# VRProp-Net: Real-time Interaction with Virtual Props

Catherine Taylor
University of Bath
Marshmallow Laser Feast
cct43@bath.ac.uk

Robin McNicholas
Marshmallow Laser Feast
robin@marshmallowlaserfeast.com

Darren Cosker
University of Bath
dpc22@bath.ac.uk

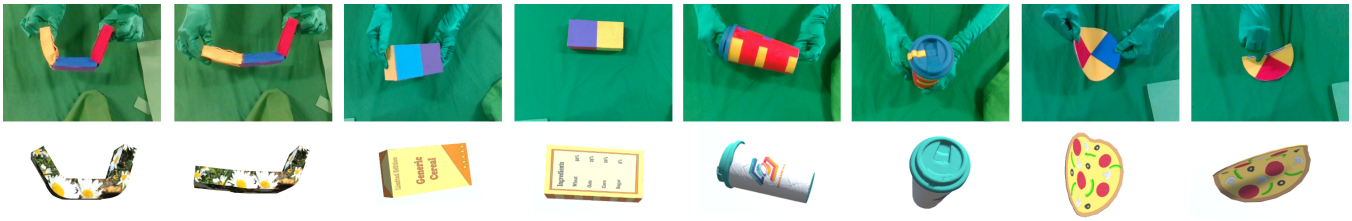

**Figure 1: Predicted shape and pose from unlabelled RGB images using our *VRProp-Net*. The predicted parameters can be used to drive the motion a computer-generated model which is rendered into a VR or AR scene.**

## CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Modeling and simulation**; **Computer vision**.

## KEYWORDS

VR Props, Non-rigid Object Tracking, Neural Networks

**ACM Reference Format:**
Catherine Taylor, Robin McNicholas, and Darren Cosker. 2019. VRProp-Net: Real-time Interaction with Virtual Props. In *Proceedings of SIGGRAPH '19 Posters.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306214.3338548

## 1 INTRODUCTION

Virtual and Augmented Reality (VR and AR) are two fast growing mediums, not only in the entertainment industry but also in health, education and engineering. A good VR or AR application seamlessly merges the real and virtual world, making the user feels fully immersed. Traditionally, a computer-generated object can be interacted with using controllers or hand gestures [HTC 2019; Microsoft 2019; Oculus 2019]. However, these motions can feel unnatural and do not accurately represent the motion of interacting with a real object. On the other hand, a physical object can be used to control the motion of a virtual object. At present, this can be done by tracking purely rigid motion using an external sensor [HTC 2019]. Alternatively, a sparse number of markers can be tracked, for example using a motion capture system, and the positions of these used to drive the motion of an underlying non-rigid model. However, this approach is sensitive to changes in marker position and occlusions and often involves costly non-standard hardware [Vicon 2019]. In addition, these approaches often require a virtual model to be manually sculpted and rigged which can be a time consuming process. Neural networks have been shown to be successful tools in computer vision, with several key methods using networks for tracking rigid and non-rigid motion in RGB images [Andrychowicz et al. 2018; Kanazawa et al. 2018; Pumarola et al. 2018]. While these methods show potential, they are limited to using multiple RGB cameras or large, costly amounts of labelled training data.

To address these limitations, we propose an end to end pipeline for creating interactive *virtual props* from real-world physical objects. As part of our pipeline, we propose a new neural network - *VRProp-Net* - based on a Wide Residual Network [Zagoruyko and Komodakis 2016], to accurately predict rigid and non-rigid deformation parameters from unlabelled RGB images. We compare the success of VRProp-Net to a basic Resnet34 [He et al. 2016] for predicting 3D pose and shape for non-rigid objects. We demonstrate our results for several rigid and non-rigid objects.

## 2 OUR APPROACH

Our pipeline begins by creating a virtual representation of an arbitrary physical object. We do this without manual sculpting or rigging. A textured triangular mesh can be obtained directly from the object using a 3D scanner. As opposed to capturing real object deformations we simulate the non-rigid behaviour of the object using finite element analysis [Cook 1989]. The large range of deformations generated from this are reduced to a few key blendshapes using Principal Component Analysis (PCA). During the simulation, a section of the object is fixed to ensure that PCA captures the variation in the dataset due to changes in shape rather than pose. The PCA eigen vectors at 2 standard deviations are used to create a rigged model.

The virtual object is used to generate a synthetic dataset, containing a variety of poses and shapes. For each frame, an RGB image is rendered and the corresponding deformation parameters (blend weights and orientation) saved. We chose to represent our non-rigid deformation using blendshapes as complex deformations can be represented by a small number of weights. The ground truth


*SIGGRAPH '19 Posters, July 28 - August 01, 2019, Los Angeles, CA, USA*

ACM ISBN 978-1-4503-6314-3/19/07.
https://doi.org/10.1145/3306214.3338548

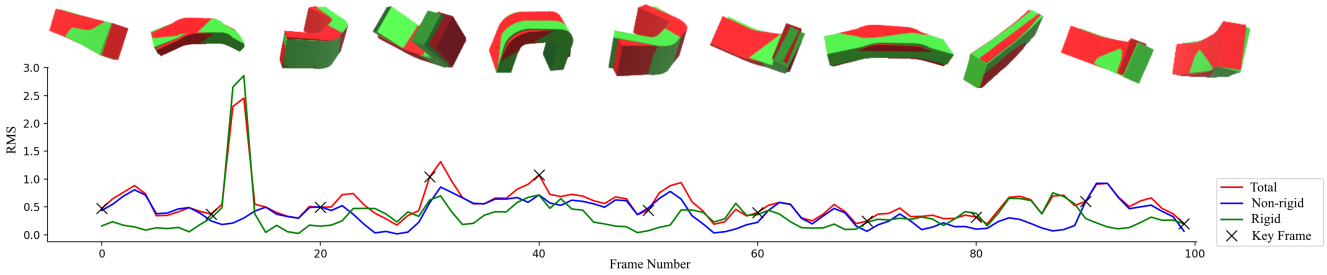


**Figure 2: Predicted shape and pose on sequence of synthetic data. The RMS error between the predicted and ground truth mesh is calculated for each frame. The ground truth mesh (green) and the predicted mesh (red) are shown for a selection of frames. The total RMS error can be divided into the contributions from rigid and non-rigid transforms.**

pairs train a convolutional neural network (CNN) to predict the deformation parameters from unlabelled RGB images.

For our network architecture, we designed VRProp-Net. This network is based on the Wide Residual Network architecture [Zagoruyko and Komodakis 2016] but the number of convolutional layers in the basic block have been extended from 2 to 4 and the kernel size has been expanded from 3 to 5. This increases the power of the blocks and allows them to better learn the deformation parameters. Though trained on synthetic images, VRProp-Net adapts to make real-time predictions on real data. The physical object is captured using a single RGBD camera and the tracked object segmented. The colours are then flattened to remove any variation due to lighting. Note that, as the user will be wearing a VR headset, we are able to use green screens in our VR environment and texture the object (see Figure 1) to assist colour tracking. The centroid of the segmented image is used to crop the object. Additionally, the centroid is back projected using the camera intrinsic matrix and the average depth to find the 3D position of the object. The cropped image is input to the network and the predicted parameters returned. These are used to update the virtual object's shape and pose with the resulting model rendered into a virtual scene.

## 3 RESULTS AND CONCLUSIONS

We explored 2 different CNNs architectures for predicting deformation parameters. We tested our pipeline and these networks on tracking sequences from 2 rigid and 2 non-rigid objects. The first network tested for prediction was Resnet34, pretrained for classification on the imageNet dataset [He et al. 2016]. This was often unstable and did not consistently make accurate predictions across all objects. Greatly improving on Resnet34, VRProp-Net better learns the deformation parameters for each of our objects, with visual performance highlighted in our results (e.g. Figures 1, 2 and supplementary material). In addition to visual comparisons, we computed the Root Mean Square (RMS) error for synthetic inputs which have known ground-truth pose and deformation using both architectures. We find that the RMS is lower overall for VRProp-Net with fewer and smaller changes in prediction parameters (noticeable as object jumps) between frames (see Figure 3). Figure 2 demonstrates our results on a synthetic sequence for a non-rigid objects and Figure 1 shows the results on real data.

We have designed a system which transports arbitrary physical objects into virtual environments to be used as interactive props and

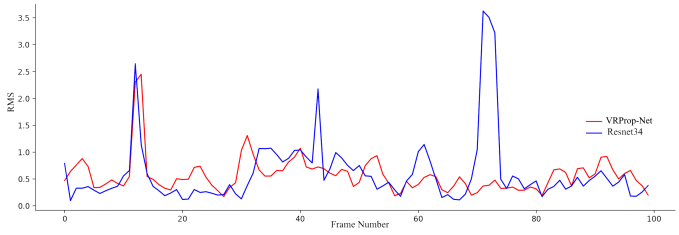


**Figure 3: Comparing the RMS for 100 frames of a non-rigid synthetic sequence between Resnet34 and VRProp-Net.**

have developed VRProp-Net for learning deformation parameters from unlabelled 2D images. We have shown successful results for this network for both rigid and non-rigid object predictions. In future work, we would like to extend our system to allow multiple props in VR, increasing the variety of experiences for which our system could be used for. Additionally, we wish to explore different representations of non-rigid objects (e.g. articulated) and adapt our network to learn the parameters of these models. Finally, we plan to make our system more robust to complex real-world environments so that we do not require gloves or green screens.

## REFERENCES

M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. W. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba. 2018. Learning Dexterous In-Hand Manipulation. *CoRR* (2018).

R. D. Cook. 1989. *Concepts and applications of finite element analysis.* (3rd ed.).

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. 630–645.

HTC. 2019. Discover Virtual Reality Beyond Imagination. https://www.vive.com/uk/.

A. Kanazawa, M. J Black, D. W Jacobs, and J. Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the CVPR*. 7122–7131.

Microsoft. 2019. Microsoft HoloLens | Mixed Reality Technology for Business. https://www.microsoft.com/en-us/hololens.

Oculus. 2019. Oculus Rift. https://www.oculus.com/rift/.

A. Pumarola, A. Agudo, A. Porzi, L.and Sanfeliu, V. Lepetit, and F. Moreno-Noguer. 2018. Geometry-aware network for non-rigid shape prediction from a single view. In *Proceedings of CVPR*. 4681–4690.

Vicon. 2019. Motion Capture Systems. https://www.vicon.com/.

S. Zagoruyko and N. Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).