Active Learning of Ensemble Classifiers for Gesture Recognition

J. Schumacher[1], D. Sakič[1], A. Grumpe[2], G. A. Fink[1], C. Wöhler[2]

[1]Lehrstuhl Informatik XII, Technische Universität Dortmund, Germany
[2]Image Analysis Group, Technische Universität Dortmund, Germany

Abstract. In this study we consider the classification of emblematic gestures based on ensemble methods. In contrast to HMM-based approaches processing a gesture as a whole, we classify trajectory segments comprising a fixed number of sampling points. We propose a multi-view approach in order to increase the diversity of the classifiers across the ensemble by applying different methods for data normalisation and dimensionality reduction and by employing different classifier types. A genetic search algorithm is used to select the most successful ensemble configurations from the large variety of possible combinations. In addition to supervised learning, we make use of both labelled and unlabelled data in an active learning framework in order to reduce the effort required for manual labelling. In the supervised learning scenario, recognition rates per moment in time of more than 86% are obtained, which is comparable to the recognition rates obtained by a HMM approach for complete gestures. The active learning scenario yields recognition rates in excess of 80% even when only a fraction of 20% of all training samples are used.

1 Introduction

It is undisputed that gestures constitute an important modality in human-human and human-machine interaction. Therefore, a multitude of approaches for recognising different types of gestural expressions – ranging from hand-arm gestures to full-body motion – have been proposed in the literature [11].

In this study we focus on the aspect of mapping the motion parameters of the relevant body parts (e.g. the gesturing hand or the upper body including head and arms) to a gesture or action category, assuming that body motion has already been captured successfully and has been converted to a stream of 3D trajectory data. In contrast to related approaches, we make use of classifier ensembles for recognition as these are known to usually improve recognition performance in a wide range of pattern recognition tasks [8]. Additionally, we rely on both labelled and unlabelled data in an active learning framework for classifier training in order to reduce the manual labelling effort.

In order to be able to use a sufficiently diverse set of classifier types we decided not to consider gesture recognition as a problem of recognising variable length trajectories, which is usually solved by applying hidden Markov models (HMMs) [11] or dynamic Bayesian networks [20]. Rather, we address the problem by classifiying trajectory segments comprising a fixed number of sampling points.

For both active learning and for building powerful classifier ensembles, considering different views on the data or the classification problem is highly beneficial. Considering the issue of diversity on two levels of abstraction, we first create different feature representations of the trajectory data obtained by applying different subspace transforms. Second, different classifier types are used for constructing heterogeneous ensemble classifiers, as the diversity of homogeneous classifier ensembles usually comes at the cost of a large number of base classifiers, which in turn requires that many different views on the data can be obtained by sampling in feature or sample space [10]. In combination, by considering different feature representations, different classifier types, and different parameterisations of the two, a considerable variety of classification approaches is available for combination within an ensemble. Eventually, the most promising configurations are selected by applying a genetic search algorithm.

2 Related Work

The field of gesture recognition in general deals with the problem of recognising meaningful expressions conveyed by human motion [11]. Many approaches focus on either the recognition of hand-arm gestures or on the interpretation of full-body motion, usually referred to as "action recognition" [14]. The recognition of hand-arm gestures, which is mostly considered for artificially crafted gesture alphabets such as sign languages [12], involves first the capturing of the dynamic movement of the relevant body parts and second the analysis of the resulting temporal trajectories. Capturing dynamic motion of human body parts constitutes a challenging computer vision problem (cf. e.g. [15]), which may be simplified using visual markers or depth cameras [19].

For analysing time-series data in general and such obtained from human gestural expressions in particular, HMMs are most widely used today [11]. Other common approaches include dynamic Bayesian networks [20] and conditional random fields [22]. Interestingly, to the authors' best knowledge the problem of dynamic gesture recognition has not been addressed yet using learning approaches based on ensemble classifiers.

The combination of multiple base classifiers within a classifier ensemble has been shown to improve classification performance for a wide range of pattern recognition problems [8, 16]. In order to build a successful classifier ensemble, a set of base classifiers has to be created that is as diverse as possible while achieving satisfactory individual performances. Then the decisions of the base classifiers have to be combined into a final classification decision of the ensemble.

Classical ensemble creation techniques like bagging and boosting combine base classifiers of the same type but with different parameterisations by using different subsets or differently weighted versions of the training data for parameter estimation [10]. Random subspace sampling applies a similar strategy to the original feature space by randomly selecting different feature subsets [6]. The quite popular random-forest technique combines dataset sampling known from bagging with the use of decision trees as base classifiers [2].

In order to obtain the final decision of a classifier ensemble from the individual classifier outputs, a variety of techniques can be applied [24], such as majority voting or boosting, where classifier outputs are usually combined as a weighted average. The most general combination method is generalised stacking, where the problem of combining classifier outputs is considered as another classification problem [23].

As fully supervised learning of gestures requires a large amount of labelled training data, it is favourable to reduce the labelling effort by employing active learning techniques. In the concept of active learning it is assumed that the classifier is trained initially based on a relatively small amount of labelled training data. It then selects from a large set of unlabelled data the most informative samples and requests their labels from the user ("oracle") by performing a query. A broad overview of active learning approaches is given in [18]. For single classifiers, the strategy to select samples close to the decision boundary, for which the classification result is uncertain, is shown to be optimal in [13]. This concept is thus termed "uncertainty sampling" [18]. For a combination of several classifiers, such as the ensemble classifiers regarded in this study, the "query by committee" algorithm is introduced in [5], which selects those samples from the unlabelled data for which the results of the individual classifiers are most dissimilar [18]. In this study we will rely on a selection strategy which combines these two approaches.

3 Ensemble Classifiers for Gesture Recognition

The classifier ensemble used for gesture recognition in this study combines different classifier types with different views on the data, i.e. different sets of features, in order to obtain a high degree of diversity within the ensemble.

The 3D positions of the head and the hands of the gesturing person were obtained from multiocular image sequences as described in detail in [15]. According to [15], various features such as the coordinates of the hand positions relative to the head, velocity values, or trajectory curvatures are extracted. As the recognition result is desired to be independent of the position and orientation of the person in 3D space, additional variants of all features based on their changes over time ("delta features") are determined. These extraction steps result in an overall number of 90 features.

Based on the extracted features, we compute different views on the data by a combination of different preprocessing steps. The feature values are normalised to the same order of magnitude using as a first approach the transformation to the interval $[-1, +1]$ based on the corresponding minimum and maximum values and as a second approach the division of each feature by its mean absolute value. The dimensionality of the feature vectors is reduced by two methods: principal component analysis (PCA) [17] and independent component analysis (ICA) [7].

For the classifier ensembles, three types of classifiers are used: a linear, quadratic, or cubic polynomial classifier (PC) [17], a multi-layer perceptron (MLP) [10], and a support vector regression (SVR) [1]. The MLP is used with

only one hidden layer, where the number of hidden neurons is chosen such that the number of network parameters does not exceed 20% of the number of training samples and also remains smaller than one-half of the dimension of the input data. The SVR is used rather than the more commonly applied support vector machine (SVM) as it allows to compute the confidence bands of the estimated class-specific probabilities, which are much less straightforward to obtain for the SVM due to its discrete decision function. For a polynomial kernel and a RBF kernel of the SVR, the γ parameter according to [1] is chosen as the inverse average scalar product between the training samples [3] and as the inverse squared average mutual RMS distance between the training samples, respectively.

In the context of ensemble learning, it is generally assumed that using classifiers with dissimilar behaviour within an ensemble leads to an increased recognition performance [16]. In this study we quantify the diversity of the classifiers based on the cross-correlation measure and the rate of double faults as defined in [9], which both consider the respective class assignments of the individual samples and decrease with increasing diversity. These diversity measures can only be computed for pairs of classifiers. For more than two combined classifiers, we estimate the overall diversity by the average of the pairwise diversities.

The selection of the most promising ensemble configurations is based on a multi-criteria optimisation using a genetic search algorithm [4], considering the correlation coefficient, the rate of double faults, and the average classification error. This optimisation results in a three-dimensional Pareto front which comprises all Pareto-optimal solutions, where a solution is Pareto-optimal if no other solution exists for which all three criteria obtain a smaller value. Classifier ensembles consisting of L base classifiers are selected from those comprised by the Pareto front based on three criteria:

1. Select the classifier ensemble with the smallest average classification error under the constraint that its cross-correlation coefficient does not exceed the minimum value by more than 5%.
2. Select the classifier ensemble with the smallest average classification error under the constraint that its rate of double faults does not exceed the minimum value by more than 5%.
3. Select the classifier ensemble with the smallest rate of double faults under the constraint that its cross-correlation coefficient does not exceed the minimum value by more than 10%.

In order to keep the computational effort of the classifier ensembles in a reasonable range, only ensembles consisting of $L = 3$, 5, and 7 base classifiers are regarded. For each number L three classifier ensembles are selected according to the above criteria.

It is assumed that each of the L base classifiers determines a class assignment $\boldsymbol{y}_l$ and a decision vector $\boldsymbol{d}_l = (d_{l1}, d_{l2}, \ldots, d_{lK})$ with $\sum_{j=1}^{K} d_{lj} = 1$. Five methods to determine the overall ensemble decision are applied to the selected classifier ensembles:

1. Majority voting.

2. Sum of decision values. The class assignment is given by the maximum of the vector $\boldsymbol{d}_{\text{ens}} = \sum_{l=1}^{L} \boldsymbol{d}_l$.
3. Weighted sum of decision values. The class assignment is given by the maximum of the vector $\boldsymbol{d}_{\text{ens}} = \sum_{l=1}^{L} \boldsymbol{w}_l^T \boldsymbol{d}_l$. The elements of the weight vector $\boldsymbol{w}_l$ are given by $w_{li} = 1/\sigma_{li}$ with σ_{li} as the renormalised confidence as defined in [3] of the i-th decision value of the l-th classifier.
4. Master classifier for decision values. The class assignment is obtained by generalised stacking, relying on the output of a "master classifier" which combines the decision values of the base classifiers. We always use a first-order PC as the master classifier, leading to $\boldsymbol{d}_{\text{ens}} = A\boldsymbol{d}_L^*$ with A as the coefficient matrix of the PC and $\boldsymbol{d}_L^* = \left[\boldsymbol{d}_1^T \ \boldsymbol{d}_2^T \ \ldots \ \boldsymbol{d}_L^T\right]^T$ as the concatenated decision vectors of the base classifiers.
5. Master classifier for decision values and renormalised confidences. The master classifier determines the ensemble decision $\boldsymbol{d}_{\text{ens}}$ based on the decision values and the renormalised confidences of the base classifiers, such that $\boldsymbol{d}_{\text{ens}} = A\boldsymbol{d}_{\sigma L}^*$ with $\boldsymbol{d}_{\sigma L}^* = \left[\boldsymbol{d}_1^T \ \boldsymbol{\sigma}_1^T \ \ldots \ \boldsymbol{d}_L^T \ \boldsymbol{\sigma}_L^T\right]^T$.

4 Active Learning of Gestures

As a first step, a supervised training of the ensemble classifier is performed based on the initial labelled training set. During active learning, a sample is selected from the set of unlabelled samples and its label is queried if at least one of the four following conditions is fulfilled:

1. The maximum decision values of all L base classifiers are below a given threshold θ_1.
2. The differences between the highest and the second-highest decision value of all L base classifiers are below a given threshold θ_2.
3. The renormalised confidence values [3] of all L base classifiers exceed a given threshold θ_3.
4. The class assignment is different for all L base classifiers.

Condition 1 corresponds to the approach of uncertainty sampling. Condition 2 also selects unlabelled samples for which no clear class assignment can be obtained. Condition 3 relies on the renormalised confidence values, which have been found in [3] to denote how closely a new sample resembles the samples already used during the training process. Condition 4 selects unlabelled samples with a high degree of dissimilarity among the class assignments of the base classifiers and is thus a variant of the query by committee approach [5, 18].

5 Experimental Evaluation

The 3D trajectory data used in this study were extracted from enblematic gestures in a multi-camera framework. The labelled data set was adopted from [15],

Fig. 1. Examples of the gesture classes (from left to right) "circle", "come here", "down", "go away", "point", "stop", "up", "horizontal wave", and "vertical wave". The extracted 3D trajectories have been reprojected into the image (from [15])

Table 1. Number of instances and samples per gesture class

	Circle	Come here	Down	Go away	Point	Stop	Up	Hor. wave	Vert. wave
# instances	95	92	92	96	86	83	88	89	79
# samples	4000	2015	2056	1941	1573	1549	1651	3497	5617

where a detailed description is provided and the data are utilised for the classification of gestures using HMMs.[1] The gestures considered are performed by 16 different persons. According to [15], the raw 3D trajectories are smoothed using impulse-based resampling, which leads to a curvature-dependent distance between the resampled trajectory points obtaining low values when the local curvature of the trajectory is high. Table 1 lists the nine different classes, the number of instances (performed gestures) per class, and the number of samples (feature vectors corresponding to overlapping windows) per class.

Based on a series of tests of a variety of classifier configurations, the most appropriate window length corresponds to 8 subsequent resampled points. For the training data, the offset between two subsequent samples amounts to 1 step, while 4 steps are used for the test data. A number of 22 favourable base classifiers defined by the utilised normalisation technique, dimensionality reduction method, and classifier type were identified. In this context, the number of PCA components was chosen such that the reconstruction error corresponds to 0.1 and 0.01, respectively. The number of ICA components was set manually [7], where the "virtual dimensionality" [21] was used as an upper limit. The degree of the PC was set to 1, 2, and 3, where higher degrees were restricted to manageable sizes of the weight matrix. The SVR approach was used with polynomial kernels of degree 2 and 3 and with RBF kernel, and the number of hidden neurons of the MLP was determined automatically (cf. Section 3).

For all three selection criteria according to Section 3, the recognition rates of ensembles of size 5 are generally better than those of ensembles of size 3 and comparable to those of ensembles of size 7. The best ensembles of size 5 obtained with the genetic search algorithm are listed in Table 2.

For the evaluation of the ensemble classifiers, an 8-fold cross-validation was performed, where for each run the samples associated with 14 persons were used for training and the samples associated with the remaining 2 persons for testing.

[1] The trajectory data set is accessible at http://patrec.cs.tu-dortmund.de.

Table 2. Determined ensembles of size 5. "BC" stands for "base classifier". The numbers after "PCA" and "ICA" denote the reduced number of dimensions. The digits after "PC" and "Poly" denote the polynomial degree

Selection criterion	Base classifiers			
(cf. Section 3)	ID	Normalisation	Dim. red.	Classifier type
Cross-correlation (1)	BC1	Division by mean	PCA59	PC2
	BC2	Min-max interval	ICA59	PC2
	BC3	Division by mean	ICA59	PC2
	BC4	Min-max interval	PCA150	SVR Poly2
	BC5	Division by mean	PCA240	MLP
Rate of double faults (2)	BC1	Min-max	PCA120	SVR RBF
	BC2	Min-max interval	ICA569	SVR RBF
	BC3	None	ICA471	SVR RBF
	BC4	Min-max interval	ICA569	SVR Poly3
	BC5	Min-max interval	PCA150	SVR Poly2
Combination (3)	BC1	Division by mean	PCA200	SVR RBF
	BC2	None	ICA471	SVR RBF
	BC3	Min-max interval	ICA59	PC2
	BC4	Division by mean	ICA59	PC2
	BC5	Min-max interval	PCA150	SVR Poly2

The recognition rates obtained are shown as box plots in Fig. 2. For all configurations considered, the median recognition rate of the best ensemble classifier is higher than that of the best base classifier. However, the difference is always smaller than the uncertainty intervals of the recognition rates. The master classifier which takes into account the base classifier decision values and renormalised confidence values yields the highest recognition rate for ensemble selection methods 1 and 3 and the second highest for ensemble selection method 2. The median recognition rates of the best ensemble classifiers are higher than 86%. Note that all recognition rates are per moment in time and not per trajectory.

5.1 Active Learning Scenario

For active learning, the data set is divided into an initial training set comprising 5% of all training samples associated with 3 different persons, a larger set of unlabelled samples associated with 11 persons used for active learning, and an independent test set consisting of samples associated with 2 persons. These data sets are permuted 8 times in order to facilitate an 8-fold cross validation.

We found that for the small training sets encountered during active learning the most favourable ensemble classifier consists of a SVR with RBF kernel, a SVR with polynomial kernel of degree 2, and a quadratic PC, combined by a weighted sum of their decision values with the inverse renormalised confidence values as weights (method 3 according to Section 3). The feature values were normalised to the interval $[-1, 1]$ based on their minimum and maximum values. Using PCA, the dimensionality of the samples was reduced to 120 and 80 for the SVR with quadratic polynomial kernel and with RBF kernel, respectively. For

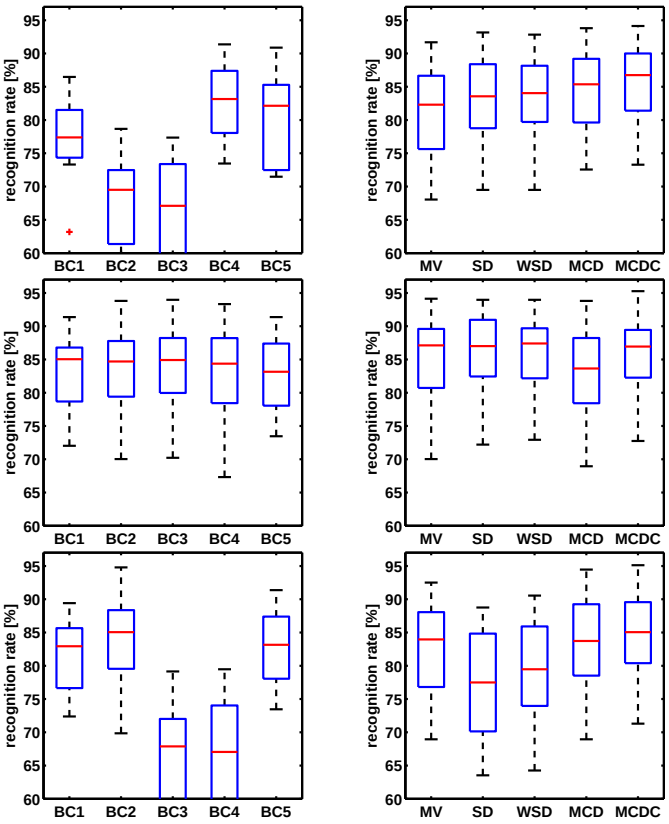

Fig. 2. Box plots of the recognition rates of the base classifiers (left column) and the ensemble decisions using different combination methods (right column) for (from top to bottom) selection criteria 1, 2, and 3. The ensembles correspond to those listed in Table 2. BC: base classifier; MV: majority voting; SD: sum of base classifier decision values; WSD: sum of base classifier decision values weighted by inverse renormalised confidence values; MCD: master classifier for decision values; MCDC: master classifier for decision values and renormalised confidence values.

the quadratic PC, the number of PCA components was adapted dynamically to the increasing number of training samples in order to ensure that the coefficient matrix did not become underdetermined for small sample set sizes, where the maximum number of PCA components was set to 59. For sample selection from the set of unlabelled samples according to the conditions listed in Section 4, the threshold values were set to $\theta_1 = 0.5$, $\theta_2 = 0.1$, and $\theta_3 = 2$. Each time when 500 samples had been queried, the ensemble classifier was re-trained.

The results of the active learning scenario are shown in Fig. 3, where the solid curve denotes the median recognition rate and the dotted curves the 25% and 75% quantiles, respectively. The recognition rate saturates at a value of about $82\% \pm 4\%$, when a fraction of approximately 30% of all training samples

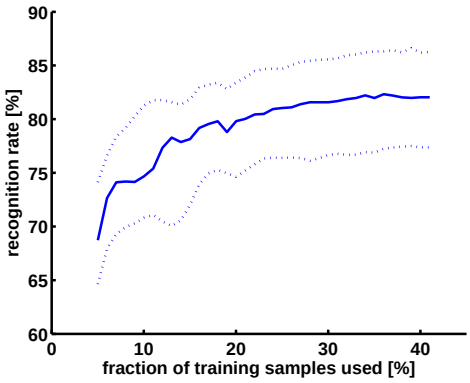

Fig. 3. Recognition results obtained in the active learning scenario. The ensemble classifier consists of a SVR with RBF kernel, a SVR with polynomial kernel of degree 2, and a quadratic PC.

(including the initial ones) have been used for training. However, the median recognition rate exceeds a value of 80% already when a fraction of 20% of all training samples have been used. This behaviour illustrates the efficiency of the employed active learning approach.

6 Summary and Conclusion

We have investigated the classification of gestures based on ensemble methods by classifying trajectory segments comprising a fixed number of sampling points. We have presented a multi-view approach in order to increase the diversity of the classifiers across the ensemble. In addition to supervised learning, an active learning framework has been employed in order to reduce the manual labelling effort. In the supervised scenario, we have obtained median recognition rates per moment in time of more than 86%. Similar recognition rates in between 84% and 90% are observed for the HMM-based approach in [15] for complete gestures. A median recognition rate of about 80% has been obtained in the active learning scenario, using an initial training set of 5% of all training samples. An amount of further 15% has been selected according to four criteria specifying those samples for which the class assignment of the ensemble classifier is most uncertain. The recognition rate saturates at a value of about $82\% \pm 4\%$. The recognition performance observed in the supervised and in the active learning scenario illustrates that gesture recognition based on the classification of trajectory segments using ensemble methods is a promising approach that may be applied in various areas, such as human-robot interaction or non-obtrusive user interfaces.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2007)

2. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
3. Cui, T., Grumpe, A., Hillebrand, M., Kreßel, U., Kummert, F., Wöhler, C.: Analytically tractable sample-specific confidence measures for semi-supervised learning. In: Proc. Workshop Computational Intelligence. pp. 171–186 (2011)
4. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. Wiley (2001)
5. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning 28, 133–168 (1997)
6. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998)
7. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)
8. Kuncheva, L.I.: Combining pattern classifiers: methods and algorithms. Wiley (2004)
9. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51, 181–207 (2003)
10. Marsland, S.: Machine Learning: An Algorithmic Perspective. CRC Press (2009)
11. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Trans. on Systems, Man, and Cybernetics, Part C 37(3), 311–324 (2007)
12. Ong, S., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(6), 873–891 (2005)
13. Park, J.M., Hu, Y.: Online learning for active pattern recognition. IEEE Signal Processing Letters 3(11), 301–303 (1996)
14. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28, 976–990 (2010)
15. Richarz, J., Fink, G.A.: Visual recognition of 3D emblematic gestures in an HMM framework. J. of Ambient Intelligence and Smart Environments 3(3), 193–211 (2011)
16. Rokach, L.: Pattern classification using ensemble methods. World Scientific (2010)
17. Schürmann, J.: Pattern Classification. Wiley-Interscience (1996)
18. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
19. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2011)
20. Suk, H.I., Sin, B.K., Lee, S.W.: Hand gesture recognition based on dynamic bayesian network framework. Pattern Recognition 43(9), 3059–3072 (2010)
21. Wang, J., Chang, C.I.: Independent component analysis based dimensionality reduction with applications in hyperspectral image analysis. IEEE Trans. on Geoscience and Remote Sensing 44, 1586–1600 (2006)
22. Wang, S.B., Quattoni, A., Morency, L.P., Demirdjian, D.: Hidden conditional random fields for gesture recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2. pp. 1521–1527 (2006)
23. Wolpert, D.H.: Stacked generalization. Neural Networks 5, 241–259 (1992)
24. Xu, L., Krzyzak, A., Suen, C.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. on Systems, Man and Cybernetics 22(3), 418–435 (1992)