Handwritten Digit Recognition by Combining SVM Classifiers

Dejan Gorgevik, *Member, IEEE* and Dusan Cakmakov

***Abstract* — Recent results in pattern recognition have shown that SVM (Support Vector Machine) classifiers often have superior recognition rates in comparison to other classification methods. In this paper, a cooperation of four SVM classifiers for handwritten digit recognition, each using different feature set is examined. We investigate the advantages and weaknesses of various cooperation schemes based on classifier decision fusion using statistical reasoning. The obtained results show that it is difficult to exceed the recognition rate of a single, well-tuned SVM classifier applied straightforwardly on all feature sets. In our experiments only one of the cooperation schemes exceeds the recognition rate of a single SVM classifier. However, the classifier cooperation reduces the classifier complexity and need for training samples, decreases classifier training time and sometimes improves the classifier performance.**

***Keywords* — classifier, decision fusion, features, statistical.**

I. Introduction

Combining features of different nature and the corresponding classifiers has been shown to be a promising approach in many pattern recognition applications. Data from more than one source that are processed separately can often be profitably re-combined to produce more concise, more complete and/or more accurate situation description. In this paper, we discuss classification systems for handwritten digit recognition using four different feature sets and SVM classifiers [1]. We start with a SVM classifier applied on all features as one set. Further, we used four SVM classifiers that work on the different feature sets for the same digit image. As the feature sets "see" the same digit image from different points of view, we examined the possibility of decision fusion using statistical cooperation schemes. An extensive number of cooperation schemes were examined and corresponding recognition results are presented. Our aim was not to compete with the recognition rates of the other handwritten digit recognition systems e.g. [2], [3], but to compare the qualities of different feature sets, corresponding SVM classifiers and their combination based on different decision fusion schemes.

The presented results show that it is difficult to achieve the recognition rate of a single optimized SVM classifier applied on the feature set that includes all features by combining the individual SVM decisions. On the other hand, the cooperation of individual classifiers designed for separate feature sets reduce the classifier complexity and need for training samples, offering better opportunity to understand the role of the features in the recognition process.

This work is supported by the Ministry of Sciences of the Republic of Macedonia, Contract 13-1556/4-02 from 26.11.2003.

D. Gorgevik is with the Department of Computer and Information Technology, Faculty of Electrical Eng., University "Sv. Kiril i Metodij", Karpos II bb, POBox 574, 1000 Skopje, Macedonia; (e-mail: dejan@etf.ukim.edu.mk).

D. Cakmakov is with the Department of Mathematics and Computer Science, Faculty of Mechanical Eng., University "Sv. Kiril i Metodij", Karpos II bb, POBox 464, 1000 Skopje, Macedonia; (e-mail: dusan@mf.ukim.edu.mk).

II. The System Architecture

The recognition system is constructed around a modular architecture of feature extraction and digit classification units. The preprocessed isolated digit images are input for the feature extraction module that transfers the extracted features toward SVM classifiers (Fig. 1).

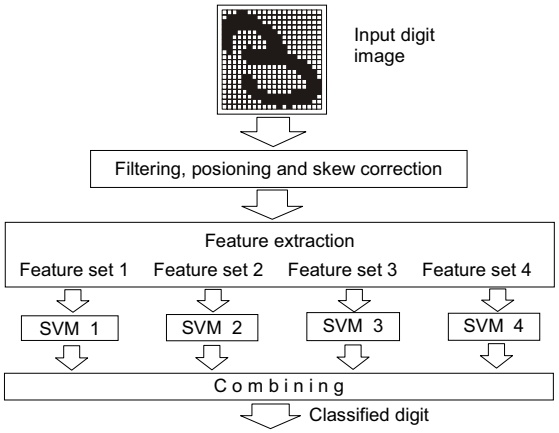

Fig. 1. The system architecture

Each image is centered in a square bounding box, and then slant correction is performed. The slant angle is estimated as the inclination of the line connecting the gravity centers of the top 25% part and the bottom 25% part of the image (Fig. 2a). Then a sub-pixel precision shear transformation is performed in order to remove the estimated inclination.

Four feature sets were extracted from each digit image:

- projection histograms,
- contour profiles,
- ring-zones and
- Kirsch features.

Feature extraction was performed on the original unscaled image, after the slant correction.

The first 23 features (FS1) are simple horizontal, vertical and diagonal projection histograms. Since not all of the character images were of the same size, the projection vectors were linearly rescaled in order to obtain 7 features from the horizontal projections, 6 features from the vertical projections, and 5 features from each of the two diagonal projections (Fig. 2b).

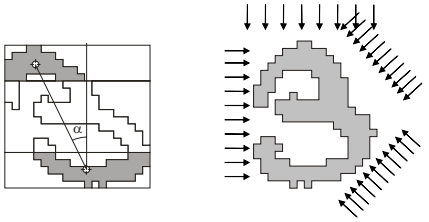

Fig. 2. (a) Slant correction; (b) Projection histograms

The second feature set (FS2) is composed of 30 contour profile features (Fig. 3). The image is scanned from left to right, top to bottom, right to left and bottom to top, respectively. The distance from the corresponding edge of the image to the first black pixel which the scanning line intersects, represent the contour profile features on the first level. The distance to the first black pixel of the second black pixel run represent the contour profile features on the second level. Since not all of the character images were of the same size, the profile vectors were linearly rescaled in order to obtain 6 features from the left and right contour profiles and 5 features from the upper and lower profiles on the first level of the digit image. Finally, 4 features were extracted from the upper and the lower contour profiles of the second level.

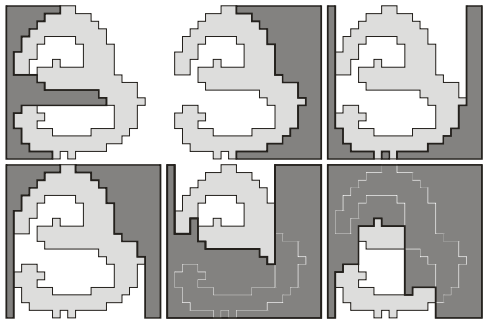

Fig. 3. Contour profiles of first and second level

The third feature set (FS3) contains 44 features extracted as pixel counts in rings zones around the gravity center of the image (Fig. 4). We have used three rings, each divided in different number of equal zones. The outermost ring has a radius r equal to the distance from the gravity center to the furthest black pixel of the image. The first ring with radius 0.2·r provides 4 features and the second ring with radius 0.5·r provides 24 features. The last 16 features of this group are provided from the outermost ring.

The last group of 72 features (FS4) use Kirsch operator [4] to detect local directional information of the edges of the input pattern. Compared to chain codes that also describe the edge direction, Kirsch edge detection is more robust even under noisy conditions.

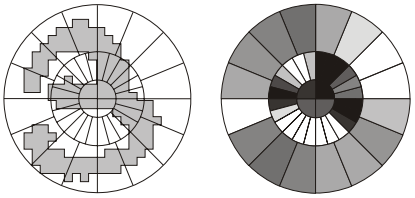

Fig. 4. Ring-zone features

The first black pixel which the scanning line intersects forms the first outermost periphery. The second black pixel which is the starting point of the second black pixel run forms the second outermost periphery (Fig. 5). When the image is scanned in horizontal direction, the vertical and both diagonal Kirsch features are extracted at the outermost periphery. When the image is scanned in vertical direction, the horizontal and both diagonal Kirsch features are extracted at the outermost and second outermost periphery. This way, 3 Kirsch directional features are provided for each periphery pixel. The feature vectors are again linearly rescaled to 15 features coming from the left and right periphery each, 12 features coming from the first outermost top and bottom periphery each, and 9 features coming from the second outermost top and bottom peripheries.

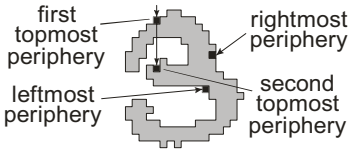

Fig. 5. Kirsch features

Kirsch feature extraction is performed on the grayscale digit images using sub-pixel precision. All parameters including the number of features by projection, the radiuses of rings for zone-pattern regions and the number of features coming from the outermost peripheries for Kirsch features are carefully chosen after several iterations using observations about their discriminative power. The features were preprocessed for zero mean and unit variance.

III. The Recognition Results

Our experiments were performed on an extract of the well-known NIST (National Institute of Standards and Technology) handwritten digit database. This database is consisted of 7 partitions denoted as: hsf_0, …, hsf_4, hsf_6 and hsf_7. Digit images from the hsf_0 partition were used for classifier training while the tuning of classifier parameters (kernel width σ and penalty C) was performed using the hsf_1 partition. The final recognition rates were estimated on most difficult partition hsf_4. So, the samples in the test set belong to different writers from those in the learning set.

We used SVMs with Gaussian kernel because it provided better recognition rates then linear, polynomial or sigmoidal kernel. Because of the large number of samples we have used SVMTorch that is a more robust variation of SVM software library [5].

The recognition rates of different classifier cooperation schemes applied on the described 4 feature sets: FS1, FS2, FS3 and FS4 are given in Table 1. In the second column the corresponding cooperation scheme is given, followed by the recognition rate and the rank of the cooperation scheme when combining classifiers are trained using 2000, 10000, 30000 and all 53449 available samples.

The first 4 rows show recognition rates of each feature set individually. The classifier dependency coefficient [6] of the four individual classifiers is given in row a). The row b) gives the recognition rates of a single optimized SVM classifier applied on the four feature sets as a whole. The row c) gives the recognition rate of a hypothetical "oracle" cooperation scheme that knows to choose the right class if it is predicted by at least one of the member classifiers. This gives the theoretical upper bound of the recognition rate achievable by classifier decision fusion.

Some of the decision fusion methods like: Product, Dempster Rule, Fuzzy Integral, and Decision Templates require possibilistic outputs. To map the original output values to [0, 1] interval we used the mapping $1/(1+e^{-x})$.

Cooperation schemes that need no extra parameters to perform the fusion of the individual classifiers' outputs into a single decision are known as fixed cooperation schemes (1,2,5-11). More advanced cooperation schemes utilize extra parameters in the process of fusion, trying to utilize individual classifier advantages and their dependencies. These parameters are usually designated in the process of so called combiner training. Such cooperation schemes are also known as trained cooperation schemes.

The cooperation schemes 1-4 are voting schemes including variations of the Borda count that is a generalization of the majority vote [7]. The cooperation schemes 5-12 use various averages, maximum and minimum selectors of the corresponding classifier outputs to make the final decision [8]. The Dempster Rule [9] and a few variations [10] are given in rows 13-22. The naive Bayes cooperation scheme given in rows 23-24 uses the confusion matrices of member classifiers to estimate the certainty of the classifier decisions [11]. The fuzzy integration 25-26 is based on searching for the maximal grade of agreement between the objective evidence (provided by the sorted classifier outputs for *i*-th class) and the expectation (the fuzzy measure values of all classifiers) [12]. We have also used a variety of decision template schemes 27-31 described elsewhere [13]. The generalized committee prediction and its variations 32-36 are based on a weighted combination of the predictions of the member classifiers [14]. Cooperation scheme 37 uses multivariate linear regression to make decision fusion. In the cooperation scheme 38 the 4 individual SVM outputs (40 features) are input to another SVM classifier. This kind of cooperation is also known as classification task [9].

Table 1 shows that the cooperation 38 (svmcmb) has unbeatable recognition rate in all cases. However, this method is most complex because it needs additional classifier and additional samples for its training.

TABLE 1: RECOGNITION RATES (%) OF COMBINING SVMS FOR 4 FEATURE SETS AND DIFFERENT SIZES OF LEARNING SET (2000, 10000, 30000 AND ALL 53449 SAMPLES); R STANDS FOR RANK.

		2000	R	10000	R	30000	R	All	R
	FS1	86.16		89.80		91.98		92.77	
	FS2	89.99		93.19		94.88		95.34	
	FS3	89.60		92.41		94.78		95.15	
	FS4	91.52		94.10		95.71		96.10	
a)	**cdc**	0.628		0.701		0.736		0.767	
b)	**Single Opt. SVM**	**94.10**		**95.84**		**97.15**		**97.27**	
c)	**oracle**	**96.88**		**97.88**		**98.47**		**98.70**	
1	**vote**	92.26	35	94.37	35	96.05	36	96.32	35
2	**borda**	92.73	27	94.85	30	96.28	30	96.62	25
3	**bks**	92.54	31	94.04	37	95.40	38	95.72	38
4	**bksv**	93.52	7	95.01	21	96.27	31	96.52	32
5	**avg**	93.01	18	95.03	18	96.38	23	96.66	22
6	**prod**	92.61	29	95.09	16	96.54	15	96.71	19
7	**harm**	92.36	34	94.99	26	96.44	21	96.61	28
8	**cprod**	92.38	33	94.89	28	96.45	20	96.69	20
9	**maxmax**	91.58	37	94.25	36	96.08	35	96.27	36
10	**minmax**	91.81	36	94.66	34	96.21	34	96.36	34
11	**med**	92.87	23	95.02	19	96.41	22	96.67	21
12	**davg**	92.93	20	94.99	25	96.31	26	96.62	26
13	**demp**	92.85	24	94.89	29	96.24	33	96.61	27
14	**dempas**	93.56	6	95.22	9	96.54	16	96.94	5
15	**dempchi**	93.10	15	95.18	12	96.54	17	96.75	13
16	**dempchi2**	93.14	14	95.20	11	96.60	8	96.79	12
17	**dempbc**	93.01	19	95.11	14	96.55	13	96.72	17
18	**demphl**	93.25	9	95.26	7	96.58	10	96.79	11
19	**dempchr**	93.68	4	95.22	10	96.53	18	96.93	7
20	**dempchr2**	93.23	12	95.25	8	96.59	9	96.80	10
21	**dempjac**	92.90	22	95.06	17	96.49	19	96.71	18
22	**dempper**	93.05	17	95.16	13	96.55	11	96.74	15
23	**pprod**	92.74	26	95.10	15	96.54	14	96.73	16
24	**bayes**	93.24	11	95.01	22	96.32	24	96.66	23
25	**fi**	92.62	28	94.78	32	96.27	29	96.55	30
26	**fic**	92.47	32	94.74	33	96.30	28	96.53	31
27	**dtp1**	93.06	16	95.00	23	96.27	32	96.48	33
28	**dtp2**	93.18	13	94.97	27	96.31	27	96.61	29
29	**dtp3**	93.26	8	95.02	20	96.32	25	96.63	24
30	**dtp4**	91.13	38	93.90	38	95.50	37	95.85	37
31	**dtm**	94.48	2	95.77	2	96.74	4	96.93	6
32	**epw**	92.59	30	95.00	24	96.55	12	96.74	14
33	**gc**	92.92	21	95.28	6	96.63	7	96.86	9
34	**mgc**	93.24	10	95.44	5	96.64	6	96.91	8
35	**ogc**	93.56	5	95.54	4	96.82	2	97.04	2
36	**omgc**	92.80	25	94.82	31	96.70	5	97.01	3
37	**mlr**	94.44	3	95.75	3	96.76	3	96.95	4
38	**svmcmb**	97.36	1	97.48	1	97.78	1	97.82	1

Increasing the number of training samples, indeed increases recognition rates of individual classifiers and their cooperation. On the other hand, increasing recognition rates of individual classifiers also increases their correlation that reduces the possibility for improvement of the cooperation recognition rates.

Voting cooperation schemes (1-4) are among worst because they use most limited information of member classifiers, ignoring useful information about second choices, reliability of the choice, distribution of the choices for different classes, etc.

The simplest cooperation schemes (5-12) as we expected, have average recognition rates and should be used in not demanding applications.

It is interesting that Dempster Rule and its variations (13-22) have in average better recognition rates than decision templates schemes (27-31).

The naive Bayes cooperation schemes (23-24) are relatively good choice while the fuzzy integration (25-26) shows weak results.

The generalized committee prediction and its variations (32-36), together with multivariate linear regression (37) are among the best methods and should be considered as serious candidates for implementation in any serious pattern recognition application based on classifier cooperation.

It is interesting to see how accuracy of classifier cooperation methods depend on the accuracies of individual classifiers (Fig. 6).

In case of trained cooperation schemes such as: decision templates (27-31), linear regression (37) or classification task (38) the cooperation accuracy is far less dependent of the individual classifiers accuracies comparing to the fixed cooperation schemes such as: voting (1-4) or averages (5-12). Our experiments show that as long as we can dispose enough samples for cooperation training, the accuracies of the individual classifiers become less important.

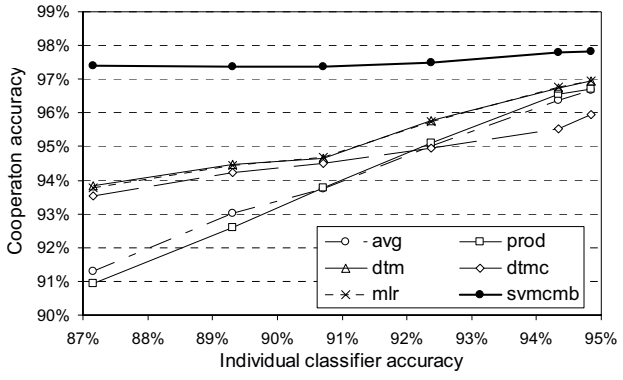

Fig. 6. Accuracy of some of classifier cooperation method as a function of the accuracies of individual classifiers

Fig. 6 shows that accuracy of classification task (38) is practically independent of the accuracies of the individual classifiers. In case of the other trained cooperation schemes this dependence is relatively week, while in case of fixed cooperation, this dependence is almost linear.

IV. Conclusion

In this paper, the cooperation of four feature sets for handwritten digit recognition using SVM classifiers is examined. We investigate an extensive number of cooperation schemes based on classifier decision fusion.

The presented results show that it is difficult to achieve the recognition rate of a single SVM applied on the feature set that includes all features by combining the individual SVM decisions. These results impose the crucial question: whether the methods for classifier cooperation are still needed [15] or pattern recognition tasks could be better solved by a single, well-optimized SVM classifier. However, the classifier cooperation schemes reduce the classifier complexity and need for samples, and sometimes can increase the classifier performance.

References

[1] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Knowledge Discovery and Data Mining*, Vol. 2, 1998, pp. 1–47.

[2] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," In F. Fogelman and P. Gallinari, editors, *International Conference on Artificial Neural Networks*, Paris, 1995, pp. 53-60.

[3] D. Gorgevik, D. Cakmakov, "An Efficient Three-Stage Classifier for Handwritten Digit Recognition," *Proc. of 16th Int. Conference on Pattern Recognition*, Vol. 4, IEEE Computer Society, Cambridge, UK, 23-26 August 2004, pp. 507–510.

[4] W. K. Pratt, *Digital Image Processing*, PIKS Inside, Third Edition, John Wiley & Sons, Inc., 2001.

[5] R. Collobert, S. Bengio, J. Mariéthoz, "Torch: a modular machine learning software library," Technical Report IDIAP-RR 02-46, Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), CH-1920 Martigny, Switzerland, 2002. Available: www.torch.ch.

[6] Catherine A. Shipp, Ludmila I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers", *Information Fusion*, Vol. 3, No. 2, 2002, pp. 135-148.

[7] T.K. Ho, J.J. Hull, S.N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 1, January 1994, pp. 66–75.

[8] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, March 1998, pp. 226–239.

[9] J. Schürmann, *Pattern Classification: A Unified View of Statistical and Neural Approaches*, Johm Wiley & Sons, Inc., 1996.

[10] D. Gorgevik, "Classifier Combining for Handwritten Digit Recognition," Ph.D. dissertation, Faculty of Electrical Engineering, Skopje, Macedonia, June 2004.

[11] L. Xu, A. Krzyzak, C.Y. Suen, "Methods of combining multiple classifiers and their application to handwritten recognition," *IEEE Transactions on System, Man and Cybernetics*, Vol. 22, 1992, pp. 418-435.

[12] S.B. Cho, J.H. Kim, "Combining multiple neural networks by fuzzy integral and robust classification," *IEEE Transactions on System, Man and Cybernetics*, Vol. 20, No. 3, 1995, pp. 380–384.

[13] L.I. Kuncheva, J.C. Bezdek, P.W. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Recognition*, Vol. 34, No. 2, 2001, pp. 299-314.

[14] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford, Clarendon Press, 1995, pp. 364–369.

[15] J. Kittler, "A Framework for Classifier Fusion - Is It Still Needed," in F. J. Ferri, J. M. Inesta, A. Amin and P. Pudil, Eds., Advances in Pattern Recognition, *Lecture Notes in Computer Science*, Vol. 1876, Springer-Verlag, 2000, pp. 45–56.