

# Writing in The Air: Unconstrained Text Recognition from Finger Movement Using Spatio-Temporal Convolution

Ue-Hwan Kim\*, Ye-Won Hwang\*, Sun-Kyung Lee, Jong-Hwan Kim  
School of Electrical Engineering, KAIST  
Daejeon, Republic of Korea

{uhkim, ywhwang, sklee, johkim}@rit.kaist.ac.kr

## Abstract

*In this paper, we introduce a new benchmark dataset for the challenging writing in the air (WiTA) task—an elaborate task bridging vision and NLP. WiTA implements an intuitive and natural writing method with finger movement for human-computer interaction (HCI). Our WiTA dataset will facilitate the development of data-driven WiTA systems which thus far have displayed unsatisfactory performance—due to lack of dataset as well as traditional statistical models they have adopted. Our dataset consists of five sub-datasets in two languages (Korean and English) and amounts to 209,926 video instances from 122 participants. We capture finger movement for WiTA with RGB cameras to ensure wide accessibility and cost-efficiency. Next, we propose spatio-temporal residual network architectures inspired by 3D ResNet. These models perform unconstrained text recognition from finger movement, guarantee a real-time operation by processing 435 and 697 decoding frames-per-second for Korean and English, respectively, and will serve as an evaluation standard. Our dataset and the source codes are available at <https://github.com/Uehwan/WiTA>.*

## 1. Introduction

As new types of technologies integrate into people’s daily lives, the need for text entry systems that suit the modern mobile devices has emerged [22]. Among various advanced text-entry methods, writing in the air (WiTA), in which people write letters with finger movement in free space, has drawn much attention [46]. Ideal WiTA systems enable people to write text without focusing on the keyboard layout on a tiny screen and implement a natural and intuitive text-entry system, while securing privacy. Applications that would benefit from WiTA by immensely improving user experience include automotive interfaces, remote

\*equal contribution

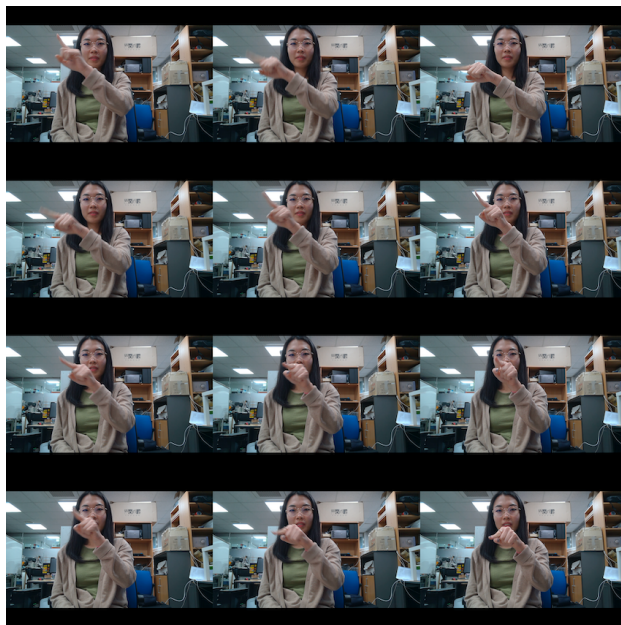


Figure 1. An example instance of the dataset collected in this work. The person in the example is writing “re” from the word “recognized”. WiTA offers a private communication tool for HCI.

signatures, and smart system controls.

Developing feasible WiTA systems is challenging due to the interdependence among the involved gestures and lack of concrete anchors or reference positions [8]. Further, understanding the correlation between various writing patterns and the corresponding characters is complicated—leading to an elaborate task bridging vision and natural language processing (NLP). As a result, contemporary WiTA systems hardly achieve satisfactory performance, which prevents their deployment into real-world applications. Conventional WiTA systems, in general, rely on traditional statistical models with hand-crafted features, which restricts their performance [3, 29]. Although researchers have attempted to apply data-driven approaches for designing WiTA systems, the current datasets available possess multiple limitations.

For instance, [8, 46] used expensive motion sensors to capture users’ writing pattern, [8, 12] forced users to follow predefined unistroke writing pattern, and [35, 12] only collected videos capturing a single English lower-case letter, which are not comprehensive enough for the development of WiTA systems. Moreover, [18] adopted an egocentric view that demands users to wear a motion capturing device.

To overcome the limitations mentioned above, we collect a benchmark dataset in this work; Fig. 1 shows an example data instance. Among multiple modalities for capturing finger movement in the air, we choose RGB cameras as the sensing device due to their superior accessibility, low cost, and generality compared to other sensing modalities such as depth or gyro sensors. In addition, we adopt a third-person view rather than an egocentric view to improve user experience by removing the possibility of attaching additional devices on users [30]. We also allow users to follow their natural handwriting patterns to maximize usability. Finally, we collect five sub-datasets—to ensure universality and actualize unconstrained text recognition from finger movement—in two languages: Korean lexical, English lexical, Korean non-lexical, English non-lexical, and the mixture of the two languages in a non-lexical format. As far as we are aware, our dataset is the most comprehensive benchmark dataset for the WiTA task, and we expect our dataset would facilitate the research on WiTA.

Next, we propose baseline models for the WiTA task, which will serve as an evaluation standard for forthcoming WiTA systems. The baseline models receive a sequence of image frames and transform the input into a sequence of characters written in the air. The proposed baseline models perform the decoding process in an end-to-end manner—performing unconstrained text recognition from finger movement. For developing the baseline models, we propose spatio-temporal residual network architectures inspired by 3D ResNet [37]. The proposed spatio-temporal residual networks effectively deal with both spatial and temporal contexts within the WiTA input signals. Furthermore, we conduct a thorough ablation study to examine the effect of each design choice and offer insights for the development of more advanced WiTA systems.

## 2. Related Works

### 2.1. Writing Recognition System

**Finger Writing.** In finger writing recognition systems, users write text in the air with right or left index finger. Then, recognition systems capture and interpret the finger movement to produce the text users have intended to write. For capturing finger movement, recognition systems integrate various types of sensors. One category of sensors get attached to users’ body and gather the finger movement information. Examples of such sensors include smartwatches

[41, 28, 43] and custom-manufactured sensors [34, 20]. This category of sensors lessen the usability since users have to carry these sensors for text-entry, and physical contacts could cause discomfort [30].

A few research groups have attempted to improve the usability of WiTA by excluding body-installed sensors. One of the approaches encodes each character or word into a set of actions and formulates WiTA as action recognition [27, 9]. Accordingly, users have to learn the new encoding systems, which in turn degrades usability. Typing in the air is another example of this approach [42]. Moreover, another group of researchers has employed Kinect (depth) [46, 6, 29] or motion sensors [8, 24] to exclude body-installed sensors. However, users do not always have access to these high-cost sensors due to their limited availability.

RGB cameras, which omit physical contacts, offer an easy-to-deploy and low-cost way for capturing finger movement. Contemporary approaches utilizing RGB cameras for WiTA focus on a fingertip tracking to formulate WiTA as handwriting recognition [2, 18, 30] or treat WiTA as gesture recognition by performing word-based recognition of written text [14, 13]. In contrast, we propose end-to-end baseline models for the WiTA task—recognizing the text written in the air on a character basis. The end-to-end architectures for unconstrained text recognition lead to simplification of the design process as well as enhancement of the performance. In addition, the proposed baselines improve usability since users are not required to slow down their writing for finger detection and tracking.

### 2.2. Convolution for Spatio-Temporal Data

One of the representative applications that utilize convolution over spatio-temporal data is video action recognition. In video action recognition, convolution deals with macroscopic semantics within a sequence of images. Among various convolution architectures [44, 26], 3D ResNet and its variants have exhibited satisfactory performance in video action recognition [37]. Moreover, the performance of a spatio-temporal convolution surpasses traditional vision methods when a simple average pooling and a multi-scale temporal window are applied [38]. In the process of taking short-term and long-term temporal contexts into account, two-path architectures have suggested [10, 11]. Deformable kernels would enable flexible reception fields and result in performance enhancement [39]. Further, [36] has shown varying the amount of channel interactions can increase the accuracy of 3D convolutional networks.

Recently, researchers have attempted to apply convolution to the hand gesture recognition task [32, 45, 25, 33]. These works concentrate on recognizing a set of pre-defined simple hand gestures. Contrary to these works, we aim to recognize the text written in the air with spatio-temporal convolution. The WiTA task involves more complex hand

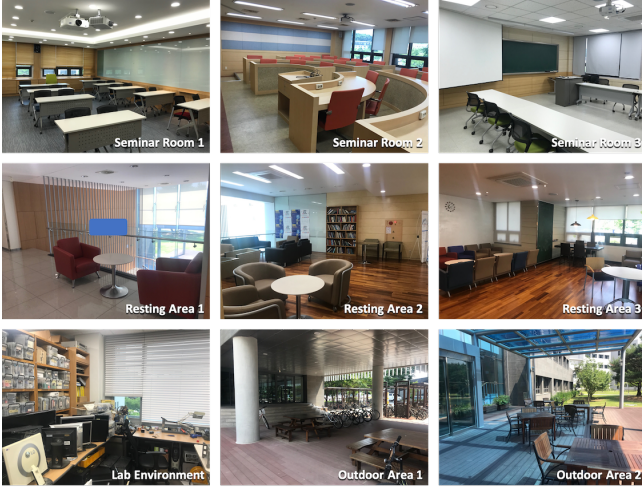


Figure 2. The data collection environments. We varied the background for each data collection process to remove the performance dependency on the background variation.

gestures than the simple hand gesture recognition task and requires unconstrained text recognition from the complex hand gestures. For the WiTA baseline models, we focus on short-term semantic context [40] and design spatio-temporal convolution architectures. The proposed spatio-temporal convolution keep the temporal structure of input sequences and generate a sequence of vectors rather than a single vector for classification.

### 3. WiTA Dataset

#### 3.1. Participants

In total, we recruited 122 participants<sup>1</sup> (74 male and 48 female). The participants aged from 19 to 42 (average = 24.33, std = 2.39). One of the participants is left-handed, two participants are ambidextrous, and the rest are right-handed. All of them use Korean as their mother tongue, and they could read and write both Korean and English without any difficulties.

#### 3.2. Environment and Apparatus

We collected our data in nine environments to ensure the robustness to background variations (Fig. 2): three seminar rooms, three resting areas, one lab environment, and two outdoor areas. Moreover, we modified the viewpoints for different data collection processes to diversify the backgrounds in our dataset. We set up a laptop (MS Surface) equipped with an RGB camera (29fps) on a desk or a table in each data collection environment. We captured image sequences with the resolution of  $224 \times 224$ .

<sup>1</sup>Table 7 in the supplementary material summarizes the statistics of the participants.



Figure 3. Examples of text for writing.

#### 3.3. Writing Interface

We implemented the data collection interface<sup>2</sup> using PyQt5<sup>3</sup> which supports cross-platform application development. The beginning page of the interface collects the demographics of participants. Next, the main page of the interface displays the text to write at the top center area, and the middle area shows the current video. The right middle area contains a group of buttons for controlling the data collection process: “start”, “stop”, “next” and “redo”.

#### 3.4. Text for Writing

To verify the generality of the proposed WiTA task among multiple languages at least in a preliminary manner, we collected five sub-datasets in two languages. The text for each type of the dataset was composed as follows (Fig. 3 shows example texts in our dataset):

- **Korean<sup>4</sup> Lexical:** We utilized the dataset<sup>5</sup> collected by the National Institute of Korean Language (NIKL).

<sup>2</sup>Fig. 8 in the supplementary material depicts the writing interface.

<sup>3</sup><https://riverbankcomputing.com/software/pyqt/download5>

<sup>4</sup>A Hangeul (Korean syllable), which is the basic building block of Korean words, consists of two to three letters: first letter, middle letter and optional last letter. Consonants can be placed at the first and last letter positions, while vowels at the middle letter position. For example, the Hangeul ‘대’ consists of two letters (‘ㄷ’ and ‘애’) while ‘한’ of three letters (‘ㅎ’, ‘ㅏ’, and ‘ㄴ’).

<sup>5</sup>[https://www.korean.go.kr/front/reportData/reportDataView.do?mn\\_id=45&report\\_seq=1](https://www.korean.go.kr/front/reportData/reportDataView.do?mn_id=45&report_seq=1)



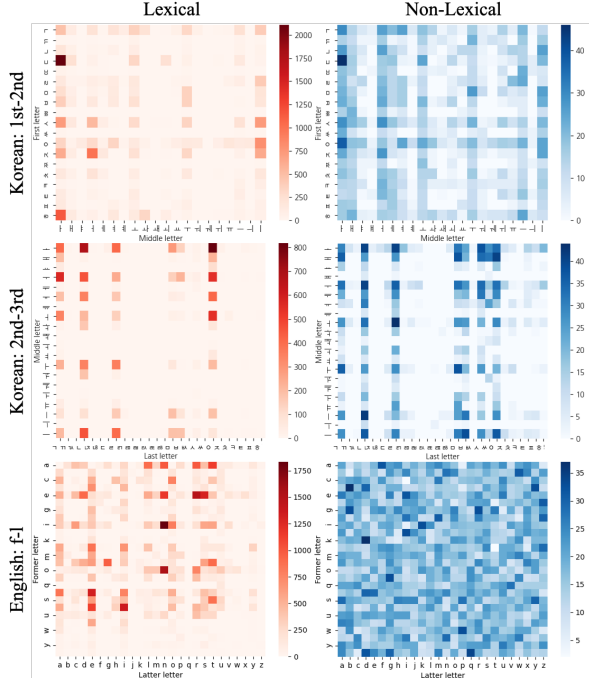


Figure 4. Co-occurrence statistics of our WiTA dataset.

Specifically, we retrieved the most frequent 6,000 Korean words dataset.

- **Korean Non-Lexical:** We randomly generated non-lexical words by sampling from the most common 1,989 syllables (Hangul) dataset<sup>6</sup>. We restricted the lengths of the generated words to range from one to three.
- **English Lexical:** We retrieved the top 6,000 most-frequent words from Google 1B dataset [7].
- **English Non-Lexical:** We randomly generated non-lexical words by sampling from 26 alphabets. The lengths of the non-lexical words are between 3 and 7.
- **Mixture Non-Lexical:** For testing multi-lingual WiTA systems, we generated non-lexical words using both Korean and English syllables<sup>7</sup>.

We randomly sampled a word at every data collection process, resulting in very few numbers of duplicated text.

### 3.5. Data Statistics and User Behavior Analysis

**Data Statistics.** Tables 1 and 2 summarize the statistics of the WiTA dataset collected in this work and compares it with those of previous studies. In respect of dimension, our

<sup>6</sup>In theory, 11,172 distinct Korean syllables (Hanguls) exist, but about 2,000 of them are practically used [21]. NIKL provides this dataset as well.

<sup>7</sup>We expect we could verify the performance of a unified WiTA model for multiple languages with this dataset in the future.

dataset is the most comprehensive compared to recent studies. Moreover, our dataset covers both Korean and English in addition to lexical and non-lexical phrases, while other datasets simply provide single-letter to less-than-three-letter videos. Since our dataset supplies videos containing semantic words, they capture the complex interdependence between gestures for different characters ( $C/V \geq 3$ ); it would foster the development of WiTA systems for real-world applications. Furthermore, our dataset is the only dataset that is accessible to the public at the moment.

Fig. 4 visualizes the co-occurrence statistics<sup>8</sup>. The lexical data is more biased than the non-lexical data in both languages. Especially, the non-lexical English shows a well-scattered distribution. In the case of Korean, the non-lexical data is more biased than that of English since only about 2,000 pairs out of 11,172 possible Hanguls are practically used—though the non-lexical Korean data shows more even distribution than that of lexical Korean data. Thus, the non-lexical data would play a vital role in the development of unconstrained text recognition from finger movement.

**User Behavior Analysis.** For the analysis of user behaviors in WiTA, we selected 12 participants for each language and analyzed the data by manually labeling the fingertips. Fig. 5 exemplifies a set of WiTA patterns. In both languages, users tend to squeeze characters to fit the whole word within the screen though not consistent for all cases. Moreover, most of the patterns are not recognizable even given the text since users were asked to freely and naturally write.

Next, Table 3 displays the quantitative analysis result. The participants wrote the Korean text faster than the English text and revealed a larger deviation in the case of Korean. We consider the difference in writing speed could have resulted from the fact that the participants were more familiar with Korean than English. Next, the scales appear distinctive for both languages since a Korean Hangul consists of two or three letters. We utilized the number of Hanguls for measuring the scale of Korean WiTA while the number of characters for English WiTA. The Korean scale is approximately 2.5 times larger than that of English, which accounts for the scale difference.

## 4. Methodology

### 4.1. Problem Formulation

We formulate the WiTA decoding for unconstrained text recognition as follows. Given a sequence of image frames that capture user’s writing in the air  $\mathcal{I} = (\mathbf{I}_1, \dots, \mathbf{I}_n)$  where  $\mathbf{I}_i$  ( $1 \leq i \leq n$ ) is an image frame, a WiTA decoding al-

<sup>8</sup>As a Hangul consists of two to three letters, we analyzed the co-occurrence between the first and the second letters, and between the second and the third letters. For English, we analyzed the co-occurrence between the former and the latter letters of every pair.

Table 1. Comparison of datasets. The proposed WiTA dataset is the most comprehensive and provides rich types of data instances. Our dataset supplies videos containing semantic text written in the air, which capture the interdependence between gestures for different characters. C/V, Sem, K, E, C and N in the table stand for character/video, inclusion of semantic words, Korean, English, Chinese and numbers, respectively.

Dataset	Year	People	Videos	Frames	Text	C/V	Sem	Sensor	View	Environment	Access
VBFR [19]	2007	69	1,794	-	E	1	-	RGB	ego	Indoor	-
VBHR [35]	2012	21	1,290	-	E	1	-	RGB	3rd	Indoor	-
ANWE [46]	2013	-	375	44,522	ECN	1	-	RGB-D	3rd	-	-
AWR [8]	2015	22	11,120	-	E	$\leq 3$	✓	Motion	-	Indoor	-
PGEI [18]	2016	24	-	93,729	EC	1	-	RGB-D	ego	Indoor+Outdoor	-
WiFi [12]	2018	5	26,000	-	E	1	-	WiFi	-	Indoor	-
FDT [30]	2019	5	1,800	-	EN	1	-	RGB	3rd	-	-
<b>WiTA (ours)</b>	2021	122	209,926	1,757,307	KE	$\geq 3$	✓	RGB	3rd	Indoor+Outdoor	✓



Figure 5. Examples of WiTA patterns. Users’ natural writing patterns are complex and challenging; most of the patterns are not recognizable even given the text. One thing to note is Korean gets written in the order of left-to-right, top-to-bottom and first-to-middle-to-last-letters.

Table 2. Summary of video and text statistics. The numbers in each cell indicate (average/std).

Language	Type	#Frames	#Characters
<b>Korean</b>	Lexical	87.82/32.72	3.05/1.08
	N-Lexical	79.36/31.02	2.00/0.82
<b>English</b>	Lexical	78.75/28.65	6.59/2.54
	N-Lexical	68.08/21.49	5.03/1.41

Table 3. Summary of User Behavior Analysis. The unit of the metrics in the table is pixel. HPS and CPS in the table stand for *Hangul-per-second* and *character-per-second*, respectively

Language	Metric	Avg.	Std.	Range
<b>Korean</b>	HPS	3.98	1.06	(2.11, 7.11)
	x-Scale	43.56	12.58	(21.96, 99.78)
	y-Scale	38.26	18.45	(11.47, 147.22)
<b>English</b>	CPS	3.57	0.86	(1.82, 5.74)
	x-Scale	18.35	7.27	(7.32, 52.78)
	y-Scale	14.39	8.64	(3.73, 57.46)

gorithm aims to find the labeling  $l^*$  with the highest condi-

tional probability:

$$l^* = \arg \max_l p(l|\mathcal{I}). \quad (1)$$

For the labeling, we adopt the concept of Connectionist Temporal Classification (CTC) [16] where there is a mapping between a labeling and paths denoted as  $\pi$ ’s. An operator  $\mathcal{B}$  maps a set of paths onto a labeling, i.e., multiple label sequence paths reduce to the same labeling by  $\mathcal{B}$ . For instance,  $\mathcal{B}(a, -, a, a, b) = \mathcal{B}(-, a, -, a, -, b, -) = (a, a, b)$ , where  $-$  indicates a blank. Thus, the conditional probability can be evaluated as follows:

$$p(l|\pi) = \sum_{\pi \in \mathcal{B}^{-1}(l)} p(\pi|\mathcal{I}), \quad (2)$$

where

$$p(\pi|\mathcal{I}) = \prod_{t=1}^T p(\pi_t, t|\mathcal{I}) = \prod_{t=1}^T y_{\pi_t}^t, \quad (3)$$

where  $\pi_t$  is the label observed at time  $t$  along path  $\pi$  and  $y_{\pi_t}^t$  is the softmax-normalized output.

In practice,

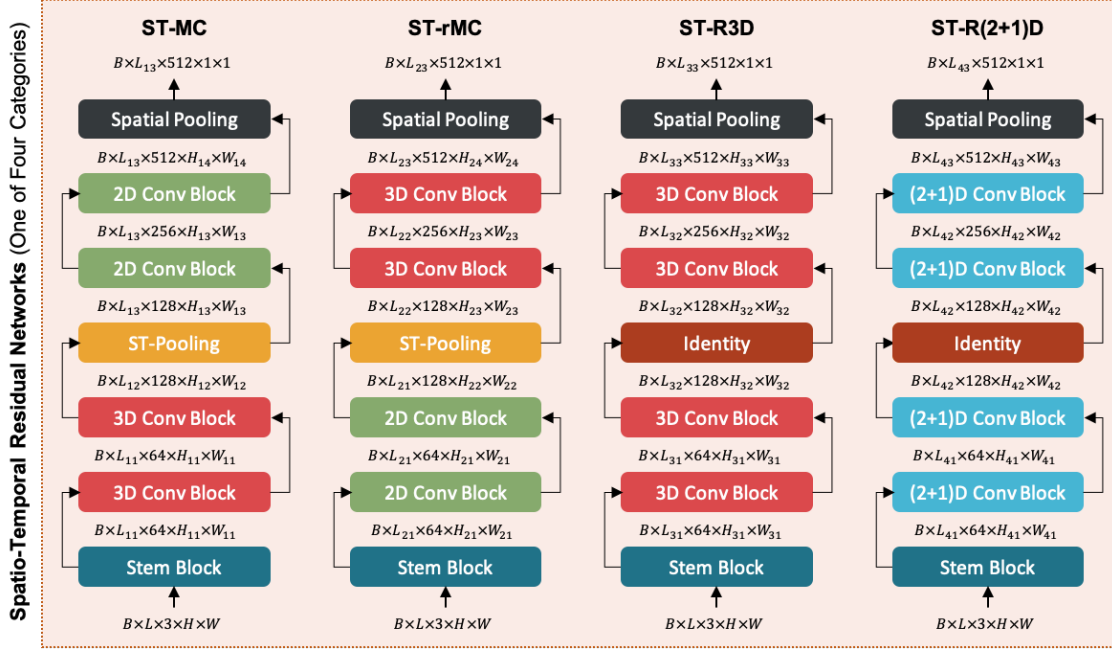


Figure 6. Overall architecture of the WiTA baseline models. We design four types of spatio-temporal residual network architectures for the WiTA task. Each model receives a sequence of image frames and the sequence gets transformed into a sequence of characters—conducting unconstrained text recognition.

$$p(l|\pi) = \sum_s^{|l'|} \alpha_s^t \beta_s^t, \quad (4)$$

where  $l'$  is a modified labeling for which blanks get added at the beginning and the end of  $l$  as well as between every pair of consecutive labels,  $\alpha_s^t$  and  $\beta_s^t$  are forward and backward variables defined for searching paths, and  $s$  indicates steps.

Finally, given pairs of input  $\mathcal{I}$  and target label  $z$  in a training set  $S$ , the objective loss function becomes

$$L_{ctc} = - \sum_{(\mathcal{I}, z) \in S} \ln p(z|\mathcal{I}). \quad (5)$$

The loss function accomplishes maximum likelihood training which simultaneously maximizes the log probabilities of all the correct labeling classifications in the training set.

## 4.2. Text Encoding

We encode text into a sequence of separate letters. Moreover, we employ a special character ‘ $\sim$ ’ to distinguish consecutive Hanguls for Korean and two identical characters that appear adjacent to each other for English. For example, “대한” and “success” becomes (ㄷ, ㄹ, ∼, ㅇ, ㅏ, ㅓ) and (s, u, c, ∼, c, e, s, ∼, s), respectively.

## 4.3. Spatio-Temporal Residual Network

We propose spatio-temporal (ST) residual network architectures (Fig. 6) inspired by convolutional residual blocks

without bottlenecks [17]. Each convolutional residual block consists of two convolution layers followed by a ReLU non-linearity [31]. The output of the  $i$ -th residual block becomes

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \mathcal{F}(\mathbf{x}_{i-1}; \theta_i), \quad (6)$$

where  $\mathbf{x}_i$  denotes the tensor computed by the  $i$ -th convolutional block and  $\mathcal{F}(\cdot; \theta_i)$  implements the composition of two convolutions with the parameters  $\theta_i$  and the application of the ReLU non-linearity. We consider four types of convolution blocks to design the proposed ST residual network architectures<sup>9</sup>: mixed 3D-2D convolutions (ST-MC), reversed MC (ST-rMC), residual 3D convolutions (ST-R3D) and 2D convolutions followed by 1D convolutions (ST-R(2+1)D).

We place a 3D pooling layer in the middle of the ST-MC and ST-rMC networks to better capture both spatial and temporal contexts. In the cases of ST-R3D and ST-R(2+1)D, we omit the 3D pooling layer since a sufficient amount of temporal contexts are captured via a number of ST convolutions. Next, we employ an adaptive spatial pooling layer at the end of each ST residual network. The spatial pooling layer preserves the temporal structure of the input tensor which gets transformed into a sequence of characters.

<sup>9</sup>Table 8 in the supplementary material analytically depicts the convolution architectures.

## 5. Experiments

### 5.1. Settings

**Data Split.** For training, validation, and testing the WiTA models, we split the collected dataset into three sets with an approximate ratio of 8 : 1 : 1. We divided the data by person to ensure robustness of the developed model towards different individuals.

**Metrics.** We evaluated each WiTA baseline model with two metrics: average decoding frames per second (D-FPS) and character error rate (CER). On one hand, we include the D-FPS as a performance metric since ensuring a real-time operation is crucial for decoders. We measure D-FPS by averaging the total number of frames decoded in a second. On the other hand, CER represents the decoding accuracy which is defined as

$$CER = \frac{MCD(S, P)}{length_c(P)} \times 100 (\%), \quad (7)$$

where  $MCD(S, P)$  is the minimum character distance (the Levenshtein measure) between the decoded phrase  $S$  and the ground-truth phrase  $P$ , and  $length_c(P)$  is the number of characters in  $P$ . The Levenshtein measure counts the number of insertions, deletions and substitutions of characters or words to transform  $S$  into  $P$ .

### 5.2. Implementation Details

We trained the WiTA models with the learning rate warm-up scheme [15] and the Adam optimizer [23] after resizing images to  $112 \times 112$ . We set the learning rate as  $1e - 3$ . We set the batch size as 4 for 18-layered models, 8 for 10-layered models and 1 for measuring D-FPS. For model selection and stopping condition of training procedures, we followed the early stopping scheme [5]. All models converged within 175 epochs of training.

To investigate the effect of each design choice, we trained WiTA models using different schemes. We controlled the following conditions: the number of layers (10 or 18), the type of pooling layers (max-pooling [4] or average-pooling [38]), data augmentation (random rotation and photometric distortions including brightness, contrast, saturation and hue), the loading of pre-trained weights (trained on the Kinetics-400 dataset) and the composition of training data.

### 5.3. Results and Analysis

**Search of Optimal Learning Configuration.** In order to identify the best learning configuration, we fixed the architecture as ST-R3D and varied the learning conditions. Most of the better performing configurations, including the best one, came from 10-layered models for English, while the best configuration for Korean came from 18-layered

models as shown in Table 4. We suspect the reason Korean requires a deeper model is due to higher complexity in writing. The general pattern in English is that the performance improves with augmentation and the pre-trained weights with a few exceptions (the 18-layered models with max pooling). For Korean, the pre-trained weights and augmentation had a different effect on the model performance; generally, the pre-trained weights boosted the performance, while augmentation did not. We presume this phenomenon occurred since some Hanguls have similarities in shape, causing ambiguity and confusion when rotated. Moreover, it is likely that the last letter was mistakenly considered as the first letter since the first and the last letters of Hangul are consonants. There are some exceptions to this pattern: when augmentation is used along with max-pooling but the pre-trained weights, it enhances the performance. Ultimately, the best configurations for Korean and English mismatched. This suggests that it is important to carefully select the design choices based on the characteristics of the language.

**Effect of Model Architecture.** In Table 5, the best learning configurations from Table 4 were adopted to compare the performance of different baseline architectures. For Korean, the pre-trained weights, 18-layers and max pooling were used for all of the four networks, whereas for English, 10 layers, average pooling, and augmentation were adopted for all four networks. For both languages, ST-R3D displayed the lowest CER, and ST-rMC outperformed ST-MC (MC failed to converge in the English dataset)—indicating that extracting temporal information in the later layers leads to better performance. However, none of the network architectures using 2D convolution could beat the performance of the ST-R3D architecture (only using 3D convolution). This implies that capturing both temporal and spatial information simultaneously throughout the entire network is crucial in the WiTA task. In both languages, D-FPS ensures real-time operations: 435.27 and 697.39 for Korean and English, respectively.

**Impact of Training Data Configuration.** Table 6 summarizes the effect of training data configuration on performance and demonstrates the increase in the amount of data prompts performance gains. It is worth noting that the total number of videos for lexical and non-lexical data are not the same. The total number of videos for the lexical data is approximately five times more than that of the non-lexical data. The performance gap between the model trained solely on the lexical data and the model on the non-lexical data is less severe in Korean than in English. We suspect this is because the Korean non-lexical data do not deviate too much from the ordinary sequence of characters that appear in the Korean lexical dataset, whereas the English data display a huge discrepancy between the non-lexical and the lexical data as shown in Fig. 4.

Table 4. Results of the ablation study for searching the optimal learning condition on the validation dataset. We controlled four factors in our study: the number of layers, the type of pooling, the application of augmentation and the usage of pre-trained weights.

Training Condition				Korean (CER)			English (CER)		
#Layers	Pooling	Agmnt	Prtrn	Lexical	N-Lexical	Overall	Lexical	N-Lexical	Overall
10	Max	-	-	49.85	64.24	51.71	29.77	42.75	31.47
10	Max	✓	-	44.55	58.66	46.37	29.07	43.40	30.95
10	Avg	-	-	45.34	62.05	47.50	27.24	<b>42.21</b>	29.20
10	Avg	✓	-	39.00	54.13	40.96	<b>27.12</b>	42.32	<b>29.12</b>
18	Max	-	-	67.28	79.08	68.81	33.10	49.35	35.24
18	Max	✓	-	29.72	40.35	31.09	82.03	87.99	82.81
18	Max	-	✓	<b>28.02</b>	<b>39.79</b>	<b>29.54</b>	84.93	90.91	85.72
18	Max	✓	✓	64.75	76.04	66.21	33.10	49.35	35.24
18	Avg	-	-	65.84	73.92	66.88	76.85	91.45	78.76
18	Avg	✓	-	68.94	77.74	70.07	41.29	60.71	43.84
18	Avg	-	✓	52.90	69.68	55.06	63.81	78.14	65.69
18	Avg	✓	✓	69.44	76.33	70.33	29.44	40.26	30.80

Table 5. Architectural impact on the performance. We measured the performance on the test dataset.

Model	Korean (CER)				English (CER)			
	Lexical	N-Lexical	Overall	D-FPS	Lexical	N-Lexical	Overall	D-FPS
<b>ST-MC</b>	60.42	69.21	61.48	704.26	-	-	-	-
<b>ST-rMC</b>	54.18	67.47	55.78	791.28	92.78	93.96	92.94	1046.67
<b>ST-R3D</b>	<b>31.62</b>	<b>44.37</b>	<b>33.16</b>	435.27	<b>28.10</b>	<b>36.46</b>	<b>29.24</b>	697.39
<b>ST-R(2+1)D</b>	-	-	-	-	86.80	91.98	87.51	588.13

Table 6. Effect of training data configuration on the performance. Each row represents a training dataset configuration and the performance on the test dataset. The numbers below the ‘Training Data Configuration’ column indicate the amount of the data consisting the each row.

Training Data Configuration		Korean (CER)			English (CER)		
Lexical	N-Lexical	Lexical	N-Lexical	Overall	Lexical	N-Lexical	Overall
100%		38.77	54.06	40.61	32.14	51.98	34.85
	100%	79.72	78.39	79.56	92.95	94.58	93.17
50%	50%	53.03	64.65	54.43	47.32	58.23	48.81
50%	100%	41.50	54.14	43.02	36.71	42.71	37.53
100%	50%	34.49	47.60	36.07	28.20	40.83	29.93
100%	100%	31.62	44.37	33.16	28.10	36.46	29.24

## 6. Conclusion

In this work, we collected a benchmark dataset for WiTA systems. To the best of our knowledge, our benchmark dataset is the most comprehensive and the only dataset enabling real-world implementation. The dataset consists of five sub-datasets in two languages including both lexical and non-lexical text to ensure universality. We captured the finger movement with RGB cameras in a third-person view from 122 participants—resulting in 209,926 videos. This data collection setting guarantees accessibility, cost-efficiency, and generality. Next, we proposed baseline mod-

els for the WiTA task. In developing the baseline models, we designed four spatio-temporal (ST) residual network architectures inspired by 3D ResNet. The proposed ST residual networks effectively handle both spatial and temporal contexts within the input sequence capturing finger movement. The proposed models exhibited 33.16% and 29.24% of CER in Korean and English datasets, respectively, with the processing speed of 435 and 697 D-FPS securing a real-time operation. We expect that our dataset and proposed baseline models would activate the research on WiTA; we make our dataset and the source codes public.



## References

- [1] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166, 2019. 14
- [2] Md Alam, Ki-Chul Kwon, Mohammed Y Abbass, Shariar Md Imtiaz, Nam Kim, et al. Trajectory-based air-writing recognition using deep neural network and depth sensor. *Sensors*, 20(2):376, 2020. 2
- [3] Christoph Amma, Marcus Georgi, and Tanja Schultz. Air-writing: a wearable handwriting recognition system. *Personal and Ubiquitous Computing*, 18(1):191–203, 2014. 1
- [4] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2559–2566. IEEE, 2010. 7
- [5] Rich Caruana, Steve Lawrence, and C Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 402–408, 2001. 7
- [6] Hyung Jin Chang, Guillermo Garcia-Hernando, Danhang Tang, and Tae-Kyun Kim. Spatio-temporal hough forest for efficient detection-localisation-recognition of fingerwriting in egocentric camera. *Computer Vision and Image Understanding*, 148:87–96, 2016. 2
- [7] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013. 4
- [8] Mingyu Chen, Ghassan AlRegib, and Biing-Hwang Juang. Air-writing recognition—part i: Modeling and recognition of characters, words, and connecting motions. *IEEE Transactions on Human-Machine Systems*, 46(3):403–413, 2015. 1, 2, 5
- [9] Han Ding, Chen Qian, Jinsong Han, Ge Wang, Wei Xi, Kun Zhao, and Jizhong Zhao. Rfipad: Enabling cost-efficient and device-free in-air handwriting using passive tags. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 447–457. IEEE, 2017. 2
- [10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019. 2
- [11] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1933–1941, 2016. 2
- [12] Zhangjie Fu, Jiashuang Xu, Zhuangdi Zhu, Alex X Liu, and Xingming Sun. Writing in the air with wifi signals for virtual reality devices. *IEEE Transactions on Mobile Computing*, 18(2):473–484, 2018. 2, 5
- [13] Ji Gan and Weiqiang Wang. In-air handwritten english word recognition using attention recurrent translator. *Neural Computing and Applications*, pages 1–18, 2019. 2
- [14] Ji Gan, Weiqiang Wang, and Ke Lu. A unified cnn-rnn approach for in-air handwritten english word recognition. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 2
- [15] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large mini-batch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 7
- [16] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 369–376, 2006. 5
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6
- [18] Yichao Huang, Xiaorui Liu, Xin Zhang, and Lianwen Jin. A pointing gesture based egocentric interaction system: Dataset, approach and application. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 16–23, 2016. 2, 5
- [19] Lianwen Jin, Duanduan Yang, Li-Xin Zhen, and Jian-Cheng Huang. A novel vision-based finger-writing character recognition system. *Journal of Circuits, Systems, and Computers*, 16(03):421–436, 2007. 5
- [20] Lei Jing, Zeyang Dai, and Yiming Zhou. Wearable handwriting recognition with an inertial sensor on a finger nail. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1330–1337. IEEE, 2017. 2
- [21] Daniel Keysers, Thomas Deselaers, Henry A Rowley, Li-Lun Wang, and Victor Carbune. Multi-language online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1180–1194, 2017. 4
- [22] U. Kim, S. Yoo, and J. Kim. I-keyboard: Fully imaginary keyboard on touch devices empowered by deep neural decoder. *IEEE Transactions on Cybernetics*, Early Access, 2019. 1
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [24] Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, and Debi Prosad Dogra. Study of text segmentation and recognition using leap motion sensor. *IEEE Sensors Journal*, 17(5):1293–1301, 2017. 2
- [25] Gongfa Li, Heng Tang, Ying Sun, Jianyi Kong, Guozhang Jiang, Du Jiang, Bo Tao, Shuang Xu, and Honghai Liu. Hand gesture recognition based on convolution neural network. *Cluster Computing*, 22(2):2719–2729, 2019. 2
- [26] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10522–10531, 2019. 2

- [27] Anders Markussen, Mikkel Rønne Jakobsen, and Kasper Hornbæk. Vulture: a mid-air word-gesture keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1073–1082, 2014. 2
- [28] Danial Moazen, Seyed A Sajjadi, and Ani Nahapetian. Air-draw: Leveraging smart watch motion sensors for mobile human computer interactions. In *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 442–446. IEEE, 2016. 2
- [29] Shahram Mohammadi and Reza Maleki. Real-time kinect-based air-writing system with a novel analytical classifier. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(2):113–125, 2019. 1, 2
- [30] Sohom Mukherjee, Sk Arif Ahmed, Debi Prosad Dogra, Samarjit Kar, and Partha Pratim Roy. Fingertip detection and tracking for recognition of air-writing in videos. *Expert Systems with Applications*, 136:217–229, 2019. 2, 5
- [31] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, pages 807–814, 2010. 6
- [32] Juan C Nunez, Raul Cabido, Juan J Pantrigo, Antonio S Montemayor, and Jose F Velez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, 2018. 2
- [33] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, page 113336, 2020. 2
- [34] Katsuyuki Sakuma, Gaddi Blumrosen, John J Rice, Jeff Rogers, and John Knickerbocker. Turning the finger into a writing tool. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1239–1242. IEEE, 2019. 2
- [35] Alexander Schick, Daniel Morlock, Christoph Amma, Tanja Schultz, and Rainer Stiefelwagen. Vision-based handwriting recognition for unrestricted text input in mid-air. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 217–220, 2012. 2, 5
- [36] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2
- [37] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018. 2
- [38] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2019. 2, 7
- [39] Junwu Weng, Mengyuan Liu, Xudong Jiang, and Junsong Yuan. Deformable pose traversal convolution for 3d action and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 136–152, 2018. 2
- [40] Zecheng Xie, Zenghui Sun, Lianwen Jin, Hao Ni, and Terry Lyons. Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1903–1917, 2018. 3
- [41] Chao Xu, Parth H Pathak, and Prasant Mohapatra. Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pages 9–14, 2015. 2
- [42] Xin Yi, Chun Yu, Mingrui Zhang, Sida Gao, Ke Sun, and Yuanchun Shi. Atk: Enabling ten-finger freehand typing in air based on 3d hand tracking data. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 539–548, 2015. 2
- [43] Yafeng Yin, Lei Xie, Tao Gu, Yijia Lu, and Sanglu Lu. Aircontour: Building contour-based model for in-air writing gesture recognition. *ACM Transactions on Sensor Networks (TOSN)*, 15(4):1–25, 2019. 2
- [44] Yong-Ho Yoo, Ue-Hwan Kim, and Jong-Hwan Kim. Convolutional recurrent reconstructive network for spatiotemporal anomaly detection in solder paste inspection. *arXiv preprint arXiv:1908.08204*, 2019. 2
- [45] Felix Zhan. Hand gesture recognition with convolution neural networks. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 295–298. IEEE, 2019. 2
- [46] Xin Zhang, Zhichao Ye, Lianwen Jin, Ziyong Feng, and Shaojie Xu. A new writing experience: Finger writing in the air using a kinect sensor. *IEEE MultiMedia*, 20(4):85–93, 2013. 1, 2, 5

# Supplementary Material

In this supplementary material, we describe the details of our study not included in the main manuscript due to space limit. We include the following additional details: statistics of the WiTA dataset, description of the model architectures, the full ablation study results on the effect of training data configuration, and the discussion on future research direction. Moreover, Fig. 7 displays the annotated result of Fig. 1. The tracking of the fingertip reveals the text written in the air—though the tracking is hardly possible for laypersons, ensuring a private communication tool.

## A. Data Collection Procedure

First, we informed the participants (see Table 7 for participant statistics) the data collection procedure and gathered the demographics (see Fig. 8 for the interface). We asked the participants to assume that a perfect AI system will decode their writing in the air and write as naturally as possible. As a warm-up, the participants familiarized themselves with the writing interface using the first ten phrases. Then, the participants wrote 75 phrases of lexical Korean and English texts, respectively and 15 phrases of non-lexical Korean, English and the Mixture texts, respectively. Each participant wrote and captured 195 ( $=75 \times 2 + 15 \times 3$ ) phrases and each data collection process took approximately 50 minutes. In total, the data we collected includes 209,926 video instances.

## B. Additional Statistics of the WiTA Dataset

Figure 9 shows the histogram of characters in each dataset split. The lexical datasets are biased towards certain characters. For example, in the English data, the character that made the most appearance (i.e., ‘e’) appeared approximately 70 times more than the character that made the least appearance (i.e., ‘z’). On the other hand, the English non-lexical data shows a well-balanced data distribution within each dataset as well as across train, validation, and test datasets. Combining all the characters in each dataset, every character appears within 300 to 400 times, and the most appeared character was approximately only 10% more than the least appeared character. Likewise, the Korean non-lexical data are more fairly distributed compared to the Korean lexical data. In particular, the first Korean non-lexical characters are well spread out, while the lexical bar graph shows a drastic difference between the most appeared character and the least appeared character. Although following the general distribution of the lexical data, the second and the third Korean non-lexical data are relatively more spread out. The drastic difference in the number of appearances of more-likely-to-appear characters and less-likely-to-appear characters in Korean non-lexical data is inevitable because



Figure 7. The annotated example instance of the dataset collected in this work. The person in the example is writing “re” from the word “recognized”. WiTA offers a private communication tool for HCI.

Table 7. Summary of the participant statistics.

Metric	Type	Value
Gender	Male	74/122
	Female	48/122
	Neutral	-
Age	Range	19 - 42
	Average	24.33
	s.t.d.	2.39
Comfort-Hand	Left	1/122
	Right	119/122
	Both	2/122
Korean Fluency	Reading	4.82/5.00 (0.47)
	Writing	4.61/5.00 (0.43)
	Overall	4.70/5.00 (0.45)
English Fluency	Reading	4.33/5.00 (0.39)
	Writing	4.16/5.00 (0.37)
	Overall	3.45/5.00 (0.30)

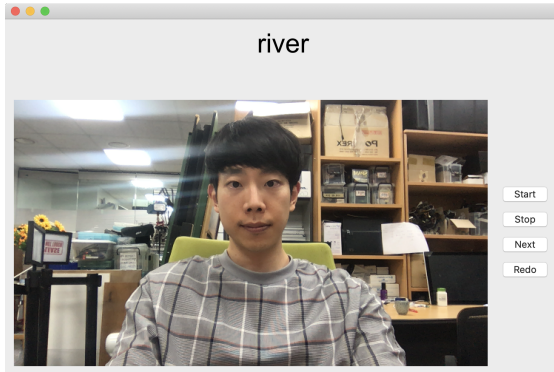
less appearing characters are simply not used frequently in the Korean language in general.

## C. Additional Description on Model Architectures

Table 8 describes the spatio-temporal (ST) residual network architectures. While ST-MC, ST-rMC, and ST-R3D contain a pair of convolutions in each convolution block, the ST-R(2+1)D architecture includes two pairs of convo-



(a) Interface of the beginning page.



(b) Interface of the main page.

Figure 8. Typing interface. The interface consists of two pages. The beginning page gathers the demographics of participants and the main page captures the videos of WiTA.

lutions in each convolution block. Except for the ST-R3D architecture, all other architectures entail 2D convolutions. The proposed ST residual networks offer a way to scale-up or scale-down the model depths. For 10-layered models,  $n$  in the table is 1 while  $n$  is 2 for 18-layered models. Though models with more than 18 layers are possible, it is highly probable that such models would hit the hardware memory limit during the training procedure.

## D. Additional Ablation Study

In order to examine how the introduction of non-lexical data affects the performance of the models, we varied the percentage of lexical and non-lexical data. First, we examined the performance of the model using the entire dataset (100% lexical, 100% non-lexical) and decreased the non-lexical data to 50% (the first group of Table 9). The performance for both English and Korean decreased when the amount of the non-lexical data was reduced. However, the lack of the non-lexical data less-affected the performance in English than in Korean. We designed a similar experiment but using only 50% of the lexical data (the third group of Table 9). In this case, however, the performances of the En-

glish and Korean models were almost equally affected by the lack of non-lexical data.

Next, we only used 100% of the lexical data and then added in the non-lexical data by 50% and 100% while removing the lexical data by the same amount of the non-lexical data that was added in so that the total amount of the used data remained the same (the second group of Table 9). Similarly, we repeated the same process using a less amount of data. We started the experiment by only using 50% of the lexical data and then added the non-lexical data by 50% and 100% while removing the lexical data by the same amount of the non-lexical data (the fourth group of Table 9). Since the former experiment used more data, the results of the former experiment show higher performance overall. However, both experiments follow similar patterns—the performance of the model for Korean decreases with the addition of the non-lexical data, while the performance of the English model increases with the addition of the non-lexical data.

We suppose the English non-lexical data being well distributed allowed the models to better understand the language. On the other hand, for Korean, although there are thousands of distinct Korean syllables (Hanguls), only a fraction of them are practically used. Therefore, removing the lexical data to account for the addition of non-lexical data led the models to get trained on less-likely-to-appear data—degrading the performance.

Finally, we compared the performance of the model using only lexical data and non-lexical data (the fifth group of Table 9). For a fair comparison, we only used 20% of the lexical data which is equivalent to the number of 100% of the non-lexical data. For both languages, higher performance was obtained when using only lexical data.

## E. Discussion

We collected a benchmark dataset for the development of WiTA systems in this work. The dataset allows accessible, cost-efficient, and general WiTA systems. Furthermore, the WiTA baselines designed with the proposed spatio-temporal (ST) residual networks implement such easy-to-deploy WiTA systems. The ST residual networks effectively deal with the spatial and temporal contexts inherent in the input image sequences. We demonstrated that the baseline models displayed moderate performance in both Korean and English with reasonable operation time. However, a few future works still exist for further improvement of the performance of the proposed baselines.

First of all, we can investigate more efficient and effective model architectures in future studies. The need for a study on model architectures that achieve higher accuracy through less computational complexity remains. We can hardly train the current baseline models with a larger batch size because of the high computational complexity. If



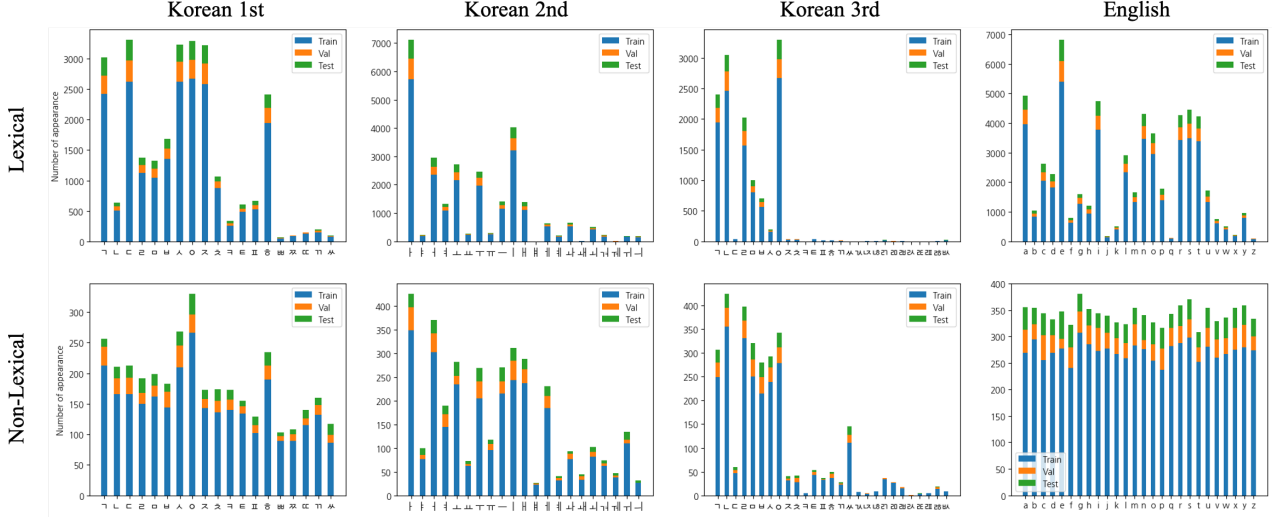


Figure 9. The histogram of each character by dataset split. The non-lexical datasets display more even distribution than the lexical datasets in both languages.

Table 8. Spatio-temporal residual network architectures.  $n$  is 1 for the 10-layered models and 2 for the 18-layered models.






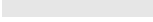










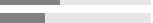



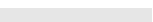



Layer Name	ST-MC	ST-rMC	ST-R3D	ST-R(2+1)D
Stem Block	$3 \times 7 \times 7$ , stride $1 \times 2 \times 2$			$1 \times 7 \times 7$ , stride $1 \times 2 \times 2$ , $3 \times 1 \times 1$ , stride $1 \times 1 \times 1$
Conv Block 1	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times n$	$\begin{bmatrix} 1 \times 3 \times 3, 64 \\ 1 \times 3 \times 3, 64 \end{bmatrix} \times n$	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times n$	$\begin{bmatrix} 1 \times 3 \times 3, 144 \\ 3 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 144 \\ 3 \times 1 \times 1, 64 \end{bmatrix} \times n$
Conv Block 2	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times n$	$\begin{bmatrix} 1 \times 3 \times 3, 128 \\ 1 \times 3 \times 3, 128 \end{bmatrix} \times n$	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times n$	$\begin{bmatrix} 1 \times 3 \times 3, 230 \\ 3 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 230 \\ 3 \times 1 \times 1, 128 \end{bmatrix} \times n$
Pooling (Middle)	Spatio-temporal pooling (maximum or average)			-
Conv Block 3	$\begin{bmatrix} 1 \times 3 \times 3, 256 \\ 1 \times 3 \times 3, 256 \end{bmatrix} \times n$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times n$	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times n$	$\begin{bmatrix} 1 \times 3 \times 3, 460 \\ 3 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 460 \\ 3 \times 1 \times 1, 256 \end{bmatrix} \times n$
Conv Block 4	$\begin{bmatrix} 1 \times 3 \times 3, 512 \\ 1 \times 3 \times 3, 512 \end{bmatrix} \times n$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times n$	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times n$	$\begin{bmatrix} 1 \times 3 \times 3, 921 \\ 3 \times 1 \times 1, 512 \\ 1 \times 3 \times 3, 921 \\ 3 \times 1 \times 1, 512 \end{bmatrix} \times n$
Pooling (Last)	Global adaptive spatial pooling (maximum or average)			-
Fully Connected	$512 \times 256$ fully connections			

future research results in a lighter and faster model architecture, we expect that the training efficiency will improve as well. In addition, the fast and accurate model architectures will maximize the usability of WiTA systems. This will foster the active utilization of WiTA in various fields.

Next, we can diversify the data collection environments in the following study. In this study, we collected the data

in several environments but used one type of device. In the following studies, we can make WiTA performance more robust by collecting data using various devices from more diverse environments. With the introduction of new devices, the data collection conditions, including FPS, image resolution, color space, and the background, will vary. In particular, we would collect data in consideration of a

Table 9. Effect of training data configuration on the performance. Each row represents a training dataset configuration and the performance on the test dataset. The numbers below the ‘Training Data Configuration’ column indicate the amount of the data consisting the each row. We designed five groups of experiments and the double lines separate each experiment group below.

Training Data Configuration		Korean (CER)			English (CER)		
Lexical	N-Lexical	Lexical	N-Lexical	Overall	Lexical	N-Lexical	Overall
100% 	100% 	31.62	44.37	33.16	28.10	36.46	29.24
100% 	50% 	34.49	47.60	36.07	28.20	40.83	29.93
100% 	0% 	38.77	54.06	40.61	32.14	51.98	34.85
90% 	50% 	59.16	72.43	60.76	28.43	40.94	30.14
80% 	100% 	64.66	74.25	65.82	30.99	39.90	32.20
50% 	100% 	41.50	54.14	43.02	36.71	42.71	37.53
50% 	50% 	53.03	64.65	54.43	47.32	58.23	48.81
50% 	0% 	53.22	63.58	54.46	83.06	91.15	84.16
40% 	50% 	69.42	79.80	70.67	62.30	71.46	63.55
30% 	100% 	70.64	79.14	71.66	48.79	48.54	48.75
20% 	0% 	66.07	74.75	67.11	88.18	92.40	88.76
0% 	100% 	79.72	78.39	79.56	92.95	94.58	93.17

dynamic background environment. As these environmental factors diversify, the reliability of the WiTA system developed through the data will enhance.

Furthermore, we can improve accuracy by integrating the WiTA system with typo correction systems. We would not be able to reduce ambiguity between some characters, no matter how much data is available. Thus, there may exist limitations in driving performance improvement with data alone. Using typo correction systems can remove apparent typos. Moreover, we expect that using the character language model (LM) [1] in WiTA systems can reduce typos by employing semantic context. We can utilize LM in WiTA systems in an end-to-end manner or a modular manner.

Last but not least, we can extend the current WiTA proposed in this work to various languages. Currently, the dataset contains Korean and English. Related researchers and we can expand the WiTA dataset using the data collection tool disclosed in this study. In the process of supporting various languages, it is necessary to consider the unique features of the language, such as designing a specific encoding method for each language. In addition, when multiple language data is collected, a single integrated WiTA system can support multiple languages at once. Then, the WiTA system can handle various types of user inputs and become versatile.