A Shape-Based Layout Descriptor for Classifying Spatial Relationships in Handwritten Math

Francisco Álvaro
Instituto Tecnológico de Informática
Universitat Politècnica de València
Valencia, Spain
falvaro@dsic.upv.es

Richard Zanibbi
Department of Computer Science
Rochester Institute of Technology
Rochester, USA
rlaz@cs.rit.edu

ABSTRACT

We consider the difficult problem of classifying spatial relationships between symbols and subexpressions in handwritten mathematical expressions. We first improve existing geometric features based on bounding boxes and center points, normalizing them using the distance between the centers of the two symbols or subexpressions in question. We then propose a novel feature set for layout classification, using polar histograms computed over points in handwritten strokes. A series of experiments are presented in which a Support Vector Machine is used with these new features to classify spatial relationships of five types in the MathBrush corpus (horizontal, superscript, subscript, below, and inside (e.g. in a square root)). The normalized geometric features provide an improvement over previously published results, while the shape-based features provide a natural representation with results comparable to those for the geometric features. Combining the features produced a very small improvement in accuracy.

Categories and Subject Descriptors

I.7.5 [**Document and text processing**]: Document Capture—*Graphics recognition and interpretation*; I.5.4 [**Pattern Recognition**]: Applications—*Computer vision*

General Terms

Design, Experimentation

Keywords

math recognition, spatial relationship classification, shape descriptors

1. INTRODUCTION

Mathematical expression recognition has three primary subproblems [4, 13]: symbol segmentation, symbol recognition and structural analysis. In this paper we focus on classifying spatial relationships between symbols and subexpressions in handwritten expressions, a critical task for structural analysis where the layout of symbols is determined.

For two sets of handwritten strokes representing a pair of subexpressions A and B, our task is to determine their spatial relationship. A or B may be comprised of one or more symbols. We consider five spatial relationships: *horizontal* (AB), *subscript* (A_B), *superscript* (A^B), *below* ($\frac{A}{B}$) and *inside* ($\sqrt{B}$, where A is $\sqrt{\ }$).

Commonly layout in math expressions is classified using bounding box geometry [13]. Simistira *et al.* [11] classify six relationships in handwritten expressions, distinguishing *above* from *below*. We use five relationships, as vertical structures are represented top-down in the MathBrush corpus [8]. They use bounding box geometry for handwritten symbols, normalizing by symbol heights and widths. Vertical centroids for symbols are shifted based on typographic categories (*ascender, descender* or *centered*). Their experiments use a much smaller data set. For typeset math, Aly *et al.* [2] distinguish just horizontal, subscript and superscript relationships using bounding box geometry normalized by virtual ascenders and descenders, with high accuracy.

In this work, we introduce a new normalization for the geometric features of Álvaro *et al.* [1]. We then propose a novel set of shape-based features. Similar shape-based features have been use to detect typographic/layout classes for symbols [10], symbol retrieval [9], symbol segmentation [7], and expression matching [6]. We are not aware of shape-based features that have been applied to spatial relation classification for math expressions.

Experimental results show that the proposed normalization and the novel shape-based descriptor provide competitive results. The combination of both sets of features resulted in a 2.7% mean classification error (10-fold cross-validation) for isolated subexpression pairs.

2. FEATURE DESCRIPTIONS

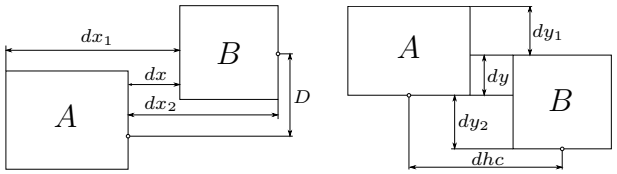
In this section, we describe geometric features based on the bounding boxes of subexpressions, and a second representation based on the actual shapes of handwritten strokes.

2.1 Geometric Features: Bounding Boxes

Álvaro *et al.* [1] define nine geometric features for spatial relationship classification using a normalization factor F, shown in Figure 1. Originally F was the height of the parent (usually, the leftmost) region A. It is particularly difficult to distinguish horizontal, subscript and superscript relationships, and the difference between the vertical centers

DocEng'13, September 10–13, 2013, Florence, Italy.

http://dx.doi.org/10.1145/2494266.2494315.



$H = \frac{\text{height}(B)}{F}$; $D = \frac{\text{cen}_\text{v}(A) - \text{cen}_\text{v}(B)}{F}$; $dhc = \frac{\text{cen}_\text{h}(A) - \text{cen}_\text{h}(B)}{F}$

$$\text{features} = [H, D, dhc, \tfrac{dx}{F}, \tfrac{dx_1}{F}, \tfrac{dx_2}{F}, \tfrac{dy}{F}, \tfrac{dy_1}{F}, \tfrac{dy_2}{F}]$$

Figure 1: Geometric features from bounding boxes of subexpressions A and B using normalization F

of A and B (feature D in Figure 1) has an important role in discriminating these layout classes.

To improve the placement of vertical centroids, symbols are divided into four typographic categories: ascendant (d, λ), descendant (p, μ), normal $(x, +)$ and middle $(7, \Pi)$. For *normal* symbols the centroid is set to the vertical centroid. For *ascendant* symbols the centroid is shifted downward to $(centroid + bottom)/2$. Likewise, for *descendant* symbols the centroid is shifted upward to $(centroid + top)/2$. Finally, for *middle* symbols, the vertical centroid is defined as $(top + bottom)/2$.

In the case of short symbols (e.g. fraction bars), using the height of A for normalization F leads to poor results. We propose a new normalization factor, the distance between the centers of the bounding boxes of the subexpressions. This is more robust against size variations in handwritten symbols.

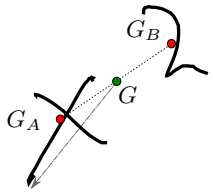
2.2 Shape Features: Polar Histograms

Many shape descriptors have been defined for image retrieval and object recognition in images [12]. In this section we define a new shape-based feature that is similar to shape contexts [3, 12]. We modify the polar shape matrix [5], which provides a powerful descriptor that is invariant under translation, rotation and scaling. However, we wish to apply this descriptor to determine the relationship between two stroke sets whose their relative position is important. As a result, we do not want rotation invariance.

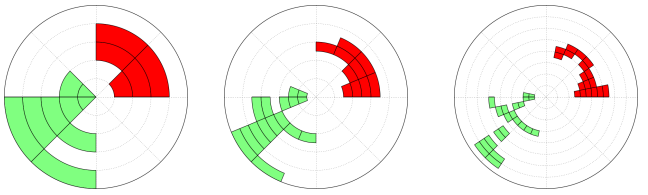
Given two sets of strokes A and B, let G_A and G_B be the center of mass of their corresponding shapes (i.e. stroke sample points). Using $G = (G_A + G_B)/2$ as a center, we draw n circles with radii equally spaced up to the maximum radius containing A and B. Moving counterclockwise, draw radii dividing each circle into m equal arcs. This descriptor is encoded as a matrix $\mathcal{M}$ such that each row represents a circle and each column represent the angle starting from 0 degrees. Each cell $\mathcal{M}(i, j)$ has one of three values obtained by majority vote of the points located in each bin:

$$\mathcal{M}(i,j) = \begin{cases} -1 & \text{more points from set } A \text{ than } B \\ 0 & \text{empty bin} \\ +1 & \textbf{tie}, \text{ or more points from set } B \text{ than } A \end{cases}$$

Figure 2 illustrates the effect of grid resolution on the polar histogram feature. We see that as the grid size is increased, the representation is more detailed, producing a warped image of the strokes.



Symbol pair (centers for x, 2 and midpoint shown)



$5 \times 8 = 40$ bins $\quad 10 \times 16 = 160$ bins $\quad 15 \times 32 = 480$ bins

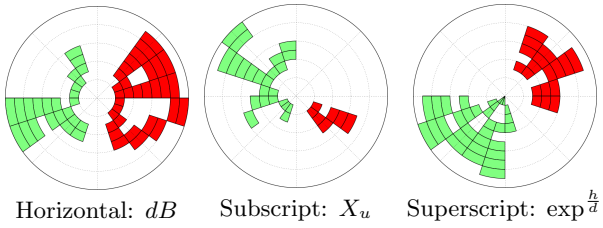
Figure 2: Varying distance (n) $\times$ angle (m) resolution in a polar histogram layout descriptor. Values shown using green (-1), red (+1), and white (0)

The $n \times m$ features are reduced using Principal Component Analysis (PCA). Figure 3 illustrates the proposed descriptor for the five spatial relations considered.

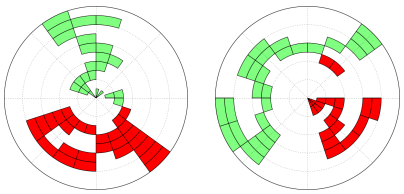
3. EXPERIMENTS

In this section we evaluate our proposed features for spatial relationship classification. The MathBrush database [8] is a public dataset containing 4,654 online handwritten mathematical expressions. Each expression has several spatial relations between symbols and subexpressions. There were 21,238 spatial relationships in the data set, classified according to the classes shown in Figure 3.

We use cross validation, splitting the dataset randomly into 10 partitions while keeping the distribution of spatial relations roughly uniform over the partitions. The training set contained 80% of the samples for each class, and the remaining 20% comprised the test set. We used a Support Vector Machine (SVM) classifier with a Gaussian kernel in our experiments. In order to tune the parameters for train-



Horizontal: dB $\quad$ Subscript: X_u $\quad$ Superscript: $\exp^{\frac{h}{d}}$



Below: $\sum_{C>h}$ $\quad$ Inside: $\sqrt{i}$

Figure 3: Polar histogram layout descriptors

ing the SVM classifier or to select the parameters of the shape-based features, we also divided the training set (80%) into 70% for training and 10% for a validation set. That split also kept the distribution of the classes, and the best parameters in the validation set were used to finally train the complete training set and compute the error using the test set for each one of the cross-validation partitions.

3.1 Geometric Feature Results

We performed several experiments to test the geometric features described in Section 2.1. First, we computed the classification error using the baseline normalization described in [1] such that F is equal to the height of the parent region A (GEO_1). Then, we classified the spatial relations using the new normalization factor, the distance between the center of the bounding boxes (GEO_2). Finally, we also extracted the geometric features GEO_2 without using the information about symbol categories (ascendant, descendant, normal, middle) in order to measure the influence of this decision (GEO_3).

The results in Table 1 show the new normalization by center point distance decreasing the mean classification error rate from 3.62% to 2.84%. The results also show a slight decrease in error when computing a vertical centroid based on symbol typographic categories, with the error decreasing from 3.48% to 2.84%.

3.2 Shape-Based Feature Results

The polar histogram-based descriptor presented in Section 2.2 has parameters that need tuning, specifically the number of circles n and angles m, and the number of principal components d to select. We performed a grid search for several sizes of the descriptor ($n = \{3, 5, 10, 15, 20\}$ and $m = \{8, 12, 16, 20, 24, 28, 32\}$), and for each size, different numbers of principal components were also tested (variance explained from 10% to 90% in increments of 10%). We used one of the 10 partitions extracted from the cross-validation experimentation to tune these parameters (see Figure 4).

We chose $n = 15$, $m = 20$ and $d = 35$ (50% of total variance) as the parameters to perform the cross validation experiments for the shape-based geometric features (SHP). For small grid sizes, results were best when high percentage of variance were accounted for in the PCA dimensions (70%-90%). However, as the grid size increased the variance in bin counts also increased, with the best results being obtained when keeping components covering roughly 50% of the variance.

Table 1: MathBrush symbol relationship classification results (10-fold cross validation). For each feature the number of features (#) and whether typographic symbol classes are used (Cat.) are shown

Feature	#	Cat.	% Error ($\mu \pm \sigma$)
GEO_3: F = dist(centers)	9	No	3.48 ± 0.39
GEO_2: F = dist(centers)	9	Yes	2.84 ± 0.16
GEO_1: F = height(A)	9	Yes	3.62 ± 0.34
SHP: $n = 15, m = 20$	35	No	3.34 ± 0.21
GEO_2 + SHP	44	Yes	2.70 ± 0.29

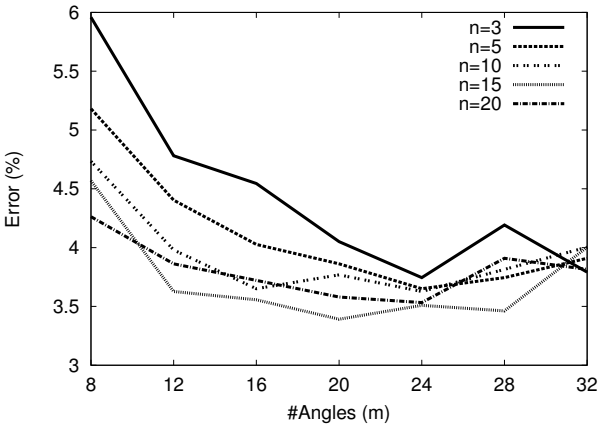


Figure 4: Fitting polar histogram parameters. Error for the best PCA dimension set for each m ($angles$) $\times$ n ($circles$) histogram is shown

The polar histogram features obtained a mean classification error of 3.34% (Table 1), without including symbol typographic classes; this is comparable to the accuracy obtained using the geometric features without symbol typographic classes, where error was 3.48%. Interestingly, the standard deviation in error was half as large as that for the bounding box-based features in this case, but within a narrow range (0.21% vs 0.39%).

Table 2 shows the confusion matrix for the GEO_2 cross validation experiments. As expected, most errors are produced in the classification of Horizontal, Subscript and Superscript relationships, whereas Below and Inside relationships have few errors.

The SVM classifier is influenced by the prior probabilities of the classes in the training data (Horizontal: 68.9%, Subscript: 5.9%, Superscript: 9.0%, Below: 12.1%, Inside: 4.1%). The Horizontal relationship represents about 69% of the samples, and its recognition error was very low. The Superscript relation had a 6.3% error, but it is the Subscript relation that is most challenging, with more than 20% error: the Horizontal/Subscript confusion is by far the most frequent.

Table 2 also shows the confusion matrix for the shape-based features. The classification errors follow a very similar distribution. Errors in Subscript and Superscript relations are slightly higher, as well as for Inside. The error rate for Below relationships in ground truth is lower, but the classifier has more false positives for the Below relationship.

We tried adding to the shape-based descriptor the information about symbols categories in the relation by displacing the centroids G_A or G_B following the methodology described in Section 2.1. However, this led to weaker results.

Given the good results for both feature types, which use quite different representations, a natural next step was to merge them. This combination led to small improvements in mean classification error to 2.7% (see Table 1). This is unlikely to be significantly different from the GEO_2 result, due to the larger standard deviation (0.29% vs. 0.16%).

3.3 Discussion

The polar histogram descriptor obtained results comparable to the geometric features when no symbol information is

Table 2: Confusion matrices for GEO_2 (geometric) and SHP (shape) descriptors (10-fold cross validation). Ground truth labels are shown along the rows (FN: false negative rate, FP: false positive rate)

	GEO_2 Output						SHP Output					
GT	Hor	Sub	Sup	Below	Inside	**FN**	Hor	Sub	Sup	Below	Inside	**FN**
Hor	28888	196	149	7	8	1.2%	28863	251	130	4		1.3%
Sub	498	1993		25	2	20.8%	581	1912		25		24.1%
Sup	239		3597	2		6.3%	322		3514	2		8.4%
Below	18	42	4	5083	9	1.4%	17	23	2	5114		0.8%
Inside	9				1707	0.5%	37	4		22	1653	3.7%
FP	2.6%	10.6%	4.1%	0.7%	1.1%		3.2%	12.7%	3.8%	1.0%	0%	

used, but was outperformed by the geometric features when typographic classes are used to move vertical centroids. One possible direction for future work is to try and incorporate this information into the shape descriptor.

From the results, the proposed descriptors are not sufficient on their own for spatial relationship classification. Language models may be needed to distinguish cases where the geometric conditions represent different relations depending on the symbols involved (e.g. the horizontal relation 'Px' vs. the subscript relation 'p_x').

However, there are opportunities to improve our descriptors, for example using continuous values for the bins in our polar histograms. For both feature types presented, it would be good to find better ways to identify the writing line, middle line (e.g. top of a lower-case 'x'), or a point between these in order to better handle the most common confusions (Horizontal vs. Subscript or Superscript).

4. CONCLUSIONS AND FUTURE WORK

In this paper we dealt with the classification of spatial relations between handwritten mathematical symbols and subexpressions. We presented a new normalization for a set of geometric features and a novel set of shape-based features, which improve upon previously published results. Our new polar histogram-based shape feature provides comparable results to geometric features when no information about symbol typographic categories (e.g. ascender) is used. The combination of both sets of features led to a small improvement in accuracy. In future work, we will consider including symbol typographic classes into the shape-based feature representation, and adding a rejection class, to detect when two subexpressions are unrelated. Finally, our features could be applied to printed expressions and compared with earlier work [2].

5. ACKNOWLEDGMENTS

This work was partially supported by the Spanish MEC under the STraDA research project (TIN2012-37475-C02-01), an FPU grant (AP2009-4363), and by the National Science Foundation (USA) under Grant No. IIS-1016815. The authors thank Lei Hu for helpful discussions.

6. REFERENCES

[1] F. Álvaro, J.A. Sánchez, and J.M. Benedí. Recognition of on-line handwritten mathematical expressions using 2D stochastic context-free grammars and hidden Markov models. *Pattern Recognition Letters*, 2012.

[2] W. Aly, S. Uchida, and M. Suzuki. Identifying subscripts and superscripts in mathematical documents. *Mathematics in Computer Science*, 2(2):195–209, 2008.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[4] K. Chan and D. Yeung. Mathematical expression recognition: a survey. *Int'l J. Document Analysis and Recognition*, 3(1):3–15, 2000.

[5] A. Goshtasby. Description and discrimination of planar shapes using shape matrices. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 7(6):738–743, 1985.

[6] N. S. T. Hirata and W. Y. Honda. Automatic labeling of handwritten mathematical symbols via expression matching. In *Int'l Conf. Graph-Based Representations in Pattern Recognition*, pp. 295–304, 2011.

[7] L. Hu and R. Zanibbi. Segmenting Handwritten Math Symbols Using AdaBoost and Multi-Scale Shape Context Features. *Int'l Conf. Document Analysis and Recognition*, to appear, 2013.

[8] S. MacLean, G. Labahn, E. Lank, M. Marzouk, and D. Tausky. Grammar-based techniques for creating ground-truthed sketch corpora. *Int'l J. Document Analysis and Recognition*, 14:65–74, 2011.

[9] S. Marinai, B. Miotti, and G. Soda. Using earth mover's distance in the bag-of-visual-words model for mathematical symbol retrieval. In *Int'l Conf. Document Analysis and Recognition*, pp. 1309–1313, 2011.

[10] L. Ouyang and R. Zanibbi. Identifying layout classes for mathematical symbols using layout context. *IEEE Western New York Image Processing Workshop*, 2009.

[11] F. Simistira, V. Papavassiliou, V. Katsouros, and G. Carayannis. Structural analysis of online handwritten mathematical symbols based on support vector machines. In *Document Recognition and Retrieval XX*, 2013.

[12] M. Yang, K. Kpalma, and J. Ronsin. A survey of shape feature extraction techniques. In P.-Y. Yin, editor, *Pattern Recognition Techniques, Technology and Applications*, pp. 43–90, 2008.

[13] R. Zanibbi and D. Blostein. Recognition and retrieval of mathematical expressions. *Int'l J. Document Analysis and Recognition*, 15(4):331–357, 2012.