Cut Digits Classification with k-NN Multi-specialist

Fernando Boto, Andoni Cortés, and Clemente Rodríguez

Computer Architecture and Technology Department, Computer Science Faculty, UPV/EHU, Aptdo. 649, 20080 San Sebastian, Spain
acbbosaf@si.ehu.es

Abstract. A multi-classifier formed by specialised classifiers for noise produced by an image is shown in this work. A study has been carried out in the case of cut images, where tree cases of specialization are considered. Classifiers based on neighbourhood criteria are used, the zoning global feature and the Euclidean distance too. Furthermore, the paper explains a modification of the Euclidean distance for classifying cut digits. The experiments have been carried out with images of typewritten digits, taken from real forms. Trying to obtain a strong database to support the experiments, we have cut images deliberately. The recognition rate improves from 84.6% to 97.70%, but whether the system provides information about the disturbance of the image, it can achieve a 98.45%.

1 Introduction

Human intervention in full scale digitization of documents is tedious because of the large amount of documents to be processed. Nowadays there are some recognition systems in the market for typewritten texts but still they create many problems. The document digitization process is usually made starting with the isolated characters and sometimes this isolation process produces some image disturbances. The noise characters can be obtained through a bad quality of digitization or as a result of a bad segmentation [1] [2] [3].

Many authors use systems based on the combination of classifiers. These systems have different aims [4]: Efficiency [5] [6], improved performance [7] [8] [9], generalisation [10]. Besides, [4] [11] makes a survey of some of the possibilities to combine classifiers and the rules to combine them. In our case, we need a multi-classifier in order to combine classifiers with a different purpose. Each classifier will specialise in a type of problem or distortion and together, by means of a decision rule, will provide a result by common consent.

Another point discussed in this item makes reference to the disturbances produced by digitization or segmentation defects, which will result in noise or blurred numbers, with thickness defects or cuts with loss of structure. Each one of these disturbances has been solved independently, that is with sub-systems which provide a good rate of success with some disturbances but with different results before other types of disturbances.

Many works found in the related bibliography, consider the problem of the cut digits as the problem where the digits must be repaired. These works use algorithms of

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 496 – 505, 2006.

tracking borders techniques, other techniques to obtain the skeleton of an image, dilatation search of rupture points, etc. [1] [2] [12] [13]. We are not going to approach the problem in this way because the needs are different.

The cut digits in the upper or lower part are produce by an improper horizontal segmentation of the code because the cell of the code is badly obtained. Figure 1 shows horizontal improper segmentation, where we can see how the segmentation process produces cut digits. The solution presented in this paper will take place in the recognition phase, specifically in the feature extraction phase.

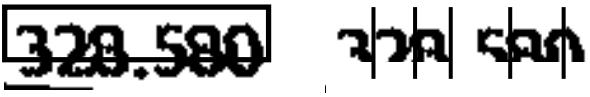

Fig. 1. Horizontal and vertical segmentation errors

Figure 2 shows the disturbances of our real problem, however in this paper we have only consider the problem of cut digits with lost of structure (type D) and more specifically, with cut digits in upper and lower part, that is, the cell of the digit is cut and the information of the digit is less than the original one. The cut digits in the real problem conform 1% of the total sample, however we have worked with synthetic and homogeneous sample.

We have present in other works, the behaviour of the system for digits with distortions like in the case A and B of the figure 2 [14]. Furthermore we have advances with all the disturbances showed in the figure and nowadays the system is already working in real applications with good results.

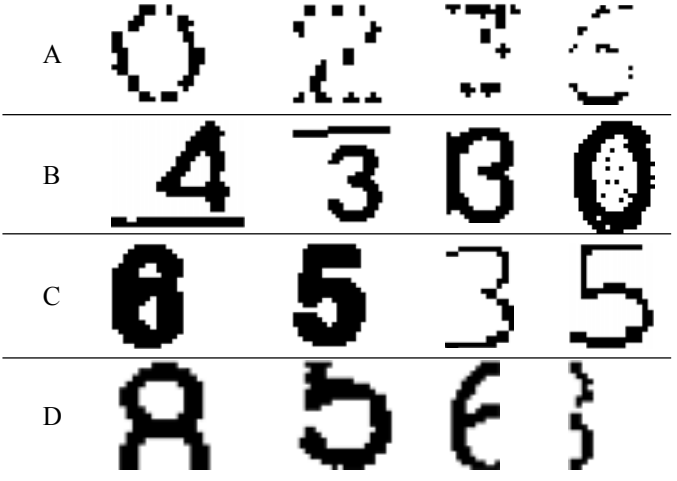

Fig. 2. Real examples of digits with distortions. A: Discontinuous lines (blurred), B: Noise added or annex lines, C: Thickness disturbance and D: Cut digits.

The point of view presented in this work is that, any image with any type of rupture will belong to the sub-space of this type of rupture. A specialised classifier in a sub-space will have a confidence level when sorting out a pattern. A set of classifiers will compete with one another and the classifier with the highest level of confidence will recognise the pattern.

The paper has the following structure. Section 2 describes the Multi-specialist scheme, section 3 shows the solution proposed for cut digits and section 4 an experimental study. Finally the conclusions in section 5.

2 Multiple Specialised Classifiers

The basic aim of the proposed Multiple Specialist Classifier (MSC), is to obtain a multi-classifier formed by specialised simple classifiers for each type of noise produced by an image. Each specialist obtains the features of the input digit, depending on what kind of disturbance is treating, so that the system manage the input patterns in a different way. This is explained in next section.

Specifically, the feature extraction process used is the *zoning* global one [5], which obtains a vector of features which corresponds to the spatial distribution of the black areas of the image. The image is divided in *mxn* cells with the same area for them all, and the percentage of blacks in this area of each cell is calculated.

The decisions adopted by each classifier will be treated jointly in a global discriminating function.

This is a general description of a Multi-classifier (Multi-specialist) [10]: $MSC(ns, \{S_1, \ldots, S_{ns}\})$, where ns is the number of specialists of the multi-classifier and $\{S_0 \ldots S_{ns}\}$ the set of specialists for it.

A specialist is defined as: $S_k(M_k, N_k, P_k)$, where M_k and N_k are rows and columns respectively of the zoning feature taken into account in this specialist. Parameter P_k is the part of the digit where the specialist supposes the lost of structure.

The reference pattern set (*RPS*) for all the specialist is the same and the feature extraction too. Zoning is used with a dimensionality of 8x5. This feature extraction generates a space dimensionality of 40 dimensions. Based on this space, the dimensionality of the *RPS* is reduced depending on the specialist specification. For instance, let a specialist specification, $S_k(6, 5, Lower)$, the system just have to take into account the first six rows of the patterns of the *RPS*, on the other hand, the system generate 6x5 features for the input pattern. Next section explains this point and how the Euclidean distance is managed.

Each specialist k returns a confidence value and its output classes ($Class_{k1}$ and $Class_{k2}$) of the two nearest patterns (2-NN classifier [15] [16]) with its corresponding distances, D_{k1} and D_{k2}, given that $Class_{k1} \neq Class_{k2}$.

A decision function determines witch classifier provides the classification result. The system used, presents an horizontal decision scheme [17], and the solution is based in a knowledge based system with a confidence index in each input pattern. So, we have determined to decided class C_{k1} of k specialist, with a confidence value V_k bigger than all the rest specialists. The confidence value for each specialist is $V_k = 1 - D_{k1}/D_{k2}$, this function is the same for all the experiments in this work.

3 Space Dimensionality Variation

This chapter will deal with the solution we have found to palliate the effects of a problem produced by an improper segmentation. This improper segmentation produces digits with a lack of structure or loss of information. The remaining information is not enough to recognise the digit. The lack of structure can be found in the upper part of the image of the digit or in the lower part generally, but sometimes side cuts appear but not so frequently in this environment.

Basically, this is the way to solve this disturbance: The reference patterns or learned set, creates a space of D dimensionality to recognise well formed and standard digits. When we present a cut digit before this dimensional space, it is placed in another point of the space, because, for lack of structure the obtainment of characteristics has created a false pattern. For instance, the more cut is the class *8*, in its upper or lower parts, the obtained vector of characteristics will show a point in the space coming nearer to the class *0*. However, we can leave aside certain characteristics of the reference pattern set (*RPS*) and create a dimensional space $D\text{-}\Phi$, and try that the Φ features not taken into consideration correspond to the structure missing in the entry pattern.

In short, we generate for the entry pattern $D\text{-}\Phi$ characteristics and we apply the distance function with the selected characteristics $D\text{-}\Phi$ of the *RPS* patterns.

Given an *RPS* pattern of class $c\ (R_1, R_2, \ldots R_D)$ within a dimensional space D and an entry pattern or test also class $c\ (P_1, P_2, \ldots P_D)$, the Euclidean distance is defined as (1) which will give as a result a value of d_1.

$$\sqrt{\sum_{i=0}^{D}(R_i - P_i)^2} \tag{1}$$

If the entry pattern is cut on the lower part which means that the lower part is missing structure, the calculation of the distance will give a value of d_2 and most probably $d_2 >> d_1$. Therefore, if we have this type of disturbance, the classifier will fail.

If we reduce the space of features $D\text{-}\Phi$ for the entry pattern, depending on the loss, the $D\text{-}\Phi$ features will appear different to those of the previous $(PC_1, PC_2, \ldots PC_{D\text{-}\Phi})$, but for the *RPS* the same will be used, even if taking into consideration some of them, that is (2) being d_3 and $d_3 << d_2$.

$$\sqrt{\sum_{i=0}^{D-\Phi}(R_i - PC_i)^2} \tag{2}$$

This is the way to suppose that the digit is cut in its lower part a $\Phi/D*100$ percent. Whether the classifier suppose that the digit is cut in the upper part, the distance is calculated as (3).

$$\sqrt{\sum_{i=0}^{D-\Phi}(R_{\Phi+i} - PC_i)^2} \tag{3}$$

Note that Φ can vary depending on how much cut is supposed the digit to be cut.

The parameter Φ will be always multiple of n, being n the columns of the zoning matrix. In the study explained in next section the parameter Φ will take integer values between n and $3n$.

4 Study of Specialization of Cuts

The experiments have been carried out with images of typewritten digits, taken from real forms and Microsoft sources. 14,750 test digits, out of a total of 100,000, of different real sources, have been considered. We have cut the 14,750 test digits with different percentages in order to have a significant database of cut digits. From this set, 6 different sets of digits have been formed: Digits with a different percentage of cutting on their upper and lower parts (10%, 20% and 30%). Therefore, the total number of digits is 103,250.

The reference set used for all specialists, has been created ad-hoc, with well defined or conformed images of typewritten digits of Microsoft sources.

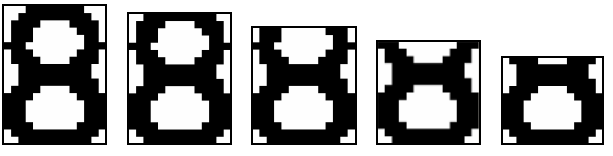

Fig. 3. Example of data base used

The simulator of the multi-classifier used has enabled us to specify the parameters of each specialist as well as their combination parameters.

A finite number of specialists has been used in the diverse experiments which are differentiated in the parameters of extraction of characteristics (M_k and N_k).

The real system used to isolate the digits, provide information about the position of the lack of structure. Therefore, the recognition process has information about the quality of the image, whether is cut or is not cut. The management of this information is important to increase the reliability of the system. We have made an experimental

Table 1. Specification of the specilists used in the study. The Parameters M_k, N_k, are the zoning matrix rows and columns respectively and P_k is the situation of the lack of structure.

S_k	Description (specialisation)	M_k	N_k	P_k
S1	Image without cuts	8	5	∅
S2	Little cut upper part	7	5	Upper
S3	Little cut lower part	7	5	Lower
S4	Moderate cut upper part	6	5	Upper
S5	Moderate cut lower part	6	5	Lower
S6	Excessive cut upper part	5	5	Upper
S7	Excessive cut lower part	5	5	Lower

study taking into account that the system has the possibility to know this information and more than this information, in order to make a general study of the system. We have considered three cases: first, the system has all the possible information of the digit (*Case A*), the position of the lack and the amount of the lack. Second, the system has only the information about the position of the lack of structure (*Case B*) and finally the system has not any information about the structure of the digit (*Case C*).

So, the knowledge of the information is used to know what is the best specialist to recognize an input pattern.

4.1 Specialisation of the Classifiers

The experiments have been carried out with the different test sets both separately and jointly. Therefore, primarily we are going to see which are the maximum percentages of success for each separate specialist, which means that each specialist will form an only independent system to observe how it reacts in front of all the disturbances. The following table shows this study.

Table 2. Results for the specialization. The last column shows the average for each specialist separately. The first column shows where the tested digit is cut and the last line shows the average covering the best results (*Case A*).

		No Cut	**Little(10%)**	**Moderate(20%)**	**Excessive(30%)**
Upper	S_1	**99,76**	98,03	76,06	43,57
	S_2	95,8	**99,23**	95,6	84,95
	S_4	73,93	93,76	**98,22**	**97,8**
	S_6	36,68	61,21	89,77	96,8
Lower	S_1	**99,76**	98,97	93,03	67,58
	S_3	94,27	**99,06**	**99,36**	94,46
	S_5	60,73	83,43	98,26	**99,27**
	S_7	37,91	51,75	80,8	96
Case A		*99.76*	*99.15*	*98.79*	*98.54*

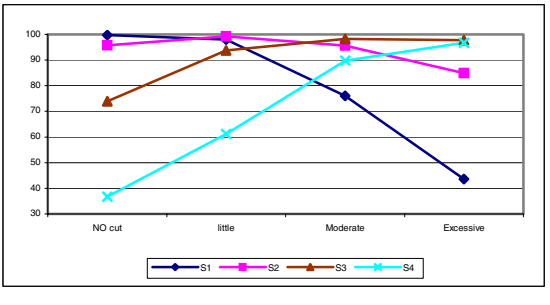

Fig. 4. The axis X: cut level. Axis Y: % recognition for each specialist classifier.

In the table 2 we can see how each system better recognises a class of images (it is a specialist). If we knew beforehand which is the best specialist for each digit, the obtained recognition would be 99.06% (the average of the row *Case A*) against 84.6% (the average of the row S_1) as obtained only with the specialist S_1 (specialised in well conformed images). It is plain to see how S_1 recognises, more or less accurately, uncut digits and not very cut digits (10%), but when the images are badly cut, the recognition rate decreases considerably. The behaviour of the other classifiers is totally different, as shown in the figure 4.

4.2 Adaptability of the System to the Loss of Information

In the following experiment, the system does not know how much cut the image is and which the best specialist for the pattern is. It only knows that the digit is cut on its upper or lower part. Therefore, the specialists for the digits cut on their upper or lower part will compete in a multi-classifier (two multi- specialists with four specialists each one), $\boldsymbol{MSC_1}$ *(1,* $\{\{S_1, S_2, S_4, S_6\}\}$*)* and $\boldsymbol{MC_2}$ *(4,* $\{S_1, S_3, S_5, S_7\}$*)*, both multi-classifiers jointly conform a unique system where the images with cuts in the upper part are classified with MSC_1 and the images cut in the lower part are managed with MSC_2. The results for this compound system is the case 2 explained previously and it is calculated like and average of $\boldsymbol{MSC_1}$ and $\boldsymbol{MSC_2}$.

Table 3. Results with lose of information. Each cell represent the percentage of patterns well recognized with a specialist and with a kind of images. The MSC_2 and MSC_1 rows are the hit rates for each type of images with MSC_2 and MSC_1 classifers repectively. The last row are the average results for MSC_2 and MSC_1 rows (*Case B*).

		No Cut	**Little**	**Moderate**	**Excessive**
MSC_2	S_1	88,01	34,41	10,93	8,88
	S_2	9,65	59,27	39,39	7,75
	S_4	1,23	4,03	43,60	56,61
	S_6	0,82	1,80	4,92	22,75
	$\boldsymbol{MSC_2}$	**99,70**	**99,53**	**98,84**	**95,99**
MSC_1	S_1	93,15	49,61	8,58	3,59
	S_3	6,06	47,50	57,20	17,79
	S_5	0,26	2,35	31,52	60,69
	S_7	0,01	0,03	1,97	13,27
	$\boldsymbol{MSC_1}$	**99,48**	**99,49**	**99,27**	**95,34**
Case B		**99,6**	**99,51**	**99,1**	**95,67**

Table 3 shows the percentages of success for each type of images, the results for each specialists and the accumulated results for both systems. A lower recognition average of the table is shown (*Case B vs Case A*), because when the digit is very cut, specialists S_1,S_2 or S_3 can confuse the test pattern with another class. For example, a digit of class 8 excessively cut resembles a non cut 0. When the digit, which by now is very cut, the specialists S_1, S_2 or S_3, classifies almost all the patterns because the certainty of the specialist for this pattern is the most adequate. If the pattern is very cut, specialists S_4, S_5, S_6, and S_7, become more important, in spite of specialists S_1, S_2,

and S_3, are still classifying some patterns. The average recognition for this experiment is 98.45% (average of the row *Case B*) against 84.6% corresponding to all digits against only specialist S_1, even if the recognition rate is lower than in the previous experiment (99.06% in *Case A*).

In the multi-specialist system of the following experiment there is a specialist S_1 specialised in uncut digits and specialised in digits cut in the upper and lower parts, S_2, S_3, S_4, S_5, S_6, S_7.

The multi-classifier is defined as $\boldsymbol{MSC_3}$ *(7, {*S_1, S_2, S_3, S_4, S_5, S_6, $S7$, *}).*

The Table 4 shows the percentages of success for each specialist, the accumulated addition for each column represents the percentage of success for each type of cut, the right column represents the average of the system.

Table 4. Results for *MSC3.* L: Little, M: Moderate, E: Excesive. Each cell represent the percentage of patterns well recognized with a specialist and with a kind of images. Last row is the percentage recognition for each tipe of images with MSC_3.

		Images cut upper			**Images cut lower**		
Specialist	**No cuts**	**L**	**M**	**E**	**L**	**M**	**E**
S1	82,62	45,44	7,58	3,27	31,90	7,81	5,59
S3	9,29	3,29	1,04	0,35	58,89	38,81	7,33
S5	1,15	1,08	0,27	0,08	3,86	43,44	56,21
S7	0,74	0,30	0,03	0,02	1,60	4,20	22,29
S2	5,53	46,92	56,70	17,27	2,78	3,59	3,07
S4	0,24	2,32	31,36	59,59	0,19	0,47	0,55
S6	0,01	0,03	1,97	12,84	0,00	0,00	0,00
MSC3	**99,57**	**99,39**	**98,95**	**93,43**	**99,22**	**98,33**	**95,04**

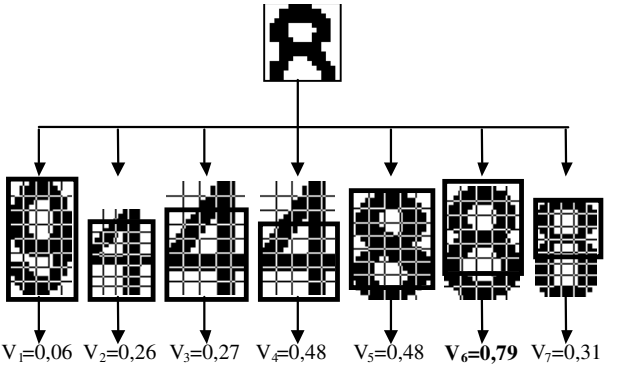

Fig. 5. System behaviour with cut digits. The image shows the reference pattern and the confidence value for each specialist.

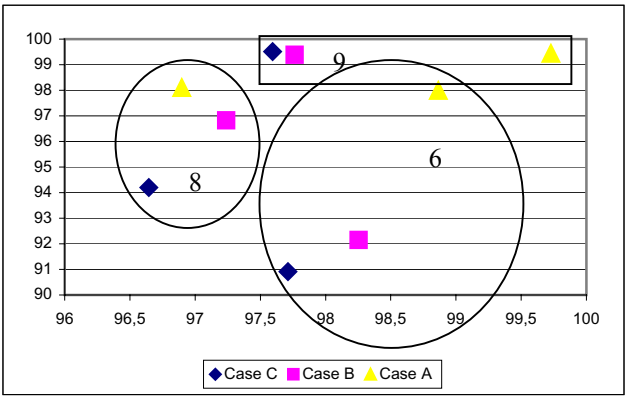

Fig. 6. Axis X is the percentage of success for digits cut in their lower part, and axis Y is the percentage of success of digits cut in their upper part. *Case A* knows where and when the digit is cut, *Case B* only knows where the digit is cut and *Case C* does not know anything, only that the digit can be cut or well conformed.

These results clearly show that each specialist is, generally speaking, a specialist in a class of patterns, even if sometimes recognition is shared with other specialist. 82.62% of the uncut images are well classified by specialist S_1. Images moderately cut in the lower part are mainly correctly recognised by specialists S_3 and S_5. The total average of the recognition rate is 97.70% (*Case C*).

Figure 5 shows how the space dimensionality variation helps to recognize cut digits. The specialist S_6 has the best confidence value because it only uses the first 6 rows of its features to calculate the Euclidean distance and the features of the entry pattern in same dimensionality space are taken into consideration.

The difficulty to recognise some classes increases with the number of specialists in the system. It is due to the interference between specialists. The figure 6 shows this behaviour.

The figure 6 represents a study made by classes of the specialisation of each system. Same as in case 8 and 6, the recognition is lower because when the digit is very cut it looks like a 0, and the same happens when it is cut on its upper part.

5 Conclusions

The noisy patterns has been one of the problems in the pattern recognition. While the human recognition have not too much problems to recognize this kind of patterns, OCR systems have poor recognition in this field.

For us the specialization in the recognition, for all kind of noise is the base. We treat each case of noise separately in parallel and then a criterion decide. So the system provide two results, the expected class of the pattern and what kind of noise has the image, depending on which specialist has responded.

We have study different specializations, the difference of each one is the knowledge of the context. If the type of noise is known the recognition is easier because the system know what is the most efficient specialist. On the other hand when the context

information is smaller the specialists have to decide between them in a multi-classifier. This is important in some segmentation systems where the knowledge of the context is possible.

We have obtained a recognition rate of 97.70% with cut digits, even though we can improve the recognition rate to 98.45% using information provided by the segmentation process, the non specialist system obtains for the same data 84.6%.

References

1. Whichello, A. P.,Yan, H.: Linking broken character borders with variable sized masks to improve recognition. Pattern Recognition Vol. 29 (8) (1996) 1429-1435
2. Rodriguez, C., Muguerza, M., Navarro, M., Zárate, A., Martín, J. I.,Pérez, J. M.: A two-stage classifer for broken and blurred digits in forms. ICPR 98 Brisbane, Australia Vol. 2 (1998) 1101-1105
3. Omachi, S., Sun, F.,H., A.: A noise-adaptive discriminant function and its application to blurred machine-printed kanji recognition. IEEE Transactions PAMI Vol. 22 (3) (2000) 314-319
4. Kittler, J., Hated, M., Duin, R. P. W.,Matas, J.: On combining classifiers. IEEE Transactions PAMI Vol. 20 (3) (1998) 226-239
5. Rodriguez, C., Soraluze, I., Muguerza, J., Martín, J. I.,Álvarez, G.: Hierarchical classifiers based on neighbourhood criteria with adaptive computational cost. Pattern Recognition Vol. 35 (12) (2002) 2761-2769
6. Alpaydin, E., Kaynak, C.,Alimoglu, F.: Cascading multiple classifiers and representations for optical and pen-based handwritten digit recognition. 7th IWFHR Amsterdam Vol. (2000) 453-462
7. Ho, T. K., Hull, J. J.,Srihari, S.: Decision combination in multiple classifier systems. IEEE Transactions PAMI Vol. 16 (1) (1994) 66-75
8. Bauer, E.,Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning Vol. 36 (1-2) (1999) 105 - 139
9. Aksela, M., Girdziusas, R., Laaksonen, J., Oja, E.,Kangas, J.: Class-confidence critic combining. 8th IWFHR Ontario, Canada Vol. (2002) 201-206
10. Ha, T. M.,Bunke, H.: Off-line, handwritten numeral recognition by pertubation method. IEEE Transactions PAMI Vol. 19 (5) (1997) 535-539
11. Erp, M. v., Vuurpijil, L.,Shomaker, L.: An overview and comparison of voting methods for pattern recognition. 8th IWFHR Ontario, Canada Vol. (2002) 195-200
12. Yu, D.,Yan, H.: Reconstruction of broken handwritten digits based on structural morphological features. Pattern Recognition Vol. 34 (2001) 235-254
13. Wang, J.,Yan, H.: Mending broken handwriting with a macrostructure analysis method to improve recognition. Pattern Recognition Letters Vol. 20 (1999) 855-864
14. Cortés, A., Boto, F.,Rodriguez, C.: Noisy digit classification with multiple specialist. Pattern Recognition and Data Mining (LNCS 3686) Vol. 1 (2005) 601-608
15. Dasarathy, B. V.: Nearest neighbor (nn) norms: Nn pattern classification techniques. I. C. S. Press (1991)
16. Devroye, L., Györfi, L.,Lugosi, G.: A probabilistic theory of pattern recognition. N. Y. Springer-Verlag (1996)
17. Rahman, A. F. R.,Fairhurst, M. C.: Multiple classifier decision combination strategies for character recognition: A review. International Journal on Document Analysis and Recognition Vol. 5 (4) (2003) 166-194