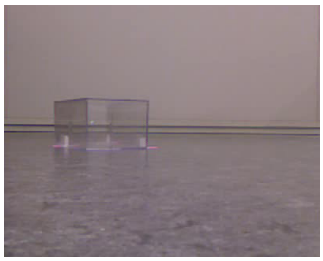
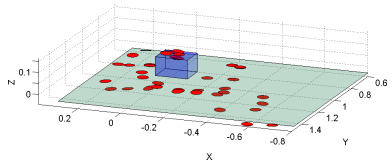
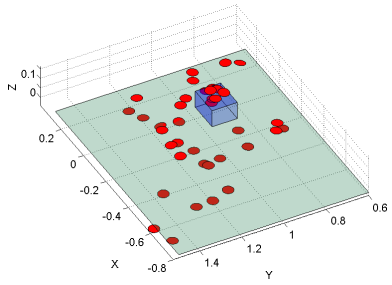


Figure 9. RGB view of the analyzed scene. The horizontal ground plane and the plastic box cannot be captured by our consumer depth camera. This makes the depth image corresponding to this scene totally blank.



(a) Side view of the reconstructed scene



(b) Top view of the reconstructed scene

Figure 10. Estimating interactions: Reconstruction of the environment through the detected interactions.

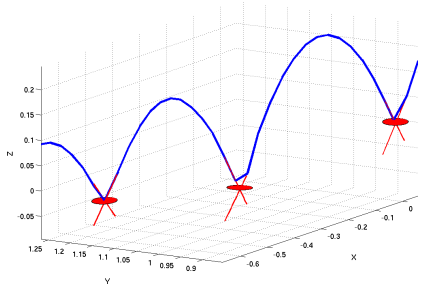


Figure 11. Estimating interactions: Trajectory of the tracked ball and detected interactions with the environment.

measured was totally blank for the entire scene, in contrast with the RGB input, shown in Fig. 9.

**Observations.** We bounced a tennis ball in the scene several times, off the ground and the transparent box. Tracking of the ball was performed as described at the beginning of the section.

**Inconsistencies.** The ball's measured position $\mathbf{T_t}$, obtained by the tracker, was continually compared to a predicted $\hat{\mathbf{T}}_\mathbf{t}$ computed through Eq. (5). If an inconsistency was observed, according to Eq. 3, it was considered to be the effect of an interaction. Such interaction must have been due to some entity which was inside the scene but at the same time not captured by the depth camera. Hence, we adapted the modeled environment accordingly: A small surface at the location of the interaction and orientated following the motion just before and just after the interaction was inferred as shown in Fig.10. For comparison the ground plane and the transparent plastic box are overlaid as well.

*Implementation detail.* Since the motion of the tracked target is very fast and the data provided by the depth camera is discrete, it is not guaranteed that the exact moment of the interaction is captured. When an interaction is detected, the system performs an extrapolation of the trajectory of the target object, by fitting a parabola to both a few measurements before and after the inconsistency. The closest 3D point between the 2 parabolas was considered as the point where the interaction actually occurred, and the orientation obtained by averaging the tangents of the 2 parabolas in their closest location to such point is assumed to be the direction of the interaction (i.e., the normal to the surface that caused the interaction). This is shown in Fig 11.

**Quantitative results.** Based on the inferred interaction surfaces two planes could be estimated: One corresponding to the ground plane and one to the top of the box. An evaluation of the obtained results is presented in Table 1.

| Quantity | Estimated | Actual | Error |
|---|---|---|---|
| Box Height | 92 mm | 90 mm | 2 mm |
| Box Top Orientation | 4.8° | 0° | 4.8° |
| Ground Orientation | 5.6° | 0° | 5.6° |

Table 1. Estimated Scene: Quantitative results

**Potential practical applications.** As demonstrated in several research works, e.g. [14, 11], automatically determining the ground plane(s) inside a scene greatly increases detection and tracking performances, and helps scene understanding. With our approach, tracking passive objects moving in space helps scene structure estimation and improves camera calibration and readjustment. Typical applications might involve sport footage, including for example soccer, tennis or basketball videos.

### 3.3. Active Object Detection

In the last set of experiments we assume a known environment and a static camera. Thus, the tracked objects must move according both to the physical laws and the structure of the environment. We can then make a distinction between the objects which can be considered as *passive*, which just react to the surrounding environment, and the *active* objects,

which have some internal source of energy that makes them move in an unpredictable way.

*Assumptions:* The scene is known and the camera is static.

*Inconsistencies:* The measured motion is not consistent with the physical model for passive objects.

*Solution:* The object is considered as active, which introduces the active *vs.* passive object classification.

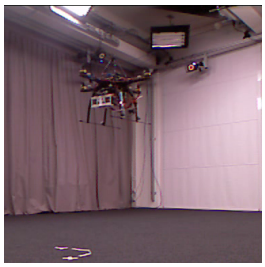
**Tracking.** A pre-processing step removes all the points belonging to the known 3D scene. To allow for objects of a generic shape, the tracking is done using frame-by-frame ICP [1] on the point-cloud provided by the depth camera. This gives a relative frame-to-frame motion which is used to estimate the object's trajectory.

**Setup.** We tested our approach on two active objects, namely an autonomous flying robot and a spring-powered toy car, and on a passive object, a tennis ball. To compute the predictions $\hat{\mathbf{T}}_\mathbf{t}$ of their trajectories, we used the formula of Eq. (5), again assuming friction to be negligible.

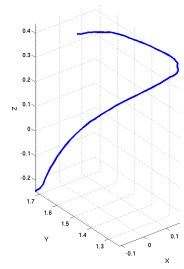
**Inconsistencies.** As in the previous experiments, at each time step $t$ we have a measured location of the object $\mathbf{T}_\mathbf{t}$, given by our tracker, and a predicted location $\hat{\mathbf{T}}_\mathbf{t}$ given by the physical model. If there is discordance (Eq. (3)) between the locations, and interactions are not possible given the known structure of the environment, then the object must have some internal source of energy that makes it act against the physics of free fall. In other words, the object is considered to be *active*. If on the contrary the object does behave following the prediction given by the model, then the object is considered to be *passive*.

**Results.** In the case of the autonomous flying robot, the difference between the measured and estimated trajectories is very clear (at all time instances) and the object can be considered *active*. Details are shown in Fig. 12. On the other hand, in the case of the tennis ball, the predicted trajectory was very consistent with the measured one, as visible in Fig. 13. This is an important cue in defining it as a *passive* object. Finally, the spring-powered toy car first moves along a straight line, then makes a turn and again moves along a line, all with nearly constant velocity. Inconsistencies with respect to the model are mostly detected when the toy car is turning, see Fig. 14.

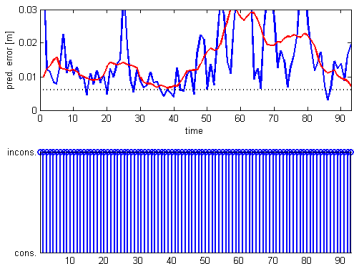
**Potential practical applications.** Estimating the nature (active vs. passive) of surrounding objects, i.e. dynamic obstacles, is one if the main tasks of autonomous path planning [24]. Knowing the behavior or a rough estimate thereof would be very beneficial in terms of saving computation time and improving accuracy. In terms of attribute based object categorization [13], knowing objects to be active or passive will help object class recognition to consider fewer options, e.g. humans vs. statues, as shown in Fig. 1(d).



(a) Setup



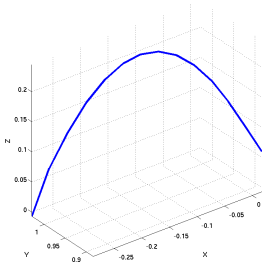
(b) Observed trajectory



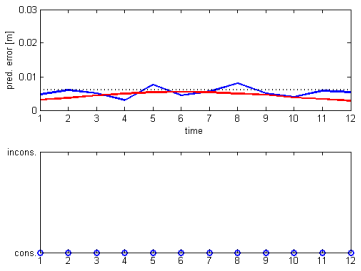

(c) Inconsistencies

Figure 12. (a) An autonomous robot flying around and the corresponding (b) observed trajectory. (c) Measured inconsistencies of the observed trajectory with the predicted one. The dotted line depicts the chosen threshold, set at $6mm$, while the red line indicates a moving average, meant to stabilize the results. The high prediction errors and hence detected physical inconsistencies give strong evidence that the observed object is *active*.

(a) Setup



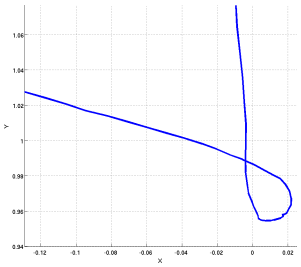
(b) Observed trajectory



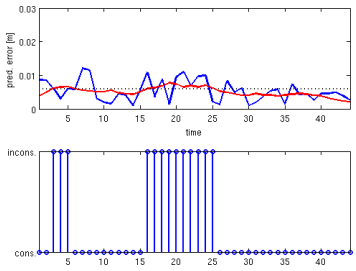

(c) Inconsistencies

Figure 13. (a) A tennis ball and the corresponding (b) observed trajectory. (c) The ball behaves following the model and is therefore consistent with our hypothesis of being a *passive* object.

(a) Setup



(b) Observed trajectory




(c) Inconsistencies

Figure 14. (a) A spring-powered toy car and the corresponding (b) observed trajectory. (c) When the car is going straight the observations match the prediction well, since we assume a constant velocity motion model on the horizontal plane. However, once the car is turning, the high prediction errors indicate physical inconsistencies and give evidence that the observed object is *active*.

## 4. Conclusion

In the presented set of proof-of-concept experiments, we have shown how a robust estimation of the 3D trajectory of an object and the detected inconsistencies between such estimation and a simple physical model, i.e. of free fall, can help in obtaining information about the scene which would not be obtainable otherwise. As noted throughout the paper, our contribution is not meant to be a deep analysis about the accuracy of the obtained results, but rather to suggest a largely unexplored research direction. A possible extension of the presented work would involve more accurate 3D measurements and an extended set of physical rules. Accurately modeling friction would for example allow to lower the tolerance threshold and thus achieve a higher confidence in the conclusions.

## References


[1] P. Besl and N. McKay. A method for registration of 3d shapes. *PAMI*, 14:239–256, 1992.

[2] K. Bhat, S. Seitz, J. Popovic, and P.-Khosla. Computing the physical parameters of rigid-body motion from video. In *Proc. ECCV*, pages 551–565, 2002.

[3] M. Brubaker, D. Fleet, and A. Hertzmann. Physics-based person tracking using simplified lower-body dynamics. In *Proc. CVPR*, 2007.

[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *IJCV*, 40(2):123–148, 2000.

[5] F. Fraundorfer, P. Tanskanen, and M. Pollefeys. A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In *Proc. ECCV*, 2010.

[6] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *Proc. CVPR*, 2011.

[7] H. Grabner, J. Matas, L. Van Gool, and P. Cattin. Tracking the invisible: Learning where the object might be. In *Proc. CVPR*, 2010.

[8] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Proc. ECCV*, 2010.

[9] A. Gupta, S. Satkin, A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *Proc. CVPR*, 2011.

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[11] D. Hoiem. *Seeing the World Behind the Image: Spatial Layout for 3D Scene Understanding*. PhD thesis, Robotics Institute, Carnegie Mellon University, 2007.

[12] N. Kyriazis, I. Oikonomidis, and A. Argyros. Binding vision to physics based simulation: The case study of a bouncing ball. In *Proc. BMVC*, 2011.

[13] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. CVPR*, 2009.

[14] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *Proc. CVPR*, 2007.

[15] L. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proc. CVPR*, 2009.

[16] D. Metaxas and D. Terzopoulus. Shape and nonrigid motion estimation through physical-based synthesis. *PAMI*, 15:580–591, 1993.

[17] I. Oikonomidis, N. Kyriazis, and A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *Proc. ICCV*, 2011.

[18] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11:520–527, 2007.

[19] E. Ozden, K. Cornelis, and L. V. Gool. Space-time-scale registration of dynamic scene reconstructions. In *Proc. ECCV*, 2006.

[20] M. Prats, P. Sanz, and A. Pobil. A framework for compliant physical interaction. *Autonomous Robots*, 28:89–111, 2010.

[21] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Proc. CVPR*, 2011.

[22] M. Salzmann and R. Urtasun. Physically-based motion models for 3d tracking: A convex formulation. In *Proc. ICCV*, 2011.

[23] A. Saxena, M. Sun, and A. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 31(5):824 –840, 2009.

[24] L. Scandolo and T. Fraichard. An anthropomorphic navigation scheme for dynamic scenarios. In *ICRA*, pages 809 –814, 2011.

[25] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proc. CVPR*, 2010.