Using Speech for Handwritten Mathematical Expression Recognition Disambiguation

Sofiane MEDJKOUNE[1,2], Harold MOUCHERE[1], Simon PETITRENAUD[2]
and Christian VIARD-GAUDIN[1]
[1]*LUNAM University, University of Nantes, IRCCyN, France*
[2]*LUNAM University, University of Le Mans, LIUM, France*
{*sofiane.medjkoune, harold.mouchere, christian.viard-gaudin*}*@univ-nantes.fr*
simon.petit-renaud@lium.univ-lemans.fr

***Abstract*—The main goal of this work is to set up a multimodal system dedicated to mathematical expression recognition. In the proposed architecture, the transcription coming out from a speech recognition system is used to disambiguate the result of a handwriting recognition module. A set of keywords is built from the transcription module and used to rescore the outputs of both the handwriting classifier and the structural analysis module. Performances evaluated on the *HAMEX* dataset show a significant improvement over a single modality system.**

***Keywords*-Mathematical expression; Handwriting recognition; Speech recognition; Data Fusion;**

I. Introduction

Most of the time, editing bi-dimensional language using common tools dedicated to the task of document formatting is more complicated than editing a standard text. Mathematical expressions (ME) are an example of such a bi-dimensionnal langage. Two successive symbols composing a ME can be arranged in many different ways, according to their spatial relationship (*left/right, up/down, subscript/superscript, inside*) giving rise to a possible complex layout. To insert a ME in a document, specialized editors like LaTeX or *MathType* are generally used. However, even using these tools, ME edition is quite time consuming. Moreover, it is very difficult with an editor like LaTeX to use the right syntax to specify the positions of the symbols and to handle the edition rules. The other widely used editor, *MathType*, offers an alternative to the previous one by giving a visual feedback to the user during the edition, but it still time consuming. The recent technological progress provides new perspectives regarding the human-machine interaction possibilities [1]. Speech and handwriting are among the modes which have most attracted researchers. Systems based on them are quite natural and do not require as much efforts as the keyboard-mouse oriented systems. As a preliminary experiment, we asked 10 persons, who are more and less familiar with ME and handle pretty well both LaTeX and *MathType* editors, to type the ME $\lim_{x_0 \to 0^+} \int_{-\infty}^{x_0} \frac{1-x^4}{2x+3} dx$, once using a pen and a sheet of paper and another time using mouse, keyboard and these two specialized editors. The average time of the pen-based edition is 4 times less than with *MathType*, and 5 times less than with LaTeX (*18 seconds in average for the pen-based edition against 75 seconds for MathType and 90 seconds for LaTeX*). In this regard, the problem of handwritten ME recognition has been widely investigated [2]. The efforts made by the scientific community led to the development of several competitive systems. Nevertheless, these systems are not hundred percent reliable. In fact, there are some drawbacks that cannot be overcome because of the nature of the handwriting signal (symbol and relation ambiguities). Most of the time, these confusions are not obvious to discern even for an experienced observer who would look at the handwritten ME layout (Fig.1). These obstacles may be crossed over only if there is an additional source of information which is able to remove the involved ambiguities.

More recently, the speech recognition community has been interested by the problem of mathematical expression recognition (MER) using automatic speech recognition (ASR) [3], [4]. Most of the works rely on an ASR system that provides the basic automatic transcription of the speech signal. Then, this latter is sent to a parsing module to convert the simple text describing the ME (1D) into its mathematical language writing (2D) [5], [4]. Here again, the systems set up are far from being hundred percent reliable. In addition to the resulting errors during the recognition step (common to all ASR systems), the transition from the textual description of the ME to its 2D writing is not obvious at all (Fig.1). The example in Fig.1 not only shows the cases where the two systems are in failure, but also that the two modalities are complementary. One can see this complementarity inasmuch that the problems encountered by both modalities are of different kinds. This leads to the fact that the missing information in one modality is generally available in the other one.

Starting from this observation, we propose in this paper to explore this track which consists of combining these two modalities (audio and handwriting) to overcome the weaknesses of each modality taken separately. Thus, the

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.288

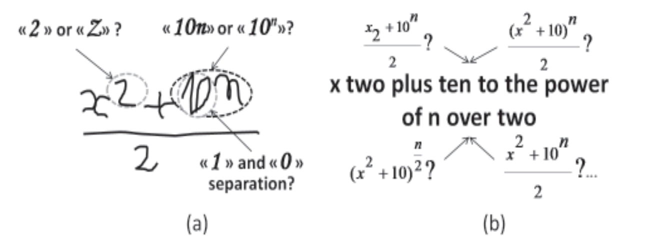

Figure 1. Examples of encountered problems in automatic MER with respect to (a) handwriting modality, (b)speech modality

remainder of this paper is organized as follows: in section II we briefly review the necessary background for the work proposed in this paper. In section III we present our system. We will devote section IV to the presentation of results and their analysis. Section V concludes this paper and gives perspectives of this work.

II. Background

We report in this section a brief overview of these new systems dedicated to math typing based either on speech or on handwriting recognition. The last subsection introduces the concept of fusion.

A. Handwritten MER

The handwriting recognition systems we consider in this paper are online ones. Therefore, the signals that are processed are composed of a set of elementary strokes. These strokes are temporally ordered according to their time of acquisition. Each stroke is defined by a certain number of points bounded by a pen-down and a pen-up points. In this work, we will consider that a pen-up is present at the end of every symbol, which can be written with several strokes. These strokes are not necessarily consecutive, since some strokes can be delayed. The number of points depends on the temporal sampling rate of the digital pen, the speed of writing and of course on the length of the stroke. Mostly, before starting the recognition process itself, the input signal undergoes a preprocessing step [6]. It consists of spatially re-sampling each stroke using a constant rate.

Recognizing a handwritten ME can be achieved thanks to three independent steps [2]. The first step is the *segmentation* process in which the possible groups of strokes are formed. This stage is far from being trivial when as supposed here, interspersed symbols are authorized. Each group is called a segmentation hypothesis (*'hs'*). Ideally, each *'hs'* corresponds to a mathematical symbol. The recognition process is the second step. It aims to assign a symbol label (or a list of possible symbols) and a recognition score for each *'hs'*. The third step is the structural analysis. All the recognized symbols are used to deduce the final ME. This is done through a spatial-grammatical analysis. Optimizing each step alone implies that the failure of one step will lead to the failure of the next one. A solution to reduce this error propagation is proposed in [7]. It consists in the simultaneous optimization of the segmentation and recognition steps. In this case, the classifier is trained separately on isolated symbols. Awal and al. proposed a more global architecture [8]. The strengths of such systems are the following. First of all, the recognition module is trained within the expressions and not longer uses an isolated symbol database. This allows a direct interaction between the different stages of the system (segmentation, recognition and 2D parsing). Secondly, during the segmentation step, a non-consecutive stroke grouping is allowed to form valid symbols. Finally, the structural analysis (2D parsing) is controlled by both symbol recognition scores and a contextual analysis (spatial costs). The ME handwriting recognition sub-part used in our architecture will be largely based on Awal and al.'s system.

B. MER using automatic speech recognition

A MER system based on speech recognition is basically composed of two main modules. The first one achieves the automatic speech transcription task. The output of this module provides a text composed of words written with alphabetic characters as they are recognized by the ASR system. This text is ideally a fair description of the ME (and it depends also on the accuracy of the speaker who speaks out the ME). The second module is a parser, which processes the previous transcription in the 2D space to deduce the associated ME.

The ASR system which is in charge of the first task in the global MER system is quiet similar to the one described in the case of handwriting modality. The main difference is the nature of the signal which is processed (acoustic one in this case). The recognition procedure involves three stages. During the first one, the acoustic signal is filtered and re-sampled, then a frame description is produced, where a feature vector is computed for each window of 25 ms with an overlap of 10 ms. The features are most of the time the cepstral coefficients and their first and second derivatives [9]. Segmentation into homogeneous parts is operated in a second step. Resulting segments are close to minimal linguistic units. The last step is to perform the decoding itself using models learned within a training step (acoustical model, pronunciation dictionary and language model). Parsing the resulting transcription from the previous module is a very hard task. In the rare existing systems [3], [4], the parsing is most of the time assisted by either introducing some dictation rules (for delimitating fraction's numerator and denominator for example) or using an additional source of information (such as using a mouse to point the position where to place the different elements). By adding such constraints, the editing process becomes less natural and far from what is expected from this kind of systems.

In the work reported in this paper, we deal with the French spoken language. The task of speech recognition in our system is carried out by a system largely based on the one developed at the LIUM presented in [9]. This latter is itself based on one of the most popular worldwide speech recognition systems (CMU-Sphinx).

C. Data Fusion

The main goal of this work is to set up a multimodal system dedicated to MER. This idea has emerged from the finding: humans interact with each other by using different interaction modes (speech, handwriting, gesture ...). To make the communication of the human being with machines almost as friendly, it is quite natural to use multiple modes of interaction at the same time to avoid the ambiguities that may arise from one of them [10]. Generally, data fusion methods are divided in three main categories [11], [10]: *early fusion* which happens at features levels; *late fusion* which concerns the intermediate decisions fusion and the last one is the *hybrid fusion* which is a mix of the two. Within each approach, three kinds of methods can be used to carry out the fusion process. Rules based approaches represent the first category, it includes methods using simple operators such as max, (weighted) mean or product. The second category is based on classifiers and the last one is based on parameter estimation.

III. The proposed handwriting and audio information fusion based system for MER

If we refer to the sub-sections II-A and II-B , it is clear that there is an imbalance between the systems based on handwriting recognition and those based on ASR. In fact, systems based on the handwriting modality are getting more mature. In the proposed architecture, the transcription coming out from the speech recognition system is used to disambiguate the result of the handwriting recognition module.

In the case of MER, the fusion methods discussed in sub-section II-C are not all relevant and applicable. Specifically, the heterogeneous nature of the signals of both modalities and their asynchrony prevent from considering an early fusion but led us to favor a late fusion. In addition, the particularity of ME offers the possibility to make on the fusion process at two different levels. In fact the fusion can be done either at the symbol level (during the recognition step) or at relational level (during the structural analysis process). A third alternative is to combine at both levels which seems to be very interesting to get better value out of the fusion process. In a previous work, we tried to check the relevance of the speech-handwriting information fusion working at the level of isolated symbols. We showed the added value of such a procedure since recognition rate was improved with respect to the mono-modality approaches (before fusion we obtained recognition rates of 81% for handwriting and 50% for speech and after fusion, we reached a recognition rate of 98%) [12]. The results obtained in this first study suggest that the fusion may bring even more benefits in the case of complete ME. Indeed, in the case of complete ME, in addition to the confusion related to the recognition of symbols, there is another major problem that is structural analysis (Fig.1). This means that even if the recognition step goes well, it is not guaranteed that the ME will be consistently well recognized. Hence, the aim of the present work is to go one step further by addressing the recognition of a complete ME. To achieve this, we propose a collaborative architecture (see Fig.2) which involves the steps described in the following sections.

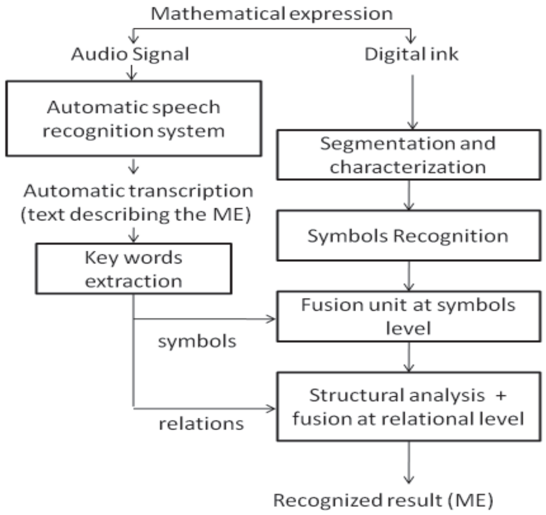
Mathematical expression
Audio Signal
Digital ink
Automatic speech recognition system
Segmentation and characterization
Automatic transcription (text describing the ME)
Symbols Recognition
Key words extraction
Fusion unit at symbols level
symbols
Structural analysis + fusion at relational level
relations
Recognized result (ME)

Figure 2. The collaborative architecture for complete MER

A. Keyword extraction from the audio transcription

The purpose of this step is to analyze the text describing the ME given by the audio system. As a result, two word categories are identified. The first one is composed of words which are useful for the MER process. They spot either symbols (such as: *'x', 'deux', 'parenthèses'*); or relations (*'indice', 'exposant'*); or both (*'intégrale', 'racine'*). The second category of words includes all the other words, they are stopwords used only to make sense from a language point of view. Here, we consider the words from the first category, as *keywords*. A dictionary is built in such a way that each symbol and each relation is associated to one or more keywords. For example if the word *'carré'* (which means squared in French) is present in the transcription, the ME we are processing could contain the symbol *'2'* and the relation *'superscript'*. If any confusion concerning these symbols appears during the handwriting recognition, the fact that they are present in the speech modality increases the confidence about them.

$$\tilde{s}(c_i) = \begin{cases} s(c_i) & \text{if (*) or } s(c_i) > s_{th} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

$$\tilde{s}(c_i) = \begin{cases} \alpha_e \times s(c_i) + \beta_e & \text{if (*)} \\ \alpha_p \times s(c_i) + \beta_p & \text{otherwise.} \end{cases} \quad (2)$$

$$\tilde{s}(c_i) = \begin{cases} \dfrac{1}{1 + e^{-\lambda_e \times s(c_i) + s_e}} & \text{if (*)} \\ \dfrac{1}{1 + e^{-\lambda_p \times s(c_i) + s_p}} & \text{otherwise.} \end{cases} \quad (3)$$

(*): symbol present in both modalities.

B. Information fusion at symbol level (IFSL)

During the recognition process within the handwriting recognition system (*cf.* sub-section II-A), the recognized hypotheses scores are adjusted according to their presence or not in the keywords list extracted previously from the ASR transcription. In other words, we perform a rescoring of the N best list of the recognized symbols that the handwriting classifier provides for each segmentation hypothesis. The fusion rules investigated to control the rescoring are defined by (1), (2) or (3). They aim to increase the scores of symbol hypotheses recognized in both modalities and to decrease those which are missing in the ASR transcription. Let $\{s(c_i), i = 1 : N\}$ be the N best list of scores corresponding to the N most probable classes (symbols) c_i assigned by the handwriting classifier for a given segmentation hypothesis. After the fusion process, the resulting list contains M symbols extracted from the initial list ($M \leq N$) where the class order could be modified as stated before. We denote the score of a class c_i after fusion by $\tilde{s}(c_i)$. The proposed rescoring transformations are of three kinds. Equation (1) defines a thresholding based approach. In this case the symbol is considered for the next step of processing (structural analysis) only if the symbol is present in both modalities or if its score is high enough (higher than a certain threshold s_{th}). The second method of rescoring is based on a linear transformation which is defined by its slope 'α' and an offset 'β' given in (2). In this case one linear transformation for score enhancement (α_e, β_e) and another one for penalization (α_p, β_p) are considered. The third kind of transformation is sigmoidal. It is similar to the previous case where the linear functions are replaced by sigmoidal ones defined by their slopes (λ_e, λ_p) and their centers (s_e, s_p) as described by (3).

C. Information fusion at relation level (IFRL)

In the same way as for the case of the symbol fusion unit, the spatial relation costs (RC) are adjusted. For rescoring the RC, we have considered linear functions, similar to (2). In that case, α_e will be less than one when a relation is found in both modalities, to decrease the cost of this solution. Conversely, α_p is taken more than one to penalize spatial relationships that are not found in the transcription. No thresholds are used in that case, i.e. β_p=β_e=0.

IV. Results and discussion

In order to train and test our proposed system, a multimodal database is required (each ME being available in its audio and handwritten forms). We used the *HAMEX* database which is mainly built for such applications [13]. Concerning the handwriting recognition system, it is the one we used to participate (as an out of competition system) to the first *Competition on Recognition of On-line Handwritten Mathematical Expressions (CROHME[1])* [14]. The vocabulary which is considered contains *56* different symbols (against 74 for the HAMEX database). In this regard, the dataset we consider in this paper is extracted from the HAMEX database taking into account the allowed vocabulary and grammar. Taking these constraints into account, our test dataset contains *519* ME extracted from the whole HAMEX test part (*1425 ME*). In addition to that, *200 ME* from the HAMEX train part satisfying the same conditions are used to tune the different parameters of the proposed system, specifically, the parameters of the rescoring functions. In fact, all the parameters involved in the fusion process (IFSL and IFRL), presented in (1), (2) or (3), are experimentaly optimized on this train database. For the ASR system, we adapted the resources used during the decoding process (prononciation dictionary and language model) using the audio transcriptions of all HAMEX train part.

A. Handwriting based MER results

We report in Table I the performances of the handwriting recognition system. Table I shows that more than 70% of the strokes are properly labeled, while close to 75% of the symbols are retrieved. At the expression level, close to 18% of the expressions are fully correctly interpreted. If we tolerate one (two) errors, either at the symbol level or at the relationship level, the recognition rate rises up to more than 31% (35%). This observation reinforces our previous statement about handwriting ambiguities, and the capacity to improve the baseline results, provided that additional information should be available. The next sub-section report results of such a procedure.

Table I
Performances of the handwriting recognition system

Evaluation level	*strokes*	*symbols*	*expressions with*		
			exact match	*1 error at most*	*2 errors at most*
Reco. rate [%]	70.77	74.45	17.92	31.21	35.84

[1] http://www.isical.ac.in/~crohme/

B. *Fusion based MER results*

The ASR system we used to provide the automatic transcription of the speech signal describing the ME has a word recognition rate of *77%* on the test database (vocabulary = *144* words). In order to estimate the impact of the errors due to the ASR system within the global architecture, we also considered the case of a perfect audio transcription. In other words, we performed an additionnal experiment where the disambiguation is done thanks to the ground-truth transcription. Table II shows the obtained results with different fusion configurations either with the transcription provided by the ASR system (white cells) or the ground-truth (gray cells), compared to the reference system based on handwriting recognition (first cell : no IFSL and no IFRL). We can observe that the recognition rates are improved when a fusion strategy is adopted whatever its configuration. The best fusion configuration is when it is performed at both symbol and relation levels and when rescoring is performed using sigmoidal transformation. In that case, the ME interpretation rate rises from *17.92%* to *23.51%* when the best fusion configuration is applied. Results obtained using the ground-truth transcription and the one given by the ASR system are quite similar. This is due to the fact that the ASR system performs well concerning the recognition of the keywords. Within the total vocabulary (144 words) encountered within the test database, *83* are keywords. The recognition rate on them is *90.06%*. However, it is worth to note that if the fusion process improves the global recognition rate, some ME which are initially well recognized and are no longer valid. Table III shows the gains and losses due to the fusion process compared to the reference system. Let H_j(*j=0, 1 or 2*) be the number of ME containing *j* errors within the Handwriting system and $F_{i/j}$ be the number of ME containing *i* errors within the fusion based system among the H_j ME. The cell located at line *i* and column *j* gives the ratio $\frac{F_{i/j}}{H_j} \times 100$. In other words, it gives the proportion of ME recognized with *i* errors after fusion and which are recognized with *j* errors before fusion. The first cell shows a minor loss in term of totally recognized expressions (around *6%* of the initially well recognized ME are not well recognized after the fusion process). Among these lost ME, half are lost because of one or two errors. These ME contain relationships which are not well expressed during the dictation which makes them more confusing. In the other hand a lot of ME with one or two errors during the handwriting recognition process are completely recognized thanks to the fusion strategy and the contribution of speech (*27.54%* and *10.53%* respectivelly). In general, Table III shows that the losses due to the fusion process are very low compared to the gains provided by this later.

Table II
Recognition rates at the expression level of different fusion methods

IFSL using	*no IFSL*	*thresholding (1)*	*linear fct (2)*	*sigmoidal fct (3)*
Reco. rate no IFRL [%]	17.92	21.23	21.00	23.70
		21.19	20.04	22.93
Reco. rate with IFRL [%]	20.04	21.59	22.16	24.47
	20.04	21.38	21.77	23.51

Table III
Gains and losses in term of ME due to the fusion process

Gains and losses in [%]		*without fusion*			
		no errors	*1 err. allowed*	*2 err. allowed*	*3 err. or more*
with fusion	*no errors*	93.55	27.54	10.53	4.49
	1 err. allowed	2.15	60.87	18.42	3.85
	2 err. allowed	1.07	1.45	47.37	3.53
	3 err. or more	3.23	10.14	23.68	88.14

In Fig.3, we show an example of a beneficial collaboration between speech and handwriting. While the handwriting recognition system fails to provide the right solution, the speech description, by giving the keyword *'de'* increases the the probability of presence of the parentheses. And by missing the symbol *'un'* penalizes the solution containing this later. This leads the recognition to end well.

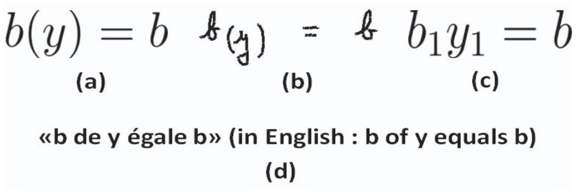

Figure 3. Real example of a contribution of the bimodal processing; (a) the ground-truth ME, (b) its handwritten version, (c) recognized result without fusion, (d) the automatic transcription of its description

V. Conclusion and perspectives

We investigated in this paper a new approach to improve the MER based on bimodal processing. We considered a primary system achieving the recognition of handwritten ME, assisted by an ASR system performing the speech recognition of the ME description provided by the user. As expected, the added value of such a processing, namely bimodal processing, is observable at both symbols and relationnal levels (*cf.* table II). This observation supports the hypothesis of the existing complementarity between the two modalities. Thanks to this processing, we increase the recognition rate from 17.92% to 23.51% corresponding to a relative gain of around *31%*.

In the light of the obtained results from this first experiment, we believe that this kind of solution is very interesting for bidimensional language processing such as ME. Thus, we plan in future work to go deeper in the definition of the language model that is attached to every recognized ME based on the transcription of the spoken ME. Instead of considering only unit at the word level, it should be interesting to work at a n-gram level to leverage the context of each uttered word.

Acknowledgment

This work is supported by the French Region Pays de la Loire under the DEPART project http://www.projet-depart.org/.

References

[1] F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab, "Human-Computer Interaction: Overview on State of the Art," *IJSSIS*, pp. 137–159, 2008.

[2] K.-F. Chan and D.-Y. Yeung, "Mathematical expression recognition: A survey," *IJDAR*, pp. 3–15, 2000.

[3] R. Fateman, "How can we speak math?" University of California at Berkeley, Tech. Rep., 2011.

[4] A. Wigmore, G. Hunter, E. Pflugel, J. Denholm-Price, and V. Binelli, "Using automatic speech recognition to dictate mathematical expressions: The development of the talkmaths application at kingston university." *JCMST*, pp. 177–189, 2009.

[5] C. Elliott and J. Bilmes, "Computer based mathematics using continuous speech recognition," in *CHI*, 2007.

[6] E. Tapia and R. Rojas, "A survey on recognition of online handwritten mathematical notation," Free University of Berlin, Tech. Rep., 2007.

[7] T. H. Rhee and J. H. Kim, "Robust recognition of handwritten mathematical expressions using search-based structure analysis," in *ICFHR*, 2008, pp. 19–24.

[8] A.-M. Awal, H. Mouchère, and C. Viard-Gaudin, "Towards handwritten mathematical expression recognition," in *ICDAR*, 2009, pp. 1046 –1050.

[9] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The lium speech transcription system: a cmu sphinx lll-based system for french broadcast news," in *Interspeech'05, ISCA*, 2005.

[10] J.-P. Thiran, F. Marqués, and H. Bourlard, *Multimodal Signal Processing - Theory and Applications for Human-Computer Interaction*. Elsevier, 2010.

[11] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, pp. 345–379, 2010.

[12] S. Medjkoune, H. Mouchère, S. Petitrenaud, and C. Viard-Gaudin, "Handwritten and audio information fusion for mathematical symbol recognition," in *ICDAR*, 2011, pp. 379–383.

[13] S. Quiniou, H. Mouchère, S. P. Saldarriaga, C. Viard-Gaudin, E. Morin, S. Petitrenaud, and S. Medjkoune, "Hamex - a handwritten and audio dataset of mathematical expressions," in *ICDAR*, 2011, pp. 452–456.

[14] H. Mouchère, C. Viard-Gaudin, D. H. Kim, J. H. Kim, and U. Garain, "Crohme2011: Competition on recognition of online handwritten mathematical expressions," in *ICDAR*, 2011, pp. 1497–1500.