ISSN 2278-3091

Volume 12, No.4, July - August 2023

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse071242023.pdf

https://doi.org/10.30534/ijatcse/2023/071242023

Recognition of Hindi Character Using OCR-Technology: A Review

Ravi Raj[1], Andrzej Kos[2]

[1, 2]Faculty of Computer Science, Electronics, and Telecommunications
AGH University of Science and Technology,
Al. Adama Mickiewicza 30, 30-059 Kraków, Poland
raj@agh.edu.pl , kos@agh.edu.pl

Received Date June 21, 2023 Accepted Date: July 29, 2023 Published Date: August 06, 2023

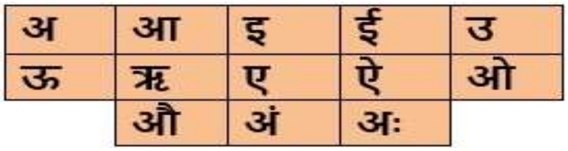
ABSTRACT

Recognition of character is a technique that enables the transformation of various kinds of scanned papers into an editable, readable, and searchable format. In the last two decades, several researchers and technologists have been continuously working in this field to enhance the rate of accuracy. Recognition of character is classified into printed, hand-written, and characters written at image recognition. Recognition of character is the major area of research in the field of pattern recognition. This paper presents an overview of Hindi character recognition by utilizing the optical character recognition (OCR) technique. We surveyed some major research breakthroughs in character recognition, especially for Hindi characters. This research article focuses to provide a deeper insight into the researchers and technologists working in the field of recognition of Hindi-character.

Key words: Classification, Feature extraction, Hindi characters, OCR, pattern recognition.

1. INTRODUCTION

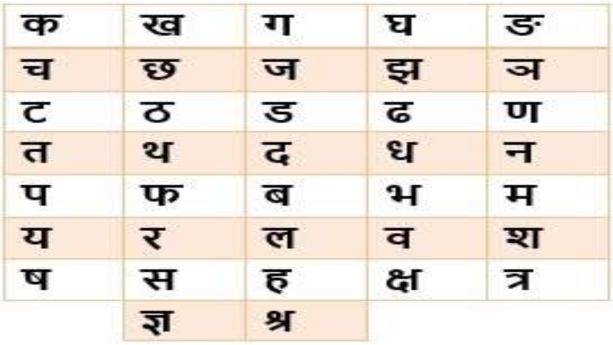
Hindi is the fourth largest language in the world, after English, Mandarin, and Spanish, if it is taken alone but, the Hindi language can be the third most-spoken if it is considered together with Urdu, after English and Mandarin. It is also known as the Devanagari language. There are 322 million people on the earth, who have Hindi as their first language, and 270 million people in the world, have Hindi as their second language. Recognition of handwritten characters is becoming a very exciting field of research because the resource of data is so big.



अ	आ	इ	ई	उ
ऊ	ऋ	ए	ऐ	ओ
	औ	अं	अः	

Figure 1: List of Hindi language vowels

Recognition of Hindi- characters is a generally difficult task as compared to the English language due to similarities in the Hindi characters, such as अ, आ, etc. [1]. Figures. 1 and 2, indicate the number of vowels and consonants in the Hindi language respectively. There are 13 vowels and 37 consonants in this language, which means there is a total of 50 alphabets in the Hindi language.



क	ख	ग	घ	ङ
च	छ	ज	झ	ञ
ट	ठ	ड	ढ	ण
त	थ	द	ध	न
प	फ	ब	भ	म
य	र	ल	व	श
ष	स	ह	क्ष	त्र
	ज्ञ	श्र		

Figure 2: List of Hindi language consonants

Various OCR systems are already augmented and utilized for several commercial purposes such as the processing of bank cheques, recognition of postal addresses, the assistance of mail reading for the blind, automatic recognition of number plates, and processing of forms. It is very tough for a system to recognize Hindi- characters because of different styles of writing, the color of ink, the actuation device, the width of the pen, and many other factors.

Thus, the development of a system for the recognition of hand-written Hindi- characters contains a bigger challenge for technologists and researchers [2]. Various methodologies are utilized for character recognition such as the convolutional neural network method, incremental recognition method, semi-incremental method, etc. Figure 3 illustrates the process of recognition through OCR technology. We present a review of the methodology of Hindi character recognition using OCR.

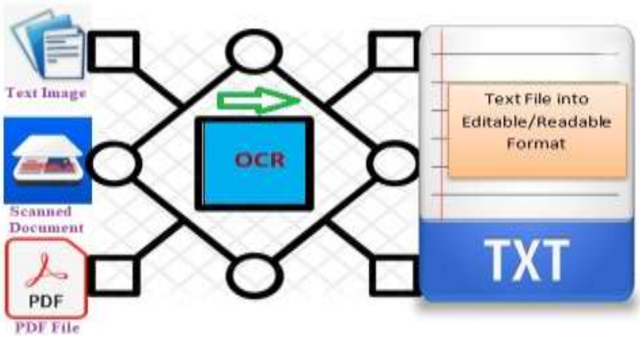


Figure 3: Process of recognition by OCR- technology.

2. LITERATURE SURVEY

In recent years, various researchers have focused on the recognition of Hindi characters with the help of several techniques of character recognition. D. Singh *et al.* [3] present a radial basis function neural network for the recognition of hand-written Hindi characters through the implementation of gradient features of eight directional values. The system of recognition of characters is trained on the collections of several styles of handwritten characters. The result shows that this technique is better with accuracy, and the time for training and classification is very less concerning the result gained by neural networks of backpropagation. V. Bansal *et al.* [4] describe a philosophy of the recognition system for the Devanagari documents. The source of knowledge utilized is mainly statistical or tailored from a word dictionary generally for optical character recognition (OCR). This article has presented a complete OCR system of Devanagari and tested it with printed documents of the real life of varying fonts and sizes. Here almost all documents are photocopies of the main document. The rate of accuracy of recognition of character is approximately 90%. G. B. Kshirsagar *et al.* [5] develop an architecture of deep learning for the Devanagari script-based speller of P300 which can recognize the target characters with better accuracy and in very few numbers of trials. Here two types of algorithms of deep learning, deep convolutional neural networks, and stacked autoencoder have been utilized. The proposed technique has been processed on 20 Devanagari words of the self-generated dataset with 79 characters gathered from 10 subjects utilizing 16 channel actiCAP Xpress of EEG recorder. The rate of accuracy for this proposed approach is around 88.22%. A. Bharath *et al.* [6] present a character recognition associated problem significantly for two main Indic scripts-Tamil and Devanagari. Two different techniques for the recognition of the word depend on Hidden Markov Models: lexicon-free and lexicon are driven. On hand-written samples of Hindi words featuring both non-standard and standard orders of writing symbols, an integration of lexicon-free and lexicon-driven recognizers remarkably outperforms both, utilized alone. The rate of accuracy for the Devanagari language in this experiment is 87.13%. D. Gupta *et al.* [7] propose an efficient approach to character segmentation for Hindi text by the utilization of polygonal approximation to obtain the digital segment straight line of the handwritten Hindi words. After that, the traversing algorithm is applied to get the character's segmentation. The rate of accuracy for this proposed approach is around 95.70%. V. K. Verma *et al.* [8] present a recognition technique for Hindi scripts that gives better accuracy in less time than the experiment. This article is mostly focused on the handling of different types of fonts such as kruti Dev 714, Alekh, and DevLys 240, and variable writing styles. The proposed system is based on OCR technology for enhancing the capabilities of recognition accuracy and identification of multiple fonts of Devanagari scripts. G. Senthil *et al.* [9] present a new methodology of creating variable images of hand-written Hindi words utilizing only characters of hand-written Hindi language and analyze the effectiveness in allowing less instance learning of documents of hand-written Hindi language. This article studies the challenges in the segmentation of Hindi characters during the recognition process.

Table 1. Description of some major hand-written Hindi character recognition systems.

Authors	Methodology	Feature	Classifier	Dataset (Size)	Rate of accuracy
N. Sharma et al. [10]	fivefold cross-validation technique	Chain code	Quadratic	11,270	80.36%
B. Shaw et al. [11]	Hidden Markov Model based system	Segments	Hidden-Markov Model	39,700	84.31%
S. Arora et al. [12]	Differential distance-based system	Structural	Feed-forward neural network	50,000	89.12%
M. Hanmandlu et al. [13]	Exponential membership model system	Vector distance	Fuzzy sets	4,750	90.65%
S. Kumar [14]	Feature extraction-based system	Gradient	Support Vector Machine and Multilayer Perceptron	25,000	94.1%
U. Pal et al. [15]	Quadratic classifier-based system	Gradient and Gaussian filter	Quadratic	36,172	94.24%
U. Pal et al. [16]	Comparative study or survey methodology	Gradient	Mirror Image learning	36,172	95.19%

Table 1 is the collection of some major techniques, research methodologies, number of samples, type of classifier, type of feature extraction, and corresponding rate of accuracy for every methodology, which are taken for the recognition of hand-written Hindi characters in the past.

3. RECOGNITION SYSTEMS

Several OCR systems have already developed since their evolution in the 1950s. There are many OCR systems available for commercial applications. Recognition of character has always been an interesting ground in testing novel concepts in pattern recognition. OCR technology is used in pattern recognition for the recognition of different numerals and characters in several languages. Recognition of character is utilized in various applications such as forensic research, banks, old document analysis, postal services, etc.

The recognition of the type of characters is classified into two categories, one is for printed characters and the second is for handwritten character recognition [17]. The process of the OCR system has mainly divided into six parts: a collection of datasets, pre-processing, feature extraction, classification, post-processing, and finally getting output into readable or editable format text, respectively [18]. Figure 4 illustrates the complete procedure of character recognition through OCR technology.

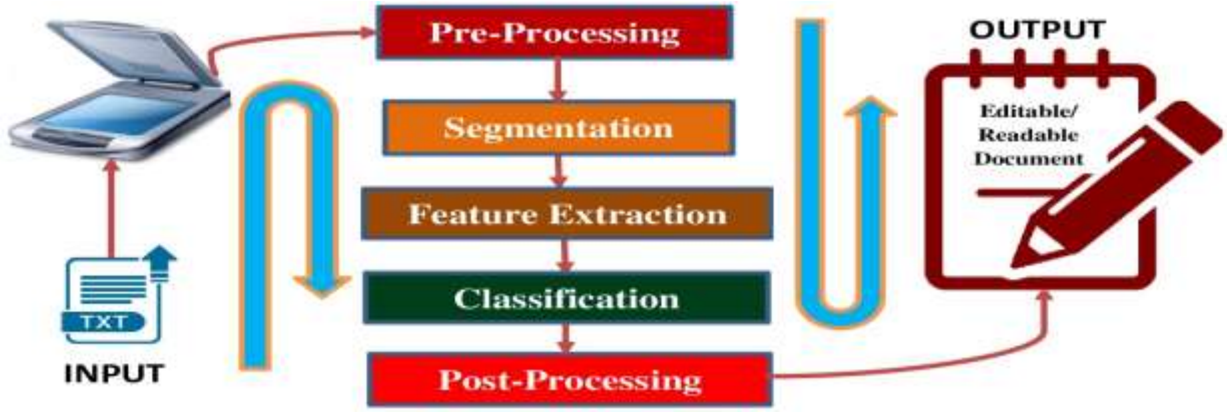


Figure 4: A complete process of character recognition through OCR-technology

A. *Input Dataset*

It is the first step in the recognition of characters procedure, which is known as data acquisition. The documents of hand-written characters are scanned, and digital images of corresponding scanned documents are generated. Further, these digital images are transferred to the next step which means preprocessing. The rate of accuracy depends upon the quantity of data that has been processed in the character recognition procedure.

B. *Preprocessing*

It is the first step in OCR technology. It transforms scanned images into the best format and then removes unwanted and noisy backgrounds. Preprocessing steps include binarization, skew corrections, noise reduction, normalization, etc. Figure 5 explains the step of preprocessing in the best way.

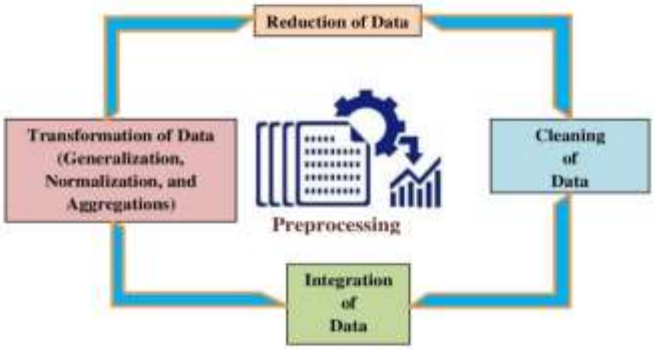

Reduction of Data
Transformation of Data (Generalization, Normalization, and Aggregations)
Cleaning of Data
Preprocessing
Integration of Data

Figure 5: A list of work that happens in the preprocessing.

C. *Segmentations*

The process of segmentation is an important step in any image processing, due to its essential problematics and the cruciality of its outcomes, which are significant for the efficiency at the global level of the image processing system [19]. The main empirical of segmentation is to characterize every variable region available in any specific text image. Therefore, characters and words are retrieved from given text images. Figure 6 represents an example of the segmentation of a Hindi word.

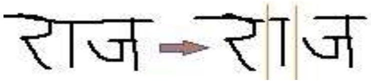


Figure 6: An example of Hindi word segmentation

D. *Feature Extraction*

The feature extraction process plays a vital role in the classification of text, it directly impacts the rate of accuracy of recognition of characters. The possible feature extraction process in the OCR technology includes mapping, filtrations, clustering, and fusion method. For every input image, particular features are retrieved and stored in the form of the feature vector. This feature vector is classified into structural and statical features.

Those features are defined because topological features are known as structural features, which reflect both the local and global properties of text images. Topological features are utilized to define object properties, elements, and structure. Some important examples of these features are extreme points, intersections, loops, endpoints, etc. [20].

Statical features are of two types: local and global features. The features which are collected from the complete image of the character are called global features, but local features are collected from the local area of the character images. Projection, histogram, moments, n-tuples, zoning, and distances are the best examples of statical features. Also, distortion and noises do not have any impact on local and global features [20].

E. Classifications

Classification methods of feature vector-based, which is structural methods, particularly in recognition of off-line character. These techniques include support vector machines, multiple classifiers, artificial neural networks, and statical methods integration. Classification is the step where characters are first recognized. Further, the characters which are recognized can be classified into the class of predefined tasks. This step is providing the final output which means it gives characters separately in their boxes. The output from this step is the digitized form of the input text, which can be edited. If any noise or errors is coming, then further steps are required otherwise not. Figure 7 represents an example of classified characters from a Hindi text.

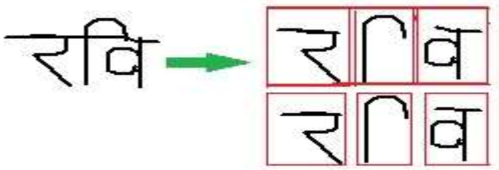


Figure 7: An example of characters classifications

F. Post Processing

Post-processing is the final part of OCR technology which involves steps for cleaning data for digitized documents getting from the classification step, such as a newspaper or a book text. The major work in this procedure is the identification of grammar and spelling errors, and further correcting them, which are created due to OCR system flaws.

G. Output

The final step towards the recognition of Hindi characters is the outcomes of the OCR technology. We will get the Hindi words or characters in the format of readable or editable text. Finally, we will get the scanned text images into the digitized format, which can be edited and further processed for the specific task assigned to that text.

The above methodology can be implemented in the real world with the help of various software such as Python, Java, C++, MATLAB, etc. The most common library tools utilized in Python are py-tesseract, Keras-OCR, Easy-OCR, OCRmyPDF, Tesseract-OCR techniques, etc. for character recognition. These all-library tools are extensively used by researchers and technologists to develop several character recognition systems.

4. FUTURE RESEARCH TRENDS

The scientist Edmund Edward Fournier d'Albe invented OCR technology 100 years ago for supporting blind people to read text. Recognition of character technology is the most trending topic for researchers nowadays. Whether it is automatically collecting data from a scanned voucher for a report of expenses or foreign language translation by using a camera on the phone, OCR technology can seem fascinating. Although it seems marvelous that we have systems that can digitize text of analog type with a higher rate of accuracy, the actuality is that the required rate of accuracy is still not sufficient [21]. OCR technique can be implemented in mobile robots to provide an autonomous navigation system for them. There are very few numbers of researchers are focusing on utilizing character recognition systems for mobile robots. We can create a predefined environment, which has Hindi characters in both directions of the path to indicate or provide assistance to mobile robots for intelligent navigation.

Text or language characters are an important part of human lives. Automatic character recognition in natural scenes has a vital experimental value. Thus, recognition of image text or scene text has become a vibrant and important area of research in pattern recognition and computer vision. In recent years, there has been exponential growth in innovation, efficiency, and practical implications of recognition techniques [22]. However, there are more research possibilities remain for future research aspects:

C. The Ability of Generalizations

The ability of generalization refers to the capacity of algorithms recognition to be effective for any range of input datasets and applications. While the algorithms of recognition and training with the limited number of datasets achieve better efficiency on various real-world datasets evaluation, they are not enough to adapt to variable inputs, such as words having smaller size characters, unknown font size, and longer characters. So, we need a system that can be applied to any general problem of character recognition.

D. Languages

There are more than thousands of languages in this world and most of them have different alphabets, different styles of writing, and variable accents of speaking. Therefore, we need an excellent system that can be suited for multilingual functionalities. As well as existing algorithms for character recognition cannot be sufficient for different languages. Thus, algorithms for progressing language-dependent recognition systems for any specific language can be a general solution.

E. Issues in Data

Most of the recent character recognition technology mostly depends upon enough input datasets to get high efficiency. The existing datasets only have thousands of sample data, which is generally small for text recognizer training. Although, collecting and annotating a huge number of data manually will need huge resources and effort. Thus, we can work efficiently to resolve the issues of data. For better utilization of this technique in the real world, researchers must develop individual datasets for every language separately. Although it is a very tough task to generate individual datasets for every language, once it is finalized the rate of accuracy will be around 100%.

F. Security

In recent years, character recognition technologies have been adapted into various scenarios of private vision applications such as ID cards, bank cards, and many more, thus the security of every recognition methodology is very important. Although, high efficiency, most artificial intelligence-based recognizers of text are mostly vulnerable to confrontational examples. Thus, there are various opportunities for future research are possible.

As well as some other areas of research such as end-to-end systems, evaluation protocols, etc. still have the possibility for future research work in that area. The main problem for the researchers is the rate of accuracy and timing so most of the researchers and technologists are focusing on this issue. Nowadays, most researchers are focusing on enhancing accuracy and reducing the time for recognition by integrating this technology with machine learning and deep learning [23]. Thus, more researchers are required in this field as well as more amount of funds is required to get better outcomes for future demand.

5. CONCLUSION

This paper provides an overview of character recognition using OCR technology. Here we have taken Hindi characters as a reference for properly showing this technology. In this paper, we illustrate a comprehensive survey on the recognition of hand-written Hindi characters. It also provides an overview of mostly utilized techniques for feature extraction and preprocessing. As well, we have demonstrated every step of OCR technology and some basic ideas of Hindi language knowledge. This work allows researchers to get basic ideas about OCR technology for Hindi or other language character recognition systems towards enhancing the accuracy of recognition systems. This paper allows early-stage researchers to get the basic knowledge of the recognition of characters. Also, we have discussed in detail the future research trends of character recognition, which can provide an outline of further research in this area of study.

ACKNOWLEDGEMENT

The research was supported financially by the AGH University of Science and Technology, Krakow, Poland, subvention no. 16.16.230.434.

REFERENCES

1. D. Chaudhary and K. Sharma, "Hindi Handwritten Character Recognition using Deep Convolution Neural Network," 2019 6th International Conference on Computing for Sustainable Global Development (INDIA.Com), 2019, pp. 961-965.
2. A. K. Singh, B. Kadhiwala, and R. Patel, "Hand-written Hindi Character Recognition - A Comprehensive Review," 2021 2nd Global Conference for Advancement in Technology (GCAT), 2021, pp. 1-5, doi: 10.1109/GCAT52182.2021.9587554.
3. D. Singh, J. P. Saini, and D. S. Chauhan, "Hindi character recognition using RBF neural network and directional group feature extraction technique," 2015 International Conference on Cognitive Computing and Information Processing (CCIP), 2015, pp. 1-4, doi: 10.1109/CCIP.2015.7100726.
4. V. Bansal and R. M. K. Sinha, "Integrating knowledge sources in Devanagari text recognition system," in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 30, no. 4, pp. 500-505, July 2000, doi: 10.1109/3468.852443.
5. G. B. Kshirsagar and N. D. Londhe, "Improving Performance of Devanagari Script Input-Based P300 Speller Using Deep Learning," in IEEE Transactions on Biomedical Engineering, vol. 66, no. 11, pp. 2992-3005, Nov. 2019, doi: 10.1109/TBME.2018.2875024.
6. A. Bharath and S. Madhvanath, "HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 670-682, April 2012, doi: 10.1109/TPAMI.2011.234.
7. D. Gupta and S. Bag, "An Efficient Character Segmentation Approach for Handwritten Hindi Text," 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), 2018, pp. 730-734, doi: 10.1109/SPIN.2018.8474047.
8. V. K. Verma and P. K. Tiwari, "Removal of Obstacles in Devanagari Script for Efficient Optical Character Recognition," 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 433-436, doi: 10.1109/CICN.2015.90.
9. G. Senthil, K. Nandhakumar, and G. R. K. S. Subrahmanyam, "Handwritten Hindi Word Generation to enable Few Instances Learning of Hindi Documents," 2020 International Conference on Signal Processing and Communications (SPCOM), 2020, pp. 1-5, doi: 10.1109/SPCOM50965.2020.9179634.
10. N. Sharma, U. Pal, F. Kimura, and S. Pal, “Recognition of off-line handwritten Devanagari characters using Quadratic classifier”, In Proc. Indian Conference on Computer Vision Graphics and Image Processing, 2006, pp. 805-816.
11. B. Shaw, S. K. Parui, and M. Shridhar, "A Segmentation Based Approach to Offline Handwritten Devanagari Word Recognition," 2008 International Conference on Information Technology, 2008, pp. 256-257, doi: 10.1109/ICIT.2008.32.

12. S. Arora, D. Bhatcharjee, M. Nasipuri, and L. Malik, "A Two Stage Classification Approach for Handwritten Devnagari Characters," International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), 2007, pp. 399-403, doi: 10.1109/ICCIMA.2007.254.
13. M. Hanmandlu, O. V. R. Murthy, and V. K. Madasu, "Fuzzy Model Based Recognition of Handwritten Hindi Characters," 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications (DICTA 2007), 2007, pp. 454-461, doi: 10.1109/DICTA.2007.4426832.
14. S. Kumar, “Performance comparison of features on Devanagari hand-printed dataset”, International Journal of Recent Trends in Engineering, vol. 1, no. 2, May 2009, pp. 33-37.
15. U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, "Off-Line Handwritten Character Recognition of Devanagari Script," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, pp. 496-500, doi: 10.1109/ICDAR.2007.4378759.
16. U. Pal, T. Wakabayashi, and F. Kimura, "Comparative Study of Devanagari Handwritten Character Recognition Using Different Feature and Classifiers," 2009 10th International Conference on Document Analysis and Recognition, 2009, pp. 1111-1115, doi: 10.1109/ICDAR.2009.244.
17. S. Srivastava, A. Verma, and S. Sharma, "Optical Character Recognition Techniques: A Review," 2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2022, pp. 1-6, doi: 10.1109/SCEECS54111.2022.9740911.
18. R. Raj and A. Kos, "A Comprehensive Study of Optical Character Recognition," 2022 29th International Conference on Mixed Design of Integrated Circuits and System (MIXDES), June 2022, doi: 10.23919/MIXDES55591.2022.9837974.
19. R. Raj and A. Kos, " Different Techniques for Human Activity Recognition," 2022 29th International Conference on Mixed Design of Integrated Circuits and System (MIXDES), June 2022, doi: 10.23919/MIXDES55591.2022.9838050.
20. U. Garain and B. B. Chaudhuri, "Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 32, no. 4, pp. 449-459, Nov. 2002, doi: 10.1109/TSMCC.2002.807272.
21. Abhinav Somani, “The Future of OCR is Deep Learning”, Forbes technology Council, Sept. 10, 2019, Available online: https://www.forbes.com/sites/forbestechcouncil/2019/09/10/the-future-of-ocr-is-deep-learning (accessed on December 1, 2022).
22. X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, “Text Recognition in the Wild: A Survey”, ACM Computing Surveys, vol. 54, issue 2, pp. 1-35, March 2022, doi: 10.1145/3440756.
23. R. Raj, and A. Kos, “Artificial Intelligence: Evolution, Developments, Applications, and Future Scope”, Przegląd Elektrotechniczny, Vol.2023, Issue 2, pp. 1-13, February 2023. Doi: 10.15199/48.2023.02.01.