2011 International Conference on Document Analysis and Recognition

Dempster-Shafer based rejection strategy for handwritten word recognition

Thomas Burger
Université de Bretagne-Sud, CNRS, Lab-STICC
F-56017 Vannes cedex, France
thomas.burger@univ-ubs.fr

Yousri Kessentini, Thierry Paquet
Université de Rouen, Laboratoire LITIS EA 4108
Site du Madrillet, St Etienne du Rouvray, France
{*yousri.kessentini,thierry.paquet*}*@univ-rouen.fr*

***Abstract*—In this paper, a novel rejection strategy is proposed to optimize the reliability of a handwritten word recognition system. The proposed approach is based on several steps. First, we combine the outputs of several HMM classifiers using the Dempster-Shafer theory (DST). Then, we take advantage of the expressivity of mass functions (the counter part of probability distributions in DST) to characterize the quality/reliability of the classification. Finally, we use this characterization to decide whether a test word is rejected or not. Experiments carried out on RIMES and IFN/ENIT datasets show that the proposed approach outperforms other state-of-the-art rejection methods.**

***Keywords*-Dempster-Shafer theory; Data fusion; Rejection strategy; Handwriting recognition**

I. Introduction

After about forty years of research in off-line handwriting recognition, the performances of current systems are still insufficient, as for many applications, more robust recognition is required. Multiple classifier combination has been intensively studied with the aim of overcoming the limitations of individual classifiers [1], [2], [3]. Most of these research works stress the real interest of the Dempster-Shafer Theory (DST) [4], [5] to combine classifiers in a manner which is both accurate and robust to difficult conditions (set of weak classifiers, degenerated training phase, overly specific training sets, large vocabulary, etc.). In this context, we have shown in previous works of ours [6], [7], that ensemble classification methods based on DST outperform the classical combination methods, as they provide higher recognition rates. However, in the overall recognition process, high recognition rates is not the only measure to characterize the quality of a recognition system. For practical applications, it is also important to look at reliability. Rejection strategies are able to improve the reliability of handwriting recognition systems. Contrarily to classifier combination, rejection strategies do not increase the recognition rate but, at least, reduce the number of errors and suggests an alternative treatment of the rejected samples [8], [9], [10]. The rejection strategies are typically based on a confidence measure. If the confidence measure exceeds a specific threshold, the recognition result is accepted. Otherwise, it is rejected. Generally, this rejection may occur as 1) more than one word appears adequate; 2) no word appears adequate.

In [10], varieties of rejection thresholds including global, class-dependent and hypothesis-dependent thresholds are proposed to improve the reliability in recognizing unconstrained handwritten words. In [9], the authors present several confidence measures and a neural network to either accept or reject word hypothesis lists for the recognition of courtesy check amounts. In [11], a general methodology for detecting and reducing the errors in a handwriting recognition task is proposed. The methodology is based on confidence modeling and its main originality is the use of two parallel classifiers for error assessment. In [12], the authors propose multiple rejection thresholds to verify handwritten word recognized hypotheses. To tune these rejection thresholds, an algorithm based on dynamic programming is proposed. It focuses on maximizing the recognition rate for a given prefixed error rate.

In this paper, we propose a new rejection strategy based on the Dempster-Shafer theory. In fact, mass functions (the central object of DST) are more complex objects than discrete probabilities, which allow for a richer description of the knowledge they encode. Thus, our aim is to exploit this additional information to derive some measures adapted to rejection strategies. More precisely, we use the DST to improve the recognition rate of a classification process, by combining several probabilistic classifiers (HMM classifiers) within the formalism of DST. The result of the combination being expressed as a mass function, we aim at using the extra available information to derive an efficient rejection strategy, and thus, improving the reliability of the recognition.

The paper is organized as follows: in section 2, we present some classical rates for the evaluation of rejection strategies, a background review on the basis of the Dempster-Shafer Theory and we recall the different steps of the DST-based ensemble classification method that we have presented in a previous work. Section 3 addresses in detail the proposed rejection strategies. In section 4, we evaluate the performance of the proposed approach. The conclusions of this paper are presented in the last section.

II. Background

In this section, we first recall the classical rates involved in the evaluation of a rejection strategy. Then, we present

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.112

the Dempster-Shafer Theory. Finally, we recall a previous work of ours on ensemble classification.

A. Evaluation of rejection strategies

Let us consider a testing set of N_{test} words. We have:

$$\begin{aligned} N_{test} &= \overbrace{N_{rec} + N_{err}}^{N_{proc}} + \underbrace{N_{rejhit} + N_{rejmiss}}_{N_{rej}} \\ &= N_{hit} + N_{mis} \end{aligned}$$

where N_{rec} is the number of correctly classified words, N_{err} is the number of incorrectly classified words, and N_{rej} is the number of words which are not classified, as they have been rejected. The latter are divided into N_{rejhit}, the number of words that would have been correctly classified if not rejected, and $N_{rejmiss}$, the number of words that would have been misclassified if processed. Finally, N_{proc} is the number of words which have been processed (i.e. not rejected), and N_{hit} and N_{mis} corresponds to the number of words that would have been respectively correctly and incorrectly classified in case of absence of rejection strategies. Then, the following rates are classically defined:

$$\begin{aligned} \text{Recognition Rate} &= \frac{N_{rec}}{N_{test}} \\ \text{Error Rate} &= \frac{N_{err}}{N_{test}} \\ \text{Rejection Rate} &= \frac{N_{rej}}{N_{test}} = \frac{N_{rej}}{N_{rej} + N_{proc}} \\ \text{Reliability} &= \frac{N_{rec}}{N_{proc}} = \frac{\text{Recognition Rate}}{1 - \text{Rejection Rate}} \\ \text{True Rejection Rate} &= \frac{N_{rejmis}}{N_{mis}} \\ \text{False Rejection Rate} &= \frac{N_{rejhit}}{N_{hit}} \end{aligned}$$

B. Dempster-Shafer theory

Let $\Omega = \{\omega_1, ..., \omega_K\}$ be a finite set, called the **frame**, or the **state-space**, made of exclusive and exhaustive classes (for instance, the words of a lexicon). A **mass function** m is defined on the powerset of Ω, noted $\mathcal{P}(\Omega)$ and it maps onto $[0, 1]$ so that $\sum_{A \subseteq \Omega} m(A) = 1$ and $m(\emptyset) = 0$. Then, a mass function is roughly a probability function defined on $\mathcal{P}(\Omega)$ rather than on Ω. Of course, it provides a richer description, as the support of the function is greater: If $|\Omega|$ is the cardinality of Ω, then $\mathcal{P}(\Omega)$ contains $2^{|\Omega|}$ elements.

It is possible to define several other functions which are equivalent to m by the use of sums or Möbius inversions. The belief function bel is defined by:

$$bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B), \qquad \forall A \subseteq \Omega \tag{1}$$

Roughly, $bel(A)$ corresponds to the probability of all the evidence which implies A. Thus, it corresponds to the lower bound of the subjective probabilities which are consistent with the available evidence. Dually, the plausibility function pl is defined by :

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \forall A \subseteq \Omega \tag{2}$$

It corresponds to a probabilistic upper bound (all the items of evidence which do not contradict A). Consequently, $pl(A) - bel(A)$ measures the **imprecision** associated to the subset A of Ω.

A subset $F \subseteq \Omega$ such that $m(F) > 0$ is called a **focal element** of m. If the c focal elements of m are nested ($F_1 \subseteq F_2 \subseteq \ldots \subseteq F_c$), m is said to be **consonant**.

Two mass functions m_1 and m_2, based on the evidence of two independent and reliable sources, can be combined into a new mass function by the use of the **conjunctive combination**, noted $\textcircled{\cap}$. It is defined $\forall A \subseteq \Omega$ as:

$$[m_1 \textcircled{\cap} m_2](A) = \frac{1}{1 - \mathcal{K}_{12}} \sum_{B \cap C = A} m_1(B) \cdot m_2(C) \tag{3}$$

where $\mathcal{K}_{12} = \sum_{B \cap C = \emptyset} m_1(B) \cdot m_2(C)$ measures the **conflict** between m_1 and m_2. $\mathcal{K}_{12}$ is called the **mass of conflict**.

The most classical way to convert a mass function onto a probability (for instance, to make a decision), is to use the pignistic transform [5]. Intuitively, it is based on the idea that the imprecision encoded in the mass function should be shared equally among the possible outcomes, as there is no reason to promote one of them rather than the others. If $|A|$ is the cardinality of the subset $A \subseteq \Omega$, the **pignistic probability** $\overline{m}$ of m is defined as:

$$\overline{m}(\omega_i) = \sum_{A \ni \omega_i} \frac{m(A)}{|A|} \qquad \forall \omega_i \in \Omega \tag{4}$$

Dually, it is possible to convert a probability distribution onto a mass function. The **inverse pignistic transform** [13] converts an initial probability distribution p into a consonant mass function. The resulting consonant mass function, denoted by $\widehat{p}$, is built as follows: First, the elements of Ω are ranked by decreasing probabilities such that $p(\omega_1) \geq \ldots \geq p(\omega_{|\Omega|})$. Second, we define $\widehat{p}$ as:

$$\begin{aligned} \widehat{p}\left(\{\omega_1, \omega_2, \ldots, \omega_{|\Omega|}\}\right) = \widehat{p}(\Omega) &= |\Omega| \times p(\omega_{|\Omega|}) \\ \forall\ i < |\Omega|,\ \widehat{p}(\{\omega_1, \omega_2, \ldots, \omega_i\}) &= i \times [p(\omega_i) - p(\omega_{i+1})] \\ \widehat{p}(.) &= 0 \qquad \text{otherwise.} \end{aligned} \tag{5}$$

It is possible to take into account the reliability of a source of information by discounting it. The **simple discounting** ${}^{\alpha}m$ of m is defined as:

$$\begin{aligned} {}^{\alpha}m(A) &= (1 - \alpha) \cdot m(A), \forall A \subset \Omega \\ {}^{\alpha}m(\Omega) &= (1 - \alpha) \cdot m(\Omega) + \alpha \end{aligned} \tag{6}$$

Given a mass function m, it is possible to compute its pignistic transform $\overline{m}$, which is a probability distribution, then, to apply the inverse pignistic transform, to compute $\widehat{\overline{m}}$, which is a consonant mass function having the same pignistic transform as m. Practically, the interest of computing $\widehat{\overline{m}}$ from m has been recently shown in [14]. As a matter of fact, the corresponding operation is interesting to discount a source of information (as an alternative to the simple discounting), and it has been named **pignistic discounting**.

C. DST-based ensemble combination method

Here, we summarize previous works of ours to derive an efficient ensemble classification technique based on the use of DST [6], [7]. We dispose of three HMM classifiers, each working on different feature sets: upper contour, lower contour and density. Our aim is to combine the outputs of these HMM classifiers in the best way.

To do so, we apply the following procedure: The first step consists of defining the frame Ω. In the case of handwritten word recognition, the set of classes (lexicon) is of a very high size with respect to the cardinality of the state space in classical DST problems: Practically, $\mathcal{P}(\Omega)$ contains $2^{|\Omega|}-1$ elements, which is intractable for a large set of classes. To face this computational issue, the state-space is dynamically defined according to the length of the list provided by each classifier. See [6] for more details on the dynamic definition of the state-space. Second, for each of the three classifiers, we normalize the log-likelihood distribution it provides, by using a sigmoid function, such as described in [7]. As a result, we have three sets of scores, which sum up to one over the set of classes. Thus they behave as three probability distributions over Ω. Third, a mass function is derived from each of the three probability distributions, by use of the inverse pignistic transform. Fourth, the Recognition Rates of the classifiers (derived from a cross-validation procedure) are used to weight each mass function according to the reliability of each classifier using a simple discounting. Then, the three mass functions are combined together using the conjunctive combination. Finally, a pignistic transform is applied, and the so-derived probability values are sorted decreasingly to provide the N best word hypotheses (the TOP N List).

This method outperforms classical combination methods which are used as references in the state of the art: We have conducted in [6] several detailed comparisons, as well as a test of significance on several datasets, and the differences of performances are always significant with immaterial p-values ($< 0.1\%$).

III. Proposed rejection strategies

In this section, we introduce two measures to *a priori* estimate the validity of the classification of a test word.

A. The measure of conflict

For a dedicated word, the first measure aims at quantifying the conflict among the evidence that has led to the classification. Intuitively, a high measure of conflict is supposed to correspond to a situation where it is sounded to reject the item, as there is contradictory information, whereas, a low measure of conflict indicates that the evidence concurs, and that rejection should not be considered. Several measures are available to quantify the conflict between several sources (such as described in [15]), among which, the **mass of conflict** from the conjunctive combination. The latter is really interesting, but in this work, we have chosen another measure, which is highly correlated with the mass of conflict, while being a bit easier to tune. Due to limited space, we do not detail the comparative theoretical and statistical studies that have led to this choice, and we focus on the description of the one that has been selected.

Let ω_* be an unknown word from the test set, and ω_1 the class that has been ranked first by the classification process (the output of which is the mass function $m_\cap$). We define $Flict$, the measure of conflict, as:

$$Flict(\omega_*) = 1 - pl_\cap(\{\omega_1\}) = bel_\cap(\{\Omega \setminus \omega_1\})$$

It corresponds to the sum of the mass of evidence which does not support the decision which has been made. This measure is really interesting, as it is easy to interpret, and as it takes its value in $[0, 1]$. On the other hand, if one wants to be really discriminative by rejecting a huge proportion of the test set, this measure is not adapted, as potentially too many test words may have a null measure of conflict.

B. The measure of conviction

For a dedicated word, the second measure aims at quantifying the conviction of the decision which has been made, i.e. whether at the end of the classification process, a class is clearly more likely than the other, or, on the contrary, whether the choice relies on a very weak preference of a class with respect to the others. Of course, we expect that a low measure of conviction corresponds to a situation where there is not enough evidence to make a clear-cut choice (and thus, rejection is an interesting option), and a high measure of conviction indicates that there is no room for hesitation, nor rejection. As with the measure of conflict, we do not detail the comparative study of several measures of conviction, and we focus on the chosen one. We define the measure of conviction as:

$$Viction(\omega_*) = \sum_{A \subseteq \Omega} \widehat{\overline{pl_\cap}}(A) - \widehat{\overline{bel_\cap}}(A)$$

i.e. the sum over $\mathcal{P}(\Omega)$ of the measure of imprecision of the pignistic discounting $\widehat{\overline{m_\cap}}$ of $m_\cap$. Contrarily to $Flict$, $Viction$ can be tuned according to the whole rejection spectrum, but its tuning is more difficult, as the values of its bounds depend on $|\Omega|$. However, **the main interest**

of $Viction$ is that it can be defined in a completely probabilistic context, without an ensemble classification based on DST. As a matter of fact, $\overline{m_\cap}$ corresponds to a probability distribution (such as the one provided by any probabilistic classifier). As a consequence, in a probabilistic case, the classifier provides a probability distribution p, and then, a consonant mass $m_p = \widehat{p}$ is derived by applying the inverse pignistic transform to p. If pl_p and bel_p are the plausibility and belief functions of m_p, we have:

$$Viction(\omega_*) = \sum_{A \subseteq \Omega} pl_p(A) - bel_p(A)$$

and this measure does not require any DST-based classifier nor any DST-based ensemble classification to be used.

C. Rejection strategies

Now, we use $Flict$ and $Viction$ to define three rejection strategies: The first and second strategies are based on each of the measures, while the third is based on a combination of the two measures. The first two rejection strategies are built in a similar way: The considered measure is compared to a threshold, which has been determined on a validation set, in order to reach a particular Rejection Rate. Depending on the sign of the difference between the measure and the threshold, the test word is classified or rejected. Of course, our two motivations for the rejection (too much conflict or too little conviction) are supposed to be independent. In practice, as the classifiers are not completely independent, and as the scores provided by the classifiers are normalized (so that they add up to one whatever the conflict and the conviction), the conviction and conflict measures are rather correlated. Hence, it makes sense to combine them, to stabilize and average the rejection performances. To do so, we simply reject a word if at least one of the two measures is beyond the threshold corresponding to the chosen Rejection Rate. As a reference method to evaluate our various strategies, we have chosen the one from [10] which provides the best result. It is sounded to choose this strategy, as it shares the same philosophy as ours: it is based on the comparison of a simple measure computed for each test word to a fixed threshold, and it does not require an extra classification process. It is based on the following measure:

$$Diff(\omega_*) = \frac{\overline{m_\cap}(\omega_1)}{\overline{m_\cap}(\omega_1) - \overline{m_\cap}(\omega_2)}$$

The $Diff$ measure varies within $[0, 1]$. Thus, a threshold in $[0, 1]$ is selected on the validation set according to the expected Rejection Rate, and the words for which the $Diff$ measure is greater than the threshold are rejected.

IV. Experimental results

Experiments have been conducted on two publicly available datasets: IFN/ENIT benchmark dataset of Arabic words and RIMES dataset for Latin words. The IFN/ENIT [16] contains a total of 32,492 handwritten words (Arabic script) of 946 Tunisian town/village names written by 411 different writers. Four different sets (a, b, c, d) are predefined in the dataset for training and one set (e) for testing. The RIMES dataset [17] is composed of isolated handwritten word snippets extracted from handwritten letters (Latin script). In our experiments, 36,000 snippets of words are used to train the different HMM classifiers and 3,000 words are used in the test. The dictionary is composed of 1,612 words.

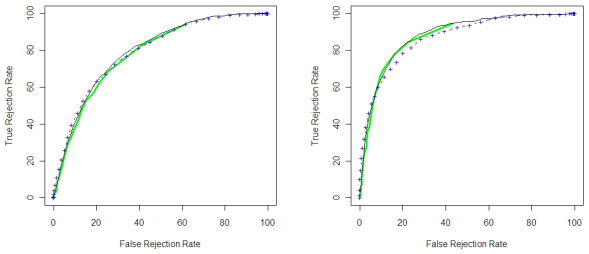

Figure 1. Comparison of the ROC curve of St1 (thick line) St2 (dotted) and St3 (thin black line) for the RIMES (left) and IFN/ENIT (right) dataset.

The ensemble classification procedure described in Section II-C has been applied to both of the test sets, following by the application of the four proposed rejection strategies: The one based on $Flict$ (St1), the one based on $Viction$ (St2), the one based on the combination of $Flict$ and $Viction$ (St3), and the reference strategy defined above (RefSt). For the experimental comparisons, we use the Receiver Operating Characteristic (ROC) curve, which is a graphical representation of the trade-off between the True Rejection Rate (TRR) and the False Rejection Rate (FRR).

It appears in Figures 1 that St1, St2 and St3 roughly behave similarly, whatever the test set. Hence, for the sake of simplicity, we consider from now only St3, which benefits from the advantages of both St1 and St2. The ROC curves, as well as the Error Rate, the Recognition Rate and Reliability with respect to the Rejection Rate are represented in Fig. 2. On the RIMES dataset, results are slightly better with St3 than with RefSt. Indeed, the value of the Area Under Curve (AUC)is 75.95% with StRef, whereas it is 79.01% with St3. On the other hand, results on the IFN/ENIT dataset are largely better with our rejection strategy than with the reference one. In fact, the value of the AUC is 72.79% with StRef, whereas it is 88.05% with St3. Moreover, we observe from this figure that for low Rejection Rates, the proposed rejection strategy produces interesting trade-offs between error and reject, which is the last important point in practical applications. Practically, the word Error Rates can be reduced from 18.50% to 6.37% on IFN/ENIT dataset and from 30.47% to 17.77% at the cost of the rejection 20% of the input words.

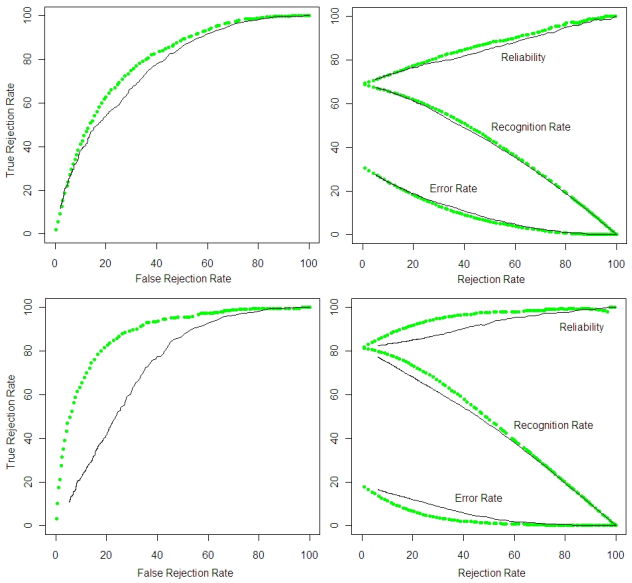

Figure 2. Comparison of the presented (dotted) and the reference (lined) methods for the RIMES (above) and IFN/ENIT (below) datasets. On the left, the ROC curve; on the right, the reliability, error and recognition rates.

V. Conclusion

We have presented a novel rejection strategy for reducing the Error Rate and improving the Reliability of the off-line handwritten word recognition system. Three different rejection strategies were investigated based on Dempster-shafer theory: The first one is based on a measure of the conflict among the evidence that has led to the choice of a particular class, while the second is based on a measure which encodes the conviction of the evidence involved in the classification process. Finally, the last strategy is based on a combination of the two previous measures. The experimental results have shown through two different publicly available datasets (one with Latin script and the other with Arabic script) that the proposed approach outperforms other state-of-the-art rejection methods. In fact, the word Error Rates can be reduced from 18.50% to 6.37% on IFN/ENIT dataset and from 30.47% to 17.77% at the cost of rejecting 20% of the input word images. Our future works will focus on alternative treatment of the rejected samples.

References

[1] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[2] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern.*, no. 3, 1992.

[3] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Trans. Systems, Man and Cybernetics, Part C: Applications and Reviews*, no. 2, pp. 216–232, 2001.

[4] G. Shafer, *A Mathematical Theory of Evidence*, Princeton, Ed. Princeton University Press, 1976.

[5] P. Smets, "The transferable belief model," *Artif. Intell.*, vol. 66, no. 2, pp. 191–234, 1994.

[6] Y. Kessentini, T. Burger, and T. Paquet, "Constructing dynamic frames of discernment in cases of large number of classes," *Submitted to the 11th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2011)*, 2011.

[7] ——, "Evidential ensemble hmm classifier for handwriting recognition," in *Proceedings of IPMU*, vol. 6178, 2010, pp. 445–454.

[8] A. Brakensiek, J. Rottland, and G. Rigoll, "Confidence measures for an address reading system," *International Conference on Document Analysis and Recognition*, vol. 1, pp. 294–298, 2003.

[9] G. Nikolai, "Optimizing error-reject trade off in recognition systems," in *ICDAR '97: Proceedings of the 4th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 1997, pp. 1092–1096.

[10] A. L. Koerich, R. Sabourin, and C. Y. Suen, "Recognition and verification of unconstrained handwritten words," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1509–1522, 2005.

[11] J. Rodrguez, G. Snchez, and J. Llads, "Rejection strategies involving classifier combination for handwriting recognition," in *Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science, vol. 4478, 2007, pp. 97–104.

[12] L. Guichard, A. H. Toselli, and B. Couasnon, "Handwritten word verification by svm-based hypotheses re-scoring and multiple thresholds rejection," *International Conference on Frontiers in Handwriting Recognition*, pp. 57–62, 2010.

[13] D. Dubois, H. Prade, and P. Smets, "New semantics for quantitative possibility theory," in *Proc. of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU 2001)*, 2001, pp. 410–421.

[14] T. Burger and S. Destercke, "The pignistic discounting: Definition and uses," *Submitted to the 11th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2011)*, 2011.

[15] W. Liu, "Analyzing the degree of conflict among belief functions," *Artificial Intelligence*, vol. 170, no. 11, pp. 909–924, 2006.

[16] M. Pechwitz, S. Maddouri, V. Maergner, N. Ellouze, and H. Amiri, "Ifn/enit - database of handwritten arabic words," *Colloque International Francophone sur l'Ecrit et le Doucement*, pp. 129–136, 2002.

[17] E. Grosicki, M. Carre, J. Brodin, and E. Geoffrois, "Results of the rimes evaluation campaign for handwritten mail processing," *International Conference on Document Analysis and Recognition*, vol. 0, pp. 941–945, 2009.