Supplementary Material to "Typenet: Towards Camera Enabled Touch Typing on Flat Surfaces through Self-Refinement"

Ben Maman
Tel Aviv University

Amit H. Bermano
Tel Aviv University

1. Dataset

We provide detailed information regarding our introduced TypingHands26 dataset. For all 26 users keyboard data was collected, and for 10 of the users surface data was collected as well. All surface sessions were collected without any real-time feedback to the user. Detailed keyboard data information is given in Table 1, and surface data information is given in Table 2. In each of the tables, for each user we list the number of sessions (column 2), total recorded time (column 3), and total length of text typed in characters (column 4).

We also provide statistics of the dataset - distribution of keypress length, distribution of number of simultaneous active keys, and gender distribution (Figure 1 and Tables 3, 4).

1.1. Keyboard Data Automatic Labelling

In order to automatically label keyboard data, we record users typing on a physical R-go keyboard. The keyboard is split and flat, in order to resemble the real use case (see Figure 2). For each key press, we log the precise keypress and release event times. Video recording and logging are done simultaneously by the same script, thus there is no alignment required for keyboard data. After the recording, labels are assigned as follows: for each key press, all frames occurring between the time of press and time of release of the key are labelled as having the key active. As can be seen in Table 3, most frames either have a single active key or none at all, but some frames have 2 active keys, and even 3, the latter being rare.

1.2. Train / Test Splits

For each of the 10 users, we used one surface session for testing. The test sessions can be seen in Table 5. Validation and early stopping were done on the remaining surface sessions (even when included in the training set for the fine tuning).

For the new user experiments (Table 3 in the main paper), testing was done on the same sessions as in Table 5.

User	KB Sessions	Time	Length C
Be.	3	01:06:41	12895
Ya.	2	40:23	6134
Or.	2	45:06	7533
Ey.	6	52:07	7373
Am.	3	01:02:03	12298
Al.	2	01:06:05	5108
Ey.2	3	51:08	7489
To.	2	51:38	5419
Ad.	3	01:08:15	9293
Om.	4	01:14:58	9303
Yo.	1	23:48	3777
Ro.	3	36:42	5126
Zi.	4	01:30:57	6982
Ma.	3	01:03:25	8843
Ya.	3	54:20	6884
Ra.	3	56:45	8259
Tu.	7	01:23:01	10376
Jor.	8	01:19:08	14548
An.	1	24:51	2986
Sh.	4	01:33:39	12728
Jo.	3	43:48	6095
Ni.	9	01:06:13	20426
Joh.	8	01:18:56	12023
Gi.	6	52:08	15056
Co.	3	29:14	4467
Da.	4	37:02	5722
All	**100**	**24:52:35**	**227,143**

Table 1. Keyboard data information. Column 2 - number of video sessions, column 3 - total video time, column 4 - total length of text entered in characters.

1.3. Test Sessions - Qualitative Results

Our system's output on the test sessions can be seen in Figures 5-14. Best results were obtained for the skilled typists: Gi., Am., Ni., Joh., Be., and Jor.. Note that the provided ground truth text in Figures 5-14 is the original provided text. This means that some of the discrepancies between the provided text and the actual typed text is in

User	Surface Sessions	Time	Length C
Jor.	20	58:20	7513
Ey.	20	01:03:20	7768
Ro.	7	22:46	2186
Jo.	2	07:47	1037
Sh.	2	07:27	1531
Am.	8	26:19	8116
Be.	7	34:49	8066
Ni.	4	49:36	17097
Gi.	3	21:36	8466
Joh.	9	58:23	14708
All	**79**	**05:50:29**	**76,488**

Table 2. Surface data information. Column 2 - number of video sessions, column 3 - total video time, column 4 - total length of text entered in characters.

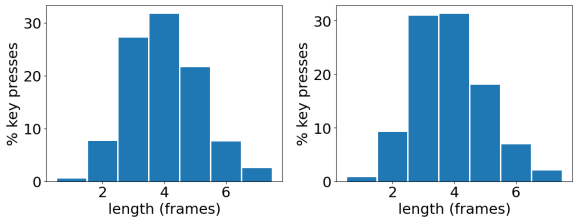

Figure 1. Key press length distribution (length in frames, frequency given in percentage of total key presses). Left - keyboard data, right - surface data. As can be seen, most key presses last 3-5 frames.

	0	1	2	3
KB	61.3	34.8	3.8	0.05
Surface	58.7	37.2	4.1	0.0
All	61.5	34.6	3.8	0.04

Table 3. Key press simultaneous active keys distribution. Frequency given in percentage of total frames. Second column - no active key, third column - single active key, etc. First row - keyboard data, second row - surface data. Most frames with an active key have a single active key, but ∼4% of the frames have 2 simultaneously active keys. 3 simultaneously active keys also exist but are rare.

Figure 2. Split keyboard used for keyboard data.

part due to user typing errors or word repetitions. Reported character-level accuracy in the main paper (93.5% on average for the 10 test sessions) is w.r.t. to the actual typed text,

	Users	Time	Len C
Male (KB)	15	14:14:40	124,238
Female (KB)	11	10:37:54	102,905
All (KB)	**26**	**24:52:35**	**227,143**
Male (Surface)	8	04:39:15	50925
Female (Surface)	2	01:11:13	25563
All (Surface)	**10**	**05:50:29**	**76,488**

Table 4. Gender distribution in the dataset. Column 2 - number of users, column 3 - total video time, column 4 - total length of text entered in characters. Rows 2-4 - keyboard data, rows 5-7 - surface data.

User	Session	Time	Length C
Gi.	gi****s9	02:48	1108
Am.	am****s2	02:02	791
Ni.	ni****s2	03:57	1252
Joh.	joh****s4	05:19	1424
Sh.	sh****s6	01:20	319
Be.	be****s5	06:05	1378
Ey.	ey****s17	02:42	437
Jor.	jor****s17	02:24	473
Ro.	ro****s6	02:46	434
Jo.	jo****s5	01:23	227
All		**30:52**	**12949**

Table 5. Test Sessions.

Method	Raw C	LSTM C
Ours	91.4	93.5
Hand Pose	54.8	49.0

Table 6. Comparison to detection based on hand pose estimation, where we use predicted keypoint locations as frame features.

which was manually adjusted after the fact according to the typist's performance as best as possible. Character-level accuracy of the 10 test sessions w.r.t the original provided text is 93.1% on average.

2. Supplementary Experiments

2.1. Alternative Approach - Hand Pose Estimation

Hand Tracking, or Hand Pose Estimation, is the task of detecting the location of the 21 keypoints of the palm, in a video. Locations can be either 2d or 3d. A natural approach for our task would be to detect key presses according to predictions of a hand tracking system. Various systems exist which perform this task from a monocular RGB video, among which is the Google Media Pipe Hand Landmark Model [6]. We tried using this system for our virtual typing task. We applied the system on all videos in our dataset, to receive per-frame 3d keypoint locations, which we regard as per-frame features. We trained an RNN-based model to classify the frames based on the keypoint locations relative

to the wrist ($3 * (21 - 1) = 60$ coordinates for each hand). The model is composed of a $(120, 512)$ linear layer, a 2-layer GRU with input and hidden size of 512 followed by a $(512, C)$ linear layer, where C is the number of classes. We used the same training data, and the same test sessions as in all the above experiments. As can be seen in Table 6, our proposed system outperformed the pose estimation based system by over 36%. Hand pose estimation is known to be a challenging task, and in our case a small error in the keypoint regression leads to false key predictions. However, the Media Pipe Hands system can be used for tasks where small regression errors have less effect, such as gesture recognition, as explained in the Google Media Pipe Documentation.

Richardson et al. [3] study the application of hand tracking for our task. However, they use marked gloves to facilitate hand tracking and to ensure high hand tracking accuracy. They also use a different camera angle which facilitates detection, but is less practical - above the hands. This highlights the fact that current accuracy of monocular RGB hand tracking systems is insufficient for the task.

2.2. Temporal Context

We have performed an ablation study for the effect of the two kinds of temporal context we apply: for frame feature extraction, and using an RNN-based classifier.

On the frame level, we denote by n the frame with offset n relative to the frame currently considered. The frame itself is denoted by 0. For example, -4 means four frames before the frame in question. We trained models with RNN classifiers, with different choices for temporal context for frame feature extraction (using sequence length of 48 for training the RNN). We have also tried classifying each frame independently, by using a single linear layer, instead of an RNN. We did this using temporal context at frame feature extraction, and without it. Results can be seen in Table 7.

We have applied two different training schemes: In the first, we train a network on keyboard data and validate on surface sessions, and fine tune on combined surface and keyboard data. We denote this scheme A in Table 7. In the second scheme, we train the network from scratch on combined surface and keyboard data. We denote this scheme B in Table 7. As can be seen, the employed design choices outperform the other options typically by a non-negligible margin.

2.3. Speed Invariance

An important property that is desired for our system is pace invariance, i.e., the system should output the same text regardless of the speed it was typed in. In order to test this property, we perform temporal subsampling on the test sessions, in order to simulate different typing speeds for each session. For each test session, we use four different subsampling rates: (i) keeping all frames (ii) omitting every 4-th frame, increasing speed by a factor of 1.33 (iii) omitting every third frame, increasing speed by a factor of 1.5 (iv) omitting every second frame, increasing speed by a factor of 2. This, of course, could have also been achieved using an interpolation based temporal retiming scheme. The difference is negligible for our task.

When using temporal context, whether for feature extraction or using an RNN classifier, typing speed will have an impact on the prediction. This is especially true when using temporal context for extracting frame features, since the frame features will depend on the typing speed. In order to create robustness to variance in typing speed, we train our system with random temporal subsampling: During training, in each epoch, for each video session we randomly select a temporal subsampling rate, either (i) none, (ii) every second frame or (iii) every third frame. we also randomly select an offset to ensure all frames are used for training.

This augmentation gives significant improvement in test accuracy when using temporal context of -4, -2, 0, 2 for frame feature extraction, with or without subsampling in test time. Table 8 shows the test results for the 4 different subsampling rates, when training with random temporal subsampling, and when training w/o random temporal subsampling.

On the other hand, when using an RNN classifier w/o temporal context for feature extraction, i.e. temporal context 0, speed invariance is better maintained even when training without random temporal subsampling, due to the fact that frame features are independent of speed. Results can be also seen in Table 8.

2.4. Camera Angle

We test our system's robustness to changes in camera angle. We show quantitative results for the users Be. and Am.: Each of the users typed the same text several times, with various camera angles: $0°$, $360°/32=11.25°$, $360°/16=22.5°$, and $360°/8=45°$, each angle used both from the left and from the right (7 angles). The different angles can be seen in Figure 3. The lengths of these sessions are 1:45-2:10 minutes. For each angle, we computed the mean text accuracy over the two users. We ran two different systems on these sessions: one was trained with random warping applied on the training data, and one was trained w/o random warping. Results can be seen in Tables 9 (with random warping) and 10 (w/o). As can clearly be seen, random warping during training provides much better robustness to changes in angle. We tested the system that was trained with random warping on the 10 test sessions used in the main paper, and appearing in Table 5, and received comparable mean accuracy of 92.9, vs. 93.5 w/o random warping. From the experiments we conclude, as can be ex-

Context	Raw C	W	LSTM C	W
RNN 0 A	87.2	53.6	91.8	75.2
RNN 0 B	88.0	56.4	90.3	69.6
RNN 0 Subsampling A	83.5	47.3	80.9	48.0
RNN -4, -2, 0, 2 A	88.5	58.6	90.9	68.7
RNN -4, -2, 0, 2 B	88.5	58.1	91.6	72.8
RNN -4, -2, 0, 2 Subsampling A	**91.4**	**68.7**	**93.5**	**78.7**
RNN -4, -2, 0, 2 Subsampling B	89.0	57.4	91.2	71.4
RNN -2, -1, 0, 1 A	87.5	57.2	90.7	70.7
RNN -4, -3, -2, -1, ..., 2 A	87.1	55.1	89.1	65.0
Linear -4, -2, 0, 2 A	87.1	55.6	88.9	71.8
Linear -4, -2, 0, 2 Subsampling A	85.0	49.6	88.8	70.6
Linear 0 B	69.8	35.0	67.9	53.8

Table 7. Temporal context experiments: character- and word-level accuracy for raw output (columns 2-3) and after beam search with LSTM model (columns 4-5). RNN denotes a system with a GRU-based classifier, and Linear denotes a system with a linear classifier, w/o an RNN. Number sequences (*e.g.*, -4, -2, 0, -2) denote temporal context used for frame feature extraction. Subsampling denotes a system trained with random temporal subsampling (see Section 2.3). *A* denotes a system trained on keyboard data and fine-tuned on combined surface and keyboard data. *B* denotes a system trained from scratch on combined surface and keyboard data.

Speed	Raw C	W	LSTM C	W
1.0, Context -4, -2, 0, 2, Subsampling	91.5	68.7	93.5	78.7
1.0, Context -4, -2, 0, 2	88.7	58.2	91.8	72.8
1.0, Context 0	87.1	53.7	92.0	75.6
1.33, Context -4, -2, 0, 2, Subsampling	90.3	64.4	91.1	75.7
1.33, Context -4, -2, 0, 2	84.8	48.4	88.5	67.6
1.33, Context 0	85.2	47.8	89.9	70.0
1.5, Context -4, -2, 0, 2, Subsampling	89.5	62.4	85.4	67.1
1.5, Context -4, -2, 0, 2	82.0	41.7	83.3	56.4
1.5, Context 0	83.6	44.2	87.5	66.6
2.0, Context -4, -2, 0, 2, Subsampling	86.0	52.0	67.0	41.5
2.0, Context -4, -2, 0, 2	67.8	20.9	65.8	32.4
2.0, Context 0	78.9	34.7	77.8	43.8

Table 8. Effect of increasing video speed in test time when training with and w/o random temporal subsampling, with temporal context of -4, -2, 0, 2, and w/o random temporal subsampling, using temporal context 0.

Angle	Raw C	W	LSTM C	W
0°	93.1	72.2	96.6	87.2
11.25° left	92.2	69.9	95.9	86.8
11.25° right	82.6	44.4	86.2	58.3
22.5° left	84.7	51.1	89.1	71.8
22.5° right	71.8	29.7	76.9	44.0
45° left	10.6	0.8	0.7	0.4
45° right	9.1	0.0	0.3	0.4

Table 9. Accuracy on sessions captured from different angles (see Figure 3), with a system trained with random warping.

Angle	Raw C	W	LSTM C	W
0°	92.6	70.7	95.4	86.1
11.25° left	89.6	64.3	91.3	75.2
11.25° right	78.8	36.8	83.4	58.6
22.5° left	81.8	43.2	87.0	67.7
22.5° right	52.0	16.5	52.3	30.5
45° left	6.1	0.8	4.8	1.9
45° right	3.3	0.0	0.2	0.4

Table 10. Accuracy on sessions captured from different angles (see Figure 3), with a system trained w/o random warping.

pected, that enriching the dataset with more captured angles would probably increase the robustness and accuracy of the method even further.

3. Architecture

We rely on a standard backbone such as Resnet to extract features. We use Resnet18, which provides real time performance, as the system is required to work on a mo-

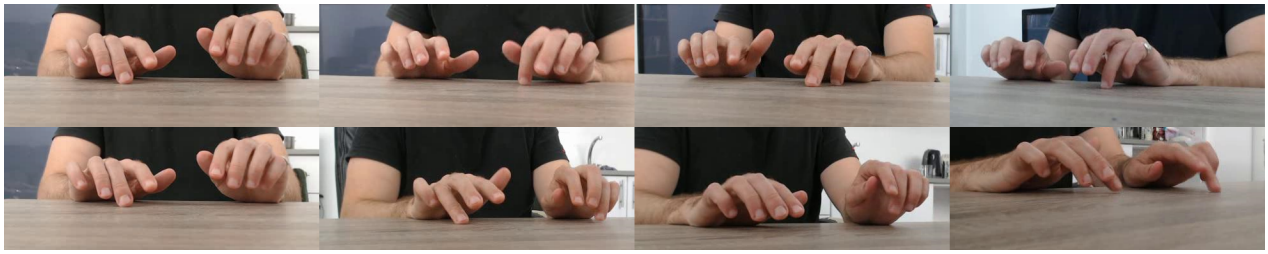

Figure 3. Different angels used for testing robustness to changing angle. Top: 0°, 11.25°, 22.5°, 45° from the left. Bottom: 0°, 11.25°, 22.5°, 45° from the right.

bile phone at a 30FPS frame rate. The output feature size is 512. As explained in the paper, to extract features for a given frame we concatenate neighboring frames to it as additional channels (all frames are converted to grayscale). The input size is (4, 150, 200) (4 channels, as the frame is concatenated to 3 neighboring frames). As a classifier we use a 2-layer Gated Recurrent Unit, with input size 512 and hidden size 512, followed by a $(512, C)$ linear layer where C is the number of classes, and a Sigmoid activation.

The output is a vector of probabilities, where each entry is a number in the interval $[0, 1]$, representing the confidence that the corresponding class is active. Multiple active classes are possible, hence the Sigmoid activation.

The classes we used for training are: a-z, space, comma, period, semicolon, quote, enter, backspace, left shift, right shift, and slash - 36 classes. The classes we used for testing are: a-z, space, comma, period, and semicolon (i.e., 30 classes of the 36).

In addition, we have the following redundancies: a class for each finger, where each index finger has two classes for its two possible keyboard columns (*e.g.*, r and t belong to different classes, but g and t are of the same one). Also, for each hand and each row (front, middle and back) a class. Lastly, the *none* class, that is active if no key is active. Altogether, $36 + 10 + 6 + 1 = 53$ classes.

4. Training

We minimize the Binary Cross Entropy Loss averaged on all classes. We use an Adam optimizer, with initial learning rate of 10^{-4}, reducing it by a factor of 0.5 upon 5 epochs without improvement in the training loss. We use weight decay of 10^{-6}, and Dropout of 0.5 on the inner layer of the GRU.

Training The RNN We train our network using batch size of 24 and a sequence length of 48. We train the network using a sticky hidden state, *i.e.*, the first hidden states are all 0, and in each of the following training iterations, the sequence of each batch receives as input the last hidden state of the corresponding batch form the previous iteration. Therefore, we implement the training scheme as a concatenation of the video sessions: We divide the training set into 24 sequential batches. For example, if the training set holds 24000 frames in total, the first 1000 frames after concatenating everything are the first batch, the next 1000 are the second batch etc.. The first training iteration contains frames 0-47 that belong to the first batch, frames 1000-1047 that belong to the second batch, etc.. The second training iteration contains frames 48-95 from the first batch, frames 1048-1095 from the second batch etc..

Full training takes $\sim$ 2 weeks ($\sim$ 150 epochs) of Nvidia GeForce RTX 2080 Ti GPU time. Real-time inference on a laptop with a webcam was done with a Quadro P3200 Max-Q GPU.

4.1. Dynamic Time Warping

In order to label surface footage, where we know the entered text but not the per-frame alignment, we find the optimal alignment between the text and the video sequence, using dynamic time warping. In order to also find the beginnings and endings of key presses, we pad the text with the *none* class between consecutive characters, *e.g.*, if we denote the *none* class by *N*, then 'one two' is padded to 'NoNnNeN NtNwNoN'. As a local cost function between a given video frame and a given character we measure the Euclidean distance between the frame's predicted vector of probabilities, and the one-hot representation of the character.

The DTW algorithm yields a monotonic mapping for the optimal alignment - each frame is mapped to a certain character and vice versa. We discard singular points, that is, frames that are mapped to more than one character. These are due to user typing errors. Typically, a key is pressed for 3-5 frames. For each non-singular frame, we define its label to be the single character it is mapped to.

This process provides a single class per frame, while our problem is multi-label. In order to obtain multi-labels, for each character, we extend it to its corresponding frames' neighbors, if the neighbors' predicted probability vectors meet certain conditions (*none* class below 0.5, the given character above 0.01, and total key activation length does

Method	Raw C	W	LSTM C	W
Pseudo-Labels	86.2	49.8	91.2	72.0
Ours	91.4	68.7	93.5	78.7

Table 11. Comparison to Pseudo-labelling.

FaX	Raw C	W	LSTM C	W
w/o	85.7	48.2	88.0	61.3
w/	**91.4**	**68.7**	**93.5**	**78.7**

Table 12. Effect of using a Focus and Expand (FaX) training scheme. Row 1 - w/o using Fax, row 2 - using Fax.

not exceed 5). This results in frame-level multi-labels.

The network is then trained on keyboard data combined with surface data, until convergence on a validation set.

4.2. Pseudo-Labelling

Pseudo-labelling is a technique for training on combined labelled and unlabelled data, where the network's predictions on the unlabelled data are used as labels for training [4, 2, 5]. We performed an ablation study to compare our proposed self-refinement to Pseudo-labelling, as proposed by Lee [2]: In the second training phase, instead of labelling the surface sessions using self-refinement, we use the network's predictions on the surface sessions as labels (on-the-fly). The loss from unlabelled data is weighted by a factor of 0.1. Results can be seen in Table 11.

4.3. Background Handling

In order to ignore the background, and in particular ignore the keyboard prior present in the keyboard data, we perform hand segmentation as preprocessing and then during training place the hands on random backgrounds, in random locations, on the fly. We follow the Focus and Expand (FaX) training scheme suggested by Arar et al. [1] in order to guide the network to focus on the hands rather than the background: For epochs $i = 0, 1, \ldots 9$, the intensity of the background in epoch i is sampled randomly from the uniform distribution on $[0, i/10]$. For epochs $i = 10, 11, 12$, the background intensity is sampled from the uniform distribution on $[0, 1]$. For epochs $i > 12$, the background intensity is 1 with probability 0.5, and with probability 0.5 it is sampled from the uniform distribution of the range $[0, 1]$. This training scheme made a significant contribution, as can be seen in Table 12.

4.4. Spatial Dropout

We performed an ablation study to test the effect on the accuracy of undetected hand parts, to examine errors that might have occurred in the hand segmentation step during training. For this, we applied spatial dropout during inference (Table 13) and during training (Table 14): We choose a factor $\alpha \in \{0.25, 0.1, 0.05, 0.025, 0.0\}$. For each of the two hands, we randomly zero out a box from the hand's original bounding box, with size ratio of α w.r.t. the original bounding box (see Figure 4). As can be seen in Table 13, α of size 0.025-0.05 have a relatively small effect on accuracy. Training with spatial dropout does not improve test accuracy, whether or not applying spatial dropout during inference (Table 14).

Figure 4. Example of spatial dropout with $\alpha = 0.25$

Spatial Dropout	Raw C	W	LSTM C	W
0.25	81.9	39.1	82.9	48.2
0.1	88.4	57.1	89.7	66.0
0.05	90.5	65.6	92.7	76.5
0.025	91.3	68.1	93.4	78.7
0.0	91.4	68.7	93.5	78.7

Table 13. Spatial Dropout during inference, when training w/o spatial dropout. Left column - factor of hand bounding box that is randomly zeroed.

Tr. Spatial Drop.	Raw C	W	LSTM C	W
0.25	87.1	54.6	87.6	59.3
inference 0.25	83.4	42.8	81.1	41.2
0.1	88.4	60.2	88.5	65.9
inference 0.1	87.4	56.6	86.1	57.7
0.0	91.4	68.7	93.5	78.7

Table 14. Spatial Dropout during training, w/o dropout during inference (rows 1, 3) and with dropout also during inference (rows 2, 4).

5. Beam Search Implementation

Threshold for Candidate Keys Since this is a multi-label problem, a possible approach for performing the Beam Search would be to consider all combinations of classes in each frame as possible candidates. This, however, is not feasible. Therefore, during inference each sequence holds one class per frame. Moreover, we select as possible candidates only keys whose vision probability exceeded a low threshold, which reduces the amount of computations per frame significantly. The threshold we used was 0.005.

Compression	LSTM C	W
1	91.6	71.9
2	**93.5**	**78.7**
3	89.7	73.0

Table 15. Effect of using super frames for beam search with LSTM based language model. Row 1 - beam search operates on original frames. Row 2 - probabilities of every 2 consecutive frames are merged using mean. Row 3 - probabilities of every 3 consecutive frames are merged using mean.

Max Sequence Length As mentioned in the main paper, we assign scores to candidate prediction sequences according to suffixes of limited length, which is the max sequence length. The vision component of the score is the geometric mean of the suffix predictions' probabilities. To compute the language component of the score, we generate the string suffix corresponding to the sequence suffix, and prepend the 5 keys that appeared before the string suffix in the entire string corresponding to the sequence. Regarding limited length suffixes rather than the entire sequences serves two purposes: (i) Maintaining sensitivity to new keys. Computing means over the entire sequences would make a single key insignificant for long sequences. (ii) Performance. The max sequence length we used in our experiments was 256.

Weighted Sum of Vision and Language The score for each sequence is a weighted score of its vision and language scores. The vision weight we used was 1.0. The language weight varied between different settings: For the LSTM based language model we used weight 1.0 for pre-recorded videos, and 2.0 for real-time usage. For the Transformer based model, we used weight 0.5.

Super Frames When using the LSTM language model, in order to further reduce the computational load, we merged consecutive frames into *super frames*. The super frame probability vector is the mean of the probability vectors of the consecutive original frames. Thus, if *e.g.* the video stream is 30 FPS and we merge every 3 consecutive frames, the beam search only has to operate on 10 FPS (while the vision model still operates on 30 FPS).

In addition to reducing the computational load, this also significantly improves text accuracy for the LSTM based language model, probably because of the smoothing effect and the reduced search space. Performing the beam search on the original per-frame probabilities gives average text accuracy of 91.4%, while merging each 2 consecutive frames gives average text accuracy of 93.5. Merging every 3 consecutive probability vectors reduces accuracy: Mean was 89.7%. Note though that merging every 3 frames gave higher word-level accuracy than using the original probabilities (73.0% vs. 71.9%).

Adjusted Probabilities	LSTM C	W
w/o	93.1	76.7
w/	**93.5**	**78.7**

Table 16. Effect of increasing probabilities for same-finger keys.

5.1. Same Finger Keys Correlation

As mentioned before, one of our main challenges is to disambiguate between keys pressed by the same finger, where only the depth distinguishes between them, eg. 't' and 'g'. This becomes even more challenging when typing on a surface, since there's no physical constraint in the form of a keyboard to force the user to make distinguishable motions. We rely on the beam search method to help disambiguate: The vision model supplies for each frame a raw probability vector P. From this vector we create an enhanced probability vector P_E, in the following manner: We define a set of pairs of keys that are likely to be confused, denote it by C. For completeness, we also add to the set the identity pairs, i.e., $(c, c) \in C\ \forall c$ — a key can trivially be confused with itself. For each key c, we define its enhanced probability $P_E(c)$ as:

$$P_E(c) = \max_{(c',c)\in C} (1-\lambda)P(c) + \lambda P(c') \tag{1}$$

where λ is a hyper parameter. In practice, we decide the value of λ according to a low threshold:

$$\lambda = \lambda(P(c)) = \begin{cases} 0.5 & P(c) >= 0.005 \\ 0.1 & P(c) < 0.005 \end{cases} \tag{2}$$

i.e., the probability enhancement is weak if the raw probability is weak. The actual probability vector we use in the beam search for each frame is its corresponding enhanced probability vector, P_E. Intuitively, the enhancement means that if a key has a high probability, then the probability of similar keys will automatically increase.

We define the set C to be the set of all pairs of keys that are pressed by the same finger and are in adjacent rows or the same row, with the exception of the pairs

$$(r, g), (f, b), (f, t), (y, j), (h, m)$$

which are less likely to be confused. *e.g.*,

$$\{(w, w), (t, g), (u, j), (u, h), (i, k), (a, z), (a, q)\} \subseteq C \tag{3}$$

but

$$\{(t, v), (e, c), (q, z)\} \cap C = \emptyset. \tag{4}$$

We measured the effect of thus adjusting the probabilities, and have seen an improvement of 2% in word-level accuracy, and 0.4% in character-level accuracy, as can be seen in Table 16.

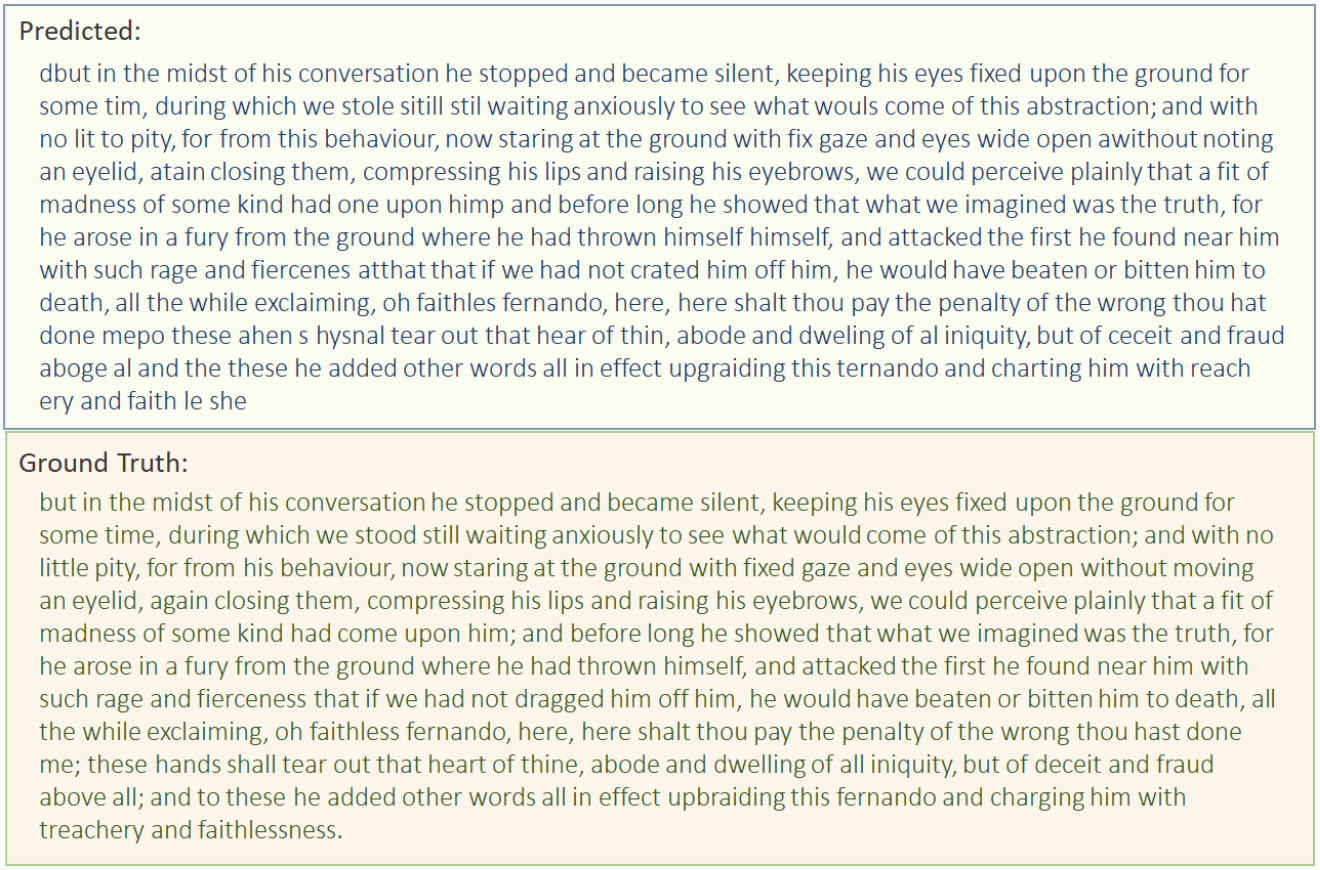

Figure 5. Test session of the user Gi.

References

[1] Moab Arar, Noa Fish, Dani Daniel, Evgeny Tenetov, Ariel Shamir, and Amit Bermano. Focus-and-expand: Training guidance through gradual manipulation of input features. *CoRR*, abs/2007.07723, 2020.

[2] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[3] Mark Richardson, Matt Durasoff, and Robert Wang. Decoding surface touch typing from hand-tracking. In Shamsi T. Iqbal, Karon E. MacLean, Fanny Chevalier, and Stefanie Mueller, editors, *UIST '20: The 33rd Annual ACM Symposium on User Interface Software and Technology, Virtual Event, USA, October 20-23, 2020*, pages 686–696. ACM, 2020.

[4] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. IEEE, 2020.

[5] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019.

[6] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *CoRR*, abs/2006.10214, 2020.

Predicted:

brothe special doctrines advocated by ant, it is very common among philosopers to regard what is a prioris in some sense mental, as concerned rather with the say we must think than with any fact of the outer world. we noted in the preceding chapter the three principles commonly called laws of thought. the view which led to their being soned is a natural one but there are strong reasons for thing that it is eroneous let us take as an illustration the law of contradiction. this is com only stated in the form nothing can both be and not be, which is intende to expres the fact that nothing can at once have and not have a tiben buality thus, for exple, if a tree is a beech it canot also be not a beech if mutable is rectangular it acannot also be not rectangular, and so on.

Ground Truth:

apart from the special doctrines advocated by kant, it is very common among philosophers to regard what is a priori as in some sense mental, as concerned rather with the way we must think than with any fact of the outer world. we noted in the preceding chapter the three principles commonly called laws of thought . the view which led to their being so named is a natural one, but there are strong reasons for thinking that it is erroneous. let us take as an illustration the law of contradiction. this is commonly stated in the form nothing can both be and not be , which is intended to express the fact that nothing can at once have and not have a given quality. thus, for example, if a tree is a beech it cannot also be not a beech; if my table is rectangular it cannot also be not rectangular, and so on.

Figure 6. Test session of the user Am.

Predicted:

eshe would not give him time toreply, but hurrying instantly to her husband, called out as she entered the library, on mr. benet, are wanted immediately we are all in an uproar. you must come and make lazy marry mr. colins, for she vows whe wil not have him, and if you do not make haste he wil change his mind and not have her.mr. bener raised his eyes from his book as she entered, and fixed them on her face with a calm unconcern which was not in the last altered by her communication.i have not the pleasure of understanding you, said he, when she had finished her seech. or what are you talkingof mr. colins and lazy lizy declares she will not have mr. coins, and mr. colins begins to say that he wil not have lizzy.and what am i to do on the occasion it seems and hopeles business.sea, to lay about it yourself. tell her that you insist upon her marrying him.le her be called own. she shal hear my opinion.mrs. bennet ran to the the bell, and mis elizabeth was summoned to the library.come here, child, cried her father as she appeared. i have sen for you on an afair of importance. i understand that mr. collins has made you and an an offer of marriage. is it trurtrue true elizabether relied that itw it it was. very well and this offer of marriage you have refusedi have, sir.

Ground Truth:

she would not give him time to reply, but hurrying instantly to her husband, called out as she entered the library, oh mr. bennet, you are wanted immediately; we are all in an uproar. you must come and make lizzy marry mr. collins, for she vows she will not have him, and if you do not make haste he will change his mind and not have her. mr. bennet raised his eyes from his book as she entered, and fixed them on her face with a calm unconcern which was not in the least altered by her communication. i have not the pleasure of understanding you, said he, when she had finished her speech. of what are you talking of mr. collins and lizzy. lizzy declares she will not have mr. collins, and mr. collins begins to say that he will not have lizzy. and what am i to do on the occasion it seems an hopeless business. speak to lizzy about it yourself. tell her that you insist upon her marrying him. let her be called down. she shall hear my opinion. mrs. bennet rang the bell, and miss elizabeth was summoned to the library. come here, child, cried her father as she appeared. i have sent for you on an affair of importance. i understand that mr. collins has made you an offer of marriage. is it true elizabeth replied that it was. very well and this offer of marriage you have refused i have, sir.

Figure 7. Test session of the user Ni.

Predicted:

a ow was the cover road that lay, on a friday night late in november, before the first of the persons with whom this history has busines the dover roadly, as to him, beyond the diver mail, as it lumbered up shooters hhilll he walked up hill in the mire by the side of the mail, as the rest of the passeners did; not because they had the least relish for walking exercies, under the circumstances, but because the hill, and the harness, and the mud, and the mail, were all so leavy, that the horses had three times already comet a stop, besides once drawing the coach across the road, with the mutinous intenet of taking it back to black heath, reins and whi and coachman and guard, however, in combination, had read that article of war which forbade a purpose othersiwise strongly in favour of the argument, that some brute animals are ended with reason, and the team had capitulated and returned to the duty. with dropping heads and tremulous tails, they mased their way through the thick ud, floundering and stunbling between whiles, as if they were falling to pieces at the larger honts. as often as the driver rested them and brought them to a stand, with a wary who sophopthenqp the near leadrer violentily shoon his head and everything uping it like an unusually emphatic hourse, denying that the coach could be got up the hill. whenever the leaeder made this ratle, the pasengers stated, as a nervous passenger might, and was disturbed in mind. m

Ground Truth:

it was the dover road that lay, on a friday night late in november, before the first of the persons with whom this history has business. the dover road lay, as to him, beyond the dover mail, as it lumbered up shooter s hill. he walked up hill in the mire by the side of the mail, as the rest of the passengers did; not because they had the least relish for walking exercise, under the circumstances, but because the hill, and the harness, and the mud, and the mail, were all so heavy, that the horses had three times already come to a stop, besides once drawing the coach across the road, with the mutinous intent of taking it back to blackheath. reins and whip and coachman and guard, however, in combination, had read that article of war which forbade a purpose otherwise strongly in favour of the argument, that some brute animals are endued with reason; and the team had capitulated and returned to their duty. with drooping heads and tremulous tails, they mashed their way through the thick mud, floundering and stumbling between whiles, as if they were falling to pieces at the larger joints. as often as the driver rested them and brought them to a stand, with a wary wo ho so ho then the near leader violently shook his head and everything upon it like an unusually emphatic horse, denying that the coach could be got up the hill. whenever the leader made this rattle, the passenger started, as a nervous passenger might, and was disturbed in mind.

Figure 8. Test session of the user Joh.

Predicted:

c a letter from him a had reacherhad reached me a wild letter whch demanded that in reply coming tosehim he wrtote of an illnes of the cle of a sidiness of the mind and of a desire to see neis best ipane indeed us only firmd it was the mathher in which all ths was said it was the heart in t which did not alow me to say ox

Ground Truth:

a letter from him had reached me; a wild letter, which demanded that i reply by coming to see him. he wrote of an illness of the body of a sickness of the mind and of a desire to see me his best and, indeed, his only friend. it was the manner in which all this was said it was the heart in it which did not allow me to say no.

Figure 9. Test session of the user Sh.

Predicted

batter some time he felt for his pipe it was not broken and that was something the he belt for his pouch and there was some tobacco in it and that was something more then he feelt for matches and he could not find any at all and that shattered his hopes completely just as well for him as he agreed when he came to his senses goodnes knows what the striking of matches and the smell of tobacco would have brought on him out of dark holes in that horrible place still at the momoht he felt verly crushed but in slapping all his pockets and feeling all round himself or matches his hand came on the uilt of his litle sword the little dagger that he bot from the trolls and that he had uite forgotten hor do the goblins seem to have noticed it as he wore it inside his breeches now he drew it out it shone pale and dim before his eyes so it is an elvish blade too he thought and goblins are not very near and yet hot far enough but omehow he was comforted it was rather splendid to be wering a blade made in gondolin for the boblin wars of which so many songs had sunt and also he had noticed that such wepons made a great impression on goblins that came upon them suddenly go back he thought no good at all to sidewys inpossible go orward only thing to do on we go so up ne got and trored along with his little sword held in front of him and one hand feeling the wall and his hert all of a patter and a pitter

Ground Truth

after some time he felt for his pipe it was not broken and that was something then he felt for his pouch and there was some tobacco in it and that was something more then he felt for matches and he could not find any at all and that shattered his hopes completely just as well for him as he agreed when he came to his senses goodness knows what the striking of matches and the smell of tobacco would have brought on him out of dark holes in that horrible place still at the moment he felt very crushed but in slapping all his pockets and feeling all round himself for matches his hand came on the hilt of his little sword the little dagger that he got from the trolls and that he had quite forgotten nor do the goblins seem to have noticed it as he wore it inside his breeches now he drew it out it shone pale and dim before his eyes so it is an elvish blade too he thought and goblins are not very near and yet not far enough but somehow he was comforted it was rather splendid to be wearing a blade made in gondolin for the goblin wars of which so many songs had sung and also he had noticed that such weapons made a great impression on goblins that came upon them suddenly go back he thought no good at all go sideways impossible go forward only thing to do on we go so up he got and trotted along with his little sword held in front of him and one hand feeling the wall and his heart all of a patter and a pitter

Figure 10. Test session of the user Be.

Predicted:

baccuracy activity actualy analysis anything anywehre aplying asenbly attorney birthday capacity carying category commonly delivery directely emeletly employee empployer enjoyin entirely everyday veryenetirely evercyda y everyone facility holidays shoefriencly ho lidays honesraly idsente eigy in teentiety indysrtry ekeyvoard manority military ninkstary normaly paumentas phusical psowsibly oroiorith robably proery ereerty recentuly recovery security slightly staratevy strongly stuching wt udying ssudaenly shnporms omsse

Ground Truth:

accuracy activity actually analysis anything anywhere applying assembly attorney birthday capacity carrying category commonly delivery directly employee employer enjoying entirely everyday everyone entirely everyday everyone facility holidays friendly holidays honestly identify identity industry keyboard majority military ministry normally payments physical possibly priority probably properly property recently recovery security slightly strategy strongly studying studying studying suddenly symptoms

Figure 11. Test session of the user Ey.

Predicted:

chavoidng the larger rooms, which were dark and mark and made fast for the night, monsierur the marquis, with his flambeau bearer going on before, went up the staircase to a dodor in a corridor..o this thrown open, admitted him to his own pricate apartment of the rooms.. his bed chamber and two others.. high vaulted room with cool iunccarpeted floors, great dogs upon the heaths for the burning of wood in winter time, and all luxuries befiting the state of a marquis in a sluxurous age and country.

Ground Truth:

avoiding the larger rooms, which were dark and made fast for the night, monsieur the marquis, with his flambeau bearer going on before, went up the staircase to a door in a corridor. this thrown open, admitted him to his own private apartment of three rooms his bed chamber and two others. high vaulted rooms with cool uncarpeted floors, great dogs upon the hearths for the burning of wood in winter time, and all luxuries befitting the state of a marquis in a luxurious age and country.

Figure 12. Test session of the user Jor.

Predicted:

cbrom on this room, nany such dots have been taken out to be hanted in the next rom my bedrom, one fellow, to our knowledge, was poinniarded in the spot for poressing gtttfrressing some insolent delicacy respecting his daughter his dahghter we have lost many pribilaeges a new philosophty has become the more and the aserion of our station, in these cats,y ,s, might tydo not bo so far as to say would but might caubut, bht might cause hs real incond gen ience I all very bad, bery bad

Ground Truth:

from this room, many such dogs have been taken out to be hanged; in the next room my bedroom , one fellow, to our knowledge, was poniarded on the spot for professing some insolent delicacy respecting his daughter his daughter we have lost many privileges; a new philosophy has become the mode; and the assertion of our station, in these days, might i do not go so far as to say would, but might cause us real inconvenience. all very bad, very bad

Figure 13. Test session of the user Ro.

Predicted:

che eto down said amoe. and ne gog jo same she make a grab at tickler, and she pramaged out. thatps what he did, sasaid joe slowly dlearning the firs betebbetween the lower gaars with the pliker and looking at it; she rampaged out pilp

Ground Truth:

she sot down, said joe, and she got up, and she made a grab at tickler, and she rampaged out. that s what she did, said joe, slowly clearing the fire between the lower bars with the poker, and looking at it; she rampaged out, pip.

Figure 14. Test session of the user Jo.