3D Gesture Recognition Applying Long Short-Term Memory and Contextual Knowledge in a CAVE

Dejan Arsić
Institute for Human-Machine Communication
Technische Universität München
Munich, Germany
arsic@tum.de

Martin Wöllmer
Institute for Human-Machine Communication
Technische Universität München
Munich, Germany
woellmer@tum.de

Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
Munich, Germany
Rigoll@tum.de

Luis Roalter
Institute for Media Technology
Technische Universität München
Munich, Germany
roalter@tum.de

Matthias Kranz
Institute for Media Technology
Technische Universität München
Munich, Germany
matthias.kranz@tum.de

ABSTRACT

Virtual reality applications are emerging into various regions of research and entertainment. Although visual and acoustic capabilities are already quite impressive, a wide range of users still criticizes the user interface. Frequently complex and very sensitive input devices are being used, although simple gestures would be preferred. While gesture recognition systems are quite common, see Nintendo's Wii mote, a CAVE has further challenges, as the person can be located in any random position and the gestures are not being performed related to a common fixpoint. Applying an infrared tracking system it is possible to reliably locate the hand and compute 3D trajectories. These are then further analyzed with a Long Short-Term Memory approach, which is able to model sequences of variable length with a higher reliability than HMMs.

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern Recognition; H.5.2 [**Information Systems**]: Information Interfaces and Presentation—*User Interfaces*

General Terms

Algorithms

Keywords

Gesture Recognition, Virtual Reality, Cave, LSTM

MPVA'10, October 29, 2010, Firenze, Italy.

1. INTRODUCTION

Virtual reality applications are nowadays emerging into various research fields. Highly immersive scenarios can be designed easily using powerful SDKs like VirTools. Nevertheless most users of CAVE Systems are not really satisfied with the current interaction schemes. The user interface is frequently either similar to desktop systems or requires special hardware with a bunch of function keys. A more natural HCI is probably based on simple gestures, which are familiar to each user and easy to use. Various vision based gesture recognition systems have already been presented in the past [7, 6, 1], but seem inappropriate for the desired CAVE scenario. The user is usually located at a random position in the CAVE, and his orientation differs as it depends on the wall the user is interacting with. Therefore it is hard to normalize the data in a similar way as a frontal interaction system this usually requires. Furthermore, the user's hands can be occluded by the user's body in inconvenient poses. Therefore it is obvious that one single 2D camera is not sufficient for this task, as it is hard to follow the hand in any constellation and a normalization of the pose is almost impossible, although current works show great advances. Visual tracking methods are further handicapped by the quite difficult lighting situations in a CAVE, as the projections are also visible in the scene and are changing constantly in a highly dynamic scenario.

Hence it seems reasonable to try other means of tracking hands and capturing motions. As state of the art virtual realities rely on marker based tracking systems to estimate the head position, which is required to generate the correct perspective in 3D, we decided to equip the user's hand with a single marker. Applying an infrared tracking system, it is possible to determine its current 3D position and orientation within the CAVE. The resulting trajectory can be further analyzed in order to detect gestures. As these are considered as dynamic and a a segmentation is desired, static classification systems cannot be applied. While most previous approaches used either recurrent neural networks (RNN) or Hidden Markov Models (HMM), we decided to

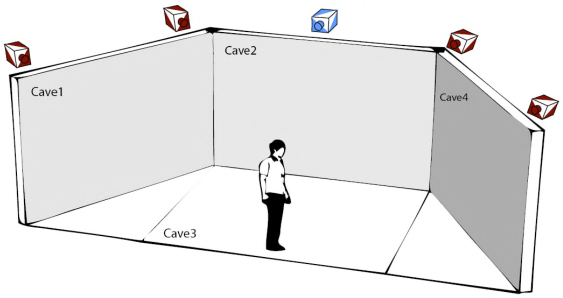

Figure 1: Scheme of the CAVE and the camera setup. Four infrared cameras (red) and one CCD camera (blue) have been installed.

use so called Long Short Term Memory (LSTM) [2] due to its various advantages compared to the previously mentioned approaches. LSTM is an RNN architecture, which is capable to deal with long time-dependencies. Long time lags are usually inaccessible in existing RNN architectures, as the backpropagated error usually blows up or decays exponentially. HMMs have shown quite high performance in various application areas, but seem to be rather complex in handling and have a suboptimal convergence of the Expectation Maximization algorithm and assume the conditional independence of observations, although this is not always provided. LSTM ,in contrast, is able to incorporate previous observations.
This paper is structured as follows: Sec. 2 will briefly describe the marker tracking and feature extraction process. Focus will be set on post processing, as this is crucial. Subsequently we will describe the classification with LSTM in sec. 4 and present first results in sec. 5. Finally we will interprete the results in the conclusion, see sec 6, and a short outlook will be given.

2. FEATURE EXTRACTION

2.1 Hand Tracking

One of the most challenging tasks in gesture recognition is the robust detection of the hands. While various vision based approaches have been presented in the past [5], experience has shown that most camera based tracking algorithms are not able to operate robustly in a CAVE, due to the difficult lighting conditions, that are frequently changing due to the projected images. As a back projection technique is applied, the person in the CAVE is illuminated with varying patterns and colors. Furtheremore, experience has shown large variations in hand pose, as the user will use it permanently in different situations. Therefore we decided to use an infrared tracking system using active infrared cameras, see fig. 1. These emit pulsed infrared light and record frames synchronously. Most of the materials located in the CAVE and the person itself absorb large parts of the infrared light. A highly reflective marker is attached to the persons hand in order to robustly detect the hand's position. Such markers can be robustly detected in the NIR image, as these are usually the lightest spots within the image. As the cameras are calibrated in respect to the CAVE world coordinate system, a triangulation between the detected points in world coordinates can be computed, allowing the exact localization of the marker in 3D space.

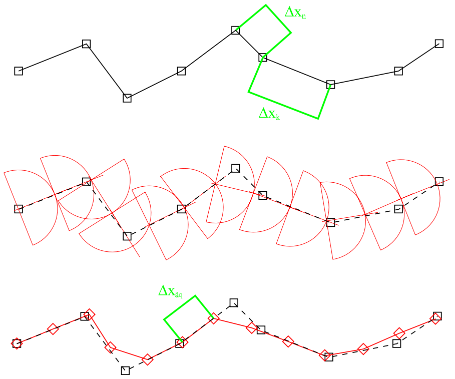

Figure 2: Resampling of the extracted trajectories. The distances of the sampling points of the original curve (top) can be reestimated with simple circles (middle), resulting in a new curve with equidistant sampling pints

2.2 Post Processing

One of the major drawbacks of the infrared tracking system is its latency in tracking. In case multiple markers are located in the scene, it tracks these randomly, and the distance between sampling points varies. This distance can therefore not be used to simply compute velocity and acceleration, as a random order is given during tracking and the time stamps cannot be exactly assigned. Therefore a quite irregular pattern can be observed within the trajectories, see fig. 2, and unequal distances

$$|x_{n+1} - x_n| \neq const. \tag{1}$$

between points. Even classification systems are frequently confused by this behavior. Therefore we decided to post process the data in a similar way hand written data is processed [8]. It has been shown that the sampling points should have equidistant distances. In order to compute a new trajectory it is possible to resample it with a predefined distance, where circles can be used to determine the new positions. Considering all extracted features, a total of 12 features is used for classification.

2.3 Features

The marker position itself can be considered as very weak feature, as no further information is available. In order to cope with this problem, we decided to compute further features based on the marker position. The probably simplest form is the computation of the first and second derivation into the direction of x,y, and z. This will give further information on the direction and the speed of movement. Furthermore turning points within the trajectories can be detected based on the computed derivations. A further feature that has been used is context, in order to cope with the problem of unwanted gestures.

3. THE VRG DATABASE

As this work is quite driven by a concrete application scenario, here interaction in a CAVE, it has been required to record a new database, due to the lack of available data. We

g	l	r	u	d	ps	pl
260	280	265	270	275	265	270

Table 1: Number of recorded gestures.

used the CAVE at our institute, which is illustrated in fig. 1. It uses three walls and a floor projection to create a virtual environment, applying stereoscopic images, which are projected onto all surfaces. Marker tracking is required to detect the position of the polarization glasses and the hand, which is performed by a total of four cameras mounted in the CAVE's corners. A CCD sensor has been mounted on top of the center wall, in order to record visual data in parallel.

29 test subjects have been asked to participate in a wizard of Oz like experiment. They have been asked to point into a particular direction or follow predefined objects. As a mouse like interface has already been implemented this task could be performed on-line. Besides moving a courser the test subjects were asked to manipulate objects and perform predefined gestures. The instructions were displayed on the screens of the CAVE. Feedback has been provided by the hidden operator of the system to provide a more realistic feeling. While recordings were conducted, following gestures were performed by the participants: wave left (l), wave right(r), up (u), down (d), push (ps) and pull (pl). To be able to discriminate between wanted and unwanted gestures pointing has also been recorded and is commonly considered as garbage. Tab. 1 illustrates the amount of recorded classes.

4. RECOGNITION OF GESTURES

4.1 LSTMs

When attempting to continuously detect and classify gestures, a large number of preceding feature frames have to be taken into account in order to capture the dynamics that characterize a certain gesture. The *number* of frames which should be used to obtain enough context for reliably detecting gestures is hard to determine. Thus, a dynamic classifier that is able to *learn* the amount of context is a promising alternative to manually defining fixed time windows for gesture recognition. Dynamic classifiers like Hidden Markov Models are often used for flexible context modeling and time warping. Yet, HMMs have drawbacks such as the inherent assumption of conditional independence of successive observations, meaning that an observation is statistically independent of past observations provided that the values of the hidden variables are known.

Recurrent neural networks can in principle map from the entire *history* of previous inputs to each output. Yet, the analysis of the error flow in conventional recurrent neural nets led to the finding that long range context is inaccessible to standard RNNs since the backpropagated error either blows up or decays over time (vanishing gradient problem). This led to the introduction of Long Short-Term Memory RNNs [4, 2]. They are able to overcome the vanishing gradient problem and can learn the optimal amount of contextual information relevant for the classification task. Thus, LSTM architectures are well-suited for our gesture recognition task.

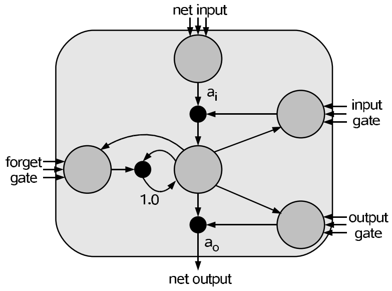

Figure 3: LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state.

An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more memory cells, along with three multiplicative 'gate' units: the input, output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate (see Figure 3). The overall effect is to allow the network to store and retrieve information over long periods of time. For example, as long as the input gate remains closed, the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate.

LSTM networks have demonstrated excellent performance in a number of pattern recognition tasks, including phoneme recognition [3], keyword spotting [9], and emotion recognition [10].

4.2 Context Based Recognition

We have described the basic principle of LSTM in the previous section, where its has been shown, that these are capable to segment longer sequences. As we are only interested in the six mentioned gestures, the system has to deal with unwanted other gestures. These can be simply pointing into any direction or moving the hand into a more comfortable position. Both types of gestures are summed up into a common seventh class, which might be considered as garbage or junk model. An overwhelming ratio of the data is usually either confused with garbage due to possibly similar templates. This is usually happening for instance if a person is pointing from the left screen to the right, as this is more or less the same movement as waving right. Having a closer look, this should not be an error, as the classifier does not know any context and only relies on the marker position. In order to eliminate these confusions, it is required to determine the position of the courser and use a trigger that initializes the gesture recognition system. This would indeed allow a pre segmentation of the data stream. Nevertheless unwanted motion will be included in the data, as start and end of the real gesture are basically unknown and can be either another gesture or simply junk. Tis unwanted data has

	SVM	HMM	LSTM
7 classes	54.63	61.94	64.69
6 classes	59.90	63.12	68.10
6 classes (train 7)	60.21	64.35	72.93

Table 2: Classification results for the 6 class and 7 class problem. The LSTM based approach performs better in all three setups.

to be removed for a reliable classification for most common approaches. Experiments have shown, that the temporal difference between triggering a gesture and performing a gesture can be quite large. Furthermore it is difficult to detect the end of a gesture reliably, as only the start is triggered. Therefore a continuous segmentation is still advantageous, as it can cope with gestures bounded by random data. It is hence reasonable to include a garbage model and additional grammar for the recognition process. This of course happens at the cost of combined gestures. Nevertheless, systems without the garbage class will be trained for comparison.

5. EVALUATION

In the following section we will discuss the current results of the LSTM based approach and compare these to other approaches such as HMM and SVM, which will allow us to demonstrate the performance of the LSTM approach. Tab. 2 shows the results for the continuous classification of the six classes plus garbage in the upper row. As can be seen, the performance of LSTM is significantly higher than the performance of SVMs and HMMs. Furthermore the performance has been tested on segmented data for six classes only, leaving garbage aside, as only the gestures required for interactions are of interest. Due to the reduced number of classes it is obvious that a high recognition rate can be achieved and LSTM performs best again. These results can be outperformed by adding context and including an additional garbage class into the training phase. Although the performance of the SVM and HMM based classification rises with considering an additional junk class, only the improvement of LSTM can be considered as significantly. This result shows, that even partially presegmented data should be classified continuously and the consideration of garbage, which is only required during the training phase.

6. CONCLUSION AND OUTLOOK

We have presented a novel gesture recognition system in 3D space for CAVE environments. Applying marker based tracking allows for high detection accuracy and concentration on the recognition process. LSTM seem to be the first choice for such a problem, as these can robustly model gestures and show a superior performance compared with HMM and SVM. One of the major advantages are the easy and short training phase, which does not require a segmentation of the data, as it is capable to train a model with continuous data streams, where only the order of gestures is known. As various transitions are already included within the data stream, no additional transition models or grammer is required for the robust classification. Future work has to include research on more robust markerless tracking techniques, which might also include other sensors. Furthermore, an alignment of the body pose will remove some confusion between the push and pull gestures, as the direction of interaction hast to be considered.

7. ADDITIONAL AUTHORS

Additional authors: Moritz Kaiser (Institute for Human-Machine Communication, email: kaiser@tum.de) and Florian Eyben (Institute for Human-Machine Communication, email: eyben@tum.de) and Björn Schuller (Institute for Human-Machine Communication, email: schuller@tum.de).

8. REFERENCES

[1] D. Arsić, B. Hörnler, B. Schuller, and G. Rigoll. Resolving partial occlusions in crowded environments utilizing range data and video cameras. In *Proceedings 16th IEEE International Conference on Digital Signal Processing, DSP2009, Santorini, Greece*, July 2009.

[2] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. Learning precise timing with lstm recurrent networks. *Journal of Machine Learning Research*, 3:115–143, 2002.

[3] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, June 2005.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[5] F. Mahmoudi and M. Parviz. Visual hand tracking algorithms. In *Proceedings of the conference on Geometric Modeling and Imaging, GMAI06*, pages 228–232, Washington, DC, USA, 2006. IEEE Computer Society.

[6] M. Moni and A. Ali. Hmm based hand gesture recognition: A review on techniques and approaches. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pages 433 –437, 8-11 2009.

[7] O. Rashid, A. Al-Hamadi, and B. Michaelis. A framework for the integration of gesture and posture recognition using hmm and svm. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, volume 4, pages 572 –577, 20-22 2009.

[8] J. Schenk, B. Hörnler, B. Schuller, A. Braun, and G. Rigoll. Gms in on-line handwritten whiteboard note recognition: The influence of implementation and modeling. In *ICDAR*, pages 877–880, 2009.

[9] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll. A Tandem BLSTM-DBN architecture for keyword spotting with enhanced context modeling. In *Proc. of NOLISP*, Vic, Spain, 2009.

[10] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie. Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks. In *Proc. of Interspeech*, pages 1595–1598, Brighton, UK, 2009.