Supplemental Material for Mixing Modalities of 3D Sketching and Speech for Interactive Model Retrieval in Virtual Reality

ANONYMOUS

ACM Reference Format:
Anonymous. 2019. Supplemental Material for Mixing Modalities of 3D Sketching and Speech for Interactive Model Retrieval in Virtual Reality. In *Proceedings of IMX '21 Talks.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this supplemental material, we describe the motivation of our choice of using a Wizard of Oz (WoZ) approach for the speech interaction (Section 2). Then we provide the sequence of user sessions for our second experiment (Section 4), show the dictionary and chair collection (shape and colors) we generated by colour permutations (Section 3).

2 WHY DO WE CHOOSE A WIZARD OF OZ APPROACH?

In this section we introduce the state of the art of Natural Language Process algorithms, we define the entire speech interaction pipeline and discuss the advantages and disadvantages of using automatic processes or the human component.

2.1 State of The Art of Natural Language Process

In the last decade, deep learning algorithms have been designed to process text data for classification tasks such as sentiment analysis or text comprehension. Recurrent Neural Networks (RNN) have been effectively utilised for text analysis [5], however they are prone to suffer from vanishing gradients [4], implying that they can forget the context and thus neglect parts of the sentence that are critical for correctly parsing it. More recent models such as Long Short-Term Memory (LSTM) [2] and its variants overcome this memory problem, making it easier to remember data. In addition, recent techniques of mass parallel processing developed by Google Brain and Google Research have helped to improve attention mechanisms, implemented in Transformer model architecture [6]. These NLP models had a considerable improvement in 2018, comparable to the impact of ImageNet [3] in 2012 for Computer Vision. These results make us consider to build a software stack for an automatic pipeline. This pipeline is described in the next section, considering the advantages and disadvantages of each step. Based on this analysis we conclude the automatic pipeline can introduce an error that percolates and/or accumulates in the subsequent steps, we chose to have an experimenter-in-the-loop approach, and so a semi-automatic. We motivate our choice in the next section.

IMX '21 Talks, June 21 — 23, 2021, New York, NY

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

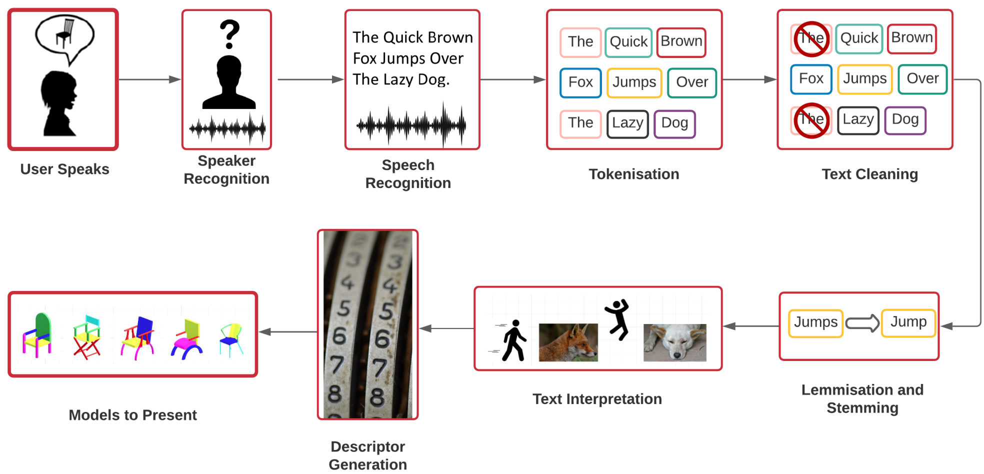

Fig. 1. Our speech pipeline includes: speaker identification, speech recognition, tokenisation, lemmatisation, stemming, text interpretation, descriptor generation, selection of the proposed results.

2.2 Speech Interaction Experiment Requirements

In the experiment, where three types of interaction are compared, while the sketch method was based on Giunchi *et al.* [1] work, the speech interaction was to be designed to be easily included in the system. We analysed the phase of speech interaction, and we split the entire process into stages. All these stages are part of a pipeline. This pipeline manipulates the speech query, as shown in Figure 1.

In our experiment, speech interaction involves a sequence of steps that starts with the user verbally describing the model, and finishes with the presentation of the results to the user. Between these two events, we can enumerate stages identified by the specific required task. This pipeline needs to last a maximum of 5 seconds with the following steps:

Speaker identification. Speaker identification or recognition is the process that allows identifying who is the active speaker. During the experiment, the user can interact with the experimenter by asking clarifications. Thus, it is essential to understand who is talking to avoid another input channel that injects words in the system.

Speech recognition. Speech recognition is the process that converts speech to plain text. This process is challenging, but efficient dictation software can be very accurate when trained with the speaker's voice. We tested three different speech recognition services without training phase, achieving poor results with audio recorded with Oculus equipment. In addition, another potential problem can be the words spoken by the user that are not relevant to the search, such as clarifications asked the experimenter during the test.

Tokenisation. Tokenisation or lexical analysis is the process that converts a sequential text in tokens. Each token is a string with its own length and meaning. Tokenising a text is a very simple operation.

Text cleaning. Text cleaning is the process that removes from the list of tokens all the items that do not contain relevant information for the training of a machine learning model. NLP libraries implement dictionaries that are used to exclude such words from the tokens list.

Lemmatisation and Stemming. Lemmatisation and Stemming are two processes that prepare the text to be used for training a model, grouping words with the same root—while Lemmatisation focuses on the morphological analysis of the word, stemming cut prefixes or suffixes of the considered word.

Text interpretation. Text interpretation is the process that maps the corpus with the meanings that are being tried to convey. In our case, each group of words represents information that describes the object or a part of it. The output of this process is a feature vector that describes the speech query.

Selection of the models. The descriptor generated by the previous step is used to measure the Euclidean distance with the feature vectors of the models in the dataset, to find the closest matches.

2.3 Motivation for the Wizard Of Oz

In this section, we enumerate the problems that each stage of the pipeline can introduce in the system. Because all the steps of the pipeline can be automatic, we analyse them, one by one and we motivate our choice of using a Wizard Of Oz instead of a fully automatic pipeline.

Speaker identification is requested because during the experiment can occur that the participant asks the experimenter some clarifications about the experiment itself, and the nature of the questions can be very different. Normally the experimenter answers and his voice is a different input that should not be used for the experiment. One possible alternative is to instruct the user to avoid any kind of interaction with the experimenter. This stage could be considered optional. An automatic step, in this case, needs to have a 100% accuracy since injecting a different and unrelated input can compromise the query. If a human handles, this phase is naive to consider the voice coming only from the participant.

Speech recognition is an essential step in the pipeline that converts all the words pronounced by the speaker to plain text. During this process, we noticed that the participants sometimes speak without providing relevant information, instead they are react to some events happening in the virtual environment, e.g., commenting on the results or also make some questions to themselves. These chunks of information need to be disregarded as they can inject incorrect information in to the pipeline. Accuracy of speech to text is the most important metric in this case. If the stage is handled in an automatic way, there are different software that can be used. We tested three different speech-to-text services, not fine-tuned with the speaker's voice, that represents a plausible condition for our experiment. The accuracy for all the three services was not satisfactory, and we present the detailed results in Section 2.4. To solve the problem of excluding uninformative chunk, one possibility is to give the user a command that enables the speech to be interpreted, but this is not an optimal solution as one of our pre-condition is to avoid the use of the hands during speech interaction. An experimenter is able to convert speech to text very quickly, but it isn't easy in our experiment configuration to write text in real-time. An efficient alternative is to provide the experimenter with a GUI and instructing the participant to follow a limited dictionary (always visible in VR). This dictionary contains all the meanings valid for the dataset that we designed. In addition, we describe a speech query in order to generate one search each 10 seconds, this is because a continuous speech is challenging to deal with for both for an automatic and a semi-automatic process.

Tokenisation stage is a very simple task to achieve from a corpus both if a computer or a human manages the stage.

Text cleaning stage is another phase of the pipeline that is simple to achieve. If handled by computer NLP libraries contains dictionaries that list all the words that do not give useful information. Also, for a human that is listening to a speaker, the task of dropping words that do not add semantic value to the sentence is easy.

Lemmatisation and stemming stage can be managed by NLP libraries, and also easily by a human with the limitation that we put in our experiment configuration.

Text interpretation and descriptor generation step can be managed automatically by one of the state-of-the-art models such as Transformer model. In this case, anyway, it is necessary to train the model, labelling all the models with many descriptions. A possible way to achieve such meta-information dataset is to use Amazon Mechanical Turk or to hire additional participants to describe a large part of the chairs and complete the colour variational dataset, replacing the colours in the description. However, this activity requires a big effort. On the other hand, the experimenter can be provided with an interface where he can clock buttons that increase the value of specific entries connected to the meanings that the user expressed. In this case, at the end of the query, a feature vector is automatically created and, moreover, can bee synchronised with the current selection in the virtual scene.

Selection of the models is a stage that anticipates the models' presentation and can be handled automatically.

Each stage can introduce errors, independently if a human or a machine manages it. These errors are also difficult to calculate for each stage, but they accumulate over the stages. We notice that speech recognition gave us very low accuracy in some cases when managed by speech-to-text software. The reasons can be found in the different component of the audio profile of each participant (tempo, rhythm, pitch, context), as well as fluency and accents which all can affect the accuracy of the transcript. In addition, we do not use software trained with the voice of each participant, and it is reasonable to experience a worse accuracy if compared with trained dictation software. However, fine-tuning would add an additional training step within the study. Moreover, the microphone we used could inject noise in the system.

On the other side, a human can deal with speech to text conversion easily, and with some limitation related to the domain of the meanings extracted from the dataset, we achieve a reliable semi-automatic system.

2.4 Speech to Text services

In this section, we present the results of some speech to text services fed with the audio files coming from some users that tested our apparatus. We tested 10 audio files in the following services:

(1) Watson IBM (https://speech-to-text-demo.ng.bluemix.net/ from heree)
(2) SONIX (https://sonix.ai/)
(3) google speech to text (https://cloud.google.com/speech-to-text)

We determined the accuracy considering groups of keywords in the speech queries. Each group conveys specific information about the style and/or colour of the chair, or the style and/or colour of a part of the chair. For example, if the speaker told "red straight arm" only that sequence of words (or eventually "straight red arm") will trigger a point in the accuracy score. We achieved 37% of accuracy for the Watson IBM, 60% for the SONIX service, and even only 16% with google speech.

3 DICTIONARY AND DATASET

We define a dictionary that contains the following characteristics describing a chair. Each feature is associated with a value that represents how much that feature impact on the description of the chair. Our dictionary contains the following concepts: Height-Length, Size, Thick, Decorated, Curvy, Modern, Antique, Slatted, Swiveling, Flexible, Stable, Reclinable, Padded, Slanted, Canvased, Missing. Each concept can be associated to the chair globally, or to each component of the segmented chairs: back, seat, arms, legs.

3.1 Chair Shapes and Colours

This section shows the shapes (45) of the chairs we include in the dataset, in Figure 2, and for one chair all the colour permutation (360) in Figure 3 for a total of 16200. This dataset is 5 time larger than the version used in Giunchi *et al.* [1] work.

Fig. 2. The 45 shapes selected for the chairs in the dataset.

Fig. 3. The 360 permutations without repetitions of colour in one type of chair.

4 SECOND EXPERIMENT SCHEDULE

This section shows the schedule of the experiment, in Figure 4.

03-00	08-00	01-00	04-00	05-00	07-00	06-00	02-00	00-00	10-01	15-01	11-01	17-01	12-01	16-01	13-01	14-01	09-01	22-02	26-02	20-02	24-02	18-02	19-02	23-02	25-02	21-02
09-00	11-00	14-00	15-00	17-00	16-00	10-00	13-00	12-00	29-02	30-02	28-02	32-02	27-02	34-02	35-02	31-02	33-02	21-01	25-01	19-01	22-01	18-01	23-01	24-01	20-01	26-01
19-00	18-00	26-00	24-00	22-00	21-00	23-00	20-00	25-00	36-02	42-02	39-02	37-02	41-02	38-02	40-02	44-02	43-02	27-01	30-01	29-01	28-01	35-01	31-01	34-01	32-01	33-01
08-02	01-02	04-02	03-02	05-02	00-02	07-02	02-02	06-02	43-01	40-01	44-01	39-01	38-01	36-01	41-01	37-01	42-01	33-00	29-00	27-00	34-00	28-00	35-00	32-00	30-00	31-00
08-01	01-01	00-01	06-01	03-01	04-01	07-01	02-01	05-01	38-00	39-00	44-00	42-00	41-00	36-00	43-00	40-00	37-00	12-02	13-02	11-02	09-02	16-02	10-02	17-02	15-02	14-02
23-02	21-02	20-02	25-02	19-02	18-02	26-02	22-02	24-02	02-00	06-00	01-00	07-00	08-00	04-00	05-00	00-00	03-00	16-01	17-01	09-01	11-01	10-01	13-01	12-01	15-01	14-01
20-01	21-01	23-01	25-01	22-01	18-01	24-01	19-01	26-01	11-00	17-00	14-00	13-00	12-00	09-00	16-00	15-00	10-00	27-02	32-02	30-02	29-02	28-02	34-02	31-02	33-02	35-02
28-01	32-01	34-01	30-01	29-01	35-01	31-01	27-01	33-01	44-02	36-02	37-02	41-02	40-02	42-02	43-02	39-02	38-02	18-00	22-00	23-00	21-00	20-00	19-00	26-00	24-00	25-00
43-01	37-01	44-01	41-01	40-01	36-01	39-01	42-01	38-01	07-02	04-02	02-02	08-02	00-02	05-02	06-02	01-02	03-02	27-00	34-00	35-00	29-00	30-00	32-00	28-00	33-00	31-00
38-00	43-00	42-00	36-00	40-00	41-00	44-00	37-00	39-00	12-02	10-02	17-02	13-02	14-02	16-02	11-02	15-02	09-02	01-01	07-01	05-01	06-01	03-01	02-01	04-01	00-01	08-01

Speech Interaction
Sketch Interaction
Sketch and Speech Interaction

First number is the chair model
Second number is the type of inteaction (same as colour)

Fig. 4. This table shows the schedule of the experiment with ten users (rows). Each column represents a searching session, where the first number is the searched shape model, while the second number and the colour reveals the interaction type. Each user searched 27 chairs, split into three sessions with 9 chairs with the same type of interaction. Both the models and the interaction are randomised.

REFERENCES

[1] Daniele Giunchi, Stuart James, and Anthony Steed. 2018. 3D sketching for interactive model retrieval in virtual reality. In *Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*. ACM, 1.

[2] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2017. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Networks Learn. Syst.* 28, 10 (2017), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[4] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th ICML (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 1310–1318.

[5] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th ICML*. 129–136.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need