Handwritten and Audio Information Fusion for Mathematical Symbol Recognition

Sofiane Medjkoune*, Harold Mouchère*, Simon Petitrenaud**, and Christian Viard-Gaudin*
* LUNAM University, University of Nantes, IRCCyN/IVC Lab., France
** LUNAM University, University of Le Mans, LIUM Lab., France
{sofiane.medjkoune, harold.mouchere, christian.viard-gaudin}@univ-nantes.fr, simon.petit-renaud@lium.univ-lemans.fr

***Abstract*—Considerable efforts are being done within the scientific community to make as easier as possible the way that the human being converses with its machine. Handwriting and speech are two common ways used to achieve this goal and are probably among those which attracted much interest. In mathematical content recognition tasks, these two modalities are used with a certain success. This paper presents an architecture based on a speech-handwriting data fusion for isolated mathematical symbol recognition. Different fusion methods are explored. The results are very encouraging since recognition rates are increased comparatively to mono modality approaches.**

Keywords- handwriting recognition; speech recognition; mathematical expressions; data fusion

I. INTRODUCTION

Speech and writing are two natural ways of communication used by humans for a long time. Both have been used with varying success in human computer interaction. In this regard, systems for handwriting recognition [1] and voice recognition [2] have emerged and have continued to generate more interest in various fields.

Particularly, in the case of mathematical equation writing, it is easier to insert expressions, with different levels of complexity, by hand or simply by dictating them than to use a keyboard and/or mouse and specialized editors like Latex or MathML. Although considerable efforts have been made to make keyboard-mouse oriented tools more user friendly, they still are quite cumbersome and time consuming in addition to the fact that they require learning a new language and writing rules.

In the case of handwriting recognition, many studies have addressed this issue [3] and the results obtained are very encouraging. However, these systems are, of course, not hundred percent reliable. The resulting errors are often due to inter symbol confusions and spacial relation ambiguities (due to the bi-dimensional nature of mathematical writing) which the writing modality alone cannot fix. Fig. 1 shows an example of this difficulty. If the three basic inter symbol relations (subscript, left/right and superscript) are well defined, at their frontiers, there still exists an ambiguity in decision making.

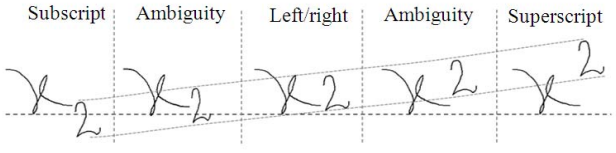

Figure 1: Ambiguity related to the spatial relationship between symbols

Concerning the speech modality, recognition systems dedicated to mathematical equations recognition also exist [4]. These tools are very helpful, however, they have their limitations since their reliability depends on the accuracy of speaker in the description of the expression. More precision he gives better the recognition result is. Also, some pauses in the speech are necessary to make the recognition accurate [4]. This is not the case in practice, during teaching for example, the lecturer does not systematically dictate all the symbols present in the expression. Fig. 2 illustrates this situation, where different possible mathematical interpretations for an audio record are given. Since all interpretations are equally likely, it is very easy to make a wrong interpretation of what is said.

As explained before, each one of the two modalities encounters problems which are of different kinds.

This suggests that setting up a system that ensures the benefits of the two modalities would allow for overcoming these drawbacks and improve the accuracy.

An attempt to use this is presented in [5], where S. Vemulapalli and al. have used audio information to disambiguate and improve the handwritten mathematical content recognition in classroom videos.

In this paper, we focus on contribution of combining an on-line handwriting recognition system and speech recognition one for isolated mathematical symbols through a fusion unit. We investigate the problem in the case of late fusion since the data coming in from both modalities is heterogeneous. In addition, available recognition systems treating each modality separately before merging can be used.

Within the next section an overview of the explored fusion approaches is given. Then, in section III, descriptions of the systems used to perform the tasks of handwriting and speech recognition are provided. In section IV, the experimental protocol is presented. Section V is devoted to preliminary results followed by analysis and discussions. The last section concludes the paper and gives perspectives of this work.

$$\frac{x^{\frac{2\eta-3}{2}+y}}{2};\ \frac{x^{\frac{2\eta}{2}-3+y}}{2};\ \frac{x^{\frac{2\eta-3+y}{2}}}{2};\ x^{\frac{2\eta-3}{2}+\frac{y}{2}};\ x^{\frac{2\eta}{2}-\frac{3+y}{2}};\ \ldots$$

Figure 2: Different possible mathematical interpretations of an audio transcription defined by: "*x to the power of two multiplied by eta minus three plus y over two*"

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.84

II. INFORMATION FUSION

The aim of information fusion is to combine information from multiple or unique sources to get better results on a given task. An example of such a procedure is the combination of multiple classifiers for handwriting recognition to improve the recognition rate [6].

The fusion process can take place at different levels [7]; principally, either at feature level (early fusion) or at decision level (late fusion). In the early fusion case, a single decision system is used. The decision is made considering all extracted features at the same time. The main advantage of this kind of architectures is that only one learning step is required. However, it needs to synchronize correctly the streams coming from all the sources, which is not always straightforward. In the case of late fusion, the problem of data normalization at the input of the fusion system is less difficult than in the case of features fusion but still exist. More than this, fusing at this level allows using suitable expert systems for each modality ensuring at the beginning the best possible results from each of them taken separately. The main disadvantage of this fusion method is that due to the fact that each modality is processed separately, the resulting architecture is not optimal. A hybrid approach also exists. In this case, feature fusion is exploited as well as the decision one, the goal of this approach is to benefit from the advantages of both early and late fusion. Fig. 3 illustrates these different levels of combination.

From another point of view, considering the method used to fuse, according to [6], we can globally distinguish three main categories:

1. Rule based systems: the goal of this kind of systems consists in using basic rules such as weighted summation or product, majority voting and so on.

2. Classification based systems: classification techniques like neural networks, support vector machines, dynamic Bayesian networks and others are used to achieve the fusion process.

3. Estimation based methods: The Kalman filter, the extended Kalman filter and particle filter are common ways to make information fusion.

In the current work, we are interested on late fusion since we are dealing with heterogeneous data. This implies that before making the fusion process, two upstream systems provide decisions corresponding to each modality. These systems accomplish the handwriting recognition and voice recognition tasks respectively.

In the following, specialized systems are described.

III. SYSTEMS DESCRIPTION

A. The on-line handwriting recognition system

The on-line handwriting recognition is performed by a recognizer similar to the one described in [8]. It is globally based on an artificial neural network. More specifically, a time delay neural network (TDNN) is used

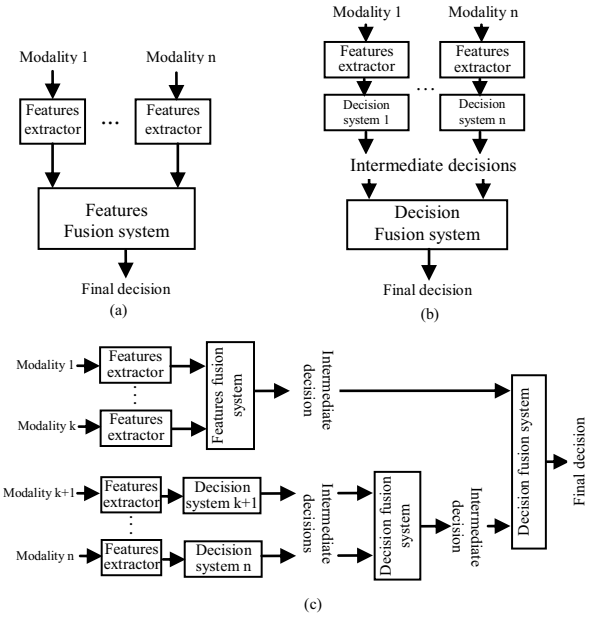

Figure 3: Different levels of fusion. (a) Features fusion level; (b) decision fusion level; (c) hybrid fusion.

since this kind of classifiers is well adapted to the sequential nature of on-line handwriting. The features used are those presented by Awal and al. in [8], they are basically the positions and their first and second derivatives sampled at a given number of points along the pen trajectory. The developed system allows not only giving the best candidate but also a list of N best hypotheses with their respective scores, where scores are normalized between 0 and 1. In [8], the normalization is performed using a softmax function.

B. The isolated word speech recognition system

Referring to the literature, many approaches are proposed to build a speech recognition system. They differ from each other either by the method used like those based on Hidden Markov Models (HMMs) or those based on template matching using Dynamic Time Warping (DTW), or by the features used such as Mel Filtered Cepstrum Coefficients (MFCC), Linear Predictive Coefficients (LPC) and so on [9].

Recognition systems based on MFCC features and HMM are known to be very reliable compared to others. However, in [9], it is reported that in the context of isolated words and small vocabulary, systems based on DTW matching are as accurate as those based on HMMs.

Since we are in the case of small vocabulary and isolated word recognition, the system in charge of the speech recognition module in our architecture is MFCC and DTW based, as described in [10].

Fig. 4 shows the global architecture of this system:

During the signal characterization, the speech signal is first filtered using the voice activity detection algorithm (VAD) [11] in order to detect the useful part of the signal,

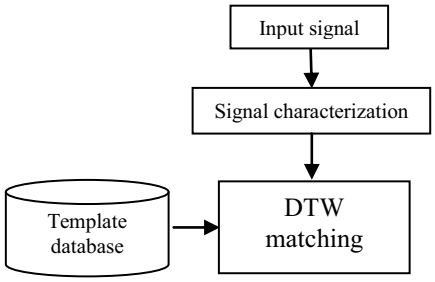

Figure 4: Global architecture of the speech recognition system

followed by spectral subtraction [12] to remove remaining noise. After that, the signal is divided in frames and MFCC coefficients are calculated for each one. The features of speech signals discrimination correspond to the 13 first MFCC coefficients and their first and second derivatives. A total of 39 features per frame is considered.

The learning template database contains features matrix corresponding to each word of the vocabulary. Each word is pronounced by several speakers.

To classify an input signal, first, its features are calculated by the signal characterization unit. Then, they are matched to all those stored in the template database. Finally, a K nearest neighbor algorithm (KNN) is used to assign a label for the spoken word.

The number of neighbors K in the KNN is fixed using the validation database. During the application of the KNN rule, if confusion appears, at most, two other neighbors are added. If confusion still exists, the class of the nearest neighbor is assigned to the signal to recognize [13].

Similarly to the handwriting case, the speech recognition system also provides, for each utterance to recognize, an N-best list of probable classes with their scores. Scores in this case also are in the range 0 to 1. They are calculated as the posterior class probabilities from the estimation of probability density functions obtained trough the K nearest neighbors and the DTW distance [14].

IV. THE FUSION APPROACHES

In this paper, we investigate six approaches of fusion, five of them concern rule based fusion methods, and the sixth one is about using a classification based fusion method. First, let us give some useful notations for the following.

Let $\boldsymbol{C} = \{c_1, c_2 \ldots c_S\}$ be the set of the S possible output classes for a new input symbol x to classify. Then, the score assigned to the hypothesis that x belongs to the class c_j for j=1…S, considering the modality i is $d_{i,j}(x)$. And we denote $d_j(x)$, the score of class c_j for x after the fusion process. Finally, x is assigned to the class c_j with the highest score $d_j(x)$.

Now, we focus on the ways these scores are obtained. In the following brief descriptions of the six used approaches are given.

A. Weighted summation

In this case, the score for each class c_j is the weighted summation of scores for that class coming from each system [7]. For M modalities this can be expressed by (1).

$$d_j(x) = \sum_{i=1}^{M} w_{i,j} d_{i,j}(x), \tag{1}$$

where, $w_{i,j}$ is the weight assigned to the score of recognition of class c_j from modality i. M is the number of considered modalities.

For the case of speech and handwriting information fusion, M is equal to 2. In this case, we replace i by h for the handwriting modality and by s for speech modality.

We used three different manners to weight:

- **By simple mean**, we trust each of the two systems in the same way, in this case: $w_{h,j} = w_{s,j} = 0.5$, for $j=1:S$. Here, no parameter has to be considered.
- **Using the global recognition rate of each system**, if R_h and R_s are respectively the global handwritten recognition rate and the global speech recognition rate, then weights considered are identical for each class (two different parameters are considered). Equation (2) gives the corresponding weights.

$$\begin{cases} w_{h,j} = \dfrac{R_h}{R_h + R_s} \\ w_{s,j} = \dfrac{R_s}{R_h + R_s} \end{cases} \tag{2}$$

- **Using the local recognition rate of each class and each system**, if $R_{h,j}$ and $R_{s,j}$ are respectively the handwritten recognition rate and the speech recognition rate corresponding to the class c_j, then:

$$\begin{cases} w_{h,j} = \dfrac{R_{h,j}}{R_{h,j} + R_{s,j}} \\ w_{s,j} = \dfrac{R_{s,j}}{R_{h,j} + R_{s,j}} \end{cases} \tag{3}$$

In this formulation, weights depend on the performance of each system for each class; so, two parameters per class are considered to adjust weights.

B. Max decision fusion [7]

The score after fusion that a given input belongs to a given class is simply the highest one of all scores coming from all the modalities for this class. In the case of bi-modality (speech and handwriting), this rule gives the final score as in (4):

$$d_j(x) = \max\left(d_{h,j}(x), d_{s,j}(x)\right), \tag{4}$$

No weight is used in this case.

C. Product

Using directly geometrical mean will penalize cases where one system makes a good decision with high score and the other one fails in this task (very low score). This is why the combination used is as described in [6], like in (5).

$$d_j(x) = 1 - ((1 - d_{h,j}(x))(1 - d_{s,j}(x))), \quad (5)$$

As in the previous case, each input is processed separately.

D. Fusion classification based

A support vector machine classifier (SVM) with a gaussian kernel is used to perform this task. We use the scores from each of the upstream systems to train an SVM classifier. For a given input of both specialized systems, if this one is chosen among *S* classes, then 2x*S* features are considered. These scores are those assigned by each specialized classifier for each class. The output is consequently one of the *S* classes, corresponding to the best result after information fusion. The train and validation databases are used to tune the parameters of the SVM (the kernel standard deviation 'σ' and the trade-off between minimizing training errors and controlling the model complexity 'C') to the optimum ones. After that, performances are evaluated on the test database. In the next section, we summarize the results obtained and give a discussion about them.

V. Experimental Results and Discussion

A. Data

This paper reports results on 74 mathematical isolated symbols. The database includes all the Latin alphabet letters, the ten digits, six letters from the Greek alphabet and the remaining classes are various mathematical symbols (integral, summation…).

The handwritten data come from CIEL database [16], where each symbol (74) is written by 279 writers.

The speech data come from HAMEX database and concern the same symbol set uttered each one by 31 various speakers (5 women and 26 men) [17].

The data of both modalities is first divided in three sub-databases. The first part is the train database, which serves to train the systems (for the mono modality classifiers or for the fusion one); the second one is the validation database, used to validate systems configurations, and the last one represents the test database. Table 1, shows the data repartition and resulting fusion possibilities in such a way that only data from the two modalities that represent the same information are fused.

Table 1: Data repartition

	# samples train	# samples validation	# samples test	# samples all
Handwriting data	8982	5323	6341	20646
Speech data	1022	515	757	2294
Fusion data	123362	38838	62203	224403

All the results reported in this paper are from experiments carried out on the databases introduced above.

In Table 2, performances of the specialized systems (handwriting and speech recognition systems) on their respective test databases are reported. The recognition systems are writer/speaker independent.

Table 2 : Specialized systems performances on the test database

	Speech recognition system	Handwriting recognition system
Recognition rate (%)	50.09	81.55

B. Results

In the following results, fusion results that will be presented use the N-best lists of both previous specialized systems. Parameter N is fixed to 5 using the validation database, in order to only consider hypotheses with significant scores.

In Fig. 5, we can see the results obtained after data fusion. The initial recognition rates are represented in dashed lines (the lowest one at 50.09% is the speech recognition rate and the other one corresponds to handwriting recognition rate 81.55%).

From left to right, bars give the recognition rates after fusion, for simple mean fusion, mean weighted by global recognition rates of the specialized systems, mean weighted by recognition rates of each class for each specialized recognition system, max rule fusion, product fusion and the rightmost corresponds to fusion classification based using SVM classifier. These results are significantly different from each other and different from the initial recognition rates (mono modality mode) according to the classical two-tailed test for difference between two proportions, with a significance level of 5%.

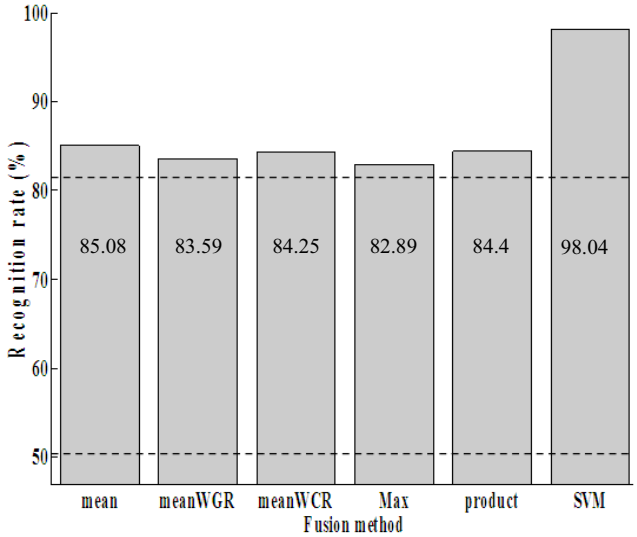

Figure 5: Recognition rates after fusion

As we can see, all the chosen fusion strategies increase the global recognition rates even slightly (the smallest recognition rate is 82.89%, which is higher than the highest recognition rate in the mono modality mode, 81.55%). The classification method based on SVM appears to be the most relevant in this case (recognition rate of 98.04%). This can be due to the difference in the dynamic of scores coming from the two systems. In other words, the difference in confidence for two hypotheses is not expressed in the same way in both systems (speech and handwriting) in term of difference in scores. In fact, in the case of fusion methods trusting the two systems in the same way (simple mean, product and max), each fusion state is considered alone and there is no weight which can compensate the difference in dynamic of scores of the hypotheses classes for a given input. Fusion methods using weights calculated from the performances of the specialized systems, either in global way or by considering each class separately, do not fix the problem and seem to be more sensitive to this difference in dynamic of scores even if these weights imply that we trust more one system comparatively to the other depending on the difference in the performances. In the case of classification based fusion, the weights are optimized during the learning step. This means that the performances of the specialized systems and the problem of the dynamic of the scores are implicitly taken into account since the weights are adjusted over all the existing examples in the training database. This makes them more significant.

VI. CONCLUSION AND PERSPECTIVES

In this paper, we have investigated the feasibility and contribution of bimodal information processing for isolated mathematical symbols recognition. This study has clearly shown that it is advantageous to use the bimodality aspect of information since this increases the global performance with respect to a single modality.

Associate audio stream and handwritten stream, may be even more interesting in the case of complete mathematical expressions. In fact, the bi-dimensional nature of mathematical expressions, make them more ambiguous than isolated symbols. In addition to the inter-symbols confusion, another source of confusion exists. It is bound to the spacial relations between symbols present in the expression. Indeed, since all symbol positions are allowed, the boundaries between the different positions are fuzzy, even for human in some cases. This implies that it is easy to make a mistake in judgment.

Starting from this statement, we plan in future work to extend these fusion strategies for the case of complete expressions. To do this, we have first to set up a system for continuous speech recognition well adapted to the case of mathematical language (adapt existing systems such as CMU Sphinx speech recognition toolkit [17]). Then, for the handwriting mathematical expression recognition issue, we can use that one described in [8], which exhibits a good behavior in terms of mathematical expression recognition.

ACKNOWLEDGEMENT

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

REFFERENCES

[1] R. Plamondon, and S. N. Srihari, “On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey”, IEEE PAMI, 22(1), pp. 63-84, January 2000.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D.Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi and C. Wellekens, “Automatic speech recognition and speech variability: a review”, Speech Commun.49, pp. 763– 786, 2007.

[3] K. F. Chan et D. Y. Yeung, “Mathematical expression recognition: a survey”, International Journal on Document Analysis and Recognition 3, no. 1: 3–15, 2000.

[4] C. Elliot et J. A Bilmes, “Computer Based Mathematics using Continuous Speech Recognition”, In Striking a Chord: Vocal Interaction in Assistive Technologies, Games, and More: CHI workshop on non-verbal acoustic interaction, San Jose, CA, April 2007.

[5] S Vemulapalli and M. H. Hayes III, “Using Audio Based Disambiguation for Improving Handwritten Mathematical Content Recognition in Classroom Videos”, DAS, 2010.

[6] A. F. R. Rahman and M. C. Fairhurst, “Multiple classifier decision combination strategies for character recognition: A review”, International Journal on Document Analysis and Recognition 5, no. 4: 166-194, 2003.

[7] Pradeep K. Atrey and al., “Multimodal fusion for multimedia analysis: a survey”, Multimedia Systems 16, no. 6: 345-379, 2010.

[8] A. M Awal, H. Mouchere, et C. Viard-Gaudin, “Towards handwritten mathematical expression recognition”, 10th International Conference on Document Analysis and Recognition, 1046–1050, 2009.

[9] L. Rabiner and B. H. Juang, “Fundamentals of Speech Recognition”, Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 1993.

[10] L. Muda, M. Begam, and I. Elamvazuthi, “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, journal of computing, volume 2, issue 3, 2010.

[11] L. R. Rabiner and M. R. Sambur, “Speaker Independent Recognition of Connected Digits”, Conference Record IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 202-205, April 1976.

[12] S. F. Boll, “Suppression of Acoustic Noise in Speech using Spectral Subtraction”, IEEE Trans. on ASSP, vol. 27(2), pp.113-120, 1979.

[13] A. Cornuéjols and L. Miclet, “Apprentissage Artificiel: concepts et algorithmes”, Eyrolles, 2002.

[14] Christopher M. Bishop, “Pattern Recognition and Machine Learning”, Springer-Verlag New York, Inc., Secaucus, NJ, USA, pp. 123-127, 2006.

[15] A. M. Awal, “Reconnaissance des structures bidimensionnelles: Application aux expressions mathématiques en ligne”, PHD thesis of the university of Nantes, 2010.

[16] S. Quiniou, H. Mouchere, S. Peña Saldarriaga, C. Viard-Gaudin, E. Morin, S. Petitrenaud and S. Medjkoune, “HAMEX – a Handwritten and Audio Dataset of Mathematical Expressions”, ICDAR 2011.

[17] http://www.speech.cs.cmu.edu.