ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 25 МАТЕМАТИКИ Выпуск 3
2018

М. В. Бродславский С. Ю. Мельников (Москва, лаб. ТВП, ООО «Линфо»). **Сравнение эффективности классификаторов в задаче текстонезависимой идентификации автора русскоязычного рукописного текста.**

Разработке автоматических методов определения авторства рукописного текста по его изображению посвящено большое количество работ. Хорошие результаты получены для английского и арабского языков [1–3]. Для русского языка исследования в этой области затруднены отсутствием доступных баз изображений рукописного текста.

Процесс автоматической идентификации автора состоит из трех этапов: выделение векторов признаков из изображения, обучение классификатора и собственно классификация. Признаки из изображений рукописных текстов выделяются различными способами. Так, в [4] применяются шаблоны отдельных символов почерка и их фрагментов. В настоящей работе признаки строятся с использованием вейвлет-преобразования Габора [2]:

$$g_{\epsilon,\eta,\lambda,\theta,\varphi}(x,y) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\delta^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\varphi\right),$$

где $x' = (x-\varepsilon)\cos\theta - (y-\eta)\sin\theta$, $y' = (x-\varepsilon)\sin\theta - (y-\eta)\cos\theta$.

Изображение просматривается скользящим окном фиксированного размера, двигающимся по горизонтали на определенный шаг. Для каждого окна применяются фильтры Габора с параметрами $\lambda = 2.7, 4.1, 5.4$, $\theta = \frac{k\pi}{8}$, $k = 1, 2, \ldots, 8$, способ вычисления $\gamma, \delta, \varphi, \varepsilon$ приведен в [1]. Вычисляются среднее и дисперсия полученных значений. Набором признаков являются получаемые вектора размерности 48.

Для классификации использовались 3 метода: модель гауссовских смесей (GMM) [3], метод опорных векторов (SVM) [5], многослойная нейронная сеть (ANN) [6].

Описание экспериментов. Сформирована коллекция изображений упрощенных рукописных текстов семи авторов (по странице на каждого автора), в которых отсутствовали знаки препинания и пробелы между словами. После приведения к одному масштабу и бинаризации по алгоритму Оцу, изображения очищались от мелких шумов с помощью медианного фильтра. Изображения разделялись на фрагмент для обучения (размером 1200*1200 пикселей) и несколько фрагментов для классификации (размером 800–1200 пикселей по горизонтали и 300–500 пикселей по вертикали).

Использовались следующие параметры классификаторов:

GMM. Количество компонент в смеси 64. Для построения модели применялся EM-алгоритм [3]. SVM. Применялось гауссовское ядро. Обучение проводилось по алгоритму Sequence Minimal Optimization [5]. ANN. Нулевой слой состоял из 48 нейронов. На первом слое число нейронов задавалось пользователем (по умолчанию 30), на конечном слое число нейронов равно числу классов. В качестве функции активации выступала биполярная сигмоидальная функция, обучение проходило по алгоритму Левенберга-Марквардта [6].

Результаты экспериментов по оценке эффективности классификаторов в закрытой задаче идентификации русскоязычного рукописного текста в зависимости от размеров скользящего окна и величины шага (в пикселях) приведены в Табл.1.

Таблица 1. Временные затраты и точность классификации

Размер окна	Шаг сдвига	Время обучения (мин)			Время классификации (мин)			Точность (классификации %)		
		GMM	SVM	ANN	GMM	SVM	ANN	GMM	SVM	ANN
64*64	32	0.85	5.7	19	16	17.4	18	90	98	98
	64	0.42	1.8	13.4	14.3	15.2	17.4	85	95	97
128*128	64	0.12	0.13	4.59	17.25	16.88	16.86	62	91	97
	128	0.08	0.05	1.31	4.96	4.89	4.83	59	87	96
256*256	128	0.08	0.03	1.52	10.76	13.21	14.6	55	75	96
	256	0.07	0.02	1.13	7.25	10.83	11.2	52	71	96

Выводы. Все три построенных классификатора могут быть использованы для идентификации автора. Наибольшее время для обучения требуется нейросетевому классификатору, наименьшее — GMM и SVM (в зависимости от размеров окна). Сравнительно длительный процесс обучения ANN компенсируется более быстрым процессом классификации при больших размерах окна и высокой точностью классификации. GMM обеспечивает хорошую точность только при маленьких размерах окна, что существенно замедляет процесс классификации. SVM по точности и временным затратам занимает промежуточную позицию между GMM и ANN.

СПИСОК ЛИТЕРАТУРЫ

1. *Shahabi F.*, *Rahmati M.* Comparison of Gabor-Based Features for Writer Identification of Farsi/Arabic Handwriting. — Int. Workshop on Frontiers in Handwriting Recognition, Oct 2006, La Baule (France), Suvisoft, 2006, p. 545–550.
2. *Said H. E. S.*, *Tan S. P G.*, *Peake G. S.*, *Baker K. D.* Writer Identification from Non-uniformly Skewed Handwriting Images. — In Proc. of the 9th British Machine Vision Conference, 1988, p. 478–487.
3. *Schlapbach A.*, *Bunke H.* Off-line Writer Identification and Verification Using Gaussian Mixture Models. — Machine Learning in Document Analysis and Recognition, Springer, Berlin, 2008, v. 90, p. 409–428.
4. *Еременко Ю. И.*, *Мельникова И. В.*, *Шаталов А. А.* Интеллектуальная система идентификации объектов с помощью алгоритмов иммунных систем. — Вестник ВГТУ, 2015, № 6, с. 38–47.
5. *Chang C.-C.*, *Lin C.-J.* LIBSVM: A library for support vector machines. — ACM Trans. on Intelligent Syst. and Technology, 2011, 2 (3), p. 1–27.
6. *Хайкин С.* Нейронные сети: полный курс, 2-е изд.: Пер. с английского. М.: Издательский дом «Вильямс», 2006, 1104 с.