UDC 004.89

I. Korzun, T. Likhouzova

AUTOMATIC SYSTEM OF EXTRACTION AND CONSOLIDATION OF DATA FROM DIGITAL IMAGES

Abstract: Optical mark recognition is the process of extracting respondent data from scanned form images by determining the state of its input fields (marks). This paper discusses the software used in solving the problem. It examines the existing applications, outlines their areas of use, and highlights the main limitations. Taking into account the pros and cons of such applications, an alternative solution is proposed. The system implements the necessary elements of automatic identification of forms and consolidation of data from them for non-profit organizations and small businesses.

Key words: optical mark recognition (OMR); image recognition systems.

Introduction

Optical mark recognition is an automatic identification technology. It is a method of capturing the information from the filled-in form based on determining the status of marks. The principle of this technology is so simple that it has led to its widespread use in management, healthcare, marketing and especially education.

Three elements take part in the OMR-problem solution:

- form generation tools;
- form scanning tools;
- the analyzer for data extraction.

Such a structure has a number of disadvantages which makes surveys content rigid and software less flexible. Most existing solutions require users to purchase specialized hardware and software. They also propose services for design, production and processing of forms. This requires significant and regular material costs and therefore it is in demand only among large organizations.

The alternative is free projects that can be used with a regular printer, scanner and computer, and very rarely a mobile device. But often they are poorly designed, involve third-party software in the process, or impose significant limitations on the variability of the form and format of the survey.

The goal of the work is to combine the necessary elements of automatic identification within an integral system. The research object is systems of optical mark recognition.

Existing recognition systems

Automatic identification and data capture refer to methods of finding objects of interest, accumulating data about them and entering it directly into the computer system without human assistance. Reading bar and QR (quick recognition) codes, radio frequency identification, biometrics, magnetic tapes, and voice recognition are related to automatic identification. Depending on the field of application, their purpose differs markedly: from the trading, transportation, organization and consolidation of information and production to security. Their main advantages: increased management efficiency, reduced time for decision-making, optimal use material and human resources [1].

Optical Mark Recognition is the technology for electronically extracting data from filled-in fields/bubbles on printed forms [2]. It is actively involved in management and education, public research and medicine. The main benefit of using this technology is the reduction of dependence on human resources by automating the data collection process.

The problem of optical form reading comprises the following steps:

- development of a template and its printing
- scanning – capturing the digital image of the filled-in form
- form image analysis for extraction of data as illustrated in Fig 1.

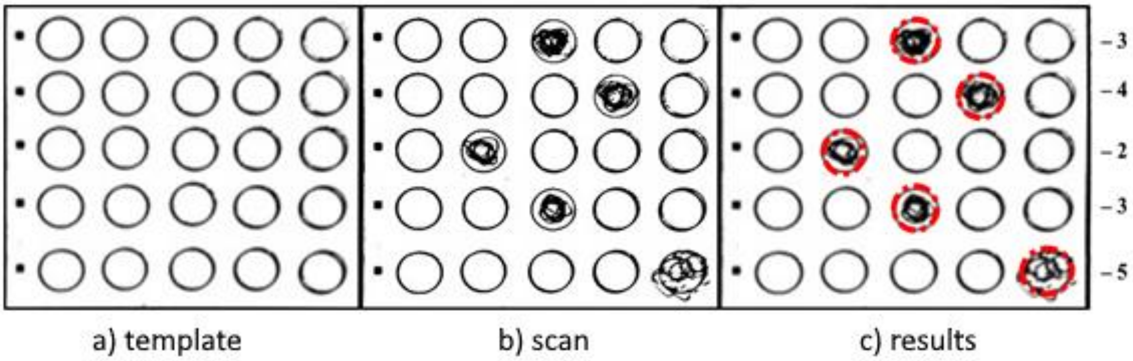

Figure. 1. The main steps of optical mark recognition

The most common approach to solving the problem involves the use of specialized equipment: OMR-scanner, appropriate material, thin transparent paper, and layout. The latter defines where input fields are and how they should be captured.

Numerous large companies apply this method in order to quickly and effectively study the issues of interest to them. Educational organizations, including the Ministry of Education and Science of Ukraine, use ORM identification to conduct state exams and competitions. Nevertheless, the method still remains omitted by non-profit and charitable foundations due to the lack or inconvenience of affordable and simple solutions.

Recently, many open-source projects have appeared. They allow the use of available office printers and scanners, but impose some restrictions on the structure of survey forms. Third-party products are often used in creation of the design of forms. This software is usually intended for different purposes like word processing and image editing. Most analogues are able to work effectively only with high-quality scanned documents with no distortion, uneven lighting or complex background. Despite the shortcomings, they are affordable.

Among the main providers of optical mark recognition services, there are Scantron, Gravic (Remark Products), and Mark Reader. These companies offer their users comprehensive solutions [3]. They include a complete toolkit of desktop software for preparing documents and equipment for printing and reading. Although these services are not free, the approaches they rely on to solve the OMR problem are common among open source application developers.

Problems of existing systems can be grouped as follows:

- high cost of paid solutions;
- limited capability of free projects when creating forms;
- the dependence of the quality and accuracy of mark recognition on the conditions of scanning or making photos of the form;
- no localization of the user interface in Ukrainian;
- lack of free-of-charge projects to solve the problem in Ukraine.

Proposed system

To accomplish the goal a desktop application “Zapit.i” was created. It is intended for automatic construction, reading and analysis of survey forms, and had the following features:

- flexible system for building form templates, with support for multi-page surveys;
- accurate and reliable document capture mechanism;
- support of languages: Ukrainian / English;
- intuitive interface;
- feedback system;
- independence from the Internet.

The application model is represented by a use case diagram in Fig 2. The initiator of inquiries is the organizer of interrogations. The actor will perform the actions of compiling answer sheets, their production, and distribution, conducting briefings on filling out questionnaires and their further collection; analysis of the obtained results and their dissemination among stakeholders. The greatest load falls on the stage of processing the respondent data. That is what the developed application is going to compensate for.

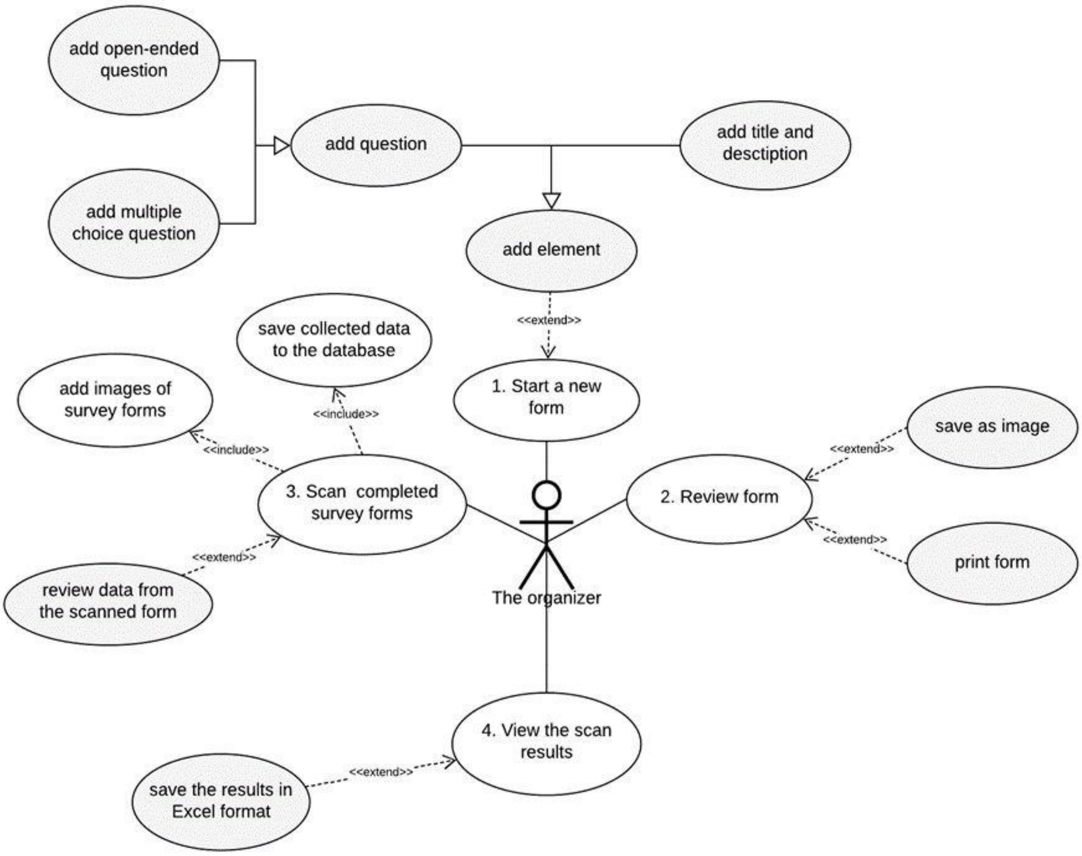

Figure. 2. The use case diagram of the proposed system

According to a recent report by the We Are Social agency, as of January 2020, 71.1 % of all-Ukrainian Internet traffic is accounted for by desktop computers and only 27.8% - by mobile devices [4]. Although the nature of the changes in this ratio indicates a decline for the former and an increase for the latter, this distribution has made it possible to choose a personal computer as the target platform.

Windows is chosen as the basic operating system of the application. According to the analytical company Net Applications, 88.14% of all personal computers in the world use it [5]. This operating system is used in educational and social organizations and government agencies of Ukraine. Therefore, this choice is completely justified.

The computer application works independently of Internet access and has a monolithic architecture. Thus, all its functionality will be located on one platform, which fully satisfies the task.

The system of image analysis and data extraction (OMR core) has the component architecture. Each of its functional elements is a separate independent module that allows you to attract the expected experience in solving other tasks.

The proposed application has several advantages over existing ones:

- unlike paid systems, the application, will be distributed under the general public license GPLv3, which guarantees the project the right to free access and modification;
- in comparison with similar applications, the program implements the complete system of optical mark recognition and contains all its functional modules:
  - the flexible form generator;
  - the filled-in forms scanner;
  - results presentation tools;
- in contrast to similar solutions, these application modules are recognizable, provide a simple and intuitive interface, and are easy to familiarize for new users.

REFERENCES

1. Ramya, A. Automatic identification and data capture (AIDC) and its technologies. (2015).

2. Smith, Andrew M. Optical mark reading - making it easy for users. SIGUCCS '81 (1981).

3. Young, Chadwick H., Glenn V. Lo, Kaisa Elizabeth Young and Alberto Borsetta. FormScanner: Open-Source Solution for Grading Multiple-Choice Exams. The Physics Teacher 54. 34-35 (2016).

4. Kemp, Simon. Digital 2020: Ukraine. (2020). https://datareportal.com/reports/digital-2020-ukraine.

5. Operating System Market Share. NetMarketShare (2020). https://netmarketshare.com/operating-system- market-share.aspx.