Vision-Based Handwriting Recognition for Unrestricted Text Input in Mid-Air

Alexander Schick∗, Daniel Morlock∗, Christoph Amma†, Tanja Schultz†, Rainer Stiefelhagen∗,†
∗Fraunhofer IOSB Karlsruhe, Germany
alexander.schick@iosb.fraunhofer.de, daniel.morlock@gmail.com
†Karlsruhe Institute of Technology Karlsruhe, Germany
{christoph.amma, tanja.schultz, rainer.stiefelhagen}@kit.edu

ABSTRACT
We propose a vision-based system that recognizes handwriting in mid-air. The system does not depend on sensors or markers attached to the users and allows unrestricted character and word input from any position. It is the result of combining handwriting recognition based on Hidden Markov Models with multi-camera 3D hand tracking. We evaluated the system for both quantitative and qualitative aspects. The system achieves recognition rates of 86.15% for character and 97.94% for small-vocabulary isolated word recognition. Limitations are due to slow and low-resolution cameras or physical strain. Overall, the proposed handwriting recognition system provides an easy-to-use and accurate text input modality without placing restrictions on the users.

Categories and Subject Descriptors
H.5.2 [User Interfaces]: Input devices and strategies

Keywords
Handwriting recognition, 3D hand tracking

1. INTRODUCTION
With current advances in technology, we see a rapidly increasing availability, but also demand, for intuitive human-machine interaction. Devices are not only controlled by mouse and keyboard anymore, but we are now using gesture-controlled devices in public areas and at our homes. Distant hand gestures in particular removed the restriction to operate a device directly; we can now interact freely with machines while moving around. In this work, we are interested in human-machine interaction that does not force users to touch a specific device or to wear special sensors, but that allows for unrestricted use.

While there is a great variety of human-computer interaction techniques to interact with distant virtual objects, e.g. to select menu entries, there is still a lack for intuitive and unrestricted text input. Although there are ways to input text by using a virtual keyboard on a display [11] or by speech recognition, there are situations where both are not suitable: the first requires interaction with a display and occupies space on it and for the latter, users must speak which is not always possible depending on the surroundings. With this work, we extend the available text input modalities by introducing an intuitive handwriting recognition system.

In the remainder of this work, we present a system that combines vision-based 3D hand tracking with 3D handwriting recognition. It allows users to write characters and words in mid-air. This extends the use of 3D hand gestures to allow for convenient and unrestricted text input. In Section 2, we discuss related work before introducing the proposed system in Section 3. In Section 4, we present the experimental evaluation before concluding in Section 5.

2. RELATED WORK
We will first discuss established approaches for handwriting recognition before focusing on approaches that are closest to the proposed system.

A comprehensive review of handwriting recognition can be found in Plamondon et al. [8]. In our work, we recognize handwriting while the text is written which is called an online handwriting recognition task. Following [8], online handwriting recognition can be classified into structural or rule-based approaches where the trajectory is analyzed by rules and grammars and into statistical methods based on linear discriminant analysis, principal component analysis, neural networks, or Hidden Markov Models (HMM) [9]. In particular, Hidden Markov Models are widely used to model motion trajectories. They can be considered a state-of-the-art technique and are also used in the proposed system.

Handwriting recognition is not limited to paper or digital surfaces, but has also been extended to the third dimension. Amma et al. introduced a wearable text input system based on a motion sensor attached to a glove [1]. Acceleration and angular rate are measured and processed in a left-to-right
HMM for recognition of whole sentences. This system is most similar to our system, but since it is intended for mobile applications it requires the user to wear a glove, which is not desirable in stationary use. Other approaches include Zhou et al. [13] and Kim et al. [4] with the same restrictions introduced by requiring to use handheld or wearable devices. Text input for large displays was also investigated by Shoemaker et al. [11], but for the Wiimote and without handwriting recognition. In contrast, our system has no such requirements and allows unrestricted text input.

In sign language recognition, hand gestures are interpreted as symbols and words. A recent overview can be found in [3]. Depending on the system, we have some techniques in common, in particular the use of computer vision in combination with HMM-based gesture recognition [12]. However, our targeted problem domain differs as in our case, we focus on recognizing continuously written words.

Hand tracking systems are of much interest in computer vision because they are a prerequisite for many forms of human-machine interaction. Examples of vision-based tracking of hands and arms in 3D were given by Azarbayejani and Pentland [2] and Nickel and Stiefelhagen [7], to name just a few. A survey for vision-based gesture recognition can be found in [6]. Gesture-based handwriting recognition with a head-mounted camera was shown by Liu et al. [5], but only for the Graffiti 2 alphabet and only for single characters. In our work, we will combine a vision-based hand tracking system with handwriting recognition to provide a text input modality that does not require additional devices. To the best of our knowledge, we are unaware of vision-based handwriting recognition of whole words based on concatenated individual character models.

3. VISION-BASED HANDWRITING RECOGNITION

The proposed handwriting recognition system consists of three modules: vision-based data acquisition, feature extraction, and handwriting recognition that will now be introduced in more detail.

3.1 Data acquisition

The proposed system recognizes characters and words that are written in mid-air with the hand acting as the pen. Therefore, the 3D trajectory of the hand is of interest. To compute this trajectory, we use a voxel carving approach to reconstruct users in 3D [10]. Based on the user’s pose, we extract the arm and track the 3D trajectory of the hand.

There are also other ways to acquire this data, e.g. with sensors based on structured light like the Microsoft Kinect. But a vision-based system has the advantage that there is a much broader variety of sensors to choose from with varying resolutions and framerates which can be important for the overall performance (Section 4.3). To deal with sensor noise and occasional misclassifications, we applied a Kalman filter to smooth the trajectory. The system setup is described in Section 4.

3.2 Feature extraction

Due to the discrete capturing process of cameras, the trajectory is formed by an ordered set of points. The features used in our system are based on the connecting vectors between each of these consecutive pairs of trajectory points. These connecting vectors are position invariant. To account for variations in writing size, we normalize the connecting vectors which makes them also scale invariant.

When writing in mid-air, users tend to write on a virtual plane in front of themselves. We exploit this by projecting the 3D trajectory of the writing hand on a 2D plane in front of the users which has the advantage that it reduces the feature space. The beginning and ending of handwriting can be easily detected because users usually move their hands forward or backward, respectively. We found that there were no pen-up and pen-down movements while writing as is the case in traditional handwriting. This leads to ligatures being part of the trajectory because we get one continuous stroke instead of alternating pen-up and pen-down movements. Figure 1 shows an example of the extracted 2D features and Figure 2b of trajectories with ligatures.

3.3 Handwriting recognition

Character and word recognition are both based on Hidden Markov Models which is an established technique in handwriting recognition [8]. Specifically, the system is based on left-to-right HMMs. First, for each character, a separate HMM is trained with training data of multiple people (see Section 4 for more details) with the extracted features described above as observations. Then, words can be modeled as a concatenation of character HMMs, i.e. for any given arbitrary word, an HMM can be constructed. A vocabulary defines the set of words that can be actually recognized. The number of words in the vocabulary equals the number of possible classes and, therefore, a larger vocabulary results in a more difficult recognition problem.

We examine two recognition tasks, the recognition of single characters and the recognition of words. Our system
allows the continuous input of isolated words, that means no artificial pauses or segmentation between characters is necessary. Words contain not only a sequence of characters but also the motion between the characters. For word recognition, we therefore perform an additional EM-training on word data to refine the beginning and end of character models in order to model the transitional motion between consecutive characters without introducing additional states.

The proposed system recognizes handwriting in mid-air based on the 3D trajectory of the writing hand. We want to note that it can also be used for general purpose gesture recognition if the gesture is based on a continuous hand trajectory.

4. EXPERIMENTS

We will first discuss the technical evaluation of the proposed system before discussing qualitative feedback that we received during data acquisition. To record the hand’s trajectory, a two-camera setup [10] was used in all experiments. The cameras captured 30 frames per second with a resolution of 640×480 pixels. While the hand tracking system is able to run with 50 fps, it was limited by the cameras to 30 fps.

We asked users to stand in front of a videowall in a distance of about 1.5 m (see Figure 2a). We instructed users to write large, block capital letters. The trajectory of the writing hand was displayed in front of the users on the videowall as visual feedback. The beginning and ending of each word was recognized by moving the hand forward or backward, respectively. Before we started collecting the data, we let users familiarize themselves with the system first. The data collection was part of a larger study; however, in this paper we focus on the handwriting recognition as it is our contribution. Details of the full experiment will be made available.

4.1 Technical evaluation

We evaluated both the accuracy of recognizing single characters and words. All evaluation was done in a person-independent manner. We asked five users to write each letter of the alphabet five times. These samples build the character recognition dataset $D_C$. For the word recognition, we selected 40 words from a list of frequent English words. The length of the words were equally distributed between 2 and 11 characters, e.g. to, about, support, and performance. We asked 16 users to write each word one time. These samples were collected in the word recognition dataset $D_W$. The datasets are shown in Table 1.

For character recognition, we used a leave-one-out cross-validation. The best performance with 86.15% recognition rate was achieved by a system with 24 states per HMM with 10 GMMs per state with diagonal covariances. We further used silence states to model no-motion at the beginning and end of each HMM. We explain the rather high number of Gaussians by two factors. First, the handwriting styles varied for each person even for block capital letters (see Figure 2b). Second, the low number of sample points for each trajectory due to limited frame rates of the cameras can lead to varying trajectories even for the same writing style. The low sample rate will also be discussed in Section 4.3. Figure 3 shows the confusion matrix for character recognition. While recognition is generally good, we observe typical confusions between N and W, and A and H which have also been observed for a glove based system [1]. Specific to our system is the confusion of letters E, F, and G with O. This can be explained by the continuous writing style of characters together with the position independence of the features: when writing each of these letters, one starts on the left and moves to the top, makes a clock-wise circle before moving the hand to the right again; hence the confusion. Figure 4 gives a visual example.

For word recognition, the HMMs were build by concatenating the character HMMs and refining them based on word examples (see Section 3.3). The character models were based on $D_C$. To evaluate the word recognition accuracy, the dataset $D_W$ was used. As mentioned in Section 3.3, the word models were build by concatenating the individual character models and additionally trained on word data of the training set to refine the transitions between characters. The training and evaluation was done in a leave-one-out cross-validation setting. For the $D_W$ dataset, the system achieved average recognition rates of 97.54%. This

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Users</th>
<th>Samples</th>
<th>Total samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_C$</td>
<td>characters</td>
<td>5</td>
<td>5 × 26</td>
<td>660</td>
</tr>
<tr>
<td>$D_W$</td>
<td>words</td>
<td>16</td>
<td>40</td>
<td>640</td>
</tr>
</tbody>
</table>

Figure 4: Similar trajectories. The images show similar trajectories for characters E and G. Ligatures are highlighted in red.
high recognition rate is due to the relatively low number of words and the high number of varying writing styles in the training set which lead to a good generalization. We expect this rate to drop with less samples and a higher number of words. However, when evaluating the same models trained with the 40 words of the \( D_w \) dataset against an increased vocabulary of 99 words, the recognition rates were still 86.60\% even though the number of classes to recognize was more than doubled.

4.2 User feedback

While recording the \( D_w \) dataset, we also gathered user feedback about the system. Handwriting in mid-air was generally perceived as intuitive and easy to understand. Users almost immediately understood the system and quickly felt comfortable using it. However, they criticised the low input speed and the physical strain. The low input speed was also confirmed by our measurements: it took on average about three seconds per character which can be explained by the hand tracking system (see Section 4.3). Low input speed leads to longer writing times which ultimately is responsible for increased physical strain.

Surprisingly, visual feedback was generally perceived as disturbing. While we first expected it to help users to see what they were writing, the continuous trajectory was generally perceived as confusing. Because our system does in no respect depend on the visual feedback, this source of disturbance can easily be avoided. Further, it would also be possible to display machine-written single characters on-the-fly when they are recognized.

4.3 Improvements

We will now discuss the major shortcomings and possible improvements. While the recognition rates were good for both character and word recognition, the system was criticized for taking too much time and being physically too straining. These problems can be explained by the hand tracking system, specifically the camera framerates of 30 \( fps \) and resolutions of 640x480 pixels. While 30 \( fps \) is generally considered real-time in computer vision, it turns out that this is not sufficient for the proposed handwriting recognition system. When writing quickly, only few points of the trajectory are sampled. (In comparison, wearable handwriting recognition devices have sample rates of over 800 samples per second [1].) These sparse samples are not sufficient to robustly recognize handwriting with the proposed system when writing quickly and, therefore, users had to write rather slowly. Further, due to the low resolution, small hand movements were not clearly visible thus forcing users to write larger letters. Both problems can be addressed by using cameras with higher framerates and higher resolutions to allow for higher writing speeds and smaller characters. While the current system is well suited to input small amounts of text, we recommend using higher framerates in future vision-based handwriting recognition systems.

5. CONCLUSION

We presented a vision-based handwriting recognition system for person-independent single character and word recognition in mid-air. The system achieved recognition rates of 86.15\% for character recognition and 97.54\% for word recognition. The only requirement of the system is that users are visible for the cameras. Apart from that, neither a display nor wearable sensors are required, thus offering users an intuitive way to insert text anywhere in a given area. This enables us to use natural text input in the fast-growing field of gesture-based interactive systems. In future work, we want to experiment with faster cameras and higher resolutions to allow for increased input speeds, smaller characters, and less physical strain.

6. ACKNOWLEDGMENTS

This work was partially supported by the FhG Internal Programs under Grant No. 092026.

7. REFERENCES