Experiments with Classifier Combining Rules in Face Expression Recognition Via Sparse Representation.

Pedro Diamel Marrero Fernandez

December 1, 2015

Abstract

The objective of this work is to evaluate the combination ruler performance in the face expression recognition, via sparse representation. A large experiment combining classifiers is reported and discussed. It also includes the combination of classifiers on different feature sets. In addition, various fixed and trained combining rules are used. The results show that the combination rules can increase accuracy of classifiers in the facial expression recognition problems.

1 Introduction

Classifier ensembles are successfully receiving great attention and accolade, not to mention the spawning wealth of research. Theoretical and empirical studies have demonstrated that an ensemble of classifiers is typically more accurate than a single classifier. Research on classifier ensembles permeate many strands machine learning including streaming data, concept drift and incremental learning [2].

The parallel combining of classifiers is computed for different feature sets. This may be especially useful if the objects are represented by different feature sets, when they are described in different physical domains (e.g. sound and vision), or when they are processed by different types of analysis (e.g. moments and frequencies). The original set of features may also be split into subsets in order to reduce the dimensionality and hopefully the accuracy of a single classifier. Parallel classifiers are often, but not necessarily, of the same type.

This work discusses ten combining methods based on proposals [7] [4] [6]. In [7] be include a common probabilistic framework for the following four combination methods: majority vote1 (MV), weighted majority vote (WMV), recall (REC) and naive Bayes (NB). Each combiner is obtained from the previous one when a certain assumption is relaxed or dropped. Proposal [6] is provided with a theoretical underpinning of many existing classifier combination schemes for fusing the decisions of multiple experts, each employing a distinct pattern representation. It has been demonstrated that under different assumptions, and using different approximations we can derive the commonly used classifier combination schemes such as the product rule (PR), sum rule (SR), min rule (RMI), max rule (RMX), median rule (RMD), and majority voting (MV).

The outputs of the input classifiers can be regarded as a mapping to an intermediate space. A combining classifier applied on this space then makes a final decision for the class of a new object. In [4] one version of constrained regression for finding the weights that minimize the variance is derived by assuming the expert's errors in approximating the posterior probability.

The notion of Sparse Representations (SRs), or finding sparse solutions to underdetermined systems, has found applications in a variety of scientific fields. The resulting sparse models are similar in nature to the network of neurons in V1, the first layer of the visual cortex in the human, and more generally, the mammalian brain [12][11]. Patterns of light are represented by a series of innate or learned basis functions, where sparse linear combinations, form a surrogate input stimulus to the brain. Similarly, for many input signals of interest, such as natural images, a small number of exemplars can form a surrogate representation for a new test image.

In SR systems, new test images are efficiently represented by sparse linear coefficients on a dictionary D of over complete basis functions. Specifically, SR systems are comprised of an input sample $x \in \mathbb{R}^m$ along with a dictionary D of n samples, $D \in \mathbb{R}^{m \times n}$. SR solves for coefficients $\alpha \in \mathbb{R}^n$ that satisfy the l_1 minimization problem $x^\star = D\alpha$.

In [13] Ouyang, Yan used the features space Histogram of Oriented Gradients (HoG)[1] and Local Binary Patterns (LBP)[14]. It is based on those two approaches which are complementary, because HOG mainly extracts contour-based shape features while LBP primarily extracts the texture information on gray level images. This feature subspace are combined using PR, SR, RMI, RMX, RMD and MV. This work also tested the features space Gabor wavelet (GW)[16] and the image vector raw (RAW). The Fig. 1 show the decision profile of SRM of the neutral expression class output in feature subspace LBP, HoG and WG of the CK dataset.

The objective of this work is to define which are the best combining methods and subset of features, for problems of the facial expression classification, via SR.

2 Probabilistic set-up

Consider a set of classes $\Omega = w_1, \ldots, w_c$ and a classifier ensemble of L classifiers. Denote by s_i the class label proposed by classifier i ($s_i \in \Omega$). We are interested in the probability:

$$P(w_k \text{ } is \text{ } the \text{ } true \text{ } class \text{ } | s_1, s_2, \ldots, s_L), k = 1, \ldots, c,$$

denoted for short $P(w_k|s)$, where $s = [s_1, s_2, \ldots, s_L]^T$ is a label vector. Assume that the classifiers give their decisions independently, conditioned upon the class label which leads to the following decomposition:

$$P(w_k|s) = \frac{P(w_k)}{P(s)} \prod_i P(s_i|w_k)$$

Once a set of posterior probabilities $p_{ij}(x), i = 1, m; j = 1, c$ for m classifiers and c classes is computed for test object x, they have to be combined into a new

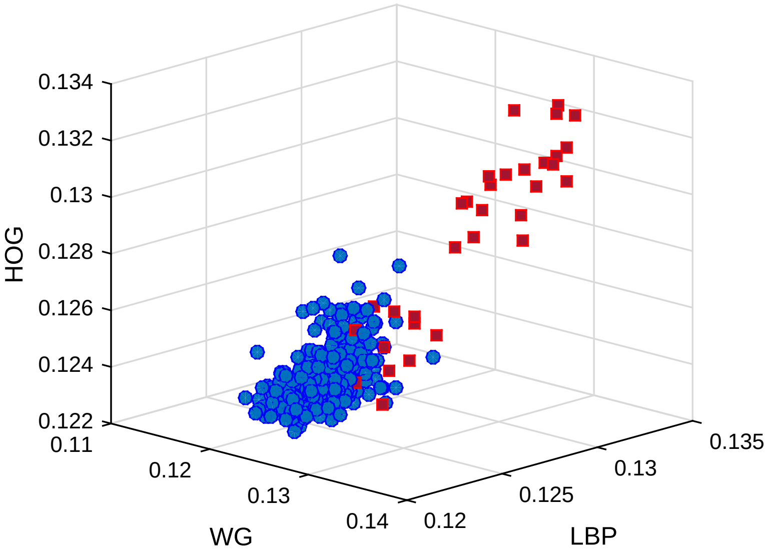

Figure 1: Decision profile of SRM of the neutral expression class output in feature subspace LBP, HoG and WG of the CK dataset. In red the true element in this class and blue not true elements in this class.

set $\mu_j(x)$ that can be used, by maximum selection, for the final classification. We distinguish two sets of rules, hard combiners and soft combiners.

3 The combining rules

Let $x \in \mathbb{R}^n$ be a feature vector and $\{1, 2, \ldots, c\}$ be the label set of c classes. We call a classier every mapping:

$$D : \mathbb{R}^n \longrightarrow [0, 1]^c – \mathbf{0}$$

where $\mathbf{0} = [0, 0, \ldots, 0]^T$ is the origin of $\mathbb{R}^c$. We call the output of D a "class label" and denote it by $[\mu_D^1(x), \ldots, \mu_D^c(x)]^T$, $\mu_D^i(X) \in [0, 1]$. The components $\mu_D^i(x)$ can be regarded as (estimates of) the posterior probabilities for the classes, give x, i.e, $\mu_D^i = P(i|x)$. Alternatively, $\mu_D^i(x)$ can be viewed as typicalness, belief, certainty, possibility, etc. Bezdek et al. [5] define three types of classifiers:

1. Crisp classifier: $\mu_D^i(x) \in \{0, 1\}, \sum_{i=1}^{c} \mu_D^i(x) = 1, \forall x \in \mathbb{R}^n$;
2. Fuzzy classifier: $\mu_D^i(x) \in [0, 1], \sum_{i=1}^{c} \mu_D^i(x) = 1, \forall x$; (Probabilistic interpretation of the outputs fall in this category)
3. Possibilistic classifier: $\mu_D^i(x) \in [0, 1], \sum_{i=1}^{c} \mu_D^i(x) > 0, \forall x$;

The decision of D can be "hardened" so that a crisp class label in $1, 2, \ldots, c$ is assigned to x. This is typically done by the maximum membership rule:

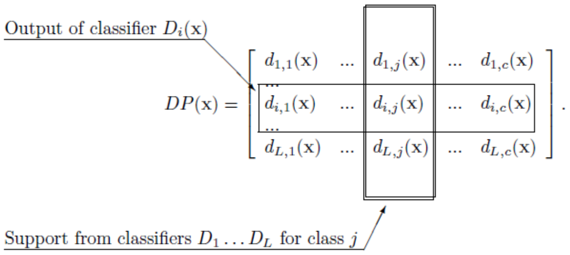
Output of classifier $D_i(\mathrm{x})$
$DP(\mathrm{x}) =$
$d_{1,1}(\mathrm{x})$ ... $d_{1,j}(\mathrm{x})$ ... $d_{1,c}(\mathrm{x})$
$d_{i,1}(\mathrm{x})$ ... $d_{i,j}(\mathrm{x})$ ... $d_{i,c}(\mathrm{x})$
$d_{L,1}(\mathrm{x})$ ... $d_{L,j}(\mathrm{x})$ ... $d_{L,c}(\mathrm{x})$
Support from classifiers $D_1 \dots D_L$ for class j

Figure 2: Decision profile.

$$D(x) = k \Leftrightarrow \mu_D^k = max_{i=1,...,c}\mu_D^i(x).$$

Let $D_1, \dots, D_L$ be the set of L classifiers. We denote the output of the ith classifier as $D_i(x) = [d_{i,1}(x), \dots; d_{i,c}(x)]^T$, where $d_{i,j}(x)$ is the degree of "support" given by classifier D_i to the hypothesis that x comes from class j. We construct $\widehat{D}$, the fused output of the L first-level classifiers as:

$$\widehat{D} = F(D_1(x), \dots, D_L(x)),$$

where F is called aggregation rule. The classifier outputs can be organized in a decision profile (DP) as the matrix of the Fig. 2.

3.1 The combining of hard classifiers

In [7] propose a common probabilistic framework for the following four combination methods: majority vote (MV), weighted majority vote (WMV), recall (REC) and naive Bayes (NB). It is shown a summary of each of the equations.

- Majority vote (MV):

  $$\log(P(w_k|s)) \propto \log(\frac{1-p}{p(c-1)})\log(P(w_k)) + |I_+^k|$$

  where $|I_+^k|$ is the number of votes for w_k.

- Weighted majority vote (WMV):

  $$\log(P(w_k|s)) \propto \log(P(w_k)) + \sum_{i \in |I_+^k|} \theta_i + |I_+^k| \times \log(c-1)$$

  where $\theta_i = \log(\frac{p_i}{1-p_i}), 0 < p_i < 1$

- Recall combiner (REC):

  $$\log(P(w_k|s)) \propto \log(P(w_k)) + \sum_i \log(1-p_{ik}) + \sum_{i \in |I_+^k|} v_{ik} + |I_+^k| \times \log(c-1).$$

  where $v_{ik} = \log(\frac{p_{ik}}{1-p_{ik}}), 0 < p_{ik} < 1$

- Naive Bayes combiner (NB):

  $$\log(P(w_k|s)) \propto \log(P(w_k)) + \sum_i \log(p_{i,s_i,k}).$$

3.2 The combining of soft classifiers

In [6] which provided a theoretical underpinning of many existing classifier combination schemes for fusing the decisions of multiple experts, each employed a distinct pattern representation. It has been demonstrated that under different assumptions and using different approximations we can derive the commonly used classifier combination schemes such as the product rule (PR), sum rule (SR), min rule (RMI), max rule (RMX), median rule (RMD), and majority voting (MV). It is shown a summary of each of the equations.

- Product Rule (PR):

$$P^{(-(R-1))}(w_j)\prod_i P(w_j|x_i) = \max_k P^{(-(R-1))}(w_k)\prod_i P(w_k|x_i)$$

  which under the assumption of equal priors, simplifies to the following:

$$\prod_i P(w_j|x_i) = \max_k \prod_i P(w_k|x_i)$$

- Sum Rule (SR):

$$(1-R)P(w_j) + \sum_i P(w_j|x_i) = \max_k [(1-R)P(w_k) + \sum_i P(w_k|x_i)]$$

  which under the assumption of equal priors simplifies to the following:

$$\sum_i P(w_j|x_i) = \max_k \sum_i P(w_k|x_i)$$

- Min Rule (MIR):

$$(1-R)P(w_j) + R\max_i P(w_j|x_i) = \max_k P^{(-(R-1))}(w_k)\min_i P(w_k|x_i)$$

  which under the assumption of equal priors simplifies to the following:

$$\max_i P(w_j|x_i) = \max_k \min_i P(w_k|x_i)$$

- Median Rule (MNR):

$$P^{(-(R-1))}(w_j)\min_i P(w_j|x_i) = \max_k med_i P(w_k|x_i)$$

- Majority Vote Rule:

$$\sum_i \Delta_{ji} = \max \sum_i \Delta_{ki}$$

  where: $\Delta_{ki} = 1$ if $P(w_k|x_i) = \max_j P(w_j|x_i)$ or 0 in otherwise.

3.3 Trainable combining of classifier

The outputs of the input classifiers can be regarded as a mapping to an intermediate space. A combining classifier applied on this space then makes a final decision for the class of a new object. In [4] one version of constrained regression for finding the weights that minimize the variance is derived by assuming the expert's errors in approximating the posterior probability.

Linear Opinion Pools (LOP):

$$P(w_k|s) = \sum_i \theta_{ki} P(w_k|s_i)$$

$$J = \sum_i \sum_k \theta_i \theta_k \sigma_{ik} - \lambda(\sum_i \theta_i - 1)$$

the solution minimizing J is:

$$\theta = \Sigma^{-1} I (I^T \Sigma^{-1} I)^{-1}$$

4 The classifiers

Consider a set of training signals $D = [D_1, D_2, \ldots, D_k] \in \mathbb{R}^{m \times p}$ from k different classes, where the columns of each sub-matrix D_j are signals from the class w_j. Ideally, giving sufficient training samples of class w_j, where $D_j = [d_1^j, d_2^j, \ldots, d_{n_j}^j] \in \mathbb{R}^{m \times n_j}$, a test signal $x \in \mathbb{R}^m$ belongs to the same class could be well approximated by a linear combination of the training samples from D_j, which can be written as:

$$x = \sum_{i=1}^{n_j} \alpha_i^j d_i^j \tag{1}$$

Then equation 1 can be rewritten in the form shown below:

$$x = D\delta_j(\alpha) \in \mathbb{R}^m \tag{2}$$

where $\delta_j(\alpha) = [0, \ldots 0, \alpha_1^j, \alpha_2^j, \ldots, \alpha_{n_j}^j, 0 \ldots, 0]^T \in \mathbb{R}^p$ is the coefficient vector in which most coefficients are zero except those associated with the class w_j. Due to the fact that a valid test sample x can be sufficiently represented only using the training samples from the same class, and this representation is the sparsest among all others, to find the identity of x is equal to find the sparsest solution of 2. This is the same as solving the following optimization problem (l_0-minimization):

$$\alpha^\star = \arg\min_{\alpha \in \mathbb{R}^p} \|\alpha\|_0 \;\; s.t \; D\alpha = x \tag{3}$$

However, solving the l_0-minimization of an underdetermined system of linear equations is NP-hard. If the solution $\alpha^\star$ sought, is sparse enough, the solution of the l_0-minimization problem 3 is equal to the solution to the following $l1$-minimization problem [**?**]:

$$\alpha^\star = \arg\min_{\alpha \in \mathbb{R}^p} \|\alpha\|_1 \;\; s.t \; D\alpha = x \tag{4}$$

Then, the estimate x using the coefficients corresponding to a given class w_j, $x \approx \widehat{x}_j = D\delta_j(\alpha^\star)$ is possible. This is consistent with previous findings and also mimics the behaviour of simple-cells in the visual cortex. The error of representation $e_j = |x - \widehat{x}_j|$ can be used to determine the class of the signal x.

5 Result and Discussion

5.1 A simulation study.

5.1.1 Protocol

Experiments with simulated classifier outputs were carried out as follows:

- Number of classes $c \in 2, 3, 4, 5, 10, 20, 50$;
- Number of classifiers $L \in 2, 3, 4, 5, 10, 20, 50$;
- Number of instances (labels) 500;
- Number of runs 100.

For each run, c classes were generated by labelling the 500 instances according to a symmetric Dirichlet distribution. To enforce class-conditional independence, the classifiers in the ensemble were constructed class by class. To form the label set of classifier i and class k, take the labels for class k and replace a percentage between 0 and 66.7% with labels randomly sampled from Ω. The c sets of labels for each classifier are concatenated to form the final output of classifier i. L such classifiers were generated and the four combination rules were calculated. The classification accuracies for each pair (L, c) were averaged across the 100 runs.

5.1.2 Result

As it is show in [7], the order of the four combiners in all simulations is as expected (from best to worst): NB, REC, WMV and MV. The NB combiner seems to have a great advantage over the other three combiners for larger number of classes. However, NB may suffer from the curse of dimensionality. In this part of the simulation study, we assumed that there is no noise in the estimates of the parameters of the combiners. Any noise could be very harmful to NB's accuracy. REC and WMV will be less vulnerable, and MV will be immune to the size of the validation data used for estimating the parameters.

The Fig. 3 shows the relationship between the combiners' accuracies for the whole ranges of L and c. Since there are 100 runs and 7 values of each parameter, there are 4900 ensemble accuracies for each combiner. The figure shows that the weighted majority and the recall combiner are similar, with the recall combiner having an edge over the WMV. They are both better than the majority vote combiners and worse than Naive Bayes combiner.

Fig. 4 gives the ensemble accuracies as a function of $log(c)$ for 2 and for 50 classifiers. The dependency of the combiner accuracies on the number of classifiers is shown in Fig. 5.

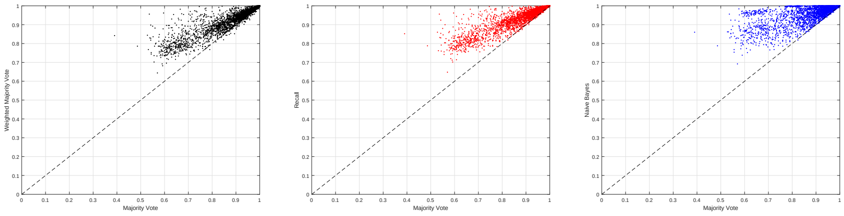

Figure 3: Relationship between the ensemble accuracies using the majority vote as the benchmark combiner. Each scatterplot contains 4900 ensembles points.

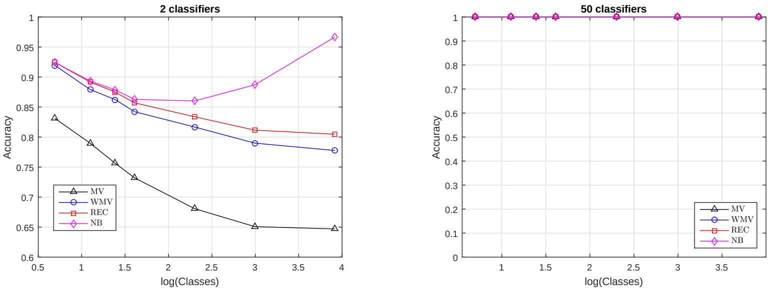

Figure 4: Ensemble accuracies of the 4 combiners as a function of log(c) (exact parameter estimates).

Table 1: Ensemble Accuracies With the 4 Combiners.

Data set	MV	WMV	REC	NB
Balance	$0.836 + 0.040$	$0.838 + 0.039$	$0.827 + 0.053$	$0.827 + 0.667$
Breast-w	$0.962 + 0.021$	$0.962 + 0.020$	$0.963 + 0.019$	$0.963 + 0.468$
Breast-y	$0.712 + 0.080$	$0.713 + 0.076$	$0.616 + 0.093$	$0.616 + 0.476$

5.2 Experiments with real data

5.2.1 Protocol

We used $L = 100$ decision tree classifiers and 10-fold cross validation. All experiments were run in Matlab. The accuracy of each ensemble is the average across the 100 testing results.

5.2.2 Result

Table 1 shows the ensemble accuracies. The best accuracies for each data set are underlined. We have indicated the winner even where, due to rounding, the values for the data set appear as identical in the table.

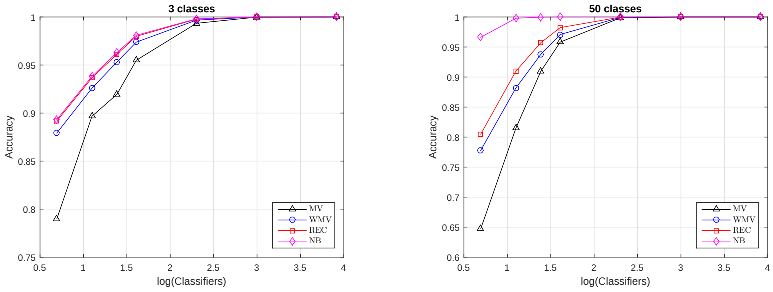

Figure 5: Ensemble accuracies of the 4 combiners as a function of log(L) (exact parameter estimates).

5.3 Experiment in application of expression recognition.

5.3.1 Protocol

To evaluate this work in FER problems, we used public DBs, i.e., Extended Cohn-Kanade(CK+) [9]. In all experiment were using person-independent FER scenarios [15], where the subjects in training set were completely different from the subjects within test set (i.e., the subjects used for training procedure cannot be used for testing phase). The CK+ DB: includes 593 image sequences from 123 subjects. From the 593 sequences, were selected 325 sequences of 118 subjects, which meet the criteria for one of the seven emotions [9]. The selected 325 sequences consist of 45, 18, 58, 25, 69, 28, and 82 sequences of Angry, Contempt, Disgust, Fear, Happy, Sadness, and Surprise, respectively [9]. In the neutral face case, was selected the first frame of the sequence of 33 random selected subjects. Similar to the methods in [3], [8], Leave-one-subject-out (LOSO) cross validation was adopted in the evaluation. LOSO selected 117 out of the 118 subjects for training, and used the remaining subjects for test. This procedure was repeated for all the 118 subjects. CK+ also included 68 landmark points for each image obtained from AAM[10]. All classes were included in this study.

5.3.2 Result

The Table 2 show the accuracy result for each ruler combination (RP, RS, RMX, RMI, RMD, RM, WMV, REC, NB and LOP) in the different feature subspaces (L/H, R/H, R/L, R/L/H, R/W, R/W/H, R/W/L, R/W/L/H, W/H, W/L and W/L/H). In bold we show the best result for each subspaces.

The Table 3 shows the comparison of the best results obtained for each combination (in bold in the Table 2) and the results of the individual methods. In bold we show the results of combination method that exceed individual methods. As can be seen, the combination of the methods increases the system efficiency.

6 Conclusion

It should be emphasized that our analysis is based in the result of the experiments for a single dataset. Conclusions will thereby at most point in a possible

Table 2: Average Test Accuracy for All Combined and All Feature Subspaces.

		Combining of Soft Classifiers					Combining of Crisp Classifiers				Trainable Classifiers
		RP	RS	RMX	RMI	RMD	RM	WMV	REC	NB	LOP
L/H	μ	**0.891**	**0.891**	0.888	0.875	0.891	0.874	0.856	0.885	0.883	**0.891**
	σ	0.216	0.216	0.226	0.228	0.216	0.240	0.240	0.230	0.229	0.216
R/H	μ	0.834	0.834	0.820	0.820	0.834	0.830	0.843	**0.869**	0.843	0.840
	σ	0.256	0.256	0.257	0.276	0.256	0.242	0.260	0.233	0.274	0.260
R/L	μ	0.842	0.842	0.826	0.827	0.842	0.827	0.832	**0.851**	0.841	0.842
	σ	0.248	0.248	0.262	0.255	0.248	0.258	0.271	0.260	0.274	0.244
R/L/H	μ	0.845	0.843	0.847	0.837	0.843	0.870	0.855	**0.876**	0.847	0.867
	σ	0.259	0.258	0.248	0.254	0.258	0.228	0.246	0.236	0.266	0.240
R/W	μ	0.901	0.901	0.893	0.862	0.901	0.866	0.920	**0.933**	0.891	0.906
	σ	0.189	0.189	0.190	0.221	0.189	0.215	0.163	0.136	0.218	0.183
R/W/H	μ	0.911	0.911	0.896	0.863	0.911	0.901	0.894	0.916	**0.925**	0.916
	σ	0.182	0.182	0.189	0.235	0.182	0.191	0.192	0.169	0.162	0.177
R/W/L	μ	0.908	0.908	0.895	0.868	0.908	0.904	0.892	0.913	0.887	**0.918**
	σ	0.183	0.183	0.188	0.226	0.183	0.194	0.203	0.191	0.224	0.177
R/W/L/H	μ	0.913	0.913	0.900	0.871	0.913	0.897	0.897	0.913	**0.917**	0.914
	σ	0.183	0.183	0.187	0.228	0.183	0.209	0.213	0.188	0.196	0.181
W/H	μ	0.924	0.924	0.925	0.902	0.924	0.932	0.882	**0.935**	0.921	0.923
	σ	0.167	0.167	0.164	0.197	0.167	0.144	0.216	0.150	0.196	0.174
W/L	μ	0.935	0.935	0.930	0.894	0.935	0.941	0.892	**0.939**	0.933	0.931
	σ	0.157	0.157	0.138	0.216	0.157	0.130	0.202	0.152	0.170	0.158
W/L/H	μ	0.919	0.919	0.930	0.908	0.919	0.918	0.898	0.911	**0.935**	0.925
	σ	0.187	0.187	0.163	0.196	0.187	0.191	0.212	0.199	0.172	0.171

Table 3: Compared to Individual Methods ($e * 1000^{-1}$).

	Combining											Individual			
Methods	L/H	R/H	R/L	R/L/H	**R/W**	**R/W/H**	R/W/L	R/W/L/H	**W/H**	**W/L**	**W/L/H**	R	L	H	W
Accuracy	891	869	851	876	933	925	918	917	939	939	935	767	914	875	892
	216	233	260	236	136	162	177	196	152	152	172	244	137	196	184

direction. They can be summarized as follows:

- The combining of crisp classifiers shows the best results for facial expression recognition problem via SR.
- Combining classifiers trained on different feature sets is very useful, especially when in these feature set probabilities are well estimated by the classifier.
- The trainable classifier combination may have good results for such problems.
- The best results were obtained using: WG + LBP + SR + REC.

References

[1] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (2005), vol. 1, IEEE, pp. 886–893.

[2] Elwell, R., and Polikar, R. Incremental learning of concept drift in nonstationary environments. *Neural Networks, IEEE Transactions on 22*, 10 (2011), 1517–1531.

[3] Huang, X., Zhao, G., Zheng, W., and Pietikäinen, M. Spatiotemporal local monogenic binary patterns for facial expression recognition. *Signal Processing Letters, IEEE 19*, 5 (2012), 243–246.

[4] Jacobs, R. A. Methods for combining experts' probability assessments. *Neural computation 7*, 5 (1995), 867–888.

[5] Keller, J., Krisnapuram, R., and Pal, N. R. *Fuzzy models and algorithms for pattern recognition and image processing*, vol. 4. Springer Science & Business Media, 2005.

[6] Kittler, J., Hatef, M., Duin, R. P., and Matas, J. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 20*, 3 (1998), 226–239.

[7] Kuncheva, L. I., and Rodríguez, J. J. A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems 38*, 2 (2014), 259–275.

[8] Li, Y., Wang, S., Zhao, Y., and Ji, Q. Simultaneous facial feature tracking and facial expression recognition. *Image Processing, IEEE Transactions on 22*, 7 (2013), 2559–2573.

[9] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on* (2010), IEEE, pp. 94–101.

[10] Matthews, I., and Baker, S. Active appearance models revisited. *International Journal of Computer Vision 60*, 2 (2004), 135–164.

[11] Olshausen, B. A., et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature 381*, 6583 (1996), 607–609.

[12] Olshausen, B. A., and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research 37*, 23 (1997), 3311 – 3325.

[13] Ouyang, Y., Sang, N., and Huang, R. Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers. *Neurocomputing 149* (2015), 71–78.

[14] Weifeng, L., Caifeng, S., and Yanjiang, W. Facial expression analysis using a sparse representation based space model. In *Signal Processing (ICSP), 2012 IEEE 11th International Conference on* (2012), vol. 3, IEEE, pp. 1659–1662.

[15] Zeng, Z., Pantic, M., Roisman, G., Huang, T. S., et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 31*, 1 (2009), 39–58.

[16] Zhang, S., Li, L., and Zhao, Z. Facial expression recognition based on gabor wavelets and sparse representation. In *Signal Processing (ICSP), 2012 IEEE 11th International Conference on* (2012), vol. 2, IEEE, pp. 816–819.