Абстракт

Препознавањето примероци како дел од интелигентното однесување е еден од главните предизвици во областа на машинската интелигенција. Оваа тема обработува некои проблеми што се заеднички за сите системи наменети за препознавање ракопис. Сите методи не се прикладни за препознавање стари кирилски букви. Затоа, овде е предложена специфична методологија. Всушност, предложени се два различни класификатори и направени се експерименти на дигитализирани примероци од оригинални историски манускрипти.

Системите за препознавање букви обично содржат пет модули: претпроцесирање, сегментација, издвојување на обележја, класификација и постпроцесирање. Тука, најмногу е посветено внимание на издвојувањето обележја и класификацијата. Би било многу корисно да се вклучи и претпроцесирање, затоа што ракописите се стари, валкани, со лош квалитет, но оваа проблематика е разгледувана кај други автори и може да се вклучи дополнително. Овде е предложена оригинална софтверска апликација.

Од битмапата на знакот-буква се издвојуваат 23 обележја. Овие обележја се меѓу статистичките и структурните или инспирирани од нив креирани се нови, сопствени обележја.

Издвојувањето обележја е направено од оригинални ракописи кои датираат во најголем број помеѓу 12-ти и 15-ти век. Ова е возможно да се направи од причини што во долг историски период стилот на знаковните графеми - букви не претрпел драматични стилски промени. Предуслов е само ракописот да биде напишан со Уставно писмо. Уставното писмо потсетува на печатен текст и од него полесно можат да се издвојат контурни линии.

Со комбинција на идеи пронајдени во литературата, овде се предложени нови обележја и анализи наречени "анализи на дамки". Рамката што може да се опише околу буквата се пресекува со два хоризонтални и два вертикални пресеци. На тој начин се формираат девет сегменти или квадранти кои можат да се групираат во: горна, средна, долна зона или: лева, централна, десна зона.

Референтниот знак-буква (прототипот) се добива од повеќе случајно избрани примероци од оригинално напишани букви. Статистичките податоци конвертирани во графички запис даваат сликовит преглед **што** може да биде прифатено како референтен знак-буква. Некогаш е тоа класичен пресек или фази пресек, но подобро е класичната унија или фази унијата да дадат добар резултат.

Направени се тастови за појавата на одделните обележја во секоја знак-буква. Изработени се два класификатори: Дрво на одлучување и Фази класификатор.

На екран можат да се споредат статистичките и графичките податоци за да се добие претстава за разијдувањето помеѓу хуманата и машинската визија. Применувајќи ги и двата класификатори истовремено исто така, можат да се споредат и резултатите добиени од нив.

Без оглед што искуството покажува дека за време на процесот на препознавање за некои букви или дрвото на одлучување или фази класификаторот даваат подобри резултати, севкупните резултати за целата азбука се речиси еднакви.

Резултатите од направените експерименти не се многу високи, но се прифатливи и охрабрувачки за натамошна работа со цел постигнување поголема прецизност и точност во препознавањето на црковнословенските букви.

Клучни зборови: Издвојување обележја, класификатори, дрво на одлучување, техники на нејасно одлучување, препознавање ракописни букви.

Abstract

Pattern recognition as a part of intelligent behavior is one of the main challenges of the machine intelligence area. This theme deals with some problems common to each Handwritten Character Recognition System. Not all methods are suitable for recognition of Old Cyrillic Letters. So, a specific methodology is proposed here. In fact, two different classifiers are proposed and some experiments are made on digitalized original historical documents.

Character Recognition Systems usually have five modules: pre-processing, segmentation, feature extraction, classification and post-processing. The attention here is put on the features extraction and classification. Pre-processing would be very useful to include, because the manuscripts are old, dirty, with poor quality, but this topic is elaborated by others authors and could be included additionally. Here, original software application is proposed.

From the character bitmap 23 features are extracted. These features are among statistical and structural ones, or were inspired by them and created as new own features.

Feature extraction was made from the original manuscripts dating mostly between 12^{th} and 15^{th} century. It is possible to do this, because in such a long historical period the style of letter graphemes have not been affected by style changes. The precondition is that manuscripts have to be written in Constitutional script. This Script looks like printed text, so that contour lines can be easily extracted.

Combining the ideas found in the literature new features and analyses call 'spot analyze' are used. Bounding box of the character image is cut with two horizontal and two vertical cuts. In that way nine segments or quadrants are combined in: upper zone, middle zone, low zone or: left, central and right zone.

Referent characters (prototypes) are obtained from many randomly chosen examples of the original characters. Their statistical characteristics, converted in graphical data, are giving a view what is acceptable as a referent prototype. Sometimes it is a crisp intersection or fuzzy intersection, but it is better if crisp or fuzzy union gives good results.

The appearance of each feature in every character is tested. Two classifiers are designed: Decision Tree and Fuzzy Classifier.

On the screen, the statistical and graphical data can be compared to get a sense concerning the incompatibility between the human and machine vision. By applying these two classifiers in the same time (simultaneously) the obtained results can also be compared.

Nevertheless, the experience shows that even through the recognition process for some characters either Decision Tree or Fuzzy classifier are giving better results, the overall results in the total set of characters are almost equal.

The results of the particular experiments are not of very high score, but are acceptable and encouraging for future work so as to achieve more accurate and precise Old Slavic Cyrillic Letters recognition.

Key words: Feature extraction, Classifiers, Decision Tree, Fuzzy Decision Techniques, Handwritten Character Recognition