INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team PAL

Pattern Analysis and Learning

Liama - Beijing - Chine

THEME COG

Table of contents

The PAL group was established in 2005. Its full name is Pattern Analysis and Learning, with research topics on the theory and methods of pattern recognition and machine learning, and their applications to document analysis, image analysis and text categorization. For more information, please visit our website http://liama.ia.ac.cn/wiki/projects:pal:home.

1. Team

Research Scientist

Cheng-Lin Liu [PhD, Professor, Head of PAL]
Xinwen Hou [PhD, Associate Professor]
Tonghua Su [Post-doctor of PAL]

Administrative Assistant

Ran Zhao [Secretary of PAL]

2. Overall Objectives

2.1. Overall Objectives

Keywords: *document analysis*, *image analysis*, *machine learning*, *pattern recognition*, *text categorization*.

The mission of the Pattern Analysis and Learning (PAL) Group at LIAMA is to study into the theory and methods of pattern recognition and machine learning, and their applications to document analysis, image analysis and text categorization. The general objective of the group is three-fold:

- **Fundamental research in pattern recognition and machine learning:** The performance of application systems depends heavily on advanced theory and methods. With the eventual goal of developing high performance pattern recognition systems, we study into various aspects of pattern recognition and machine learning. The main challenges in this area include the high dimensionality, the noisy, heterogeneous and unstationary nature of data, as well as the insufficiency of labeled data. To partially solve these problems, we study into probability density estimation, feature extraction and selection, generative and discriminative learning, semi-supervised learning, ensemble learning, manifold learning, etc. We aim to develop new and effective methods in some of these aspects, and publish scientific papers in distinguished international journals and conferences.
- **Competitive technology development in document analysis:** The area of document analysis concerns the automatic reading of machine-printed and/or handwritten documents. It finds wide applications such as document conversion and retrieval, bank form processing, census form processing, postal mail sorting, handwritten note recognition and retrieval. The automatic recognition of freely (online and offline) handwritten documents is the most challenging problem in this area. We aim to develop effective methods for handwritten character recognition (especially, large set Chinese character recognition), page segmentation, text line segmentation, character segmentation and recognition, and eventually, develop high performance recognition systems for real applications.
- **International cooperation:** The PAL Group has undertaken cooperative projects with Microsoft Research Asia and Hitachi Ltd. (Japan), and has established academic connections with researchers in France, Japan, Korea, USA, Canada, etc. In this year we have applied the scholarship of Sino-French joint doctorate, and will send graduate students to INRIA through this scholarship. Our next step is to enhance such connections, and build new connections with researchers in Europe, especially those in France. We also plan to organize one or two international conference or workshop.

2.2. Highlights

Our research goal is centered on solving typical problems in character recognition and document analysis. For Chinese character recognition, there are over ten thousands of character classes need to be recognized. Current machine learning methods degenerate with the increasing number of classes and the relaxed constraints of handwriting, so new methods of feature extraction and classifier learning are deserved to partially solve this problem. For document analysis, character and line segmentation is often interwoven with character recognition through some sequential graphical models, such as Hidden Markov Models, Markov Random Fields, Conditional Random Fields, etc. How to learn such models more efficiently and introduce new graphical models to document analysis are our research emphases.

3. Scientific Foundations

3.1. Fundamental Research in Pattern Recognition and Machine Learning

Keywords: *Density Estimation*, *Ensemble Learning*, *Feature Selection*, *Neural Networks*, *Support Vector Machines*, *classifier design*.

Participants: Cheng-Lin Liu, Xinwen Hou, Xiaobo Jin, Xiaohua Liu, Guoqiang Zhong, Yifeng Pan, Yanming Zhang, An Lu, Heng Wang, Mingbo Wang.

3.1.1. Pattern Classification

The basic task of pattern recognition and machine learning is to classify an object into several classes. Two-class problem is more often referred because it is simple and fundamental. In fact, multi-class problem is often converted into two-class problem for computation efficiency. Suppose $\mathbf{x} \in \mathcal{R}^{\mathbf{n}}$ is the datum and $X = \{\mathbf{x}\}$ is the dataset, a classifier H is a mapping assigning each $\mathbf{x}$ a label 1 or -1

$$H : X \rightarrow \{1, -1\} \tag{1}$$

The simplest classifier is linear classifier where the data are linearly mixed and threshold by a sign function

$$H(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}'\mathbf{x} + \mathbf{b}) \tag{2}$$

where $\mathbf{w}$ is a vector and $\mathbf{b}$ is a scalar. Another simple and useful classifier is quadratic classifier

$$H(\mathbf{x}) = \mathrm{sgn}(\mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{w}'\mathbf{x} + \mathbf{b}) \tag{3}$$

where $\mathbf{A}$ is a matrix. It can be verified that when the data distribution is Gaussian, the optimum classifier is quadratic. For character recognition, quadratic classifier can get satisfied accuracy. But for other applications, we need more powerful classifiers.

3.1.1.1. Neural Networks

Neural Networks mimic the behavior of human nerve: mixing the inputs then output by stimulant

$$H(\mathbf{x}) = \sigma(\mathbf{A}'\mathbf{x} + \mathbf{a}) \tag{4}$$

where $\mathbf{A}$ is a vector, $\mathbf{a}$ is a scalar and $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function. Multi-layer Neural Networks can be constructed sequentially

$$H(\mathbf{x}) = \sigma[\mathbf{B}'\sigma(\mathbf{A}\mathbf{x} + \mathbf{a}) + \mathbf{b}] \tag{5}$$

where $\mathbf{A}$ is a matrix, $\mathbf{B}$ and $\mathbf{a}$ are vectors, and $\mathbf{b}$ is a scalar. The training of Neural Networks is often aimed at minimizing the square error by gradient descending

$$\min_{\mathbf{A},\mathbf{a},\mathbf{B},\mathbf{b}} \sum_i ||H(\mathbf{x_i}) - y_i||^2 \tag{6}$$

where y_i is the labels of $\mathbf{x_i}$. Because of the non-convexity of the objective function, gradient descending often falls into local minima.

3.1.1.2. Support Vector Machines

Support Vector Machines (SVMs) [23] is a linear classifier aiming at maximizing geometrical margin

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{b}} ||\mathbf{w}||^{\mathbf{2}} \\ &subject\ to\ y_i[\mathbf{w}'\mathbf{x_i} + \mathbf{b}] \geq 1 \end{aligned} \tag{7}$$

Vapnik converted the above optimizing problem into its dual problem by K-T theorem

$$\begin{aligned} &\max_{\alpha} \sum \alpha_i - \frac{1}{2} \sum y_i y_j \alpha_i \alpha_j < \mathbf{x_i}, \mathbf{x_j} > \\ &subject\ to\ \sum y_i \alpha_i = 0, \quad \alpha_i \geq 0 \end{aligned} \tag{8}$$

The above quadratic program problem has only one maximum and can be solved accurately. SVMs can be easily extended to nonlinear case by kernel trick, i.e., substituting linear inner product $< \mathbf{x_i}, \mathbf{x_j} >$ by kernel function $K(\mathbf{x_i}, \mathbf{x_j})$

3.1.1.3. Ensemble Learning and AdaBoost

There is another way of improving the classification accuracy by integrating the outputs of several classifiers. The simplest ensemble learning is majority vote

$$H(\mathbf{x}) = \text{argmax}_c \#\{H_t(\mathbf{x}) = c\} \tag{9}$$

where $\#$ is the number of elements in the set. Freund and Schapire proposed to ensemble linear combination of classifiers and gave a very effective method called AdaBoost [25], [30]

$$\begin{aligned} &\min_{\alpha_t} \sum_i e^{-y_i \sum_t \alpha_t H_t(\mathbf{x_i})} \\ &H(\mathbf{x}) = \text{sgn} \sum_t \alpha_t H_t(\mathbf{x}) \end{aligned} \tag{10}$$

3.1.2. Feature Selection

Classification systems often need to deal with high dimensional feature vectors, which contain many irrelevant or redundant components. In order to achieve high performance, we must select the most relevant components, while eliminate the irrelevant or redundant ones. This task is known as feature selection, which can be categorized into three different approaches [27]. The first is embedded approach, where the selection algorithm is embedded within a basic induction algorithm, such as ID3 , C4.5 and CART. These algorithms carry out a greedy search through the space of decision trees, at each stage using an evaluation function select the feature that has the best discriminative ability. The second is filter approach, where an independent selection process occurs before the basic induction step, such as FOCUS, Branch and Bound and Relief. These algorithms evaluate the features based on their correlation with the target function (e.g., mutual information or Fisher information) and then select the top k features with the highest values. The third is wrapper approach, where the feature subsets are evaluated by the classification accuracy of the induction algorithm on the training data. Wrapper usually performs better than filter, but has heavy computational costs for running the induction algorithm repeatedly for each feature subset considered. In feature selection, only exhaustive search can guarantee a global optimal feature subset, but it is a NP-hard problem. Therefore we could only get suboptimal solution in practice. There have been a large variety of searching techniques concerning this, ranging from Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) to Sequential Forward Floating Selection (SFFS) and Genetic Algorithm (GA). These searching strategies are also computational cost since they evaluate every candidate feature subset, and are hard to be applied to those cases where the feature set is enormous, e.g., over 100,000. Recently Boosting [29], [26], [24], [28] has attracted much attention in ensemble learning and practical applications, for its elegant nature combining feature selection and classifier fusion. Boosting acts like Sequential Forward Selection in wrapper approach, but it only need to evaluate the feature to be selected, instead of the feature subset. Therefore, Boosting can be applied to those cases where the feature set is enormous, e.g., from 40,000 to 120,000 Harr-like features in face detection [31].

3.1.3. Density Estimation

Many pattern recognition algorithms, especially those based on Bayesian theory, require accurate probability density estimation. There are two types of density estimation: non-parametric and parametric. Non-parametric density estimation approximates the distribution by smoothing the data

$$\mathbf{p}(\mathbf{x}) = \sum_{\mathbf{i}} \mathbf{k}(\mathbf{x} - \mathbf{x_i}) \tag{11}$$

where $\mathbf{k}$ is a window function centered around 0. If we have prior knowledge that the data obey a distribution, the density can be approximated by estimating the parameters of the specific distribution. For Gaussian distribution, only the average vector and covariance matrix are needed to estimate the distribution. Many distributions can be approximated by Gaussian Mixture Models (GMMs)

$$\mathbf{p}(\mathbf{x}) = \sum_{\mathbf{i}} \alpha_{\mathbf{i}} \mathbf{g}(\mathbf{x}; \mu_{\mathbf{i}}, \Sigma_{\mathbf{i}}) \tag{12}$$

where $\mathbf{g}$ is the Gaussian function. The parameters of GMMs can be estimated through the famous Expectation Maximization algorithm. For high dimensional data, the inversion of covariance matrices in Gaussian distribution and GMMS often results in computation unstability. To overcome this phenomenon, the data vector can be decomposed into two subspaces. In the prime subspace the density is estimated by the corresponding vector and covariance matrix projection. In the complemental subspace the density is estimated by the corresponding vector projection but the covariance matrix is diagonal. The full density is the product of the two densities in the two subspaces.

$$\mathbf{p}(\mathbf{x}) = \sum_{\mathbf{i}} \alpha_{\mathbf{i}} \mathbf{g}(\mathbf{U}'\mathbf{x}; \mathbf{U}'\mu_{\mathbf{i}}, \mathbf{U}'\Sigma_{\mathbf{i}}\mathbf{U}) \mathbf{g}(\widetilde{\mathbf{U}}'\mathbf{x}; \widetilde{\mathbf{U}}'\mu, \sigma^{\mathbf{2}}\mathbf{I}) \tag{13}$$

where U and $\widetilde{U}$ are respectively the projection matrices of the two subspaces.

3.2. Character Recognition

Participants: Cheng-Lin Liu, Xiangdong Zhou, Tianfu Gao, Fei Yin, Longlong Ma, Bo Xu, Jinlun Yu, Qiufeng Wang, Dahan Wang, Liang Xu, Heng Zhang.

We mainly concern the recognition of handwritten Chinese characters, as well as the related techniques for developing document analysis systems: document image processing, text line extraction, character segmentation and recognition, and the related feature extraction and classifier learning issues.

Handwritten Chinese character recognition (HCCR) encounters some challenges: the large number of classes, complexity of character structures, large variability of handwritten shapes, confusion between similar characters. Since the first work of printed Chinese character recognition (PCCR) was published in 1966, many research efforts have been contributed to both printed and handwritten Chinese character recognition. Research on online HCCR was started as early as PCCR, whereas offline HCCR was started in late 1970s, and has attracted much attention from the 1980s. Since then, many effective methods have been proposed to solve this problem, and the recognition performance has advanced significantly.

The approaches of HCCR can be roughly grouped into two categories: feature matching (statistical classification) and structure analysis. Based on feature vector representation of character patterns, feature matching approach usually computes a simple distance measure (correlation matching), say, Euclidean or city block distance, between the test pattern and class prototypes. Currently, sophisticated classification techniques, including parametric and non-parametric statistical classifiers, neural networks and support vector machines (SVMs), can yield higher recognition accuracies. Nevertheless, the selection and extraction of features remains an important issue. Structure analysis is an inverse process of character generation: to extract the constituent strokes and compute a structural distance measure between the test pattern and class models. Due to its resembling of human cognition and the potential of absorbing large deformation, this approach was pursued intensively in the 1980s and is still advancing. However, due to the difficulty of stroke extraction and structural model building, it is not widely followed.

Currently, the recognition of constrained handwritten (handprinted) characters can achieve very high accuracies, say, over 99%. However, for freely written (unconstrained) characters, the recognition accuracy is far from the desired level for practical applications. In addition to the shape classification of handwritten shapes, the segmentation of text lines and the segmentation of characters in handwritten documents are very difficult to solve. The text lines in handwritten documents may be skewed, curved, and may interfere with each other. The characters in a line vary significantly in the size, location, and between-character gap, and may touch each other. Currently some sequential graphic models, such as Hidden Markov Models, Markov Random Fields, Conditional Random Fields, etc., can integrate the confidence scores in text line extraction, character segmentation, character recognition and linguistic information. How to seek a global optimization framework so as to reduce the ambiguities in these processing stages is our concern.

4. Application Domains

4.1. Pen-Based Memorandum Analysis and Recognition

Pen interface is widely and increasingly used in (desktop and notebook) personal computers, PDA, mobile phones, and so on. In the status of pen-input (online) handwriting recognition, isolated characters can be recognized with fairly high accuracy. However, for entering documents more conveniently and efficiently, automatic analysis and recognition of free-format documents is desired. Handwritten documents containing text, formulas, figures and tables are the target of this research. Via automatic layout analysis, character segmentation and recognition, we aim to convert a trajectory-based handwritten document into a well-formatted electronic document. Handwritten document analysis encounters some difficulties: the segmentation of text lines (probably arbitrary orientation) and each region of formulas, figures and tables, and the segmentation

and recognition of each line or region. Currently, we aim to attack two problems: layout analysis and text line recognition. In layout analysis, the document is segmented into text regions and non-text ones via stroke classification and clustering, and then further classified into text lines and formula/figure/table regions. In text line segmentation and recognition, we will integrate over-segmentation and candidate character recognition in a framework of path search. Eventually, the recognized regions will be converted into text codes while the un-recognized regions remain stroke trajectories in the converted electronic document.

4.2. Offline Handwritten Document Analysis

Offline handwritten document analysis is aimed at examining a set of scanned document images and extracting semantic information from them. The task includes image processing, layout analysis, text line segmentation, character segmentation and character recognition. There is a huge demand of automatical analysis of historical documents, personal documents (letter and diary), and the document analysis technology is expected to partially satisfy this need. There are mainly three difficulties: ambiguity between text lines, ambiguity between the words and characters in a line, and the large shape variation of the characters. Therefore, line segmentation, character segmentation and character recognition are the main research issues of offline document analysis.

4.3. Human Identification in Video

Human identification in video is an important problem for video surveillance and multimedia retrieval. Unlike other biometrics like iris recognition and fingerprint recognition that need contact or close view of body part, human identification in video poses many challenges: uncontrolled illumination, cluttered background, low resolution due to capturing from distance, and so on. People mostly recognize a person from his/her face. In video sequence, however, face is visible only when the person has a frontal or nearly frontal view. When the face is not visible, we can instead recognize the person from the body action: evidences in psychophysics show that humans can be discriminated from the walking pattern (gait pattern). Though both face recognition and gait recognition has been widely studied in computer vision, they are rarely combined. We aim to combine the facial and gait patterns to improve the reliability of human identification from video sequence. The technical issues include: background subtraction, face/body detection and tracking, facial feature extraction and classification, body shape representation and sequence classification, fusion of facial and gait information, fusion of multiple image frames, etc. We will put more emphasis on face recognition from low-resolution images, gait recognition using sequence analysis, and the fusion of facial and gait information.

5. New Results

5.1. Fundamental Research in Pattern Recognition and Machine Learning

The classification performance of nearest prototype classifiers largely relies on the prototype learning algorithms, such as the learning vector quantization (LVQ) and the minimum classification error (MCE). In [13], we proposes a new prototype learning algorithm based on the minimization of a conditional log-likelihood loss (CLL), called log-likelihood of margin (LOGM). A regularization term is added to avoid over-fitting in training. The CLL loss in LOGM is a convex function of margin, and so, gives better convergence than the MCE algorithm. Our empirical study on a large suite of benchmark datasets demonstrates that the proposed algorithm yield higher accuracies than the MCE, the generalized LVQ (GLVQ), and the soft nearest prototype classifier (SNPC).

Density estimation in high-dimensional data spaces is a challenge due to the sparseness of data which is known as “the curse of dimensionality”. Researchers often resort to low-dimensional subspaces for such tasks, while discard the distribution in the complementary subspace. In [14], we propose a new mixture density model based on pooled subspace. In our method, the Gaussian components of each class share a subspace and the complementary subspace is incorporated in the density function. The subspace and Gaussian mixture density are estimated simultaneously in EM iteration steps. We apply the density model to pattern classification in experiments on UCI datasets and compare the proposed method with previous ones. The experimental results demonstrate the superiority of the proposed method.

5.2. Offline Character Recognition

Pattern classification methods based on learning-from-examples have been widely applied to character recognition from the 1990s and have brought forth significant improvements of recognition accuracies. This kind of methods include statistical methods, artificial neural networks, support vector machines, multiple classifier combination, etc. In [22], we briefly review the learning-based classification methods that have been successfully applied to character recognition, with a special section devoted to the classification of large category set. We then discuss the characteristics of these methods, and discuss the remaining problems in character recognition that can be potentially solved by machine learning methods.

The technology of handwritten Chinese character recognition (HCCR) has seen significant advances in the last two decades owing to the effectiveness of many techniques, especially those for character shape normalization and feature extraction. In [15] we review the major methods of normalization and feature extraction and evaluates their performance experimentally. The normalization methods include linear normalization, nonlinear normalization (NLN) based on line density equalization, moment normalization (MN), bi-moment normalization (BMN), modified centroid-boundary alignment (MCBA), and their pseudo-two-dimensional (pseudo 2D) extensions. As to feature extraction, we focus on some effective variations of direction features: chaincode feature, normalization-cooperated chaincode feature (NCCF), and gradient feature. The features are compared with various resolutions of direction and zoning, and are combined with various normalization methods. In experiments, the current methods have shown superior performance on handprinted characters, but are insufficient applied to unconstrained handwriting.

The recognition of Indian and Arabic handwriting is drawing increasing attention in recent years. To test the promise of existing handwritten numeral recognition methods and provide new benchmarks for future research, we [17] presents some results of handwritten Bangla and Farsi numeral recognition on binary and gray-scale images. On proper pre-processing, feature extraction and classification, we achieved very high accuracies on three databases: ISI Bangla numerals, CENPARMI Farsi numerals, and IFHCDB Farsi numerals. The benefit of recognition on gray-scale images is justified.

To improve the accuracy of handwritten Chinese character recognition (HCCR), we [10] propose linear discriminant analysis (LDA)-based compound distances for discriminating similar characters. The LDA-based method is an extension of previous compound Mahalanobis function (CMF), which calculates a complementary distance on a one-dimensional subspace (discriminant vector) for discriminating two classes and combines this complementary distance with a baseline quadratic classifier. We use LDA to estimate the discriminant vector for better discriminability and show that under restrictive assumptions, the CMF is a special case of our LDA-based method. Further improvements can be obtained when the discriminant vector is estimated from higher dimensional feature spaces. We evaluated the methods in experiments on the ETL9B and CASIA databases using the modified quadratic discriminant function (MQDF) as baseline classifier. The results demonstrate the superiority of LDA-based method over the CMF and the superiority of discriminant vector learning from high-dimensional feature spaces. Compared to the MQDF, the proposed method reduces the error rates by factors of over 26%.

The accuracy of handwritten Chinese character recognition can be improved by pair discrimination of similar characters. In [12], we propose a new method for combining the baseline classifier with incomplete pair discriminators to better exploit their complementariness. The outputs of the baseline classifier and pair discriminators are transformed to two-class probabilities, which are then fused by pairwise coupling (PWC) for final decision. In our experiments using the modified quadratic discriminant function (MQDF) as baseline classifier and LDA-based pair discriminators, the PWC method outperforms the filter method. At best, the error rate of MQDF was reduced by factors over 28%.

For character recognition in document analysis, some classes are closely overlapped but are not necessarily to be separated before contextual information is exploited. For classification of such overlapping classes, either discriminating between them or merging them into a metaclass does not satisfy. Merging the overlapping classes into a metaclass implies that within-metaclass substitution is considered as correct classification. For such classification problems, we [11], [16] propose a partial discriminative training (PDT) scheme for neural

networks, in which, a training pattern of an overlapping class is used as a positive sample of its labeled class, and neither positive nor negative sample for its allied classes (classes overlapping with the labeled class). In experiments of handwritten letter recognition using neural networks and support vector machines, the PDT scheme mostly outperforms crosstraining (a scheme for multi-labeled classification), ordinary discriminative training and metaclass classification.

Separating text lines in handwritten documents remains a challenge because the text lines are often un-uniformly skewed and curved. In [20], we propose a novel text line segmentation algorithm based on Minimal Spanning Tree (MST) clustering with distance metric learning. Given a distance metric, the connected components of document image are grouped into a tree structure. Text lines are extracted by dynamically cutting the edges of the tree using a new objective function. For avoiding artificial parameters and improving the segmentation accuracy, we design the distance metric by supervised learning. Experiments on handwritten Chinese documents demonstrate the superiority of the approach.

5.3. Online Character Recognition

We present an effective approach [21] for grouping text lines in online handwritten Japanese documents by combining temporal and spatial information. Initially, strokes are grouped into text line strings according to off-stroke distances. Each text line string is segmented into text lines by dynamic programming (DP) optimizing a cost function trained by the minimum classification error (MCE) method. Over-segmented text lines are then merged with a support vector machine (SVM) classifier for making merge/non-merge decisions, and last, a spatial merge module corrects the segmentation errors caused by delayed strokes. In experiments on the TUAT Kondate database, the proposed approach achieves the Entity Detection Metric (EDM) rate of 0.8816, the Edit-Distance Rate (EDR) of 0.1234, which demonstrates the superiority of our approach.

In [18] we propose a new radical-based approach for online handwritten Chinese character recognition. The approach is novel in three respects: statistical classification of radicals, over-segmentation of characters into candidate radicals, and lexicon-driven recognition of characters. Currently, we have applied the approach to Chinese characters of left-right structure and are extending to other structures. Preliminary results on a sample set of 4,284 characters consisting of 1,118 radicals demonstrate the superiority of the proposed approach.

5.4. Image Analysis

Detecting texts from natural scene images is a challenge due to the variations of size, font, color and alignment and it is often affected by complex background, light shadow, image distortion and degrading. In [19], we present a robust system to accurately detect and localize texts in natural scene images. For text detection, a region-based method utilizing multiple features and cascade AdaBoost classifier is adopted. For text localization, a window grouping method integrating text line competition analysis is used to generate text lines. Then within each text line, local binarization is used to extract candidate connected components (CCs) and non-text CCs are filtered out by Markov Random Fields (MRF) model, through which text line can be localized accurately. Experiments on the public benchmark ICDAR 2003 Robust Reading and Text Locating Dataset 1 show that our system is comparable to the best existing methods both in accuracy and speed.

Insect has the richest species diversity on the earth. The task of insect species identification is one of the most fundamental works for many entomological fields and usually being difficult and complex. At present, reliable insect identification is mainly carried out by taxonomists, but this approach can't meet practical needs at all. Therefore, many scientists have tried to resolve this problem by using modern computing technology. Pattern recognition technology has developed rapidly and made automated insect images identification feasible. The history, underlying theory, general process and the prospects for developing and automated insect image identification system are briefly reviewed and discussed in [9].

6. Contracts and Grants with Industry

6.1. Radical-Based Handwritten Chinese Character Recognition for Tablet PC Applications

Participants: Cheng-Lin Liu, Xiangdong Zhou, Longlong Ma, Jinlun Yu.

Cheng-Lin Liu contracted with Microsoft Research Asia in 2007. This project is aimed at radical-based or component-based Handwritten Chinese character recognition. Microsoft Research Asia is the industrial leader of computer software. For handwriting recognition in mobile computers, the character recognizer is desired to be compact, accurate, and adaptable to new writing styles. To achieve this goal for Asian languages (such as Chinese) with thousands of characters, we can utilize the hierarchical structure of characters to reduce the number of classes. Such radical-based or component-based recognition approach encounters the difficulty of radical segmentation from characters. Previous works mostly relied on stroke analysis and resulted in low recognition accuracies. In this project, we propose a new radical-based recognition approach which shows several features: re-definition of radicals to facilitate segmentation, statistical classification of radicals and integrated segmentation-classification for higher accuracy. The radical-based recognizer will have much smaller storage space and needs small number of labeled samples for adapting to a new writer.

6.2. National Funds

Cheng-Lin Liu won the National Outstanding Youth Foundation from the National Natural Science Foundation of China (NSFC) in 2008. The Project title is "Pattern Recognition Theory, Methods and Applications".

6.3. Professional Activities

1. Cheng-Lin Liu serves as an associate editor of Pattern Recognition from 2006.

2. Cheng-Lin Liu serves as a member of editorial board of Image and Vision Computing from 2008.

3. Cheng-Lin Liu was a Program Co-Chair of the 2008 Chinese Conference on Pattern Recognition.

4. Cheng-Lin Liu was a member of program committee for international conferences ICFHR 2008 (11th International Conference on Frontiers of Handwriting Recognition) and DAS 2008 (8th International Workshop on Document Analysis Systems).

6.4. Academic Exchanges and Visits

1. June 1-6, 2008, Xiaohua Liu Participated in the International Joint Conference on Neural Network (IJCNN) in Hong Kong, and presented her work on subspace learning [14]

2. July 2-4, 2008, Cheng-Lin Liu participated in the 3rd IAPR Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR) in Paris, France, and presented his work on partial discriminative training [16].

3. August 19-21, Cheng-Lin Liu and Fei Yin participated in the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR) in Montreal, Quebec, Canada, and presented their works on character recognition [17] and document analysis [20].

4. September 16-19, 2008, Xiangdong Zhou and Yifeng Pan participated in the 8th IAPR Workshop on Document Analysis Systems (DAS) in Nara, Japan, and presented their works on online document analysis [21] and text detection [19].

5. November 13-14, 2008, Cheng-Lin Liu attended the 3rd Korea-Japan Joint Workshop on Pattern Recognition (KJPR) in Seoul, Korea, and gave an invited talk "Handwritten Document Analysis Using Machine Learning techniques"

6. December 8-11, 2008, Tianfu Gao, Xiaobo Jin and Longlong Ma participated in the 19th International Conference on Pattern Recognition in Tampa, Florida, USA, and presented their works on character recognition [12], [18] and prototype learning [13].

7. June 13-September 13, Adrien Delaye, a PhD student at the IRISA, France, worked with the PAL Group as an internship.

7. Dissemination

7.1. Teaching

From October, 2008, Cheng-Lin Liu opened the course of Pattern Recognition for doctorial candidates at the Institute of Automation, CAS.

8. Bibliography

Major publications by the team in recent years

[1] X. HOU, C.-L. LIU, T. TAN. *Learning Boosted Asymmetric Classifiers for Object Detection*, in "Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR06)", vol. 1, 2006, p. 330-338.

[2] C.-L. LIU, S. JAEGER. *Online handwritten Chinese character recognition: The state of the art*, in "IEEE Trans. Pattern Analysis and Machine Intelligence", vol. 26, 2004, p. 198-213.

[3] C.-L. LIU. *Classifier combination based on confidence transformation*, in "Pattern Recognition", vol. 38, 2005, p. 11-28.

[4] C.-L. LIU. *Normalization-cooperated Gradient Feature Extraction for Handwritten Character Recognition*, in "IEEE Trans. Pattern Analysis and Machine Intelligence", vol. 29, 2007, p. 1465-1469.

[5] C.-L. LIU, K. MARUKAWA. *Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition*, in "Pattern Recognition", vol. 38, 2005, p. 2242-2255.

[6] C.-L. LIU, H. SAKO, H. FUJISAWA. *Discriminative Learning Quadratic Discriminant Function for Handwriting Recognition*, in "IEEE Trans. Neural Networks", vol. 15, 2004, p. 430-444.

[7] C.-L. LIU, H. SAKO, H. FUJISAWA. *Effects of Classifier Structures and Training Regimes on Integrated Segmentation and Recognition of Handwritten Numeral Strings*, in "IEEE Trans. Pattern Analysis and Machine Intelligence", vol. 26, 2004, p. 1395-1407.

[8] C.-L. LIU, H. SAKO. *Class-specific feature polynomial classifier for pattern classification and its application to handwritten numeral recognition*, in "Pattern Recognition", vol. 39, 2006, p. 669-681.

Year Publications

Articles in International Peer-Reviewed Journal

[9] X. CHEN, X. HOU, C.-L. LIU, X. LIU, Z. ZHANG. *Advances in the Automated Insect Image Identification*, in "Chinese Bulletin of Entomology", vol. 45, 2008, p. 317-322.

[10] T. GAO, C.-L. LIU. *High accuracy handwritten Chinese character recognition using LDA-based compound distances*, in "Pattern Recognition", vol. 41, 2008, p. 3442-3451.

[11] C.-L. LIU. *Partial discriminative training for classification of overlapping classes in document analysis*, in "International Journal of Document Analysis and Recognition", vol. 11, 2008, p. 53-65.

International Peer-Reviewed Conference/Proceedings

[12] T. GAO, C.-L. LIU. *Combining Quadratic Classifier and Pair Discriminators by Pairwise Coupling for Handwritten Chinese Character Recognition*, in "Proceedings of the International Conference on Pattern Recognition", 2008.

[13] X. JIN, C.-L. LIU, X. HOU. *Prototype Learning with Margin-Based Conditional Log-likelihood Loss*, in "Proceedings of the International Conference on Pattern Recognition", 2008.

[14] X. LIU, C.-L. LIU, X. HOU. *A Pooled Subspace Mixture Density Model for Pattern Classification in High-Dimensional Spaces*, in "International Joint Conference on Neural Network", 2008, p. 2467-2472.

[15] C.-L. LIU. *Handwritten Chinese character recognition: Effects of shape normalization and feature extraction*, in "Arabic and Chinese Handwriting Recognition", Lecture Notes in Computer Science, vol. 4768, Springer-Verlag, 2008, p. 104-128.

[16] C.-L. LIU. *Partial discriminative training of neural networks for classification of overlapping classes*, in "Artificial Neural Networks in Pattern Recognition: Third IAPR Workshop", Lecture Notes in Artificial Intelligence, vol. 5064, Springer-Verlag, 2008, p. 137-146.

[17] C.-L. LIU, C. SUEN. *A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters*, in "Proceedings of the 11the International Conference on Frontiers in Handwriting Recognition", 2008, p. 278-283.

[18] L. MA, C.-L. LIU. *A New Radical-Based Approach to Online Handwritten Chinese Character Recognition*, in "Proceedings of the International Conference on Pattern Recognition", 2008.

[19] Y. PAN, X. HOU, C.-L. LIU. *A robust system to detect and localize texts in scene images*, in "Proceedings of the 8th IAPR Workshop on Document Analysis Systems", 2008, p. 35-42.

[20] F. YIN, C.-L. LIU. *Handwritten Text Line Segmentation by Clustering with Distance Metric Learning*, in "Proceedings of the 11the International Conference on Frontiers in Handwriting Recognition", 2008, p. 229-234.

[21] X. ZHOU, W. DAHAN, C.-L. LIU. *Grouping text lines in online handwritten Japanese documents by combining temporal and spatial information*, in "Proceedings of the 8th International Workshop on Document Analysis Systems", 2008, p. 61-68.

Scientific Books (or Scientific Book chapters)

[22] C.-L. LIU, H. FUJISAWA. *Classification and learning in character recognition: Advances and remaining problems*, Machine Learning in Document Analysis and Recognition, Springer-Verlag, 2008, p. 39-161.

References in notes

[23] C. J. C. BURGES. *A tutorial on support vector machines for pattern recognition*, in "Data Mining and Knowledge Discovery", vol. 2, 1998, p. 121–167.

[24] N. DUFFY, D. P. HELMBOLD. *A geometric approach to Leveraging weak learners*, in "EuroColt", 1999, p. 18–33.

[25] Y. FREUND, R. E. SCHAPIRE. *A decision-theoretic generalization of on-line learning and an application to Boosting*, in "International Journal of Computer and System Sciences", vol. 5, 1997, p. 119–139.

[26] J. FRIEDMAN, T. HASTIE, R. J. TIBSHIRANI. *Additive Logistic Regression: a statistical view of Boosting*, in "Annals of Statistics", vol. 28, 2000, p. 337–407.

[27] I. GUYON, A. ELISSEEFF. *An introduction to variable and feature selection*, in "Journal of Machine Learning Research", 2003, p. 1157–1182.

[28] J. KIVINEN, M. K. WARMUTH. *Boosting as entropy projection*, in "The twelfth annual conference on Computational learning theory", 1999, p. 134–144.

[29] R. MEIR, G. RÄTSCH. *An introduction to Boosting and Leveraging*, in "Advanced Lectures on Maching Learning", LNAI, vol. 2600, 2003, p. 118–183.

[30] R. E. SCHAPIRE, Y. SINGER. *Improved Boosting algorithms using confidence-rated predictions*, in "Machine Learning", vol. 37, 1999, p. 297–336.

[31] P. VIOLA, M. JONES. *Rapid object detection using a Boosted cascade of simple features*, in "Proc. International Conference on Computer Vision and Pattern Recognition", 2001, p. 1063–6919.