# Data-driven haptic feedback utilizing an object manipulation data-set

Athanasios Ntovas, Lazaros Lazaridis, Alexis Papadimitriou, Athanasios Psaltis,
Apostolos Axenopoulos, Petros Daras
The Visual Computing Lab - Centre for Research and Technology Hellas/Information Technologies Institute
{atdovas@, lazlazari@, alexp@, at.psaltis@, axenop@, daras@} iti.gr

## Abstract

*This study presents an ongoing work on a new large-scale, user-object interaction data-set incorporating visual, sensorial and positional modalities, which can potentially be used for (a) assessing vision-related machine learning models for different tasks targeting scene understanding, such as activity recognition, visual object affordances and object detection; (b) providing realistic interactions in the Virtual Reality (VR) world; (c) enhancing 3D perception in robotic applications such as manipulation.*

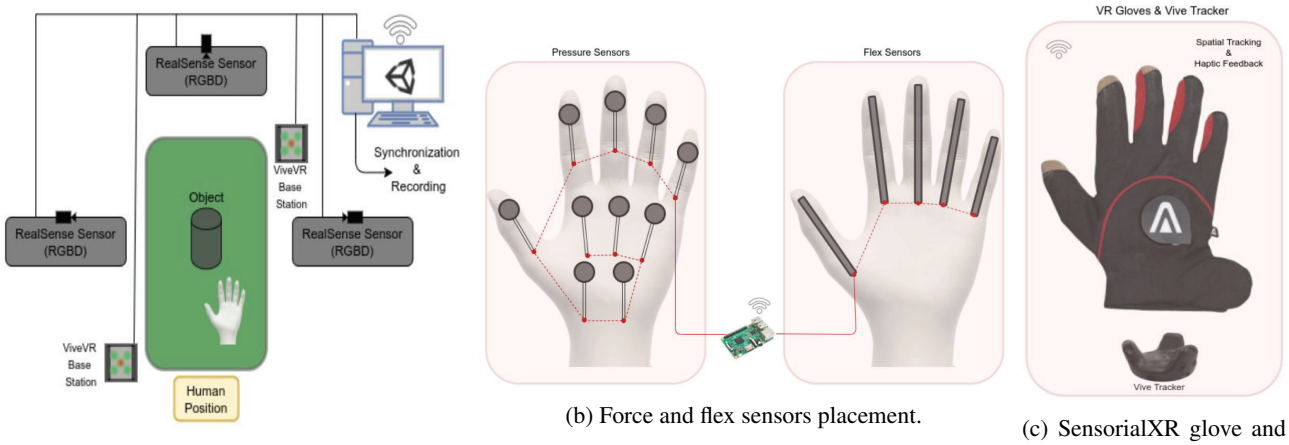
*The aim is to provide a large and diverse set of stereo video sequences, filmed from multiple cameras and involving multiple actors, together with sensorial and positional data recorded in our lab's premises. The data-set is utilized as a first effort to provide realistic haptic feedback to a user interacting with a 3D object in a virtual environment. This data-set is expected to bridge the aforementioned gap between theory and application and facilitate the development of techniques which allow robots to better understand their surroundings. A set of experiments and a preliminary analysis show promising results and demonstrate the particular characteristics of the involved representation schemes.*

## 1. Introduction

Over the last decade, deep learning techniques in the computer vision domain have succeeded in understanding and performing tasks such as object detection[2][14] and image segmentation[15], among others, however, they have yet to show a significant progress when it comes to bridging the gap between theory and application and in particular when required to transfer this knowledge to robotics. The problem of scene understanding[16] has therefore increasingly gained attention in the computer vision community[7] who now focus on developing visual methods which can be used in real robots. Two example domains that have attracted attention in the last few years and are considered to be essential to robots understanding and interacting with their environment are visual affordances[8][5] and haptic feedback[13].

Besides visual and physical properties, objects can also be characterized by their functional aspects, appropriately named object "Affordances". Affordances are widely used for action prediction and anticipation. There are multiple methods in the literature that predict future actions based on the affordance information[10][11]. This happens because affordances indicate a set of possible actions that can be potentially performed in a given environment. This property of affordances can be extended and used in recognizing actor's activities in a way to completely understand a scene. Many studies used affordances in order to recognize human activity[19][18]. Affordances have considerable impact on the accuracy of object recognition vision systems. The contextual detail provided by affordances in a scene offers important cues in object classification tasks[4][17]. Affordances take object detection one step further. They offer extra knowledge providing an intuition about the object's role in the scene[20]. Last but not least, there is a vast amount of literature on detailed scene understanding via categorizing objects based on their specific functionality[3][9].

With regards to haptic feedback, and to the best of our knowledge, the haptic sensation obtained through robotic or virtual interaction is severely poor compared to the sensation obtained through physical interaction. In our physical life, the haptic channel is pervasively used, such as perception of stiffness, roughness and temperature of the objects in external world, or manipulation of these objects and motion or force control tasks such as grasping, touching or walking etc. In contrary, both in terms of haptic feedback received during robotic system teleoperation[21] and during interactions in a virtual world[1], haptic experiences are fairly poor in both quantity and quality as it is difficult for robots to learn complex representations by combining modalities, such as vision and touch[12]. With the booming of Virtual Reality (VR) in many areas such as medical simulation, robotic teleoperation and product design, there is an urgent requirement and a long-term goal of the respective industry to provide an as much as possible immersive experience to the user, aiming to improve the realism of haptic feedback




(a) Spatial setup.

(b) Force and flex sensors placement.

(c) SensorialXR glove and HTC's Vive tracker.

Figure 1: (a) Spatial setup. (b) Force and flex sensor placement. (c) VR data glove and Vive's positional tracker.

for VR systems, and thus to achieve equivalent sensation comparable to the interaction in a physical world. So far, existing VR systems have managed to provide fairly realistic visual and auditory feedback, however they are still lacking in providing realistic haptic feedback, especially when considering the fact that human users can perceive the physical world through its abundant haptic properties.

In the following sections we present our ongoing effort of building a user-object interaction data-set, which incorporates data recorded through visual, sensorial and positional modalities.

## 2. Experimental setup and components

In order to obtain the data-set that will ultimately be used to train our deep learning models, capable of learning the intrinsic properties of real-life human-object interactions and thus be able to infer corresponding realistic haptic feedback values to a user who is interacting with various VR objects within a VR environment, various components have been employed to both measure values deriving from different modalities, e.g., visual or sensory. Unity3D, an open-source game development tool, was used to collect positional data in its VR environment, while a Raspberry Pi was connected to the adopted flex and pressure sensors to enable the collection of sensorial data. The complete hardware and spatial setup is shown in Figure 1(a) and individual components are further elaborated in the following subsections.

### 2.1. Visual data

In order to collect data from the visual modality RealSense D415 [1] stereo imaging cameras were used, placed around the table where the user-interaction was performed, allowing capturing of a complete user-object manipulation process in both RGB and Depth streams. Three cameras were placed in the scene to record the interactions from multiple perspectives.

### 2.2. Positional data

To obtain reliable pose sensing information, the commercially available SensorialXR[2] VR gloves were used, which utilize 15 Inertial Measurement Units (IMUs) sensors to provide pose sensing capabilities, and also 10 haptic actuators capable of delivering 1024 levels of vibrational intensity. Moreover, HTC's Vive Tracker[3] was also utilized by attaching it to the VR glove, which allowed for accurate live tracking of a hand's position, rotation, speed and acceleration within Unity3D's VR environment during hand-object interaction and in essence provide global hand positioning. The VR glove and HTC's Vive tracker are shown in Figure 1(c).

### 2.3. Sensorial data

Attached to the aforementioned VR gloves are 5 flex[4] and 10 force FSR X 402[5] [6] sensors, shown in Figure 1(b), capable of measuring flex and force values. Flex sensors have the property that as the sensor is flexed, the resistance across the sensor increases and they can therefore offer measurements regarding how much each finger was bent during the interaction. Force sensitive resistors on the other hand, offer a small active sensing area from 0.3N to 50N which will vary each resistance depending on how much pressure is being applied to the sensing area. The harder the force, the lower the resistance. They can therefore be

[1] https://www.intelrealsense.com/depth-camera-d415/

[2] https://sensorialxr.com/

[3] https://www.vive.com/us/accessory/vive-tracker/

[4] https://cdn.sparkfun.com/datasheets/Sensors/ForceFlex/FLEXSENSORREVA1.pdf

[5] https://www.interlinkelectronics.com/fsr-x-400

[6] https://www.sparkfun.com/datasheets/Sensors/Pressure/fsrguide.pdf

used to measure the amount of force a user is applying to each object through his fingers during an interaction with an object.

## 3. Real-world manipulation data-set

Initially, 70 participants were invited to take part in multiple discrete real-life human-object interactions. Unfortunately, at the time of writing and due to Covid-19 restrictions and lock-down measures, we were only able to perform 18% of total envisaged recordings. The users were specifically asked to enter a recording room, shown in Figure 2, wear the custom data glove and perform several affordances provided by several objects, grouped into four main categories, namely 'Tools', 'Kitchen tools', 'Office objects' and 'Other'. Male and female participants were equally distributed among the subjects and all of them were right-handed to match the right-hand glove used in our experiments. Measures were taken to minimize exposure to possible Coronavirus infections and as a precaution a protective mask was worn by all present individuals at all times.

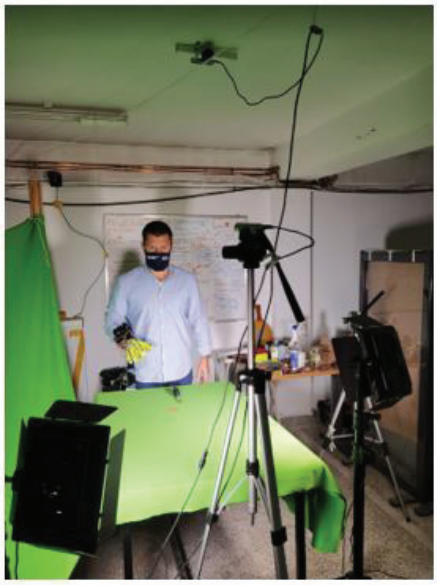


Figure 2: Recording room.

For each discrete recording, which lasted on an average of 15 seconds, the data collected composed the following:

1. Three .bag files, produced by each of the three Realsense cameras placed in the scene, containing both RGB and Depth streams recorded at 15 FPS and a resolution of 640px width by 480px height each.
2. One json file, produced by Unity3D at a rate of 15 frames per second, which stored all data relating to the glove and is as follows:
   - Positional coordinates have been recorded for each of the 5 fingertips and 5 palm positional sensors from SensorialXR gloves. Also accurate positional cordinates were given through HTC VIVE Tracker. All these coordinates were translated into real-world coordinate system using the camera's intrinsic values. Respectively to every positional coordinate part, velocity and acceleration have also been recorded to json files.
   - Flex values produced by the flex sensors at a rate of 15 values per second throughout the duration of the experiment and temporarily stored on the Raspberry Pi for synchronization purposes before being forwarded to Unity3D. As shown in Figure 1(b), 5 flex values were recorded per sample by flex sensors that have been attached to SensorialXR glove.
   - Force values produced by the force sensors at a rate of 15 values per second throughout the duration of the experiment and temporarily stored on the Raspberry Pi for synchronization purposes before being forwarded to Unity3D.As shown in Figure 1(b), 10 force values were recorded per sample by force sensors that have been attached to SensorialXR glove.

In total, 13 users were able to record 69 discrete affordance interactions using 57 objects and overall produce 317 GB of data. Further statistical analysis of the data-set will be provided upon completion of the recordings.

## 4. Realistic haptic feedback

In this section the formal methodology is described to successfully infer haptic feedback signals to the user while interacting with a VR object. We consider the data collected as time series data, i.e., obtained through repeated measurements over time, and thus, several sequence learning methods, such as hidden Markov models and more recently established Recurrent Neural Networks were considered for implementation. Recently however, Long Short-Term Memory (LSTM) models, a type of Recurrent Neural Network (RNN) architecture that possesses the property of remembering values over arbitrary intervals, have proven to be well-suited to classify, process and predict time series data which encompass time lags of unknown duration.

### 4.1. Architecture

The architecture adopted in our scenario is that of an LSTM model. The main characteristic of an LSTM model and its main advantage over other competitors is its ability to remember time series data due to its insensitivity to gap length. Compared to a traditional RNN which has the form

of a chain of repeating modules of neural network, and the actual repeating module has a very simple structure, such as a single tanh layer, the LSTM's repeating module has a different structure and comprises of four NN layers instead of an RNN's single one. This is the key difference that allows them to avoid the long-term dependency problem that is inherent to RNNs.

In order to examine the fundamental functionality of an LSTM in depth, let $\mathbf{X}(t)$ be an input sequence and $\mathbf{P}(t)$ the corresponding target output. An LSTM then maps $\mathbf{X}(t)$ to $\mathbf{P}(t)$ through a series of intermediate representations [6]:

$$\mathbf{I}(t) = \sigma[\mathbf{W}_{xi}\mathbf{X}(t) + \mathbf{W}_{hi}\mathbf{H}(t-1) + \mathbf{B}_i] \quad (1)$$

$$\mathbf{F}(t) = \sigma[\mathbf{W}_{xf}\mathbf{X}(t) + \mathbf{W}_{hf}\mathbf{H}(t-1) + \mathbf{B}_f] \quad (2)$$

$$\mathbf{O}(t) = \sigma[\mathbf{W}_{xo}\mathbf{X}(t) + \mathbf{W}_{ho}\mathbf{H}(t-1) + \mathbf{B}_o] \quad (3)$$

$$\mathbf{G}(t) = \tanh[\mathbf{W}_{xc}\mathbf{X}(t) + \mathbf{W}_{hc}\mathbf{H}(t-1) + \mathbf{B}_c] \quad (4)$$

$$\mathbf{C}(t) = \mathbf{F}(t)\mathbf{C}(t-1) + \mathbf{I}(t)\mathbf{G}(t) \quad (5)$$

$$\mathbf{H}(t) = \mathbf{O}(t)\tanh[\mathbf{C}(t)] \quad (6)$$

$$\mathbf{P}(t) = \mathbf{W}_{hp}\mathbf{H}(t) + \mathbf{B}_p \quad (7)$$

where $\sigma(.)$ is a non-linear scaling factor, $\mathbf{C}(t)$ is the 'internal memory' of the LSTM and the gates $\mathbf{I}(t)$, $\mathbf{F}(t)$ and $\mathbf{O}(t)$ control the degree to which the memory accumulates new input $\mathbf{G}(t)$, attenuates its memory and influences the hidden layer output $\mathbf{H}(t)$, respectively. The LSTM is parametrized by the learnable weight matrices $\mathbf{W}$ and biases $\mathbf{B}$.

**Model Parameters:** As said before, the fist try is to predict realistic haptic feedback inside VR environment. VR glove has vibrators at the same position force sensors have been placed. The goal is that haptic feedback of a specific vibrator must be analogous to the force that has been predicted via LSTM training. So, these 10 force values were defined as the output of LSTM. The input consisted of all the above-mentioned hand-parts positional coordinates, velocity and acceleration from inside the Unity's VR environment, and the 5 flex values from flex sensors that have been recorded during experiments.

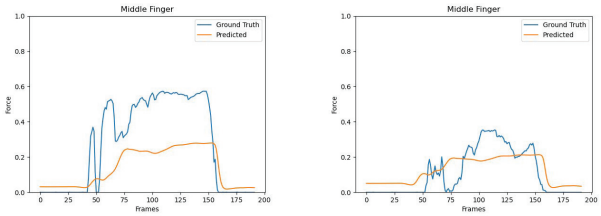
### 4.2. Model evaluation

We performed extensive experiments, fine-tuning our LSTM model using a variety of combinations of loss metrics, learning rate (lr), hidden layers, number of units of each layer of our LSTM model. Specifically, we have conducted experiments examining (a) Mean squared error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) as regression loss functions, (b) values $10^{-3}$, $10^{-4}$, $10^{-5}$ as learning rates, (c) one to three number of hidden layers and, (d) 16, 32, 64 number of units per layer, however, due to page limitations only the best performing model training combination is reported. Preliminary results (Table 1) indicate that the best results that managed to reduce our test set's loss were produced when the MSE loss function was used combined with a learning rate of $10^{-4}$, 3 hidden layers, each one containing 16 hidden units.

| Metrics | lr | Layers | Units | Error |
|---|---|---|---|---|
| MSE | $10^{-4}$ | 3 | 16 | 0.01204 |
| MAE | | | | 0.01552 |
| RMSE | | | | 0.01340 |

Table 1: Experimental results.

Moreover, in order to depict the qualitative aspect of our results, we provide Figure 3, which shows a correlation between the force values collected as ground truth and the predicted force values, to be delivered as haptic signals to the user, provided by our LSTM during inference. Prediction diagrams do not seem to perform well, but this is something reasonable since less than 20% of total lab experiments have already been recorded. Performance improvement is expected by the completion of the recordings.




(a) Middle finger force values. (b) Ring finger force values

Figure 3: Ground truth vs predicted force values (normalized to [0,1]) for the (a) middle finger, and (b) ring finger.

## 5. Conclusion and future work

In this work, we presented our ongoing work on a novel, large-scale data-set, which can potentially be used for a variety of tasks, such as scene understanding, action recognition, 3D perception, among others. We presented the existing status of our collected data-set, the equipment utilized to record the data, the types of data recorded, as well as a preliminary analysis of evaluation metrics considered for the task of providing realistic haptic-feedback in a virtual environment, conceptualized by Unity3D game development software.

Due to Covid-19 related restrictions, we were not able to complete the development of the data-set, however we aim to do so in the near future. In order to evaluate the model in real-life scenarios we plan on conducting user tests and upon completion asking users to complete a questionnaire using a 7-point Likert scale.

## References


[1] Wang Dangxiao, Guo Yuan, Liu Shiyi, Yuru Zhang, Xu Weiliang, and Xiao Jing. Haptic display for virtual reality: progress and challenges. *Virtual Reality & Intelligent Hardware*, 1(2):136–162, 2019. 1

[2] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014. 1

[3] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR 2011*, pages 1529–1536. IEEE, 2011. 1

[4] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR 2011*, pages 1961–1968. IEEE, 2011. 1

[5] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *arXiv preprint arXiv:1807.06775*, 2018. 1

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[7] Derek Hoiem, James Hays, Jianxiong Xiao, and Aditya Khosla. Guest editorial: Scene understanding. *International Journal of Computer Vision*, 112(2):131–132, 2015. 1

[8] Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, David Abel, and Doina Precup. What can i do here? a theory of affordances in reinforcement learning. In *International Conference on Machine Learning*, pages 5243–5253. PMLR, 2020. 1

[9] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011. 1

[10] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 1

[11] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015. 1

[12] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950. IEEE, 2019. 1

[13] Xiaoguo Li, Anthony Meng Huat Tiong, Lin Cao, Wenjie Lai, Phuoc Thien Phan, and Soo Jay Phee. Deep learning for haptic feedback of flexible endoscopic robot without prior knowledge on sheath configuration. *International Journal of Mechanical Sciences*, 163:105129, 2019. 1

[14] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3367–3375, 2015. 1

[15] Xin Liu, Bin Dai, and Hangen He. Real-time object segmentation for visual object detection in dynamic scenes. In *2011 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 423–428, 2011. 1

[16] Anh Nguyen. Scene understanding for autonomous manipulation with deep learning. *arXiv preprint arXiv:1903.09761*, 2019. 1

[17] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011. 1

[18] Siyuan Qi, Siyuan Huang, Ping Wei, and Song-Chun Zhu. Predicting human activities using stochastic grammar. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1164–1172, 2017. 1

[19] Tuan-Hung Vu, Catherine Olsson, Ivan Laptev, Aude Oliva, and Josef Sivic. Predicting actions from static scenes. In *European Conference on Computer Vision*, pages 421–436. Springer, 2014. 1

[20] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2855–2864, 2015. 1

[21] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2855–2864, 2015. 1