Projections Based OMR Algorithm

Dragan Ivetic
Faculty of Engineering
Universuty of Novi Sad
Trg Dositeja Obradovica 6, 21000 Novi Sad
Serbia & Montenegro
ivetic@uns.ns.ac.yu

Dinu Dragan
Faculty of Engineering
Universuty of Novi Sad
Trg Dositeja Obradovica 6, 21000 Novi Sad
Serbia & Montenegro
dinud@uns.ns.ac.yu

Abstract – **The paper presents the results on Optical Mark Recognition (OMR) of hard-copy questionnaire for Faculty/University evaluation. The solution is discovered in simple scalar region description based on horizontal and vertical projections. Both projections are used for hard-copy questionnaire preparation while the vertical projection is used in marked answer extraction process. The horizontal projection is used to identify regions with answers for every question of questionnaire. Inside the recognized region, sub-regions (answer fields of question) are further identified. Sub-region identification is carried out by the vertical projection. The same projection is used for identification of marked sub-region of filled questionnaire. Digitalization of questionnaire papers and its influence on OMR algorithms efficiency is commented at the end of the paper.**

I. INTRODUCTION

Results presented in this paper arise as a part of the SIZENA (System for Faculty/University Evaluation) program developed for a Tempus project [1]. System for Faculty/University evaluation, SIZENA, is described by the block diagram in Fig. 1. This system uses e-copy and hard-copy questionnaires. In the first case, students fill e-copy questionnaire using their own computers outside the teaching facilities or using computers in faculty computer centre. In the second case, students fill hard-copy questionnaires during lecturing hours. Later on, the questionnaire papers are scanned and marked answers are extracted from digitalized pictures of scanned questionnaires into data base. Gathered data are used to generate e-copy or hard-copy reports.

It is fair to ask: why to use hard-copy questionnaire in an age of E-documents? There are certain facts that advocate hard-copy: the number of active users of computer and Internet in our country is small, the number of computer skilled persons is even smaller, ration between number of available computers and number of students on our Faculty is about 1:60. Furthermore, the hard-copy questionnaire form is far more familiar in our society.

The questionnaire participators can answer on questions either by marking offered answers or writing open answers. In SIZENA, answering question in hard-copy questionnaire is restricted on marking offered answers. Restriction is made because Optical Character Recognition (OCR) of handwritten text in our verbal region is a huge problem and a great challenge. Optical recognition of handwritten text is difficult because there are two official alphabets, Latin and Cyrillic. Optical Mark Recognition (OMR) algorithms are logical choice for solution of marked answers extraction problem.

There is surprisingly small number of publications on OMR algorithms. Another problem is acronym overlay. Acronym OMR is also used for Optical Music Recognition. Precisely this acronym found in [2] led to horizontal and vertical projections as possible solution for OMR algorithms.

System for Faculty/University evaluation works with documents, precisely with digitalized document images. This makes it one kind of a Document Management System (DMS). DMS are systems for document management (fetching, processing, storage and contents printing). They can manipulate with digitalized images of entire documents or with some particular data extracted from given documents.

The OMR algorithms will be introduced in II section of the paper. The third section defines the vertical and horizontal projections. The fourth section describes regions with answers extraction process. The fifth section describes sub-region extraction and optical recognition of marked sub-regions. The concluded section summarizes results and problems and gives topics of future research.

II. OMR

Optical Mark Recognition (OMR) is the process of automatic recognition if some field (in this case sub-region) is marked or not. Sub-region can be rectangular, circular or some other predestined area on paper document. Sub-region is marked when it is crossed, Fig. 2.a, checked, Fig. 2.b, or filled. All of these predestined sub-regions have their meaning. If sub-region is marked then the meaning of sub-region is confirmed, else it is denied. OMR algorithm detects if sub-region is marked or not and the information about that is converted in some internal data which is used in further processing. OMR algorithm could be used in DMS whenever it is needed to store only some certain data from the document, or as addition to Imaging Document System[1]. Algorithm is often used for data extraction from various questionnaires, paper exams with offered answers [4], voting sheets, etc.

There are two phases of OMR algorithm. The first phase is document preparation when its regions are detected. Regions contain sub-regions for marking offered answers on given question, Fig. 3. This phase involves two sets of activities on every page of digitalized document. The first one, the region with sub-regions is detected and its position (coordinates) is calculated. The next one, sub-regions are detected inside every region, and their number and position is calculated. The first phase of OMR algorithm is carried

[1] DMS that stores the digitalized images of entire document.

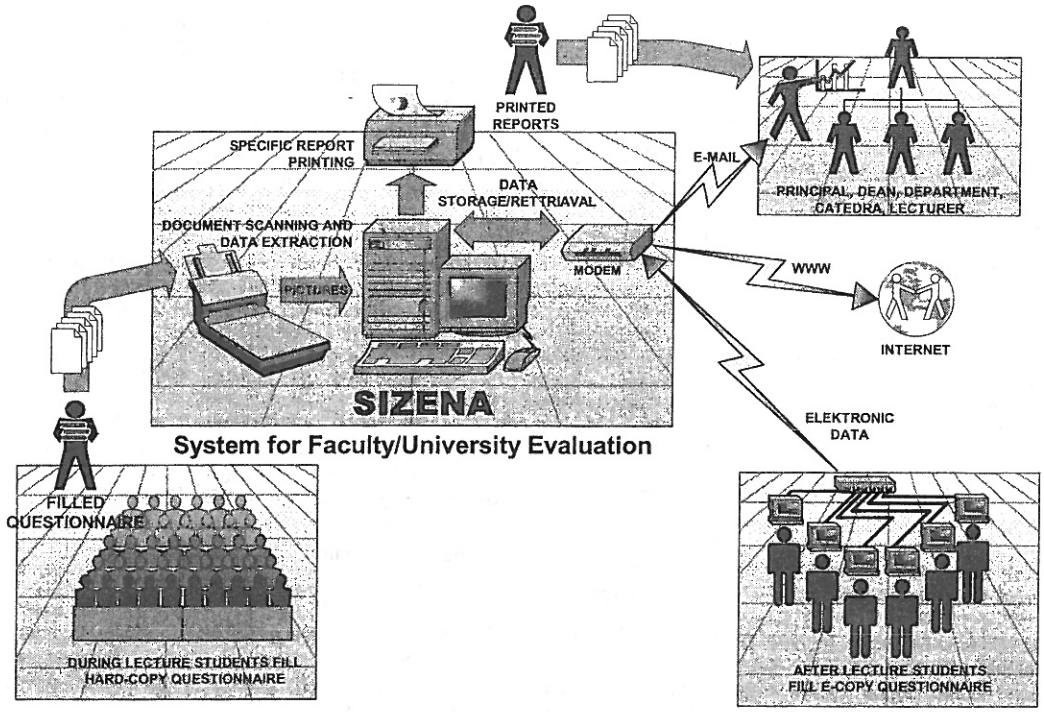

Fig. 1. System for Faculty/University evaluation based on hard-copy and e-copy questionnaire

out only once per document form (profile). This phase is executed on empty (not filled) digitalized document. Document preparation is used to speed up process of marked answer extraction.

In the second phase marked answers are extracted from filled document. Information gain in previous phase are now used to access regions with sub-regions on every page of digitalized document. Inside of this regions, OMR algorithm identifies which of the sub-regions is marked as answer on question.

Presented OMR algorithm is a base for building rather automated system then automatic one, since it leaves open a manual correction of omissions in marked sub-region extraction.

III. PROJECTIONS

Projections represent simple scalar description of image region without holes [2, 5]. Region is considered as collection of connected image pixels with certain semantic meaning. Image can be black and white, grayscale or colored. Pixels from the same region could have the same (component) color value, or value different from the background, etc.

Image projections or projections of some image part can be done along any of two orthogonal axis in 2D plane. If projection is done along vertical axis, then it is a vertical projection, Fig. 4c, and if it is done along horizontal axis, then it is a horizontal projection, Fig. 4a. It is possible to project image pixels in arbitrary direction, and then it is called skewed projection [2, 3].

If digitalized page is defined by a function f with two parameters: x - offset on horizontal axis, and y - offset on vertical axis, then vertical (Pv) and horizontal (Ph) projection can be defined with expressions (1) and (2), respectively.

$$Pv(x) = \sum_{y} f(x, y), \tag{1}$$

$$Ph(y) = \sum_{x} f(x, y), \tag{2}$$

Two types of projections can be distinguish depending on interpretion of the function f in expressions (1) and (2). Projection (vertical or horizontal) by value implies that f defines the value (level of grey or component color) of image pixel on (x, y) coordinates. In this case the expressions (1) and (2) give pixels value addition along vertical or horizontal axis. Projection (vertical or horizontal) by number implies that f is binary function that defines if pixel on coordinates (x, y) is part of the region

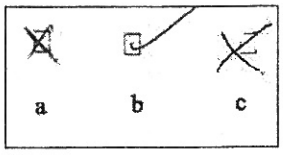

Fig. 2. Modes for sub-region marking

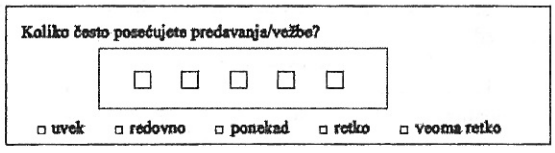

Fig. 3. Region of interest with mark sub-regions

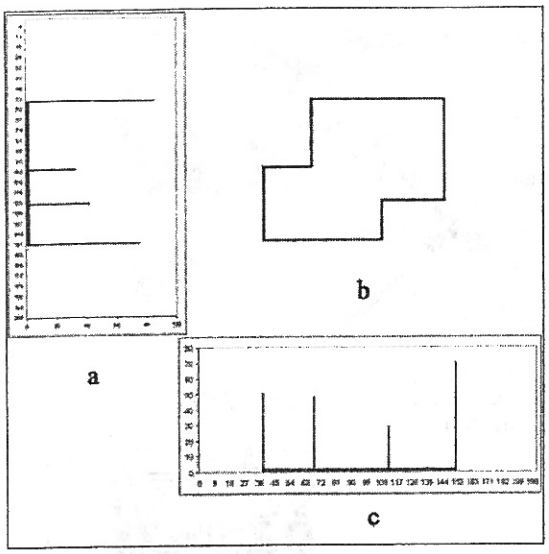

Fig. 4. Vertical and horizontal projections of one simple region

(border) or it is not. Now the expressions (1) and (2) give the number of pixels along vertical or horizontal axis that are part of the region or its border. The vertical and horizontal projections by number are used in the paper because regions of interest are not colored and they are defined by their border.

The horizontal and vertical projections by number can be used for simple region description. The maximum of region vertical projections represents region height, and the maximum of region horizontal projections represents region width. On Fig. 4b. is represented one simple region defined by its border. It is black/white image. Fig. 4.a. is showing its horizontal projections by number, and Fig. 4.c. its vertical projections by number. Peaks of projections indicate region borders along the projection directions.

IV. REGION EXTRACTION

The first phase in OMR algorithm is document preparation.

Data collected in this phase is the same for all digitalized images of filled questionnaires that have the same form: regions with answers, number of sub-regions in that region, question index for that region and page number. This phase is executed only once on digitalized image of the empty (not filled) questionnaire.

Questionnaire text usually consists of: questions, regions with answers (sub-regions for answer mark) and each sub-region explanation. Segment of digitalized document with sub-regions for answer mark is only one of interest for marked answer optical recognition while question segment and explanation segment are ignored. One region of interest is shown in Fig. 3. where none of its sub-regions is marked.

Region of interest can be detected in two ways: the first one is completely manual and the other way is to automatically detect regions of interest and their sub-regions with small possibility of user correction.

In the first case, a user should mark region of interest on digitalized questionnaire image. Easiest way is to draw rectangle round the region of interest, Fig. 3. and to calculate drawn rectangle coordinates, or to enter its coordinates, and to enter the number of sub-regions.

In the second case, region of interest are detected by horizontal projections, expression (2). A questionnaire document form does not have to be predefine or in some way marked, but the sub-region explanation should be over or under mark field but not in the same line. Explanations in region shown in Fig. 3. are beneath sub-regions. When explanations are beside mark fields (sub-regions), as it is shown in Fig. 5, optical recognition (data extraction) is far more complicated because there must be a way to detect projections sub-regions that correspond to explanation text and this projections sub-regions should be ignored during data extraction.

The horizontal projections of a part of digitalized questionnaire, Fig. 6, is shown in Fig. 7. When question, answer and explanation areas are clearly separated by line space, it is easy to distinguish them and it is easy to find where are the regions of interest on the digitalized image of the questionnaire. The results of horizontal projections are separated and precisely defined projections regions. Projections of mark sub-regions are much narrower then the projections of questions and explanations because the latest have a lot of pixels. Ratio between them is approximately 3:1. Their order is not important in region of interest detection.

Regions of interest can be detected in the following steps:

1. All projections regions are detected by examining values of horizontal projections on vertical axis. The coordinates on which groups of connected pixels start and end are vertical coordinates of projections regions.

2. Regions of interest are detected among the projections regions. Size of some questions and explanations vertical projections are not larger then size of answer regions vertical projections, Fig. 7. Therefore regions of interest can not be detected comparing the size of projections regions.

All the detected projections regions of interest have the same shape, and it differs from questions and explanations projections regions. The solution is to move the vertical axis and to track down how it is intersecting with pixels inside detected projections regions. This is described in the following pseudo Pascal code:

```
while (not (All Region Examined)) do
begin
  Move Vertical Axis By One Pixel;
  if (Number Of Pixels In Region That Intersect Vertical
  Axis Drastically Change) then
  begin
    Find Region Starting And Ending Coordinates On
    Current Vertical Axis;
    Move Horizontal Axis By One Pixel;
    if (There Are Pixels On Region Starting And Ending
       Coordinates) and
```

c2) Da li nastavnik/asistent početku svakog časa daje jasan koncept sadržaja predavanja/vežbi?

☐ uvek ☐ ponekad ☐ nikad

Fig. 5. Sub-regions with explanation beside them

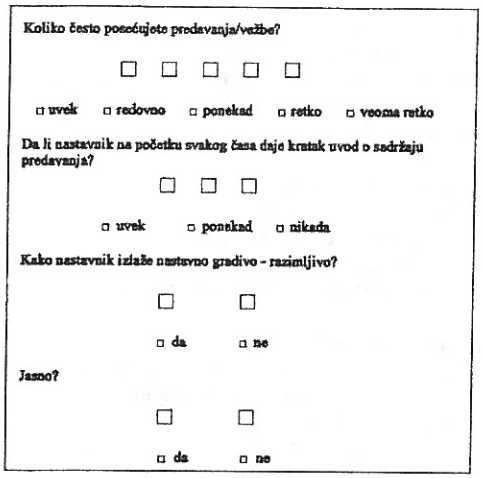

Fig. 6. Part of a digitalized questionnaire

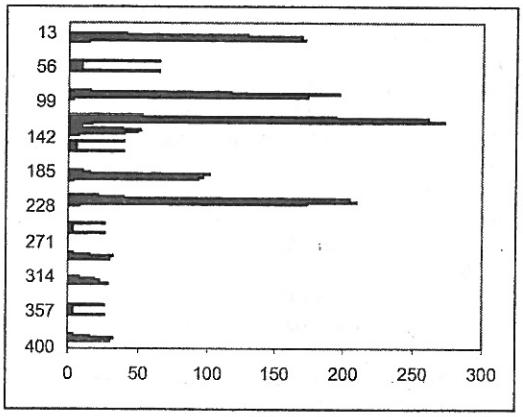

Fig. 7. Horizontal projections of digitalized questionnaire from Fig. 6.

(Pixels Between Region Peaks Do Not Intersect Vertical Axis) ***and*** *(Region Neighbors Are Not Regions Of Interest)* ***then***
This Is Region of Interest;
else
This Is Not Region of Interest;
end
end

Projections region of interest vertical coordinates are defined with coordinates of its peaks. Nevertheless, it is a good practice to take higher region than detected because digitalized questionnaire may be dislocated or skewed.

There is a way to speed up the region of interest detection, by marking the area with answers. Left margin of the paper is empty everywhere except where the regions of interest are, Fig. 8. Now, horizontal projections are applied only on left margin in this way. The vertical coordinates of projections regions of interest are easy to extract and define.

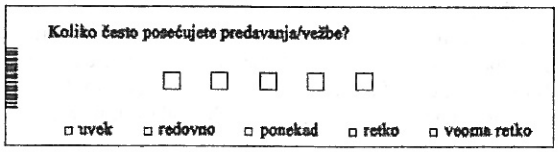

Fig. 8. Signed region of interest

Every region of interest has to be related with the corresponding question by question index. The easiest way is to count detected regions of interest, following digitalized document from top to bottom, first detected region corresponds to first question on that page, second region corresponds to second question and so on. The other solution is to mark every answer region with some kind of bar-code which contains question index, Fig. 8. In this way, errors that could occur because question is on one page and region with answers on other, are avoided.

There are two ways to detect the page number of the digitalized questionnaire that the region is on. First way is to count the pages from beginning of the region extraction process. The first processing page is the first page of digitalized questionnaire, and so on. This way is not good enough to detect page numbers since there is no guarantee that digitalized page processing order is the same as page order in questionnaire, or that pages that are going one after another are from the same filled questionnaire. Thus each of the page should be marked with some kind of bar code that contains page number and unique questionnaire identification. Before data extraction, bar-code of the digitalized questionnaire is extracted and analyzed. The other solution would be to extract the page number using OCR algorithm. Depending on page number, the corresponding set of answer region coordinates are used for data extraction.

V. SUB-REGIONS AND MARKED ANSWERS EXTRACTION

This section describes a detection of sub-regions inside the region of interest, and detection of marked answers. In both cases, the vertical projections, expression (1), are used.

After the regions of interest are detected, it is proceeded to sub-regions detection in document preparation phase. Vertical projections by number of region defined in Fig. 3. are shown in Fig. 9. The projections pixels are now grouped in projections sub-regions. The number of projections sub-regions is equal to the number of offered answers on certain question.

The optical recognition of marked sub-region could be done by simple algebra. In marked sub-region extraction

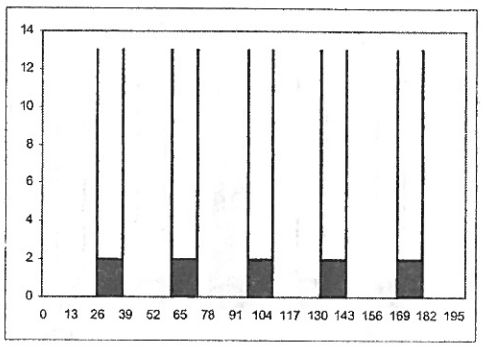

Fig. 9. Vertical projections of region from Fig. 3.

phase, the values taken from vertical projections applied on digitalized image of filled questionnaire are compared to vertical projections values of empty questionnaire digitalized image. If subtraction of projections sub-region extracted from filled questionnaire and projections sub-region extracted from empty questionnaire differs (is not zero) than this sub-region is marked. Drawback of this approach is that there is strong pixel-position and pixel-value dependence. If scanned questionnaire paper is dislocated or skewed the subtraction will not be accurate and that would bring to corrupted data extraction. Besides, if histogram of digitalized images of empty and filled questionnaire distinguish, the subtraction will not be accurate and extracted data will be corrupted as well.

Another approach is to store sub-region vertical projections peaks coordinates. These coordinates are also projections sub-region horizontal coordinates inside the region of interest that covers them. Only projections values within these coordinates are examined in marked sub-region extraction phase. Marked sub-region is detected comparing values within these coordinates. This approach is strong pixel-position dependence as well. If scanned questionnaire paper is dislocated, sub-region projections peaks could get inside area that is examined.

The reminder of this section describes two algorithms for marked answer extraction based on the vertical projections that are independent of dislocation during questionnaire scanning and are independent of digitalized image histogram changes.

Marked sub-region optical recognition process is the same for all defined regions of interests no matter how many sub-regions are there and it is executed in the following two steps.

1. This step is the same in both algorithms and it executes vertical projections, expression (1), inside defined region. The vertical projections of region with one marked sub-region, Fig. 10, is shown in Fig. 11.

2. In this step, the vertical projections values are examined and marked sub-region is detected, if there is

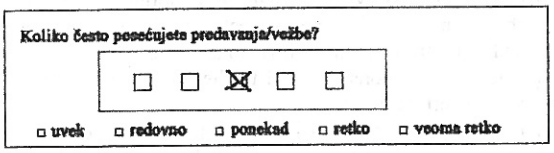

Fig. 10. Region with marked answer

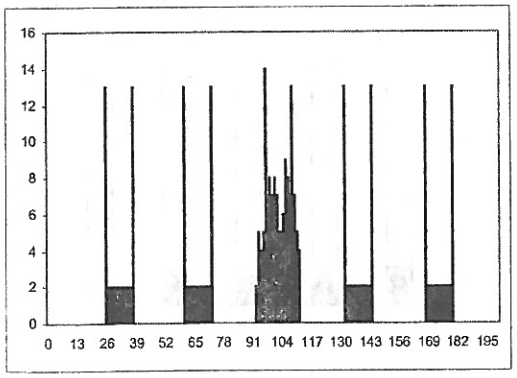

Fig. 11. Vertical projections of region from Fig. 10.

any. Both algorithms are described in pseudo Pascal code.

The first algorithm is simpler and it is based on projections sub-region size comparison. The projections sub-region size is space covered by sub-region vertical projections quantified by number of pixels within sub-region projections.

```
Calculate Sub Regions Size;
Sort Sub Regions By Their Size;
if (The Largest Sub Region Size / The Second
  Largest Sub Region Size > Minimum Value) then
  The Largest Sub Region Is Marked Sub Region
else
  There Is No Marked Region;
```

Second algorithm is based on moving of the horizontal (null) axis. It is examined how sub-regions projections values intersect the null axis, Fig. 12.

```
Previous := Number of Pixels That Intersect Axis;
Number of Intersect Pixels Decreased := false;
Number of Intersect Pixels Stabilized := false;
while (not((End Of Image)or(Number of Intersect Pixels
      Decreased))) do
begin
  Move Horizontal Axis By One Pixel;
  Current := Number of Pixels That Intersect Axis;
  if (Previous – Current > Minimum Value) then
  begin
    Previous := Current;
    while (not(End Of Image)or(Number of Intersect
          Pixels Stabilized)) do
    begin
      Number of Intersect Pixels Decreased := true;
      Move Horizontal Axis By One Pixel;
      Current := Number of Pixels That Intersect Axis;
      if (Previous – Current < Small Value) then
        Number of Intersect Pixels Stabilized := true;
    end
  end
  else
    Previous := Current;
end
if (Number of Intersect Pixels Stabilized) then
begin
  for (All Sub Regions) do
    if (All Pixels In Sub Region Intersect Null Axis) then
      This Is Marked Sub Region
  if (Marked Sub Regions Number > 1) then
    Error: To Many Marked Regions;
end
```

In a case that number of detected sub-regions is smaller than expected (their accurate number was detected in the first phase of OMR algorithm), some error occurred. The error could occurred either during questionnaire scanning or a questionnaire participator answered marking two (or more) sub-regions and/or the whole space between them. If the number of detected sub-regions is bigger then expected, questionnaire participator put marker somewhere between sub-regions and it has no contact with nearby sub-region borders. In a case that marker is connected with sub-region like in Fig. 2c, the sub-region size would be only bigger and the third described OMR algorithm would detect it as marked.

The fourth solution has an advantage over the third one since it detects the marked sub-region even when the

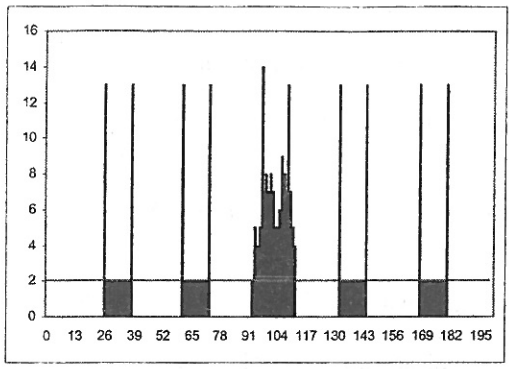

Fig. 12. Marked sub-region detection moving the Null Axis

difference in projections size of the marked and non marked sub-region is small (because of noise, error during scanning or questionnaire participator made to small marker).

VI. CONCLUSION

The paper presents an extraction of answers from hard-copy questionnaire. The vertical and horizontal projections were used in digitalized document preparation and marked answers extraction. Algorithms based on vertical and horizontal projections represent qualitative and flexible technique for a marked answer extraction. Position and number of mark sub-regions do not have to be predefined.

However, OMR algorithm based on projections is not without flaws. It is noise sensitive. Noise is the consequence of questionnaire paper scanning. Noise pixels can produce some extra regions or sub-regions during horizontal or vertical projections. Fortunately, this projections regions or sub-regions are small in size and contain small number of pixels, so they can be ignored. Smoothing filters are seen as a natural solution for this problem, but they could also corrupt valid pixels, making projections regions and sub-regions of interest blurred and making their borders harder to detect, as well.

The OMR algorithms described in this paper where designed for black/white and grayscale images. Dilemma in marked answer extraction from grayscale digitalized questionnaire is which value to take as threshold in vertical and horizontal projections by number. All pixels with value smaller from background value are region pixels. Background value is defined with threshold. Threshold value can be found using algorithms based on digitalized questionnaire image histogram (brightness threshold) [2]. When digitalized questionnaire image is black/white this problem is eliminated. However, pixel value depends on quality of the scanner and its internal logic in deciding if scanned pixel value is 1 or 0. In this case border of projections region/sub-region can be easily corrupted.

Digitalized questionnaire image may become skewed during scanning (although better scanners detect skewing). If that happens, results from vertical and horizontal projections drastically change and differ from results showed in the paper. There are efficient algorithms that can detect skew orientation for digitalized document [3]. One solution is to mark the area with answers on questionnaire papers, Fig. 8. If horizontal projections of this areas deviates from expected values the digitalized questionnaire image is skewed. The next step is to detect skew orientation and intensity, and to use projections along skew direction [3].

Described OMR algorithms are sensitive on image resolution. If image resolution of digitalized questionnaire is higher there are more pixels that define projections region/sub-region and difference in projections size of marked and non marked sub-region is bigger and easier to detect. However, there are more pixels that represent digitalized questionnaire and that means more work and needs more computing power. If image resolution of digitalized questionnaire is smaller there are fewer pixels that define projections region/sub-region and difference in projections size of marked and non marked sub-region is smaller and hard to detect. But in this case there is less work and needed computing power is smaller. In system for Faculty/University evaluation questionnaires were scanned in 300x300 resolution. It made a good balance between system requirements, answer extraction speed and OMR algorithm efficiency.

System for Faculty/University evaluation can be improved in several ways. This improvements should include OCR algorithms to automatically extract questions and sub-regions explanation. Sub-region context would be automatically detected and linked to them. Another goal is to expand questionnaire with open answers. Developing OCR algorithm for Latin and Cyrillic handwritten text extraction could provide this functionality. This improvements are topics of future research.

VII. REFERENCES

[1] TEMPUS project No UN_JEP-16079-2001 *Institutional Evaluation at the Agricultural Universities*

[2] Milan Sonka, Vaclav Hlavac, Roger Boyle, *Image Processing, Analysis and Machine Vision*, 2nd Ed., PSW, 1998.

[3] Diego Nehab, *Staff Line Detection by Skewed Projection*, PHD thesis, Princeton University, May 2003.

[4] Andrew M. Smith, "Optical Mark Reading – make it easy for users," *Proceedings of the 9th annual ACM SIGUCCS conference on user services*, October 1981.

[5] Bilan Zhu, Masaki Nakagava, "Document and images analysis: Information encoding into and decoding from dot texture for active forms," *Proceedings of the 2003 ACM symposium on Document engineering*, November 2003.