IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Correlating Image and Surrounded Text from Web Pages Using Convolution Neural Network

Nirmala[1] Gopalkrishna Joshi[2] Prakash Hitremath[2]

1Department of CSE Nitte Meenakshi Institute of Technology Bengaluru India, 2Department of CSE BVB college of Engineering and Technology Hubballi India, 3Department of CSE BVB college of Engineering and Technology Hubballi India

Abstract

Associating images with text has drawn much attention in the recent years. This helps extraction of relationship images with text in a meaningful fashion, store, and retrieve information instantly. Besides Search Engines and online shopping platforms, areas like the Medical Industry, Forensic Analysis, Social Media etc necessitate correlation of images with appropriate text data. For example, correlating patients' images with appropriate unstructured reports will help a doctor retrieve the patient's information instantly. Similarly, correlating forensic images and text will help the forensic analysis team with their analysis. Correlating sentiments from customer images with text will help social media and online shopping platforms analyze customer data and provide useful recommendation This paper focuses on extraction of relationship between the images and their corresponding text, and introduces a framework that combines Image Segmentation with Natural Language Processing to correlate images with appropriate text.

Keywords: Image, Text ,Recognition, correlation,Object

Introduction

The advancement of internet has become a vital source of information this may consist of different components like textual descriptions, captions, images etc. For example, Headlines, article text, images, and image captions are all components of news articles Headlines, article content, pictures, and image captions are all components of news stories. It is important to comprehend the relationship between these components in order to create a system that can automatically extract information from online data. Students of human anatomy must learn technical terms of domain specific terminology from several scientific and technical areas. Anatomic text book focus on the geometric properties and spatial relations for example on osteology chapter explains the features of structure and complex shape of the bones, myology chapter describes the structure and function of muscles along with the bone features, syndesmology describes the joints direction of movements. The examples explained above shows the major problems for medical students. (i) to learn the denotations of a large number of semantic concepts, (ii) to relate information from various sub-disciplines that are scattered throughout large documents or even over several publications, and (iii) the need to mentally reconstruct complex spatial arrangements.

Anatomy's ultimate objective is to provide a complete understanding of spatial relationships between objects and distinguishing characteristics that allow specialists to identify objects. Several evocative images complement textual descriptions in anatomy textbooks and can effectively convey visual characteristics and

spatial relationships. The expense of printing would make it impossible to include illustrations of every possible spatial configuration. This absence of legitimates in a specialized class of textbooks like anatomic atlases that only containing descriptions. Several numbers of same overlapping, tiny, thin or complex shaped objects have to be represented is the another characteristic illustration in the human anatomy. Because of the complexity of these spatial arrangements, a large variety of abstraction techniques have been developed, which also aid in focusing the learner's attention.

The social media helps in extracting tweets and analyzing the sentiments in tweets has attracted most of the researchers. People post photos of their daily lives and share short messages on social media platforms such as Twitter and Facebook. Sentiment polarity classification is a basic task for obtaining a better understanding of social behavior and providing service to users. Most articles in such platform tweets may consist of two sections, an image and a short text. A sentiment classifier must have the two sections, where multimodal methods or cross modal methods can be applied. The different modalities have individual semantic features which is one of the major challenges in the multimodal or cross-modal sentiment analysis. The images and the text may not be correlated in the tweet, which will have a major impact on the classification accuracy.

Image Segmentation

In digital image processing and computer vision, image segmentation is the process of partitioning a digital image into multiple segments (objects). The objective of segmentation is to make an image's representation more meaningful and easier to analyze by simplifying it. Objects and boundaries (lines, curves, etc.) in images are typically located using image segmentation. More precisely, Image segmentation is the process of assigning a label to each pixel in an image so that pixels with the same label share similar properties. Image segmentation generates either a set of segments that cover the whole image or a set of contours derived from the image (Edge Detection). In terms of some characteristic or computed property, such as color, intensity, or texture, each pixel in a region is similar. In terms of the same characteristic, adjacent regions have significantly different colors . When image segmentation is applied to a stack of images, as is common in medical imaging, the resultant contours may be used to create 3D reconstructions using interpolation algorithms like marching cubes.

Background and related work

Various methods on the correlation of text and images have been developed by the researchers. Few of them are discussed in this section. Nelleke Oostdijk et al. [1] proposed a case study on text and images that revels the inadequacy of assumptions about their connections and interaction. Their main goal is to develop automated systems that can extract event information from online news articles about flood disasters. They performed a manual analysis of 1000 articles containing flood-related keywords. They used an automatic classifier to validate the manual analysis, demonstrating the technological feasibility of multimedia analysis methodologies that allow for more realistic text-image interactions. Finally, case study demonstrates that paying more attention to the relationship between text and visuals can improve the collecting of multimodal data from news articles. Timo Gotzelmann et al. [2] presents the paper which talks about the concept and evaluation of a novel technique that uses coordinate textual descriptions and graphics to guide the students to understand complex spatial relations and acquire unknown concepts of a domain-specific terminology. This paper's main approach is to transform user interactions into queries to an information retrieval system. Here, the developed system uses pre-computed multi-level representations of the text and for suggesting textual descriptions on 3D models which support the present learning task. Finally, the study reveals that the results which are obtained are matched with the preferences of the users. Ruoxu Ren et al. [3] proposed an experimental study which is used to explore two problems that are often occurred in image retrieval tasks. The first problem is whether text similarity consistently implies image similarity. If it is no then the second problem is to find out the conditions for the implication. In this paper, they have used text mining techniques and an open source image recognizer to extract the text and images from the articles. To find out answers for the above two problems they have correlated. Ke Zhang et al. [4] proposes a cross model technique to classify image sentiment polarity that consider both images and captions. The correlation between textual information to image is transferred using this method. First, The image and caption are sent to a mapping-model, where they are transformed into vectors

and their labels are calculated using the MMD (Maximum Mean Discrepancy).Finally, LSTM classifies the sentiment polarity. Stephane Clinchant et al. [5] proposes a method in which the main goal is to introduce a set of techniques in order to efficiently fuse text and image retrieval systems in the context of multimedia information access. These techniques overcome a conceptual barrier rather than a technical one. Using four different image CLEF datasets, they test the proposed techiques against late and cross-media fusion. Juan C. Caicedo et al. [6] proposed a technique to fuse visual characteristics and unstructured text data in a medical image retrieval system. The main goal of this research is to investigate if semantic information from text descriptions can be translated into a visual similarity metric. A medical image collection from the ImageCLEFmed08 challenge is used to test the proposed technique. Finally, the results shown the improvement of the visual test fused approach with respect to only using visual information. Xin Zhou et al. [7] proposed classical approaches such as maximum combinations, sum combinations and multiplication of the sum and the number of non–zero scores were employed. Different types of normalization strategies were studied. The findings reveal that fused runs outperform the best original runs, and multimodality fusion exceeds single modality fusion statistically. Hironobu Takahashi et al. [8] proposes a method to make relationship between images and words. They have used two processes in this method. One method is to uniformly partition each image into sub-images with key-words, and the other method is to perform vector quantization on the sub-images. The result of these processes demonstrate that each sub image can be correlated to a set of words each of which is chosen from the list of terms assigned to full images. J Jeon et al. [9] proposed an automatic approach to annotating and retrieving images from the training set of images. Blobs can be used to describe those images. Clustering is used to generate these blobs using image features. They showed that probabilistic models allow us to predict the probability of generating a word given the blobs in an image from the set of images with annotations in the training data. Finally, results shows that the performance of the cross-media relevance model is almost 6 times as good when compared to the model based on the word-blob occurrence model. Henning Muller et al. [10] proposed a review of content -based image retrieval systems in medical applications and it provides clinical benefits and future directions. It gives an introduction to the image retrieval information and the technologies behind used. It explain the various approaches and propositions for the use of image retrieval in medical practice. It identifies clinical benefits of image retrieval system in clinical practice as well as in research. Trong-Ton Pham et al.[11] proposes a study of Latent Semantic Analysis (LSA) on different tasks. Multimedia document retrieval (MDA) and automatic image annotation (AIA). It deals with the study of the influence of LSA on the retrieval of a significant number of a multimedia documents and it shows how different image representations can be combined by LSA to improve automatic image annotation. Caroline Lacoste et al. [12] proposed a system based on the support vector machine (SVM) to enable the construction and learning of medical semantics from images. They have presented two different visual approaches within the framework: Global indexing to access image modality and local indexing to access semantic features. The goal of these two fusion techniques is to improve textual retrieval utilizing UMLS-image indexing.

Methodology

The purposed work to address the problem of correlation by extracting the details from images and text separately and then finding the intersection between the two. More than 1500 webpages like Britannica and other scientific journals were scraped using beautiful soup to extract the information. The Figure 1 depicts the flow diagram used in the system.

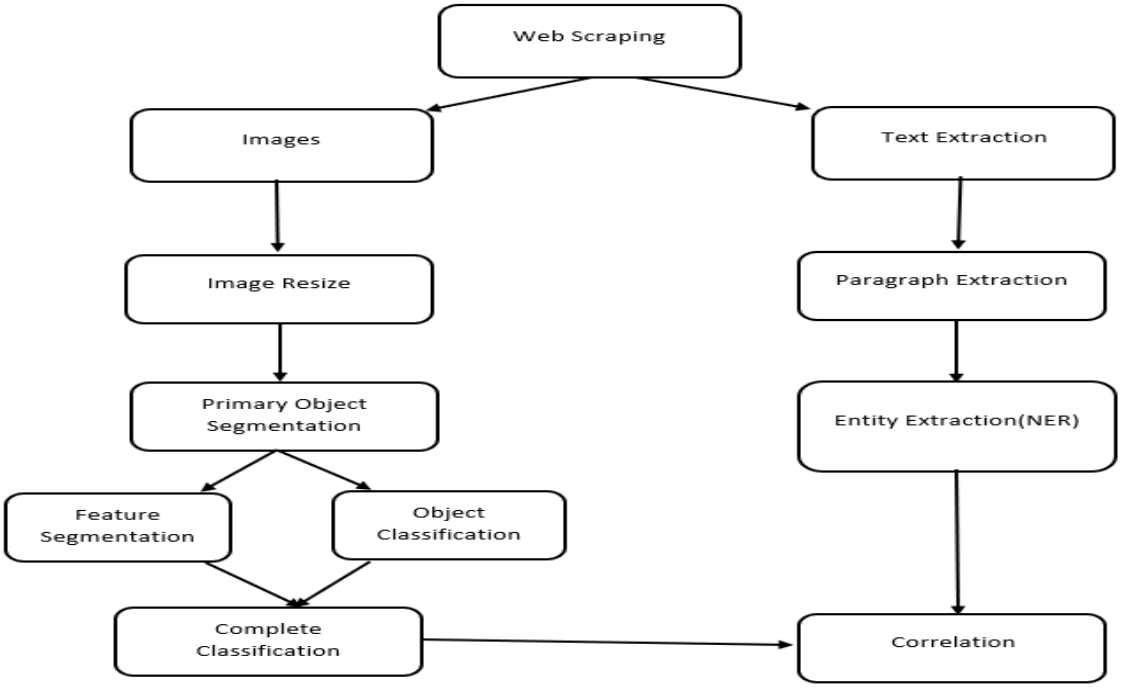


Figure1: Identification and classification of images correlated with text:

Animal and flower data was obtained from kaggle dataset and manual annotations are performed as depicted in the Figure 2 below

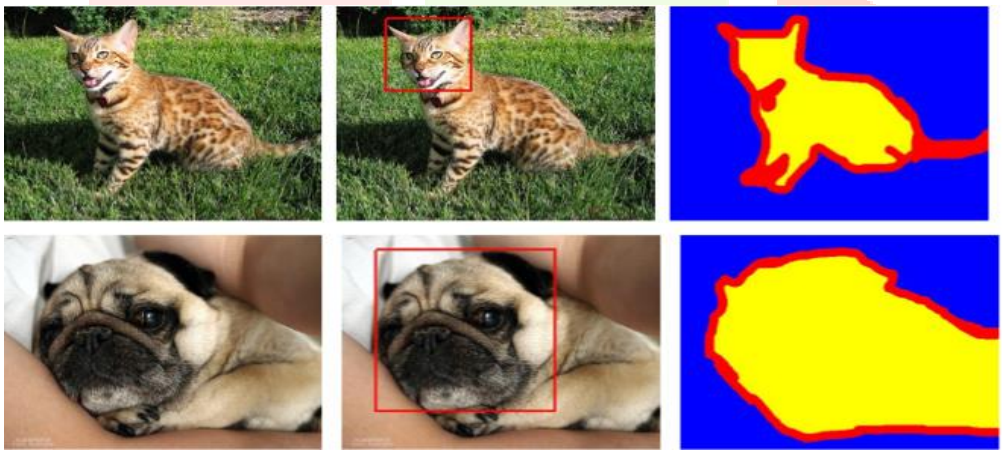


Figure 2 : Annotation of animal image

Steps to find the correlation between image and text 1) Image Segmentation using YOLO:

2) Named entity recognition (NER) (Extracting

entities from the text using SPACY)

Image Segmentation using YOLO

YOLO is a new algorithm for object detection. Object detection is framed as a regression problem to spatially separated bounding boxes and class probabilities. It processes images in real time at 45 frames per second. Fast YOLO processes an 155 frames per second which is smaller version of YOLO. The workflow of

YOLO is it runs on a single CNN (predicts multiple bounding boxes and class probabilities for those boxes). YOLO improves detection performance by training on full images and it is fast. To detect the object from new images ,The model trained using deep neural network(variant of CNN) at test time . On a Titan X GPU, it runs at 45 frames per second with no batch processing.Secondly, When making predictions, YOLO considers the images as a complete image. When compared to Fast R-CNN, it produces less than half the amount of background errors. Third, Object representations that are generalizable are learned by YOLO. The fully connected layers predict the output probabilities and coordinates while the initial convolutional layers extract the features from the image. The Google Net model for image classification inspired the YOLO architecture. It contains 24 convolutional layers followed by two fully connected layers. Instead of inception modules, 1*1 reduction layer followed by 3*3 convolutional layers are used. Final layer predicts both class probabilities and bounding box coordinates. Final layer uses linear activation function and hidden layer uses leaky Relu.

Images from the scarped web pages preprocessed and transformed into grey scale image to maintain uniform contrast with the boundaries. Since RBG color image contains so much redundant information which is not necessary. The images were resized to 256 * 256 ,converted into grey scale, normalized , and glitches were eliminated. The learning rate was set to 0.01 and the model trained up to 256 epochs to detect the primary objects and their features Multiple levels. These details were used to classify the objects into the appropriate categories (eyes/ears/petals/sepals etc.).The actual objects were classified into respective specie like lion, tiger etc and entity labels were obtained.

Named entity recognition(NER) (Extracting entities

from the text using SPACY **)**

Paragraphs of text were extracted from scraped webpages, annotated for entity labels that pertain to various features and species description. An NER model was trained using SPACY , which has been used to extract the entities from new web pages. Entities are the words or groups of words that represent information about common things such as persons, locations, organizations, etc.SPACY is used to perform several NLP related tasks, such as part-of-speech tagging, named entity recognition, and dependency parsing. The first step for a text string, when working with SPACY, is to pass it to an NLP object. This object is essentially a pipeline of several text pre-processing operations through which the input text string has to go through. The Figure3 depicts the processing operations

Correlation

Entity labels extracted from images and entity text extracted from text normalized using lemmatization and word vecterization using SPACY.. The vectors from images and text paragraphs were compares to arrive at the intersection

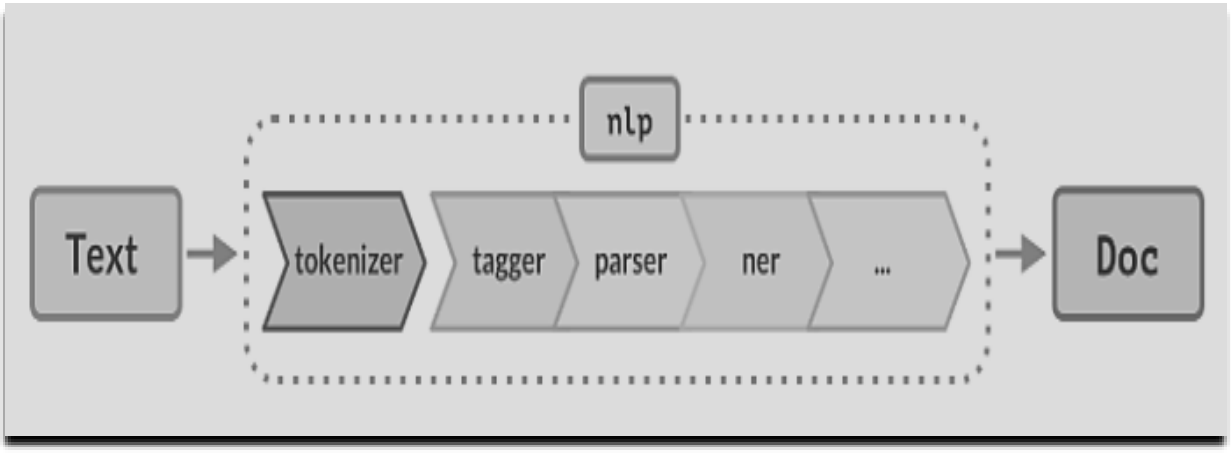
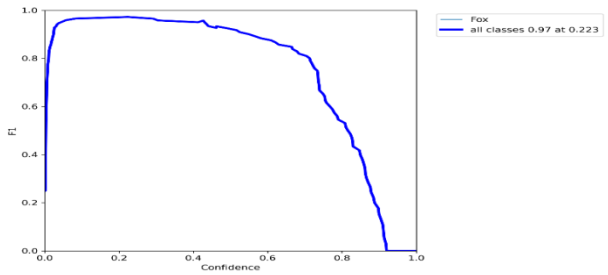


Figure3 : the processing operations

Results:

The YOLO model provided precision of 97%, recall of 98% and F1-score of 97% when classifying image entities and Correlation of net precision of 95%, recall of 92% and F1-score of 93.5% The following ar Figures 4,5,6 gives graphs of precision, recall and f1-score from image classification.



. ***Figure 4: F1-score v/s confidence***

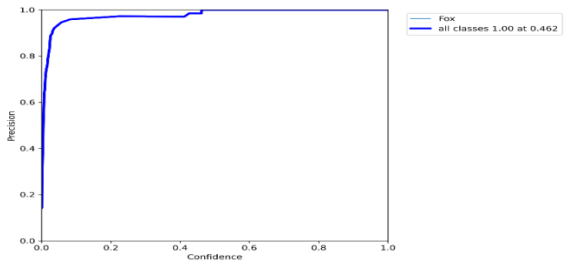


Figure5:Precision v/s confidence

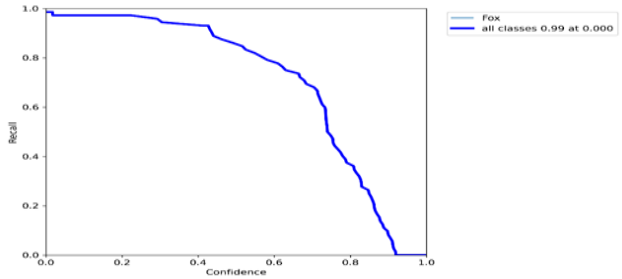


Figure 6:Recall v/s confidence

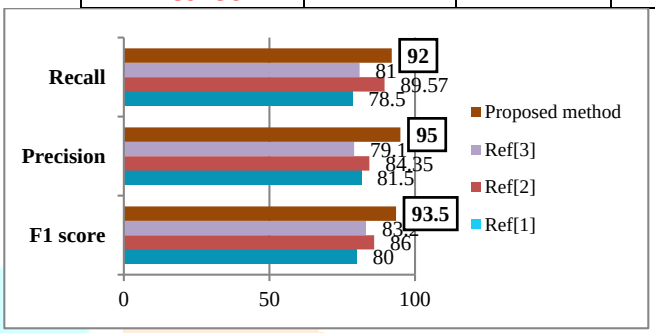
Correlation

Entity labels extracted from images and entity text extracted from text normalized using lemmatization and word vecterization using SPACY.. The vectors from images and text paragraphs were compares to arrive at the intersection. This correlation technique yielded the a net precision of 95%, recall of 92%, and f1-score of 93.5%.

Table IV provides the comparison of the proposed method with three other methods in the literature, namely, Oostdijk et al. [1], Gotzelmann et al. [2] and Zhang et al. [4], which is graphically depicted in the Fig. 8. It is observed that the methods in [1], [2] and [4] do not extract image objects, instead these methods match only text entities from paragraphs and captions.

TABLE I. COMPARISON WITH OTHER METHODS

	F1 Score	Precision	Recall
Oostdijk et al. [1]	80.0	81.5	78.5
Gotzelmann et al. [2]	86	84.35	89.57
Zhang et al. [4]	83.2	79.1	81.0
Proposed Method	93.5	95	92



Conclusion

The proposed framework uses a combination of, multiple image segmentation levels(using YOLO) and named entity recognition using SPACY to correlate the image and the text . This method (pipeline) is highly effective and gives a precision of 95%, recall of 92% and F1-score of 93.5% .The proposed methodology uses deep learning for entity extraction from both images and text. While most papers from the references did not handle the entities from images. They handle the entities from text using vectors and compared entities from text paragraphs and caption.

References:

[1] Nelleke Oostdijk, Hans Van Halteren, Erkan Basar, Martha Larson: The Connection between the Text and Images of News Articles: New Insights for Multimedia Analysis, *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4343-4351, 2020.

[2] Timo Gotzelmann, Pere-Pau Vazquez, Knut Hartmann, Andreas Nurnberger, Thomas Strothotte: Correlating Text and Images: Concept and Evaluation, *Conference Paper* . July 2007, DOI: 10.1007/978-3-540-73214-3_9.

[3] Ruoxu Ren, Chee Khiang Pang, Li Ma, ParthaDutta: An Experimental Study Of Correlation Between Text And Image Similarity By Information Fusion Approach, *Springer International Publishing- Proceedings in Adaptation, Learning and Optimization*, Vol. 2, DOI: 10.1007/978-3-319-13356-0_53, 2015.

[4] Ke Zhang, Yunwen Zhu, Wenjun Zhang, Weilin Zhang, Yonghua Zhu: Transfer Correlation Between Textual Content to Images for Sentiment Analysis*: SPECIAL SECTION ON INTEGRATIVE COMPUTER VISION AND MULTIMEDIA ANALYTICS- IEEE Access*, Vol. 8, 2020.

[5] Stephane Clinchant, Julien Ah-Pine, Gabriela Csurka: Semantic Combinations of Textual and Visual Information in Multimedia Retrieval, *Proceedings of the 1st International Conference on Multimedia Retrieval-Cross-modal Image and Video Processing-ICMR*, 2011.

[6] Juan C. Caicedo, Jose G. Moreno, Edwin A. Nino, Fabio A. Gonzalez: Combining Visual Features and Text Data for Medical Image Retrieval Using Latent Semantic Kernels, *Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval-MIR*, 2010, DOI:10.1145/1743384.1743442.

[7] Xin Zhou, Adrien Depeursinge, Henning Muller: Information Fusion for Combining Visual and Textual Image Retrieval, *20th International Conference on Pattern Recognition (ICPR)*, pp. 1590-1593, IEEE Press (2010)

[8] Yasuhide Mori Hironobu , Hironobu Takahashi , Ryuichi Oka: Image-to-Word Transformation Based on Dividing and Vector Quantizing Images With Words, *First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.*

[9] J. Jeon, V. Lavrenko, R. Manmatha: Automatic image annotation and retrieval using cross-media relevance models, *Proceedings of the 26th annual international ACMSIGIR conference on Research and development in information retrieval*, pp. 119-126,2003.

[10] H. Muller, N. Michoux, David Bandon, A. Geissbuhler: A review of content-based image retrieval systems in medical applications—clinical benefits and future directions, *International Journal of Medical Informatics*, Vol. 73, Issue. 1, pp. 1-23, 2004.

[11] Trong-Ton Pham, Nicolas Eric Maillot, Joo-Hwee Lim, Jean-Pierre Chevallet: Latent Semantic Fusion Model for Image Retrieval and Annotation, *Proceedings of the Sixteenth ACM -Conference on Information and Knowledge Management*, pp. 444,439

[12] Caroline Lacoste, Joo-Hwee Lim, Jean-Pierre Chevallet, Diem Thi Hoang Le: Medical-Image Retrieval Based on Knowledge-Assisted Text and Image Indexing, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 17, Issue. 7, 2007, DOI: 10.1109/TCSVT.2007.897114.

[13] Nagy G, Prateek Sarkar, "Document style census for OCR", First International Workshop on Document Image Analysis for Libraries, pp. 134-147, 2004.

[14] Sarfraz M, Zidouri A, Shahab S.A, "A novel approach for skew estimation of document images in OCR system", International Conference on Computer Graphics, Imaging and Vision, pp. 175- 180, July 2005.

[15] Shaolei Feng; Manmatha R, "A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books", Joint Conference on JCDL, pp.109-118, June 2006.

[16] Desrochers D, Qu Z, Saengdeejing A, "OCR readability study and algorithms for testing partially damaged characters", International Symposium on Intelligent Multimedia, Video and Speech Processing, pp.397-400, 2001.

[17] Leija L, Hernandez P, Santiago S, "Reader instrument of basic texts to the teaching of blind people", 21st Annual Conference of the Biomedical Engineering Soc, vol.1, pp.588, 1999. [17] Bazzi I, Schwartz R, Makhoul J, "An omni font open-vocabulary OCR system for English and Arabic", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.21, no.6, pp.495-504, Jun 1999.

[18] Sun-Hwa Hahn, Joon Ho Lee, Jin-Hyung Kim, "A study on utilizing OCR technology in building text database", Tenth International Workshop on Database and Expert Systems Applications, pp.582- 586, 1999.

[19] Jaehwa Park, Govindaraju V, Srihari S.N, "OCR in a hierarchical feature space", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, no.4, pp.400-407, Apr 2000. [20] Ishitani Y, "Model-based information extraction method tolerant of OCR errors for document images", Sixth International Conference on Document Analysis and Recognition, pp.908-915, 2001.

[21] Manoj T H, A Santha Rubia ―A Survey and Evaluation of Edge Operators: Application to Text Recognition‖, International Journal of Computer Technology and Application, Vol 3, Issue 4, 1481- 1484.