METHODS OF COMBINING SVMs APPLIED TO HANDWRITTEN DIGIT RECOGNITION

Dejan Gorgevik[1], Dusan Cakmakov[2]

[1] Faculty of Electrical Engineering – Skopje, dejan@etf.ukim.edu.mk

[2] Faculty of Mechanical Engineering – Skopje, dusan@mf.ukim.edu.mk

***Abstract* – In this paper, the cooperation of four feature families for handwritten digit recognition using SVM (Support Vector Machine) classifiers is examined. We investigate the advantages and weaknesses of various cooperation schemes based on classifier decision fusion using statistical reasoning. Although most of presented cooperation schemes are variations and adaptations of existing ones, such an extensive number of investigated classifier decision fusion schemes have not been presented in the literature until now. 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 families by combining the individual classifier decisions. In our experiments only one of the cooperation schemes managed to exceed the recognition rate of a single SVM classifier. However, using classifier cooperation reduces the classifier complexity and need for training samples, decreases classifier training time and sometimes improves the classifier performance.**

***Keywords* – structural, statistical, features, classifier, decision fusion**

1. 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 families and SVM classifiers [2]. We start with a SVM classifier applied on all feature families as one set. Further, we used four SVM classifiers each working on different feature family from 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 statistical 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. [12], [6], but to compare the qualities of different feature families, corresponding SVM classifiers and their combination based on different classifier decision fusion schemes.

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

2. 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 (see *Figure 1*).

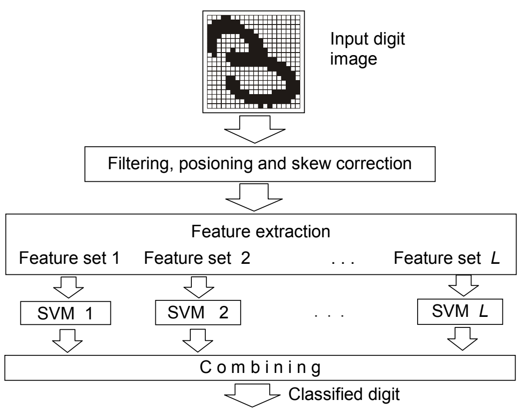

Figure 1. The system architecture

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 (see *Figure 2*). Then a sub-pixel precision shear transformation is performed in order to remove the estimated inclination.

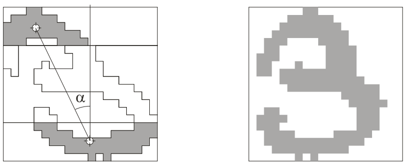

Figure 2. Skew detection and correction

This approach showed to provide more reliable slant correction than the "standard" approaches e.g. [7].

3. FEATURE EXTRACTION

Four feature families are extracted from each digit image: projection histograms, contour profiles, ring-zones and Kirsch features.

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 (Figure 3).

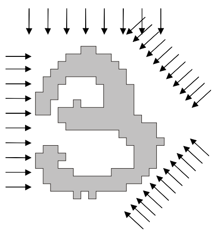

Figure 3. Projection histograms

The second feature family (FS2) is composed of 30 contour profile features (see Figure 4). 6 features were extracted from the left and right contour profiles, while 5 features were extracted from the upper and lower profile of the digit image. Finally, 4 features were extracted from the upper and the lower contour profiles of the second level.

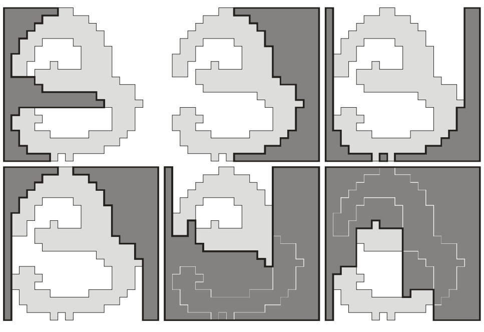

Figure 4. Contour profiles of first and second level

The third group contains 44 features (FS3) extracted as pixel counts in rings zones around the gravity center of the image (see *Figure 5*). 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 $0.2 \cdot r$ provides 4 features and the second ring with radius $0.5 \cdot r$ provides 24 features. The last 16 features of this group are provided from the outermost ring.

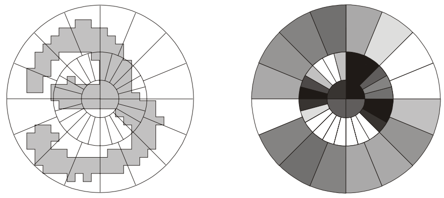

Figure 5. Ring-zone features

The last group of 72 features (FS4) use Kirsch operator to detect local directional information of the edges of the input pattern [13]. Compared with chain code which also describes the edge direction, Kirsch edge detection is more robust even under noisy conditions.

The image is scanned from left to right, top to bottom, right to left and bottom to top, respectively. 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 (see *Figure 6*). 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.

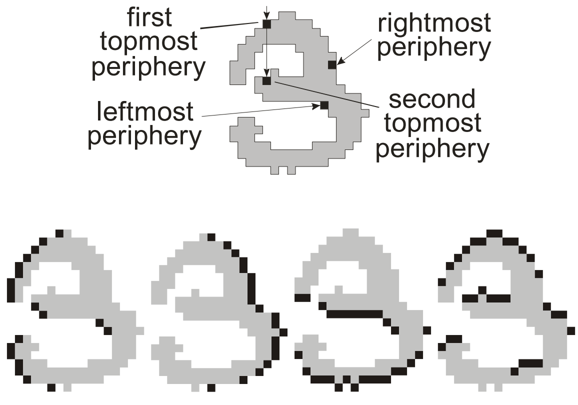

Figure 6. Kirsch features

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 were carefully chosen after several iterations using observations about their discriminative power. The features were preprocessed for zero mean and unit variance.

4. 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 for validation. 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. On Figure 7 a fragment of the NIST database is given.

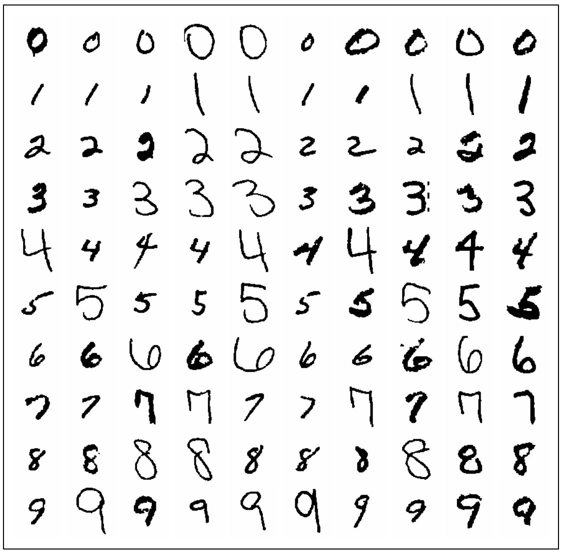

Figure 7. A fragment from the NIST database

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 training software library [4].

Some of the decision fusion methods: 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 $1/(1+e^{-x})$.

The recognition rates of different classifier cooperation schemes applied on selected 4 feature sets: FS1, FS2, FS3 and FS4 are given in Table 1.

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

The cooperation schemes 1-4 are voting schemes including variations of the Borda count that is a generalization of the majority vote [8]. The 5-12 cooperation schemes use various averages, maximum and minimum selectors of the corresponding classifier outputs to make the final decision [9]. The Dempster Rule [14] and many variations [5] are given in rows 13-39. The naive Bayes cooperation scheme given in rows 40-41 uses the confusion matrices of member classifiers to estimate the certainty of the classifier decisions [15]. The fuzzy integration 42-43 is based on searching for the maximal grade of agreement between the objective evidence (provided by the sorted classifier outputs for the *i*-th class) and the expectation (the fuzzy measure values of all classifiers) [3]. We have also used a variety of decision templates schemes 44-70 described elsewhere [11][5]. The generalized committee prediction and its variations 72-75 are based on a weighted combination of the predictions of the member classifiers [1]. Cooperation scheme 76 uses multivariate linear regression to perform decision fusion. In the cooperation scheme 77 the 4 individual SVM outputs (40 features) are input to another SVM classifier. This kind of cooperation is also known as classification task [14].

Table 1 shows that cooperation 77 (svmcmb) has unbeatable recognition rate in all cases. However, this method is most complex because it uses additional classifier and additional samples for its training. Because of additional number of samples used in the training process, this method sometimes outperforms even the “oracle method”.

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 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 (13-39) have in average better recognition rates than decision templates schemes (44-70).

Table 1. *Recognition rates (%) of combining SVM classifiers for 4 feature families (FS1, FS2, FS3, FS4) 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)	**Single Opt. SVM**	**94.10**		**95.84**		**97.15**		**97.27**	
b)	**oracle**	**96.88**		**97.88**		**98.47**		**98.70**	
1	**vote**	92.26	68	94.37	68	96.05	67	96.32	64
2	**borda**	92.73	54	94.85	58	96.28	49	96.62	40
3	**bks**	92.54	61	94.04	72	95.40	75	95.72	75
4	**bksv**	93.52	12	95.01	31	96.27	52	96.52	53
5	**avg**	93.01	34	95.03	26	96.38	34	96.66	31
6	**prod**	92.61	59	95.09	21	96.54	22	96.71	26
7	**harm**	92.36	67	94.99	43	96.44	31	96.61	45
8	**cprod**	92.38	65	94.89	56	96.45	30	96.69	28
9	**maxmax**	91.58	75	94.25	71	96.08	65	96.27	65
10	**minmax**	91.81	70	94.66	67	96.21	63	96.36	63
11	**med**	92.87	46	95.02	27	96.41	32	96.67	29
12	**davg**	92.93	41	94.99	42	96.31	41	96.62	41
13	**demp**	92.85	47	94.89	57	96.24	58	96.61	44
14	**dempp1**	93.03	33	95.11	17	96.56	15	96.73	22
15	**dempp2**	92.97	39	95.10	20	96.56	15	96.74	20
16	**dempp3**	92.66	55	95.05	24	96.57	13	96.74	18
17	**dempp4**	92.94	40	95.13	15	96.62	8	96.81	10
18	**dempi1**	93.19	19	95.24	9	96.58	11	96.77	14
19	**dempi2**	92.75	52	94.96	50	96.48	29	96.64	34
20	**dempi3**	91.80	71	94.84	59	96.52	26	96.65	33
21	**dempi4**	92.37	66	94.83	60	96.36	36	96.61	43
22	**dempi5**	92.14	69	94.74	64	96.34	37	96.59	48
23	**dempc**	92.75	51	95.12	16	96.57	14	96.73	23
24	**dempmnk**	92.93	43	95.06	23	96.51	27	96.72	24
25	**dempch**	92.78	50	95.04	25	96.55	20	96.76	15
26	**dempcan**	92.81	48	94.28	69	94.88	76	94.77	77
27	**dempas**	93.56	10	95.22	10	96.54	23	96.94	5
28	**dempchi**	93.10	25	95.18	13	96.54	24	96.75	16
29	**dempchi2**	93.14	23	95.20	12	96.60	9	96.79	13
30	**dempbc**	93.01	35	95.11	18	96.55	18	96.72	25
31	**demphl**	93.25	15	95.26	7	96.58	12	96.79	12
32	**dempchr**	93.68	7	95.22	11	96.53	25	96.93	7
33	**dempchr2**	93.23	18	95.25	8	96.59	10	96.80	11
34	**dempjac**	92.90	45	95.06	22	96.49	28	96.71	27
35	**dempper**	93.05	30	95.16	14	96.55	17	96.74	19

		2000	R	10000	R	30000	R	All	R
36	**dempse**	93.55	11	94.98	45	96.29	48	96.55	49
37	**dempfr**	93.69	6	94.95	52	96.24	58	96.49	55
38	**dempm**	92.64	57	94.68	66	96.38	33	96.62	42
39	**dempmc**	93.92	5	94.94	53	96.06	66	96.24	66
40	**pprod**	92.74	53	95.10	19	96.54	21	96.73	21
41	**bayes**	93.24	17	95.01	31	96.32	38	96.66	31
42	**fi**	92.62	58	94.78	63	96.27	50	96.55	50
43	**fic**	92.47	64	94.74	65	96.30	47	96.53	51
44	**dtp1**	93.06	26	95.00	37	96.27	53	96.48	56
45	**dtp2**	93.18	20	94.97	47	96.31	45	96.61	46
46	**dtp3**	93.26	14	95.02	30	96.32	38	96.63	35
47	**dtp4**	91.13	76	93.90	76	95.50	73	95.85	73
48	**dti1**	93.34	13	95.02	28	96.31	41	96.63	35
49	**dti2**	92.49	63	94.91	54	96.08	64	96.20	67
50	**dti3**	92.65	56	94.95	51	96.37	35	96.62	39
51	**dti4**	91.79	72	94.25	70	95.97	69	96.17	68
52	**dti5**	91.69	73	94.00	74	95.75	71	96.11	71
53	**dtc**	91.69	73	94.01	73	95.76	70	96.12	70
54	**dte**	93.05	30	95.01	34	96.32	40	96.67	30
55	**dtmnk**	93.18	20	94.97	47	96.31	45	96.61	46
56	**dtch**	91.13	76	93.90	76	95.50	73	95.85	73
57	**dtcan**	92.50	62	93.99	75	94.77	77	94.87	76
58	**dtas**	92.98	37	95.00	35	96.31	43	96.63	37
59	**dtchi**	93.05	32	95.02	28	96.27	57	96.50	54
60	**dtchi2**	93.11	24	95.01	33	96.27	50	96.52	52
61	**dtbc**	93.06	26	95.00	37	96.27	53	96.48	56
62	**dthl**	92.93	42	94.99	44	96.23	60	96.45	62
63	**dtchr**	92.98	37	95.00	35	96.31	43	96.63	37
64	**dtchr2**	92.98	36	94.98	46	96.22	62	96.48	56
65	**dtjac**	93.06	26	95.00	37	96.27	53	96.48	56
66	**dtper**	93.06	26	95.00	37	96.27	53	96.48	56
67	**dtse**	93.17	22	94.79	62	96.03	68	96.16	69
68	**dtfr**	93.68	8	94.91	55	96.23	61	96.48	56
69	**dtm**	94.48	2	95.77	2	96.74	4	96.93	6
70	**dtmc**	94.23	4	94.96	49	95.54	72	95.96	72
71	**epw**	92.59	60	95.00	41	96.55	18	96.74	17
72	**gc**	92.92	44	95.28	6	96.63	7	96.86	9
73	**mgc**	93.24	16	95.44	5	96.64	6	96.91	8
74	**ogc**	93.56	9	95.54	4	96.82	2	97.04	2
75	**omgc**	92.80	49	94.82	61	96.70	5	97.01	3
76	**mlr**	94.44	3	95.75	3	96.76	3	96.95	4
77	**svmcmb**	97.36	1	97.48	1	97.78	1	97.82	1

The naive Bayes cooperation schemes (40-41) are relatively good choice while the fuzzy integration (42-43) shows weak results.

The generalized committee prediction and its variations (72-75), together with the multivariate linear regression (76) are among the best methods and should be considered as serious candidates for implementation in any serious pattern recognition application based on classifier cooperation.

5. CONCLUSION

In this paper, the cooperation of four feature families 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 feature families by combining the individual SVM decisions. In our experiments only one of the cooperation schemes exceeded the recognition rate of a single SVM classifier. These results impose again the question if the methods for classifier decision fusion are still needed [10] or pattern recognition tasks could be better solved by single, well-optimized e.g. SVM classifier. However, the classifier cooperation schemes reduce classifier complexity and need for samples and sometimes can increase the classifier performance.

6. REFERENCES

[1] C.M. Bishop: *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

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

[3] 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.

[4] Ronan Collobert, Samy Bengio, Johnny 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. (www.torch.ch)

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

[6] Dejan Gorgevik, Dusan 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.

[7] P. J. Grother: Karhunen Loève Feature Extraction for Neural Handwritten Character Recognition, *Proceedings of Applications of Artificial Neural Networks III*, SPIE, Orlando, Florida, 1992, pp. 155-166.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] S.W. Lee: Multilayer Cluster Neural Networks for totally unconstrained handwritten Numeral Recognition, *Neural Networks*, Vol. 8, No. 5, 1995, pp. 783-792.

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

[15] 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.