International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 11 | Nov -2017 www.irjet.net p-ISSN: 2395-0072

Recognition of Handwritten Mathematical Equations

Prachi Gupta[1] , Neelam Pal[2], Lavanya Agrawal[3]

[1,2,3] *B tech, IMS Engineering college, Uttar Pradesh, India*

Abstract– *The aim of this project is to develop an assistance system that analyses the handwritten mathematical equations based on handwriting recognition algorithms. The system will be able to recognize images of handwritten equations and output the corresponding characters in LATEX.*

Being able to change handwritten equations into LATEX has applications for consumers and academics. While large number of researches have been done for digits and characters recognition, much less progress has been made surrounding handwritten equation recognition. While typing is much faster than writing by hand, math equations have the opposite property: writing mathematical equations by hand is more efficient than typesetting them. Furthermore, whereas handwritten equations are human readable, the "code" of typesetting languages are often highly nested and difficult to edit. These problems of typesetting present an opportunity for improving the workflow of writing math equations digitally.

Handwritten character or symbol recognition is one of the application in pattern classification. It is generally easy for anyone to recognize handwritten or characters and symbols but it is difficult for a computer to recognize them. This difficulty can be overcome by adopting machine learning approach by designing a system that recognizes the patterns. Pattern classification involves features extraction, concept behind the observation and classifier. For this purpose, character geometry as feature extraction technique and two classifiers Support Vector Machines (SVM) and K-nearest neighbor (KNN) are used. Two classifiers are used for the comparative analysis.

Key words: SVM, KNN, Character geometry, Machine learning

1. INTRODUCTION

Pattern recognition forms the basis of learning for everyone. It is generally easy for a person to differentiate a handwritten number "7" from an "9"; However, it is difficult for a programmable device(computer) to solve this kind of emotive problem. The problem is difficult because each pattern usually contains a large amount of processed data, and the recognition problems typically have an indistinct,

high-dimensional, structure. Pattern recognition is the science of making conclusion from perceptual data, using tools from statistics, probability, computational geometry, machine learning, and algorithm design. Pattern classification involves features extraction and classification. Pattern is defined as combination of features that are characteristic of an individual. In classification, a pattern is a pair of variables {x, w} where x is a collection of observations or features (feature vector) and w is the concept behind the observation (label). The quality of a feature vector is related to its ability to bias examples from different classes. Examples from the same class should have similar feature values and while examples from different classes having different feature values. Feature can be defined as any distinctive aspect, quality or characteristic which, may be symbols or numerals . The combination of d features is represented as a d-dimensional column vector called a feature vector. The dimensional space defined by the feature vector is called feature space. Objects are represented as points in feature space. The goal of a classifier is to separate feature space into class-labeled decision regions.

2. ALGORITHMS

2.1 K-Nearest Neighbour

neighbor. KNN is a type of instance-based learning, or lazy learning, where the function is only nearly local and all computation is delayed In pattern recognition, the k-nearest neighbor algorithm (KNN) is the input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, generally small). If k = 1, then the object is simply assigned to the class of that single nearest until classification. The KNN algorithm is among the easiest of all machine learning algorithms. Both for classification and regression, can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more far ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for KNN classification) or the object property value (for KNN regression) is known. This can be thought of as the training set for the algorithm, though no obvious training step is required.

2.2 Support Vector Machine

(SVMs, also support vector networks) are supervised learning models with associated learning. In machine learning, support vector machines algorithms that inspect the data used for classification and regression analysis. Given a set of training examples, each related to one or the other of two categories, an SVM training algorithm builds a model

that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of distinct categories are divided by a clear gap that is as broad as possible. New examples are then mapped into that same space and forecast to belong to a category based on which side of the gap they fall.

3 CHARACTER GEOMETRY FEATURE EXTRACTION TECHNIQUE

Character geometry technique follows the steps described below Universe of Discourse: At first, the universe of discourse is selected because the features extracted from the character image include the positions of different line in the character image.

Zoning: The image is divided into windows of equal size and feature extraction is applied to each individual zone rather than the whole image. In our work, the image was divided into 9 equal sized windows. Starters, Intersections and Minor Starters: To extract different line segments in a particular zone, the whole skeleton in that zone should be extended across. For the above reason, particular pixels in the character skeleton are treated as starters, intersections and few starters.

Character traversal: Character traversal starts after zoning by which line in each zone are extracted. The first step is the starters and intersections in a zone are identified and then filled in a list. Then the algorithm starts by considering the starter list. Once all the starters are processed, minor starters find along the course of traversal are processed. The positions of pixels in each of the line segment obtained during the process are stored. After traversing all the pixels in the image, the algorithm ends.

Distinguishing the line segments: After all the line segments in the image are extracted, they are classified into any one of the following lines – Horizontal line, Vertical line, Right diagonal line, or Left-diagonal line.

Feature Extraction: After the line type of each segment is resolved, feature vector is formed based on this processed data.

3.1Architecture of the Proposed System

Input image: Input picture can be a picture comprising of either a character, image or expressions. The information picture ought to be in a .png augmentation necessarily.

Preprocessing of input image: Picture pre-dealing is compulsory for any photo based applications. The precision and consolidating rate of such frameworks must be on a very basic level high in order to ensure the achievement of the resulting steps. In any case, more often than not, the noteworthiness of these systems stay ignored which brings about second rate results. The objective of the pre-handling step is to construct necessary information.

Feature extraction: This is a sort of dimensionality decreasing that successfully addresses interesting parts of a photo as a little component vector. This approach is useful when picture sizes are huge and a reduced segment depiction is required to quickly complete assignments, for instance, picture organizing and recovery. Character Geometry Feature Extraction Technique is used for the component extraction since it is one which support the component extraction strategies which does not have any mistakes.

Performing SVM and KNN classification: KNN count is a flexible and clear estimation which arranges the given get ready cases in perspective of its neighbors. For collecting the new case this count figures the Euclidian division with the new representation and recognizes its neighbors after that it consigns the class in light of the k regard .k regard will be customer portrayed. This estimation reconstruct the class that addresses the most outrageous of the k cases. SVMs (Support Vector Machines) are an important technique for data classification. A classification task generally incorporates disengaging data into get ready and testing sets. Each event in the planning set contains one "target regard" (i.e. the class names) and a set of "attributes" (i.e. the components or watched factors). The goal of SVM is to explain a model (in light of the arrangement data) which predicts the target estimations of the test data given only the test data properties. SVM requires that each data case is addressed as a vector of real numbers. Scaling before applying SVM is tough. The rule of SVM depends upon a straight separation in a high estimation feature space where data are mapped to consider the possible non-linearity of the issue.

Printing the recognized symbol as Output: The expression, symbol or character given as the input picture is behold and imprinted in the command window. Recognized image as bestow is a consequence of the considerable number of procedures clarified . The perceived image will be the form which can be behold by the PC, human and machines.

4. EXPERIMENTAL RESULTS

Here we have used MATLAB as our PC language, here dataset contains both printed and handwritten images. Mathematical Symbol dataset not only contains different symbols but also contains alphabets of different languages. For the handwritten mathematical symbol/character recognition system, a private dataset is used. This own dataset is prepared for Mathematical Symbols by considering all possible constraints such as variations in writing styles, representatives consisting with some noise, etc.

5. CONCLUSIONS

Among different feature extraction techniques and strategies, Character geometry is selected as the feature extraction technique. Character geometry feature extraction technique is one which supports other different feature extraction techniques. SVM and KNN classifiers are used for the classification. After numerous executions it has been found that classifier has great correctness compared to SVM as classifier. The efficiency of KNN decreases with the increase in dataset.

S.NO	INPUT IMAGE	SVM	KNN
1	1 0 1 2 3 4 5 6 7 8 9	100%	100%
2	A B C D E	100%	100%
3	F G H I J	100%	100%
4	K L M N O	100%	100%
5	P Q R S T	100%	100%
6	U V W X Y Z	100%	100%
7	a b c d e	40%	60%
8	f g h i j	100%	100%
9	k l m n o	100%	100%
10	p q r s t	100%	100%
11	u v w x y z	100%	100%
12	$\beta_u * \lambda -$	100%	100%
13	( ) + ^ γ	100%	100%
14	σ → ◯ ,	100%	100%
15	∏ ⁿ ∫ ∞ Δ	100%	100%
16	. ↔ Θ X V	80%	80%
17	A + B * C	100%	100%
18	< k (ⁿ)	100%	100%
19	u E { 1, . . , n − 3 }	100%	100%
20	a + b * c	100%	100%
21	2 n − 1	100%	100%
22	t a n Θ	75%	75%
23	a + c	100%	100%
24	! @ # ; : =	33%	33%

Table (a)- Classification of input images for printed images dataset

S.NO	INPUT IMAGE	SVM	KNN
1	1 0 1 2 3 4 5 6 7 8 9	100%	100%
2	A B C D E	100%	100%
3	F G H I J	100%	100%
4	K L M N O	100%	100%
5	P Q R S T	100%	100%
6	U V W X Y Z	100%	100%
7	a b c d e	40%	60%
8	f g h i j	100%	100%
9	k l m n o	100%	100%
10	p q r s t	100%	100%
11	u v w x y z	100%	100%
12	$\beta_u * \lambda -$	100%	100%
13	( ) + ^ γ	100%	100%
14	σ → ◯ ,	100%	100%
15	∏ ⁿ ∫ ∞ Δ	100%	100%
16	. ↔ Θ X V	80%	80%
17	A + B * C	100%	100%
18	< k (ⁿ)	60%	60%
19	u E { 1, . . , n − 3 }	100%	100%
20	a + b * c	100%	100%
21	2 n − 1	25%	100%
22	t a n Θ	25%	100%
23	a + c	100%	100%
24	! @ # ; : =	33%	33%

Table (b)- Classification of input images for handwritten images dataset

6. FUTURE WORK

In future, this above technique can be attempted against standard databases. Likewise more tests could be directed with extra benchmark datasets. By using efficient segmentation technique, offline handwritten mathematical equations with respect to superscript and subscript could be recognized efficiently.

7. ACKNOWLEDGEMENT

We are earnestly grateful to our mentor **Dr. UPASANA PANDEY ma'am,** Associate Professor Department of computer science and engineering, IMS Engineering College for guiding us at each and every step to make this project a successful one. Finally we will like to express heartiest gratefulness to almighty and to our parents for supporting us.

8. REFERENCES

[1]Richard Zanibbi & Dorothea Blostein, "Recognition and retrieval of mathematical expressions", IJDAR, SpringerVerlag, 2011.

[2] Zhao Xuejun, Liu Xinyul, Zheng Shenglingl, Pan Baochang and Yuan Y.Tang, "On-line Recognition Handwritten Mat hematical Symbols", IEEE, 1997.

[3] Dorothea Blostein & Ann Grbavec, " Handbook on Optical Character Recognition and Document Image Analysis", (Chapter22)-"Recognition of Mathematical Notation", World Scientific Publishing Company, 1996

[4] J. S. Raikwal & Kanak Saxena ," Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set " In International Journal of Computer Applications (0975 – 8887) Volume 50 – No.14, July 2012M.

[5] http://www.inftyproject.org