Deformable Objects for Virtual Environments

Catherine Taylor*
University of Bath
Marshmallow Laser Feast

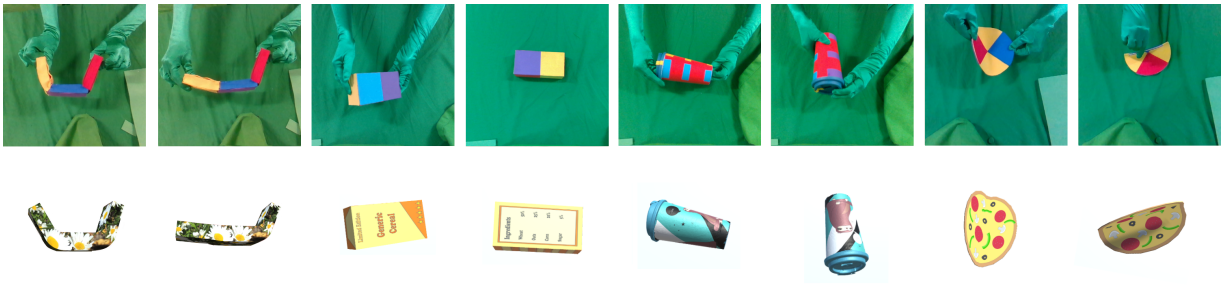

Figure 1: The pose and shape of a physical object can be predicted from an RGB or RGBD image using a convolutional neural network. The predicted parameters can be used to drive the motion a computer-generated model which is rendered into a VR or AR scene.

Abstract

An important aspect of a good virtual or augmented reality (VR/AR) experience is the feeling of immersion and the inability to easily distinguish between the real and the virtual worlds. The perceived immersion is greatly impacted by the manner in which a user interacts with the virtual environment and computer-generated objects in the scene. Traditional methods of interaction, such as controllers or hand gestures, do not accurately model real world behaviour and offer limited tactile feedback. To address these limitations, my research will explore how the behaviour of physical objects can be used to control virtual objects and investigate the influence this has on immersion within VR and AR applications. In particular, I will focus on tracking *deformable* (i.e non-rigid) objects for VR.

Keywords: Virtual Reality, Non-Rigid Object Tracking, Virtual Objects, VR Props.

Index Terms: Computing methodologies—Neural networks Computing methodologies—Modelling and simulation Computing methodologies—Computer vision

1 Introduction and Motivation

Virtual and augmented reality are becoming increasingly common mediums in both academia and industry, particularly within the fields of entertainment, health and art. A successful VR or AR application is immersive, keeping the user captivated and focused on the virtual environment. While traditional methods of interacting with virtual elements, such as controllers or hand-gestures, have been used frequently as useful tools for connecting the real and virtual worlds, they do not accurately model real-world interactions with physical objects.

My doctoral research will focus on object tracking, in particular non-rigid object tracking, for virtual reality. The tracked real-world objects will control the behaviour of virtual objects and so be used as a more immersive alternative to controllers or hand gestures.

*e-mail: c.taylor3@bath.ac.uk

Taking motivation from the recent successes that have arisen from combining neural networks with traditional computer vision tasks, I propose, in my work, using neural networks to track rigid and non-rigid objects in RGB or RGBD data. Current notable works in this area are limited by requiring substantial amounts of labelled data or multiple camera inputs [1,4,11,19] and so there is much area for improvement. In contrast, I will use unlabelled, synthetic training datasets, allowing my approach to be scaled easily to different arbitrary objects.

This paper will continue in Section 2 with a review of the key pieces of related work. I will then discuss, in Section 3, my current work, which has focused on designing an end-to-end pipeline for creating interactive virtual props from real-world objects. This pipeline features a custom neural network based approach for tracking rigid and non-rigid objects in unlabelled RGB images. Section 4 will outline several areas of potential future work and, finally, Section 5 will conclude.

2 Related Work

Conventionally, sequences of button presses from hand-held controllers or consoles have been used as a means of interacting with virtual objects. As VR gained popularity, through systems such as the HTC Vive [3] and Oculus Rift [9], controllers became more advanced, with their position and orientation being tracked and used as additional ways of interacting with virtual environments. Moreover, VR systems often offer external sensors (e.g. the Vive Tracker [3]) as additional attachments which are able to track the 3D position and orientation of whatever object they are appended to. In contrast, hand gestures, for example as used by Microsoft's HoloLens [8], can be used to control the behaviour of a virtual object. Both controllers and hand gestures have been used frequently in a range of novel experiences but they do not model the intuitive way to interact with a physical object and offer limited tactile feedback.

On the other hand, these limitations can be addressed by tracking the behaviour of a physical object and using the results to drive the motion of a virtual model. To do this, a virtual model can be fit to a set of feature points detected in RGB or RGBD data [10, 15]. Rigid object tracking has been an extensively researched area [12, 15], however, tracking non-rigid objects is still a complex problem [5, 6, 16]. Motion capture systems, such as Vicon [17, 18],

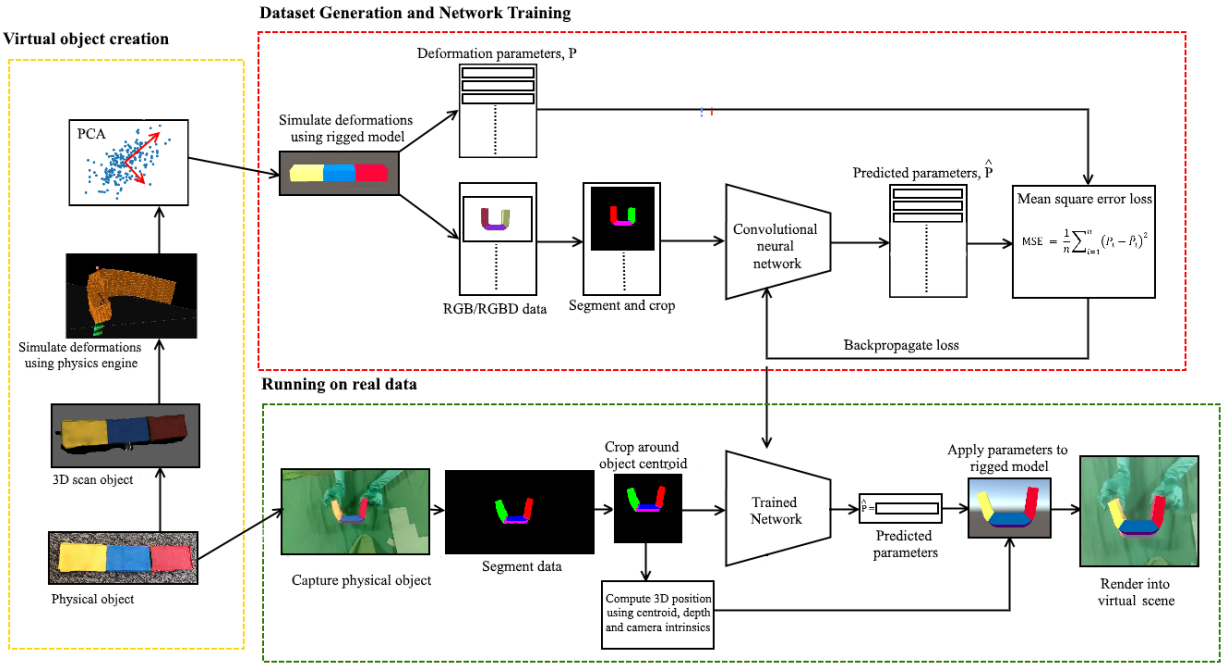

Figure 2: Our proposed end-to-end pipeline for creating an interactive virtual prop from a physical object. To begin, a virtual object is created by 3D scanning the physical item. A wide range of deformations are generated using an FEM simulation and these reduced down via PCA into blendshapes. The rigged model is used to build a synthetic dataset which trains a CNN to predict deformation parameters from unlabelled RGB images. Finally, the trained network predicts deformation parameters from RGB images from a single RGBD sensor and these are used to update the pose and shape of the virtual model [14]

track sparse markers on the surface of an object and are often used for capturing non-rigid motion. However, these systems are very sensitive to marker position and small drifts or occlusions due to hand interaction can cause tracking to fail. Additionally, these systems often involve costly, non-standard hardware.

There have been several notable works which use neural networks for tracking rigid and non-rigid objects [1, 4, 11]. While these have shown positive results, they must be trained with costly large labelled datasets or require multiple RGB cameras. Thus, they do not adapt well to track arbitrary physical objects for which no labelled dataset exists.

3 Current Progress

At the time of the consortium, I will have just begun the third year of my PhD. This section will outline my work to date over the first two years.

In order to use physical objects to control elements of a virtual scene, a robust tracking method must be designed for capturing the behaviour of rigid and non-rigid real-world objects. In turn, the captured behaviour of the physical objects will be used to drive the motion of the virtual objects. As tracking rigid objects is a well researched problem, my initial researched focused on capturing the non-rigid behaviour of deformable objects. An extensive literature into tracking non-rigid objects highlighted the potential of using neural networks tracking and this has become a central focus of my research.

My research to date has resulted in an end-to-end pipeline for transporting real objects into virtual and augmented environments to be used as interactive props [13, 14]. An overview of this pipleline can be seen in Figure 2. The pipeline starts with the automatic generation of a rigged blendshape model. In contrast to manually sculpting and rigging a model, the automatic generation is fast and does not require any 3D modelling expertise. To begin, the chosen object is 3D scanned, obtaining a textured polygon mesh. A finite element simulation is applied to this mesh, generating a large set of shapes representing different non-rigid deformations. Finally, Principal Component Analysis (PCA) is used to reduce the dimensions of the datatset. The PCA eigen vectors at 2 standard deviations are used to create a rigged model.

The rigged model can be used to generate a large synthetic dataset that can, in turn, be used to train a neural network. The model is imported into Unity and the deformation parameters (i.e the PCA weights, position and orientation) randomly varied between frames. In each frame, as well recording the deformation parameters, a 2D RGB image of the deformed object is rendered. The parameters of the virtual camera, for rendering the image, are set equal to the real camera intrinsics to make the synthetic images appear as close as possible to real-world captures of the object. The images are pre-processed before training the network by segmenting, flattening the colour and cropping the images. The segmentation step uses simple colour thresholding to identify pixels which do not belong to the object of interest and these pixels are masked out. As the main application of the tracking algorithm is for virtual reality, where the user will be wearing a headset, the physical environment can be controlled through the use of green screen, green gloves and texturing the object with distinct colours, to simplify segmentation. Once the image has been segmented, the coloured sections are flattened by extracting pixels in a similar range and setting them to the same colour. This flattening step is important as is removes shading and colour and lighting variations in the synthetic data so that that the network can easily adapt to make prediction on real-world data. Finally, the centroid of the object is calculated and cropped around, creating a square image that is used as a input to a neural network.

The synthetic dataset is used to train a convolutional neural network (CNN) to predict the deformation parameters from unlabelled RGB images. In the initial version of this pipeline [14], the CNN

chosen was a Resnet34 [2], pre-trained on the imageNet dataset. A simpler network (e.g. AlexNet [7]) may also be able to make fast and accurate predictions, however, this has not been tested in the initial work. While the Resnet34 produced compelling results and highlighted the potential of the pipeline, the predictions were often unstable, especially for objects which were rotated greatly. Taking this into account, a custom neural network - *VRProp-Net* - was designed for the pipeline [13]. VRPRop-Net is an extension of a Wide Residual Network (WRN) Architecture [20] which has been adapted to make more accurate predictions of the deformation parameters. The number of convolutional layers in basic blocks of the WRN were doubled from 2 to 4 and the kernel size for each of these changed from 3 to 5, in order to increase the prediction accuracy. As the networks were trained on synthetic data, the root mean square (RMS) error between a ground truth sequence of meshes and the predicted meshes can be evaluated to explore the accuracy of the tracking network (Figure 3 and Figure 4). Figure 3 compares the RMS error between each network for a synthetic sequence. While both networks have a similar frame rate, VRProp-Net produces more accurate predictions, with lower RMS error, and is more consistent, seen as fewer and smaller jumps in error between frames. This highlights the advantage of the custom neural network over a Resnet34.

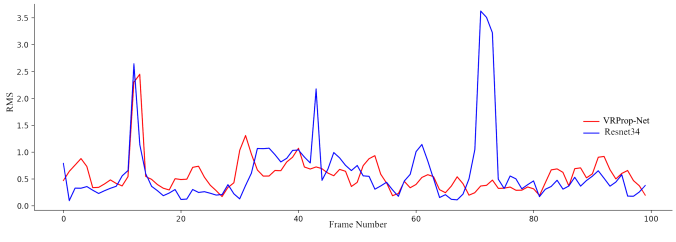

Figure 3: Comparing the RMS for 100 frames of a non-rigid synthetic sequence between Resnet34 and VRProp-Net [13]

As well as making predictions on synthetic sequences, the network adapts to predicting deformation parameters on real-world images. The physical object is captured using a RGBD camera and the image segmented, flattened and cropped as with the synthetic data. The 3D position of the object is calculated by back projecting the 2D centroid of the object using the camera intrinsic matrix and the average depth. The cropped real image is used as an input to the CNN and the predicted deformation parameters, which are returned by the network, used to update the shape and pose of the virtual object. Finally, the deformed model can be rendered into a virtual scene. The object can be rendered with the real world material, however, it could also be rendered with a contrasting texture or augmented with visual effects as suited to VR application. The pipeline was able to make acceptable predictions (as shown in Figure 1) for a several rigid and non-rigid objects. Again, VRProp-Net performed better and produced more visible pleasing results than the Resnet34. Both CNNs had an interactive frame rate of around 15fps.

4 Future Work

There are several possible areas of future work, that could build on my initial research, to create more immersive interactions in VR experiences. Taking these into account, this section will discuss three potential area that I will explore during remainder of my doctorate. These are exploring different representations of non-rigid objects, developing the tracking algorithm so that it is suitable for less controlled environments and, finally, investigating the interaction with multiple VR props.

The current pipeline has been tested on a range of simple real-world objects, which are good demonstration of props that may be used in a VR application, and have fast and accurate enough performance for VR. However, these are simple objects and so future work could focus on different objects which can undergo complex or intricate deformations. Moreover, while blendshapes work well for our chosen objects, different representations (e.g. bones or physics-based models) may provide more accurate results and may adapt better to different objects. Thus, VRProp-Net could be adapted to predict the parameters of different non-rigid models rather than just PCA weights. The current networks were trained on purely RGB data. The addition of depth into the training set was experimented with initially, with no significant improvement, and so was not included in the pipeline. However, for different representations of non-rigid objects, depth may provide useful information and may be able to capture more subtle deformations. Thus, depth could be combined with the RGB images in a future synthetic training dataset.

The current tracking network has been designed for VR applications where the user will be wearing a headset. In these setups, the physical environment can be controlled using green screens and the tracked objects coloured brightly without effecting the immersion of the experience. However, at present the tracking system is limited to these environments. A potential area of work is developing the tracking system so that it can make predictions in less controlled, in-the-wild scenes. This extends the functionality of the tracking algorithm to a larger range of applications, such as tracking for AR. As a starting point, the synthetic dataset could be augmented to include real-world complexities such as lighting variations as well as diverse backgrounds and object textures. Additionally, with a more varied and richer dataset, the network would be able to learn features more robustly and may be able to carry out the image segmentation without the need for a pre-processing step.

Finally, the current system tracks a single physical object and uses the parameters to control a single virtual object, limiting the interaction to one VR prop. Richer, more immersive experiences could be developed which allow a user to control multiple virtual objects which can interact with each other in the real and virtual worlds. I foresee two main tasks to extend the pipeline to make it suitable for multiple tasks. Firstly, the tracking algorithm must be adapted to track the behaviour of more than one physical object, while ensuring that the frame rate remains fast enough for VR. As well as this, the interaction between a user and multiple objects should be explored, for example through an immersion or perception study, to make the system feel as intuitive and natural as possible.

5 Conclusion

In a virtual reality application, the method which facilitates interaction between a participant and the surrounding virtual environment must be considered carefully to maximise the feeling of immersion. In my doctoral thesis, I will address this problem by researching methods for tracking non-rigid objects for virtual reality. In my work to date, I have built an end-to-end pipeline for transporting real objects into virtual and augmented environments and, as part of this pipeline, designed a novel neural network based tracking approach for rigid and non-rigid objects. This pipeline has been used successfully to create interactive props for several objects with different sizes and appearances. Following on from the initial research, several areas of possible future work have been outlined: non-rigid object representations, less controlled tracking environments and multiple VR props. These areas are potential future directions which I plan to investigate during the remainder of my doctoral studies.

6 Acknowledgements

My doctoral research is a collaboration between the University of Bath and Marshmallow Laser Feast. I would like to thank my academic supervisor, Darren Cosker, my industrial supervisor, Robin

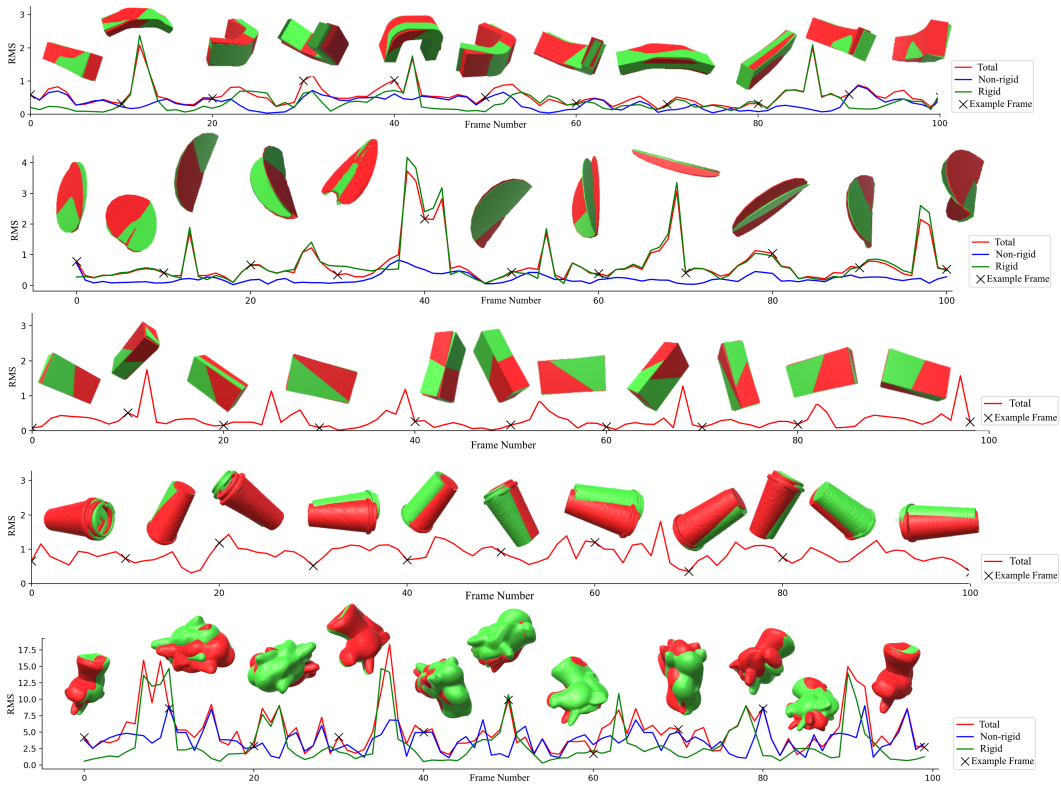

Figure 4: Predicted shape and pose on sequence of synthetic data using VRProp-Net. The RMS error between the predicted and ground truth mesh is calculated for each frame. The ground truth mesh (green) and the predicted mesh (red) are shown for a selection of frames. The total RMS error can be divided into the contributions from rigid and non-rigid transforms.

McNicholas, and the team at MLF for their help and support through my doctoral studies so far.

References

[1] 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. Learning dexterous in-hand manipulation. *CoRR*, abs/1808.00177, 2018.

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

[3] HTC. Discover virtual reality beyond imagination. `https://www.vive.com/uk/`.

[4] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

[5] A. Kanazawa, S. Kovalsky, R. Basri, and D. Jacobs. Learning 3d deformation of animals from 2d images. In *Computer Graphics Forum*, vol. 35, pp. 365–374. Wiley Online Library, 2016.

[6] L. Kausch, A. Hilsmann, and P. Eisert. Template-based 3d non-rigid shape estimation from monocular image sequences. In *Proceedings of the conference on Vision, Modeling and Visualization*, pp. 37–44. Eurographics Association, 2017.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[8] Microsoft. Microsoft hololens — mixed reality technology for business. `https://www.microsoft.com/en-us/hololens`.

[9] Oculus. Oculus rift. `https://www.oculus.com/rift/`.

[10] Y. Park, V. Lepetit, and W. Woo. Multiple 3d object tracking for augmented reality. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pp. 117–120. IEEE Computer Society, 2008.

[11] A. Pumarola, A. Agudo, L. Porzi, A. Sanfeliu, V. Lepetit, and F. Moreno-Noguer. Geometry-aware network for non-rigid shape prediction from a single view. 06 2018.

[12] J. Rambach, A. Pagani, and D. Stricker. [poster] augmented things: Enhancing ar applications leveraging the internet of things and universal 3d object tracking. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pp. 103–108. IEEE, 2017.

[13] C. Taylor, R. McNicholas, and D. Cosker. Vrprop-net: Real-time interaction with virtual props. In *ACM SIGGRAPH 2019 Posters*. ACM, 2019.

[14] C. Taylor, C. Mullanay, R. McNicholas, and D. Cosker. Vr props: An end-to-end pipeline for transporting real objects into virtual and augmented environments. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2019.

[15] H. Tjaden, U. Schwanecke, and E. Schomer. Real-time monocular pose estimation of 3d objects using temporally consistent local color histograms. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 124–132, 2017.

[16] A. Tsoli and A. A. Argyros. Tracking deformable surfaces that undergo topological changes using an rgb-d camera. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 333–341. IEEE, 2016.

[17] Vicon. Motion capture systems. `https://www.vicon.com/`.

[18] Vicon. Origin by vicon. `https://www.vicon.com/press/2018-08-13/origin-by-vicon`.

[19] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[20] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.