Articulation-Based Retrieval of Stroke Gestures

Inês Cardoso Oliveira
University of Luxembourg
Luxembourg
i.oliveira@uni.lu

Nuwan T. Attygalle
Université catholique de Louvain
Belgium
nuwan.attygalle@uclouvain.be

Réjean Plamondon
Polytechnique Montréal
Canada
rejean.plamondon@polymtl.ca

Luis A. Leiva
University of Luxembourg
Luxembourg
luis.leiva@uni.lu

Abstract

A detailed understanding of gesture articulation, i.e. how gestures are produced, can provide designers with valuable insights for developing more intuitive graphical user interfaces. Over the years, several models have been proposed to model and understand human movements, and in particular handwriting movements. Recently, Deep Learning (DL) models have become a popular approach to handwriting synthesis, being able to produce high-quality realistic samples. However, it is unclear whether these models encode articulation features. To bridge this gap, we conduct a retrieval task. As the ground truth, we employ the Sigma-lognormal ($\Sigma\Lambda$) model, a well-established descriptor of human movements. Experimental results reveal that DL embeddings and $\Sigma\Lambda$ features are mostly unrelated, suggesting a gap between these two.

Index Terms

Stroke gestures, Information Retrieval, Kinematic Theory, Deep Learning

I. Introduction

The ubiquitous adoption of touchscreen-based devices, such as smartphones and tablets, has established "stroke gestures" as a critical input modality for interacting with graphical user interfaces (GUIs). Essentially, stroke gestures are representations of a two-dimensional trajectory of one or more contact points on a sensitive surface. This type of interaction has also been extended into other domains, such as Augmented Reality glasses [4] and head-mounted displays [16], which reinforces its versatility and continued importance in the future.

Human-Computer Interaction (HCI) research has shown that stroke gestures can be employed as efficient shortcuts to access different applications or functions in a system [1]. For GUI designers, understanding gesture articulation,[1] i.e., how stroke gestures are produced, is paramount when making design decisions, for example, knowing which gestures are easier to perform, and thus easier to recall, or identifying gestures with greater variability within a population, which might not make good candidates for shortcuts. This information can be obtained through controlled studies with real users or, alternatively, by using computational models of human handwriting, saving considerable time and effort. Thus, having a effective articulation-based gesture retrieval system could provide valuable insights for the design of intuitive and user-friendly commands for GUIs.

Modeling human movements has long been a subject of interest, and many methods have been proposed; e.g., based on behavioral models [13], kinematic models [9], or minimization principles [5]. The Kinematic Theory [11] and its latest instantiation, the Sigma-Lognormal ($\Sigma\Lambda$) model [12], describes movements by modeling velocity profiles using lognormal functions, and is among the most accurate approaches so far [3]. Specifically for stroke gesture data, the Kinematic Theory has been successfully used for synthesis and analysis [6], [7]. Recently, Deep Learning (DL) approaches to handwriting synthesis have become increasingly popular, due to their capacity to generate high-quality and realistic samples [2], [8], [14]. Still, it is currently unclear whether these models encode any relevant information about articulation execution that can be leveraged to build a workable retrieval system.

In this paper, we explore whether DL models for handwriting synthesis do effectively capture articulation-based features. We analyse three different DL models: Style-disentangled Transformer (SDT) [2], Diffusion model for Handwriting Generation (DHG) [8], and DeepWriteSYN (DWS) [14]. To collect accurate articulation information, we employ the $\Sigma\Lambda$ model [12]. Our hypothesis is that similar velocity patterns mean similar gesture articulations.

II. Methodology

To evaluate the extent to which DL models for handwriting synthesis encode features related to the production process of gestures, we design and conduct a retrieval task. First, we use the $\Sigma\Lambda$ features to perform retrieval and treat these results as ground truth. We then repeat the same retrieval task using the embeddings extracted from the DL models (SDT, DHG, and DWS) and assess their ability to retrieve samples similar to the ground truth. Embeddings are lower-dimensional representations of the input data that encode complex meaningful information in a compact form. We focus on evaluating the presence of articulation information in the latent space, as it reflects the abstract features the model has learned during training.

[1]The term 'articulation' refers to the characterization of the kinematics, that is, the motion, of the hand or utensil used to produce the strokes, encompassing features such as trajectory, velocity and length.

A. Dataset

We perform experiments on $1-GDS, the most popular unistroke gesture dataset in HCI. The dataset contains 16 gesture classes executed by 10 users. Each user provided 10 samples per gesture at 3 articulation speeds (slow, medium, fast) using an iPAQ Pocket PC. Considering that a user likely executes the same symbol in a consistent way within their preferred articulation speed, for our study we use only one sample per gesture per user for slow and fast speeds. As such, the total number of samples in our experiments is 320.

B. Ground Truth Construction

The $\Sigma\Lambda$ model assumes that a complex handwritten trace can be decomposed into a series of primitives[2] that connect a sequence of virtual targets. The velocity profile of the ith primitive is described by a lognormal-shaped function, which is scaled and time-shifted by parameters D_i and t_{0_i}, and characterized by the logtime delay μ_i and the logresponse time σ_i of the neuromuscular system that produces the trace:

$$||\vec{v}_i(t)|| = D_i\Lambda(t; t_{0_i}, \mu_i, \sigma_i^2) = \frac{D_i}{\sigma_i\sqrt{2\pi}(t-t_{0_i})}\exp\left(\frac{-[\ln(t-t_{0_i})-\mu_i]^2}{2\sigma_i^2}\right) \tag{1}$$

Additionally, each primitive is characterized by its starting and ending angle, θ_{s_i} and θ_{e_i}, and its angular variation can be described as:

$$\phi_i(t) = \theta_{s_i} + \frac{\theta_{e_i}-\theta_{s_i}}{2}\left[1+\text{erf}\left(\frac{\ln(t-t_{0_i})-\mu_i}{\sigma_i\sqrt{2}}\right)\right] \tag{2}$$

The trajectory that produces the handwritten trace $\vec{v}(t)$ is computed as the temporal overlap of each primitive's velocity $\vec{v}_i(t)$:

$$\vec{v}(t) = \sum_{i=1}^{N}\vec{v}_i(t) = \sum_{i=1}^{N}\begin{bmatrix}\cos\phi_i(t)\\ \sin\phi_i(t)\end{bmatrix}D_i\Lambda(t; t_{0_i}, \mu_i, \sigma_i^2) \tag{3}$$

To create our ground truth, each sample was reconstructed with the $\Sigma\Lambda$ model, and the lognormal parameters of each primitive were extracted, using the Sigma-Lognormal extractor proposed by [10]. Since each sample can be composed of a variable number of lognormal functions, we summarized the six parameters of each pen-down trajectory using four statistical measures: mean, standard deviation, maximum, and minimum.

C. Extraction of Deep Learning embeddings

1) Style-disentangled Transformer: SDT consists of three main components: a dual-head style encoder, a content encoder, and a Transformer decoder. The style and content encoders extract features from offline images, and the style decoder produces both character-wise and writer-wise style features through two separate heads. The Transformer decoder passes these features through a multi-head attention layer, whose outputs is fed to a Gaussian Mixture Model (GMM) that models the pen movement distribution autoregressively. Additionally, the multi-head attention layer updates at each iteration with the previous GMM output. We extracted the features at the final iteration of the GMM after it has passed through the multi-head attention layer.

2) Diffusion-Handwriting-Generation: DHG consists of two main components: a text-style encoder, and a diffusion probabilistic model. The text-style encoder, built on a pretrained MobileNetV2 model, extracts both visual and textual features from offline image representations. These extracted features are then fed into a diffusion U-Net model, which is composed of multiple downsampling blocks followed by upsampling blocks, using convolutional skip connections. We extracted features after the output of the last convolutional encoding layer of the diffusion model.

3) DeepWriteSYN: DWS is a Variational Autoencoder that uses Bidirectional Recurrent Neural Nets (RNNs) for both the encoder and decoder. Each RNN cell is built using HyperLSTM units, which not only predict the next token in a sequence but also dynamically generate the weights for the subsequent LSTM cell. We extracted the features from the bottleneck layer (mean and variance of the latent distribution) from the encoder output, after the bidirectional HyperLSTM layers.

D. Retrieval and Evaluation

Data retrieval was performed using KD-trees, an algorithm for efficient nearest-neighbor search. The data was split into 70% training and 30% testing. In total, four trees were created using the training data: one tree using the $\Sigma\Lambda$ ground-truth features, and three trees for the embeddings generated by SDT, DHG, and DWS. After construction, retrieval was performed by querying each tree with the test samples.

The retrieved examples from the embeddings of the three models were then compared against the retrieved examples from the $\Sigma\Lambda$ ground truth. To evaluate retrieval performance, we computed Precision@K, which measures how many relevant items are in the top K positions, and Rank-biased Overlap (RBO), which quantifies the similarity of two ranked lists, considering also their order.

[2]As it was done in the early development of the Kinematic Theory [15], we use "primitive" to refer to these decomposed units, as opposed to "stroke", to avoid confusion with the "stroke gestures" term which is very popular in HCI.

III. Experimental Results

Figure 1 shows a t-distributed stochastic neighbor embedding (t-SNE) projection of the ΣΛ ground-truth features of the $1-GDS dataset. Samples that are spatially close tend to exhibit similar velocity profiles, but they do not necessarily belong to the same symbol class. Although there is a noticeable separation between slow and fast samples, this distinction is not absolute. In some cases, slow samples are clustered closely with fast samples. This can occur when a complex symbol, such as a right curly brace drawn at high speed (as seen in sample 6), resembles a simpler shape, like a checkmark, when drawn slowly (as shown in sample 5).

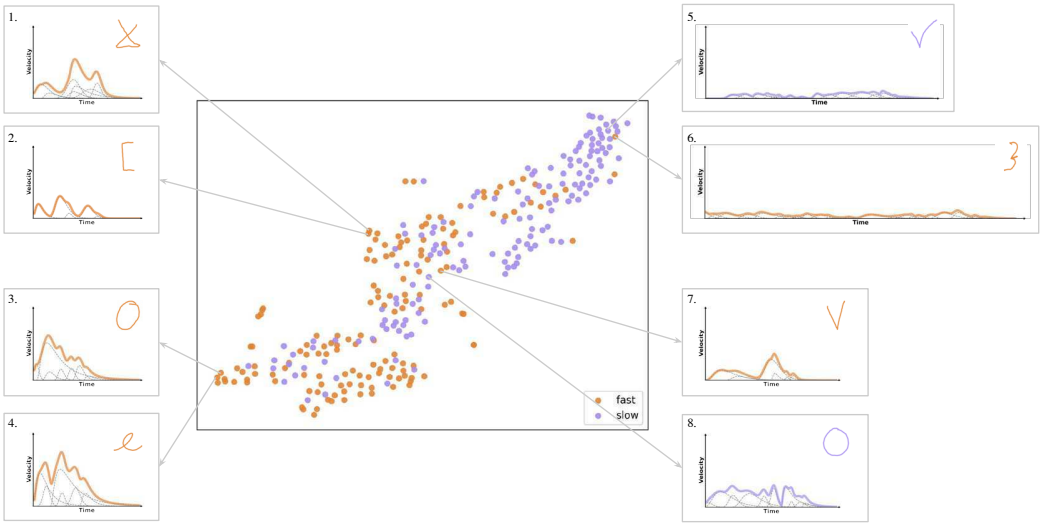

Fig. 1. t-SNE projection of the $1-GDS dataset using ΣΛ features as ground-truth. Selected samples are highlighted with their corresponding velocity profile. We can see that different gesture shapes may have similar articulation characteristics, as indicated by their velocity profiles, further showcasing the relevance of our work.

The retrieval experimental results are summarized in Table I. In addition to the results of the DL models, we consider two different baselines: *Random*, which represents the performance of a random retrieval system, and *Simple*, which retrieves samples based on one elementary feature, the number of points in each gesture.

TABLE I
RETRIEVAL PERFORMANCE RESULTS. TOP RESULT FOR EACH METRIC IS MARKED IN BOLD.

Model	Precision@5	Precision@10	Precision@20	RBO@5	RBO@10	RBO@20
SDT	0.049 ± 0.096	0.073 ± 0.083	0.145 ± 0.082	0.037 ± 0.087	0.049 ± 0.073	0.083 ± 0.062
DHG	0.056 ± 0.094	0.073 ± 0.083	0.118 ± 0.070	0.033 ± 0.073	0.049 ± 0.067	0.074 ± 0.059
DWS	**0.058 ± 0.100**	**0.084 ± 0.089**	**0.150 ± 0.089**	**0.045 ± 0.106**	**0.060 ± 0.086**	**0.092 ± 0.071**
Random baseline	0.015 ± 0.055	0.048 ± 0.066	0.095 ± 0.063	0.016 ± 0.057	0.026 ± 0.050	0.050 ± 0.047
Simple baseline	0.044 ± 0.088	0.077 ± 0.079	0.139 ± 0.082	0.029 ± 0.082	0.046 ± 0.062	0.077 ± 0.061

All DL models are able to outperform the Random baseline, by a considerable margin. However, all metrics have relatively low values across the three models, indicating weak relationship between DL embeddings and ΣΛ ground-truth data. It is worth noting that the standard deviations are high in all cases, suggesting there is considerable performance variability across different queries.

Among the models, DWS achieves the best performance in both retrieving relevant results and ranking them effectively. The performance differences between the three models are minor when retrieving 5 samples, but the gap increases slightly when retrieving 20 samples or more, especially when comparing DWS and DHG. Notably, the Simple baseline performs comparably to DL models, further reinforcing their weak relationship in terms of learned articulation features.

Figure 2 show examples of retrievals obtained with the ΣΛ ground-truth features, together with DWS, DHG, and SDT embeddings. The samples retrieved using the ground truth exhibit greater shape diversity, showcasing their focus on more than visual appearance. In contrast, the DL models appear to retrieve samples based on their visual shape, although this is less pronounced in DWS.

Figure 3 compares the velocity profiles of the queries with those of the retrieved results using the ground truth features. On average, the retrieved profiles visually align well with their corresponding queries.

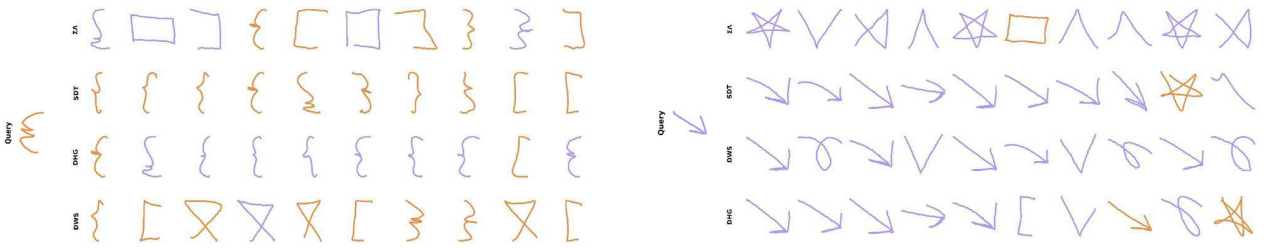

Fig. 2. Top-10 retrieval for two different queries, using ΣΛ, SDT, DHG and DWS features. Purple color indicates slow speed and orange indicates fast speed.

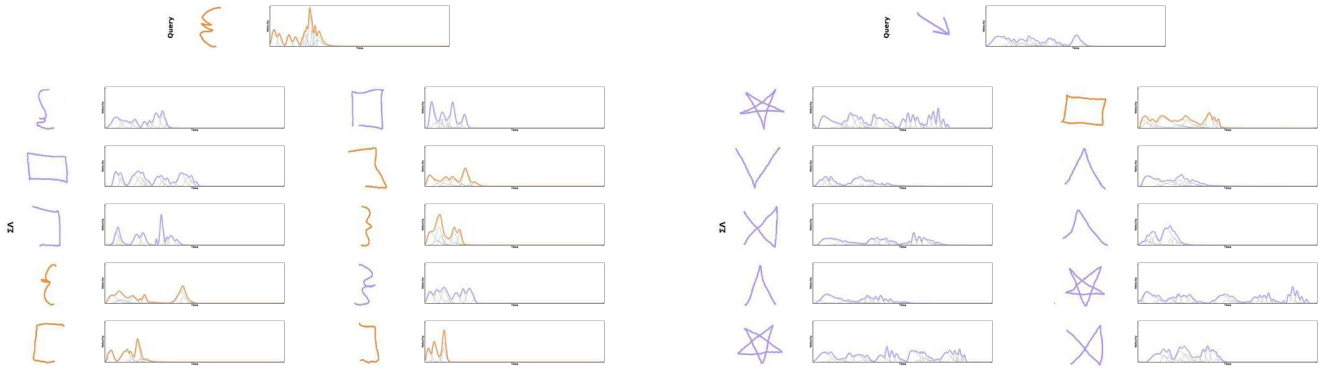

Fig. 3. Velocity profiles of the top-10 retrieval samples for two different queries, using ΣΛ ground truth features. Purple color indicates slow speed and orange indicates fast speed.

IV. Conclusion

In this work, we designed a retrieval task to investigate whether DL models for handwriting synthesis effectively capture articulation features. The results show a weak relationship between DL embeddings and ΣΛ features, suggesting that DL models do not adequately encode execution-related features. This could be due to them prioritizing other aspects of handwriting, like the spatial structure, but it could also be that they do encode articulation features but they differ from the ΣΛ features. One limitation is that we are comparing latent features which lie in entirely different feature spaces. A natural next step would be to use the DL models to fully reconstruct each sample, instead of extract the latent embeddings, then compute their corresponding ΣΛ parameters, and finally assess whether articulation features are preserved.

Acknowledgments

Research supported by Fonds National de la Recherche Luxembourg - FNR (SCRIPTOR project, grant AFR/22/17177001) and the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147).

References

[1] C. Appert and S. Zhai. Using strokes as command shortcuts: Cognitive benefits and toolkit support. *Conference on Human Factors in Computing Systems - Proceedings*, 2009.

[2] G. Dai, Y. Zhang, Q. Wang, Q. Du, Z. Yu, Z. Liu, and S. Huang. Disentangling writer and character styles for handwriting generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5977–5986, 2023.

[3] M. Djioua and R. Plamondon. Studying the variability of handwriting patterns using the kinematic theory. *Human movement science*, 28, 2009.

[4] F. Fang, H. Zhang, L. Zhan, S. Guo, M. Zhang, J. Lin, Y. Qin, and H. Fu. Handwriting velcro: Endowing ar glasses with personalized and posture-adaptive text input using flexible touch sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4), 2023.

[5] T. Flash and N. Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. In *Journal of Neuroscience*, 1985.

[6] L. A. Leiva, D. Martín-Albo, and R. Plamondon. Gestures à go go: Authoring synthetic human-like stroke gestures using the kinematic theory of rapid movements. *ACM Trans. Intell. Syst. Technol.*, 7(2), 2015.

[7] L. A. Leiva, D. Martín-Albo, and R. Plamondon. The kinematic theory produces human-like stroke gestures. *Interacting with Computers*, 29(4), 2017.

[8] T. Luhman and E. Luhman. Diffusion models for handwriting generation, 2020.

[9] D. Meyer, J. Keith-Smith, S. Kornblum, R. Abrams, and C. Wright. Speed-accuracy tradeoffs in aimed movements: Toward a theory of rapid voluntary action. *Attention and Performance XIII*, 1990.

[10] C. O'Reilly and R. Plamondon. Development of a sigma-lognormal representation for on-line signatures. pattern recognit spec issue front handwrit recognit. *Pattern Recognition*, 42:3324–3337, 12 2009.

[11] R. Plamondon. A kinematic theory of rapid human movements. *Biol. Cybern.*, 72(4):309–320, 1995.
[12] R. Plamondon and M. Djioua. A multi-level representation paradigm for handwriting stroke generation. *Human Movement Science*, 25(4):586–607, 2006. Advances in Graphonomics: Studies on Fine Motor Control, Its Development and Disorders.
[13] A. Thomassen, G. van Galen, P. Keuss, and C. Grootveld. *Motor Aspects of Handwriting: Approaches to Movement in Graphic Behavior*. Acta psychologica. North-Holland, 1984.
[14] R. Tolosana, P. Delgado-Santos, A. Perez-Uribe, R. Vera-Rodriguez, J. Fierrez, and A. Morales. Deepwritesyn: On-line handwriting synthesis via deep short-term representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:600–608, 2021.
[15] A. Woch and R. Plamondon. Characterization of bi-directional movement primitives and their agonist-antagonist synergy with the delta-lognormal model. *Motor control*, 14:1–25, 01 2010.
[16] L. Zhan, T. Xiong, H. Zhang, S. Guo, X. Chen, J. Gong, J. Lin, and Y. Qin. Toucheditor: Interaction design and evaluation of a flexible touchpad for text editing of head-mounted displays in speech-unfriendly environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(4), 2024.