Álvaro Muñoz, F.; Sánchez Peiró, JA.; Benedí Ruiz, JM. (2013). An Image-Based Measure for Evaluation of Mathematical Expression Recognition. En Pattern Recognition and Image Analysis. Springer. 682-690. doi:10.1007/978-3-642-38628-2_81.

An image-based measure for evaluation of mathematical expression recognition

Francisco Álvaro, Joan-Andreu Sánchez, and José-Miguel Benedí

Instituto Tecnológico de Informática, Universitat Politècnica de València.
{falvaro,jandreu,jbenedi}@dsic.upv.es

Abstract. Mathematical expression recognition is an active research field that is related to document image analysis and typesetting. In this study, we present a novel global performance evaluation measure for mathematical expression recognition based on image matching. Using an image representation for evaluation tries to overcome the representation ambiguity as human beings do. The results of a recent competition were used to perform several experiments in order to analyze the benefits and drawbacks of this measure.

Keywords: Performance evaluation, Image-based modeling, Mathematical expression recognition

1 Introduction

Automatic recognition of Mathematical Expressions (ME) is an important problem for scientific document analysis and scientific document typesetting [3,10]. ME recognition techniques have been studied for both handwritten [10] and printed ME [3]. For recognition of handwritten ME, most of the works have concentrated on online recognition [10]. Online recognition of ME makes use of stroke information that is not present in offline recognition. Offline techniques [10] must be considered for handwritten images and printed ME recognition.

ME recognition comprises mainly two problems, that is, the recognition of mathematical symbols of the ME, and the recognition of the structural relation between these mathematical symbols [10]. As a pattern recognition problem, a fundamental issue in ME recognition is the definition of automatic evaluation techniques. Given that the recognition of mathematical symbols can be stated as a regular classification problem, the classification error rate of individual symbols is usually provided as a performance measure. However the recognition of the structural relation between mathematical symbols, which can be seen a parsing problem, requires more sophisticated evaluation methods [8,11].

When the ground-truth structural information is fully available, a representation (for example LaTeX format or MathML format) that allows automatic evaluation is needed. Evaluation techniques are usually based on tree-matching [8], but these techniques may report non-existent recognition errors due to the representation ambiguity of the coded ME [1].

In this paper, we present an automatic global performance evaluation measure of ME recognition systems when the ground-truth information is available as a coded string in LaTeX. Given a recognition result and its ground-truth, we generated the images from their LaTeX string and then we compared both images. This way we avoided most of ambiguity representation problems by comparing ME as human beings do, but the comparison between the images should be tackled in order to obtain a normalized error value.

We review some proposals for the evaluation of ME in Section 2. Section 3 describes the proposed measure. Section 4 presents the experiments performed to validate this approach and conclusions are presented in Section 5.

2 Evaluation of ME recognition systems

Several metrics have been proposed in the past to evaluate mathematical expression recognition systems. Some metrics have be defined at symbol level when the ground-truth is available [6]. However, these values only take into account the evaluation of a specific part of the ME recognition problem. Another measure that is often used is the expression recognition rate [10,6].

Given that the previous methods only provide a partial vision of the possible errors, additional measures have been developed. Chan and Yeung [2] proposed an integrated performance measure, which was a simple combination of symbol recognition and operator recognition rates. Garain and Chaudhuri [4] presented a global performance index that combined symbol and structural errors according to the complexity of the ME. Sain *et al.* [8] proposed EMERS, a tree matching-based performance evaluation measure. EMERS computes an edit distance using the tree representation of the ME. Zanibbi *et al.* [11] defined a set of performance metrics at different levels based on bipartite graph representation: that different metrics seem to provide a canonical representation, but it is not detailed in the paper and no experimentation is reported.

One important problem of the global metrics is the representation ambiguity that is present in the ME ground-truth. Given a ME, it is usually coded as a string in LaTeX or MathML. However, the same ME can be represented (*ground-truthed*) in several correct ways using these codifications [1]. Actually, in competition CROHME 2012 the organizers added a section with normalization guidelines for the output tree of recognized ME, although the expression-level reported metrics were expression recognition rate and structure recognition rate [6]. Therefore, an automatic global performance evaluation measure that can tackle the representation ambiguity problem seems appropriate.

3 Image-based ME global error

Given a recognition result of a certain expression and its ground-truth (both usually coded as a string in LaTeX or MathML), we wanted to evaluate the quality of this result. Since there can be several string representations of the same ME, and the image obtained should be unique, we propose comparing the images directly instead of their string representation.

As the image representation of a ME can be generated from its string codification, the idea was to compute a matching between the recognized expression image (test image) and the ground-truth label (reference image). In the following subsections we explain how by using an image-matching model (Section 3.1), we defined the evaluation algorithm (Section 3.2) that is used to finally compute the recognition error (Section 3.3).

3.1 Image-matching model (IDM)

In order to obtain a matching between two images, the initial idea was to compute a 2-dimensional warping between them. Keysers *et al.* [5] presented several deformation models for image classification, and the Image Distortion Model (IDM) represented the best compromise between computational complexity and evaluation accuracy. For this reason, we chose the IDM to perform a matching between two images.

The IDM is a zero-order model of image variability [5]. This model uses a mapping function with absolute constraints; hence, it is computationally much simpler than a 2-dimensional warping. Its lack of constraints is compensated using a local gradient image context window. This model obtains a dissimilitude measure from one image to another such that if two images are identical, their distance is equal to zero.

The IDM has two parameters: warp range (w) and context window size (c). The algorithm requires each pixel in the test image to be mapped to a pixel within the reference image not more than w pixels from the place it would take in a linear matching. Over all these possible mappings, the best matching pixel is determined using the $c \times c$ local gradient context window by minimizing the difference with the test image pixel.

3.2 The evaluation algorithm (BIDM)

Once we had a model that was able to detect similar regions of two images, we wanted to use this information to compute an error measure between them. Starting from the IDM-distance algorithm presented in [5], we proposed the Binary IDM (BIDM) evaluation algorithm shown in Fig. 1. First, instead of calculating the vertical and horizontal derivatives ('ver_der' and 'hor_der') using Sobel filters, these derivatives are computed using the method described in [9]. Next, the double loop computes the IDM distance for each pixel, and these values are stored individually. After that, the difference between each pixel of the test image and the most similar pixel found in the reference image can be represented as a gray-scale image (Fig. 2c-1). At this point, we have a dissimilitude value for each pixel of the test image. However, rather than knowing how different a pixels is, we want to know whether or not a pixel is correct. This is achieved by normalizing the distance values in the range $[0, 255]$ and then performing a binarization process using Otsu's method [7] (Fig. 2c-2). Finally, we intersect the foreground pixels of the test image with the binarized mapping values (like an error mask), and, as a result, we know which pixels are properly recognized and which are incorrectly recognized (Fig. 2c-3). Therefore, since the background

Input: test image A ($I \times J$), warp range w
reference image B ($X \times Y$), context window size c
Output: BIDM(w, c) from A to B

$A^v = \text{ver\_der}(A)$; $A^h = \text{hor\_der}(A)$; $B^v = \text{ver\_der}(B)$; $B^h = \text{hor\_der}(B)$
for $i = 1$ to I **do**
 for $j = 1$ to J **do** {
  $i' = \left\lfloor i\frac{X}{I} \right\rfloor$, $j' = \left\lfloor j\frac{Y}{J} \right\rfloor$, $z = \left\lfloor \frac{c}{2} \right\rfloor$
  $S_1 = \{1, \ldots, X\} \cap \{i' - w, \ldots, i' + w\}$
  $S_2 = \{1, \ldots, Y\} \cap \{j' - w, \ldots, j' + w\}$
  $$map(i,j) = \min_{\substack{x \in S_1 \\ y \in S_2}} \sum_{m=-z}^{z} \sum_{n=-z}^{z} (A^v_{i+n,j+m} - B^v_{x+n,y+m})^2 + (A^h_{i+n,j+m} - B^h_{x+n,y+m})^2$$
}
normalize_depth(map, 255)
binarize(map) *//Otsu's method*

fg $= \{(x,y) \mid A(x,y) < 255\}$ *//Foreground pixels*
cp $=$ fg $\cap \{(x,y) \mid map(x,y) = 0\}$ *//Correct pixels*

return $\frac{|cp|}{|fg|}$ *//Correct pixels ratio*

Fig. 1. Binary IDM (BIDM) evaluation algorithm.

pixels do not provide information, the number of correct pixels is normalized by the foreground pixels.

The time complexity of the algorithm is $O(IJw^2c^2)$, where $I \times J$ are the test image dimensions, w is the warp range parameter, and c is the local gradient context window size. It is important to note that in practice both w and c take low values compared to the image sizes as we will show in the experiments.

3.3 Recognition error (IMEGE)

The BIDM algorithm computes the number of pixels of an test image that are correctly allocated in another image according to the IDM model. The algorithm that we used followed the concepts of precision and recall to compute the Image-based Mathematical Expression Global Error (IMEGE)[1]. First, we compute the BIDM value from the test image to the reference (precision p). Second, we compute the same value from the reference image to the test image (recall r). Finally, both values are combined using the harmonic mean $f_1 = 2(p \cdot r)/(p + r)$, and we obtain the final error value. Fig. 2 illustrates an example of this process.

4 Experiments

In order to validate our approach we performed several experiments using the recognition results of a recent competition. We used them to tune the BIDM parameters. Finally, we added our evaluation measure to the competition results and the values reported are analyzed.

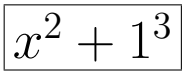
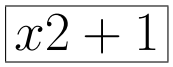
[1] Software available at `http://users.dsic.upv.es/~falvaro`

a) Mathematical expression recognition result

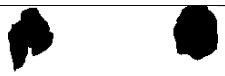
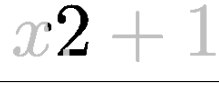
ground-truth = {`x^2 + 1^3`}

recognition = {`x2 + 1`}

b) Image generation from ground-truth and recognition

img1 =  img2 = 

c) BIDM computation in both directions

	img2 → img1	img1 → img2
1		
2		
3		
4	precision = $\frac{1489 \text{ correct}}{2197 \text{ foreground}} = 0.6777$	recall = $\frac{1429 \text{ correct}}{2338 \text{ foreground}} = 0.6112$

d) Recognition global error

$f_1(\text{precision}, \text{recall}) = 0.6427$

error $= 100(1 - 0.6427) =$ **35.73**

Fig. 2. Example of the procedure for computing the IMEGE measure given a ME recognition and its ground-truth in LaTeX.

4.1 CROHME 2012 Database

As discussed in Section 2, competitions on ME recognition have taken place recently. The organizers of CROHME 2012 [6] released the recognition results of the contestants, which provided a great resource for testing evaluation methodologies. The competition had 3 different parts each one having a different number of samples. Part-I was composed of 296 ME for training and 108 ME for test; Part-II had of 921 ME for training and 301 ME for test; and Part-III was composed of 1336 ME for training and 488 for test. Seven systems participated in CROHME 2012, hence, the total number of ME for evaluation were 756 for Part-I, 2107 for Part-II and 3416 for Part-III.

4.2 Parameter tuning

The BIDM has two parameters to be tuned: warp range (w) and context window size (c). We carried out several experiments to see the influence of each parameter on the IMEGE metric behavior. BIDM algorithm complexity is proportional to w^2 and c^2, so smaller parameter values are preferred. It is important to note that these values depend on the image resolution.

The resolution of the images had an important effect on the error computation due to renderization problems. A single error in a ME can produce displacements that slightly change the shape of other symbols in the expression. We made several decisions in order to tackle this problem. First, after performing some previous experiments we selected a $600dpi$ resolution to generate the ME images because lower resolution values could make the measure fail due to renderization problems. Furthermore, the horizontal and vertical derivatives were computed as described in [9], where a Gaussian function is applied to smooth the image. Taking this into account, we tuned the amplitude of this function to properly smooth the derivatives and alleviate the renderization problem. Moreover, the binarization process performed over the IDM-distance mapping (Fig. 2b-2) discarded small values that could occasionally appear due to slight displacements caused by a recognition error in another zone of the ME.

We tuned the BIDM parameters using the results of Part-I and Part-II of the CROHME 2012 competition. ME misrecognitions often produce displacements in the image generated. For this reason, larger w values are needed to give enough freedom to the BIDM to match correct recognized displaced regions. Consequently, tuning experiments showed that as the warp range increases, the error decreases. For larger values, the error remained almost invariant. As lower values are preferred, we chose $w = 40$ as a good compromise between correctness and performance.

As the IDM is less constrained than a 2D warping, the context window combined with the horizontal and vertical derivatives alleviate this lack. Therefore, larger context window sizes leaded to more reasonable and homogeneous mappings. However, excessive large context window sizes could cover areas that are to large. This caused misrecognized zones to extend their values to properly recognized regions and it required more computation time. On the other hand, small context windows could produce undesired results as, for example, reporting no error when comparing expressions '$(y+1)^2$' and '$(y+1^2)$', because the algorithm could match every symbol despite one was wrongly placed. For that reason, we fixed $w = 40$ and after varying the context window size, we selected $c = 27$ because it was the lower value that avoided these unlikely type of errors.

Finally, the IMEGE($w = 40, c = 27$) measure represented a good compromise between computation time and correctness. We tuned these parameters using many results produced by several systems in a competition on ME recognition. Thus, the metric IMEGE($w = 40, c = 27, 600dpi$) should also be a good measure for evaluating other ME recognition results, and then it would not be necessary to tune them again.

4.3 CROHME 2012 Evaluation

The recent competition CROHME 2012 reported several metrics in order to compare the performance of each system [6]. These measures were: stroke classification rate (ST_{rec}), symbol segmentation rate (SYM_{seg}), symbol recognition rate (SYM_{rec}), structure recognition rate (Struct) and expression recognition rate (EXP_{rec}). The systems were ranked according to the EXP_{rec} rate obtained on Part-III test dataset.

Expression recognition rate is a pessimistic metric because a single recognition error causes that a whole ME is considered as misrecognized. IMEGE metric computes a global error value that ranges from 0 to 100, hence, despite the sources of errors are not identified (segmentation, symbol recognition or structural), it reports a value that measures more precisely the quality of a ME recognition process. Table 1 shows the CROHME 2012 results including the average IMEGE($w = 40, c = 27$) value for each system. As reported metrics are in terms of recognition rates, we also reversed the IMEGE value to be consistent with them.

Table 1. CROHME 2012 results adding the IMEGE average.

	System	ST_{rec}	SYM_{seg}	SYM_{rec}	Struct	EXP_{rec}	IMEGE
Part I	I	80.74	90.74	89.20	62.04	35.19	79.78
	II	59.14	73.31	79.79	21.30	8.33	58.08
	III	90.05	94.44	95.96	70.37	57.41	86.58
	IV	78.24	92.81	86.62	50.93	28.70	76.46
	V	61.33	72.11	87.76	37.04	22.22	63.95
	VI	89.00	97.39	91.72	78.70	51.85	85.86
	VII	97.01	99.24	97.80	91.67	81.48	94.24
Part II	I	85.05	90.66	91.75	50.17	33.89	80.90
	II	58.53	72.19	86.95	12.29	6.64	48.06
	III	82.28	88.51	94.43	49.83	38.87	77.44
	IV	76.07	89.29	91.21	27.57	14.29	67.01
	V	49.06	61.09	88.36	17.61	7.97	47.07
	VI	90.71	96.67	94.57	69.44	49.17	80.28
	VII	96.85	98.71	98.06	88.37	75.08	91.98
Part III	I	79.85	91.95	86.25	42.21	22.75	73.60
	II	55.75	71.21	84.97	9.84	3.69	44.61
	III	78.94	87.75	91.38	36.89	25.61	72.47
	IV	72.12	87.51	87.62	23.77	9.43	60.61
	V	45.42	59.20	84.27	14.75	4.92	44.49
	VI	86.41	95.56	91.17	61.27	40.16	79.37
	VII	95.75	98.84	96.85	80.33	62.50	88.84

It can be seen that the proposed measure is coherent with the ranking of the systems. Moreover, as IMEGE considers a range of errors in ME recognition, there are some interesting results. For example, according to EXP_{rec} in Part-I, system V obtains significantly better results than system II, but their IMEGE values are not such different. It seems reasonable because their results at symbol level are similar, and the distribution of the errors among the ME recognition results can produce substantial variations in EXP_{rec}.

5 Conclusions

In this work, we have presented IMEGE, a novel performance evaluation measure of ME based on image matching. On the one hand, the image representation solves many of the ambiguity representation problems and this measure provides a value in $[0, 100]$ than can be interpreted as a visual error (as human beings do). On the other hand, IMEGE can not distinguish the source of the errors although it can identify the misrecognized zones of the ME. Given that this measure takes the global recognition information into account, it can be very helpful to complement EXP_{rec} and symbol related metrics in order to assess system's performance.

Acknowledgments. Work supported by the EC (FEDER/ FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), the MITTRAL (TIN2009-14633-C03-01) project, the FPU grant (AP2009-4363), and by the Generalitat Valenciana under the grant Prometeo/2009/014.

References

1. F. Álvaro, J.A. Sánchez, and J.M. Benedí. Unbiased evaluation of handwritten mathematical expression recognition. In *Proceedings of ICFHR*, pages 181–186, Italy, 2012.
2. K.F. Chan and D.Y. Yeung. Error detection, error correction and performance evaluation in on-line mathematical expression recognition. *Pattern Recognition*, 34(8):1671 – 1684, 2001.
3. P. A. Chou. Recognition of equations using a two-dimensional stochastic context-free grammar. In W. A. Pearlman, editor, *Visual Communications and Image Processing IV*, volume 1199 of *SPIE Proceedings Series*, pages 852–863, 1989.
4. U. Garain and B.B. Chaudhuri. A corpus for OCR research on mathematical expressions. *Int. Journal on Document Analysis and Recognition*, 7:241–259, 2005.
5. D. Keysers, T. Deselaers, C. Gollan, and H. Ney. Deformation models for image recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(8):1422–1435, 2007.
6. H. Mouchère, C. Viard-Gaudin, U. Garain, D.H. Kim, and J.H. Kim. ICFHR 2012 – Competition on Recognition of On-line Mathematical Expressions (CROHME 2012). In *Proceedings of ICFHR*, pages 807–812, Italy, 2012.
7. N. Otsu. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.
8. K. Sain, A. Dasgupta, and U. Garain. EMERS: a tree matching-based performance evaluation of mathematical expression recognition system. *International Journal of Document Analysis and Recognition*, 2010.
9. A.H. Toselli, A. Juan, and E. Vidal. Spontaneous Handwriting Recognition and Classification. In *Proceedings of ICPR*, pages 433–436, England, UK, 2004.
10. R. Zanibbi, D. Blostein, and J.R. Cordy. Recognizing mathematical expressions using tree transformation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(11):1–13, 2002.
11. R. Zanibbi, A. Pillay, H. Mouchere, C. Viard-Gaudin, and D. Blostein. Stroke-based performance metrics for handwritten mathematical expressions. In *Proceedings of ICDAR*, pages 334 –338, 2011.