Using Automatic Features for Text-image Classification in Amharic Documents

Birhanu Belay[1,2], Tewodros Habtegebrial[1], Gebeyehu Belay[2] and Didier Stricker[1,3]
[1]*Technical University of Kaiserslautern, Kaiserslautern, Germany*
[2]*Bahir Dar Institute of Technology, Bahir Dar, Ethiopia*
[3]*DFKI, Augmented Vision Department, Kaiserslautern, Germany*

Keywords: Amharic Document Image, Automatic Feature, Binary SVM, CNN, Handwritten, Machine Printed, OCR, Pattern Recognition.

Abstract: In many documents, ranging from historical to modern archived documents, handwritten and machine printed texts may coexist in the same document image, raising significant issues within the recognition process and affects the performance of OCR application. It is, therefore, necessary to discriminate the two types of texts so that it becomes possible to apply the desired recognition techniques. Inspired by the recent successes CNN based features on pattern recognition, in this paper, we propose a method that can discriminate handwritten from machine printed text-lines in Amharic document image. In addition, we also demonstrate the effect of replacing the last fully connected layer with a binary support vector machine which minimizes a margin-based loss instead of the cross-entropy loss. Based on the results observed during experimentation, using Binary SVM gives significant discrimination performance compared to the fully connected layers.

1 INTRODUCTION

Nowadays, there are may documents that contain mixed handwritten and machine printed texts in the same page image. In most modern and historical printed documents, writers may add handwritten texts due to multiple reasons such as text correction, inclusion of notes and instruction while others may exist for the purpose of bank cheque processing, application forms, questionnaire, examination correction, mailing address and receipts (Sahare, 2018; Kavallieratou et al., 2004).

In document image analysis and Optical character recognition (OCR) applications, separation between machine printed and handwritten texts is an important task because recognition process are different in both contexts (Mozaffari and Bahar, 2012; Peng et al., 2009). In addition, we may be interested in either of the text documents for further processing. For example, in case of bank cheque processing and handwritten recognition system the handwritten parts are interesting while the machine printed texts are important for other document processing tasks.

Therefore, to facilitate and enhance the overall efficiency of document processing and OCR application, separation between machine printed and handwritten texts is necessary (Peng et al., 2009; Xiao-Hui Li and Liu, 2018).

Multiple intensive works, for multiple scripts, have been done in the area of document image processing and several of them are available in the market with a better recognition accuracy of both handwritten and machine printed text images. However, there are underprivileged scripts in the area of document analysis and Natural Language Processing (NLP), such as Amharic, which are still untouched and open application area for research.

In both historical and modern printed document images, separation of handwritten from machine printed texts can be at line, word (Mozaffari and Bahar, 2012; Guo and Ma, 2001) and character (Fan et al., 1998) level.

Discrimination of handwritten and machine printed texts has been done for most Non-Latin (Mozaffari and Bahar, 2012; Echi et al., 2014) and Latin (Peng et al., 2009; Medhat and et al., 2018) scripts. Due to the complexity of structural layout in various scripts, different researchers use different statistical methods to find best representative features for text separation and recognition (Sahare, 2018).

Several attempts have been done for Amharic script recognition (Teferi, 1999; Meshesha, 2008;

Belay, B., Habtegebrial, T., Belay, G. and Stricker, D.
Using Automatic Features for Text-image Classification in Amharic Documents.
DOI: 10.5220/0008940704400445
In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020)*, pages 440-445
ISBN: 978-989-758-397-1; ISSN: 2184-4313

Assabie and Bigun, 2008; Belay et al., 2019b; Belay et al., 2019a) considering different contexts including handwritten, machine printed, character and text line recognition. To the best of our knowledge, no attempt has been done to discriminate handwritten from machine printed for Amharic documents.

In this regard, we propose two classifiers which use the same feature. A fully connected network with soft-max and binary SVM are the proposed classification algorithms for the discrimination of handwritten texts from machine printed Amharic texts based on automatic feature extracted by Convolutional Neural Network (CNN) (Banerjee and Chaudhuri, 2012).

The remainder of the paper is organized as follows: In section 2 the review of related works. Our proposed method and dataset preparation techniques is presented in Section 3. Experimental results are discussed in section 4. Finally, section 5, summarizes the conclusions drawn from this study and presents future work directions.

2 RELATED WORK

For discrimination of handwritten from machine printed document, various methods have been proposed and reported the discrimination accuracy for different scripts including English (Kavallieratou et al., 2004; Peng et al., 2013), Kannada (Pardeshi et al., 2016), Devnagari and Bangael (Pal and Chaudhuri, 2001), Faris/Arabic (Mozaffari and Bahar, 2012), Hindi (Srivastava et al., 2015; Shalini Puri, 2019) and Bangla (Banerjee and Chaudhuri, 2012). Since there are multiple effective OCR application for most Latin and non-Latin scripts, many researchers think as OCR is a solved research area. However, still there are multiple indigenous scripts, like Amharic, that needs better and workable OCR technologies.

Researchers attempted to address the problems of Amharic script OCR including both machine printed (Teferi, 1999; Meshesha, 2008) and handwritten (Assabie and Bigun, 2008) character recognition and tried to show how to improve the recognition accuracy and noted the challenges in developing the OCR of Amharic Script.

A binary morphological filtering algorithm was proposed by (Teferi, 1999) for OCR of Amharic typewritten Text and he recommended that adopting recognition algorithms which are not very sensitive to the features of the writing styles of characters helps to enhance recognition rate of Amharic OCR. The other work done by (Belay et al., 2019b), employed a CNN based method called factored convolutional neural network so as to recognize basic Amharic characters in terms of rows and column arrangement in "Fidel gebeta" and reported a state-of-the-art character recognition accuracy. An adaptive segmentation technique proposed by (Kassa and Hagras, 2018) for ancient Ethiopian handwritten manuscripts called Geez and reported a promising segmentation accuracy.

In a separate line of research works, attempts are done to discriminate handwritten and machine printed texts written with different scripts. Based on textual appearance and geometrical features that are extracted by wavelet transform has been proposed (Sahare, 2018) for English documents discrimination in to three classes (handwritten, machine printed and noise text) using Tobacco-800 and IFN-ENIT datasets. Handwritten and machine printed text discrimination has been also done using horizontal histogram on height distribution of characters in text-line of English document images from IAM-DB and GRUHD datasets (Kavallieratou et al., 2004).

A Markov Random Field (MRF) based method (Medhat and et al., 2018) was presented to classify texts in to machine printed, handwritten and noise. Initially they use a modified K-means clustering algorithm to cluster each class individually and then those centers are used as hidden nodes of MRF. They also done a transcription of handwritten and machine printed text using different evaluation metrics based on character and word level on IAM database and achieved about 80% recognition accuracy. Finally, they recommended that enhancing the preprocessing stages of an input image and further optimization is needed for the models which are going to be used for machine printed and or handwritten text line image recognition.

The same techniques were employed to classify image documents in three class as handwritten, machine printed and noise using a dataset collected from tobacco archive (Zheng et al., 2004). The other researchers (Mozaffari and Bahar, 2012) proposed a new feature, based on the properties of Arabic language, called baseline profile feature and separate handwritten from machine printed Arabic document at word block level.

The work of (Banerjee and Chaudhuri, 2012), presented an SVM based approach to separate handwritten from machine printed Bangal texts. The researchers noted that handwritten and machine printed document discrimination is a useful application in deleting unnecessary parts and cleaning the document images as well. Similar work presented by (Pathak and Tewari, 2015) using structural feature and threshold value to discriminate handwritten from machine printed Bangal text line images. An Other model pro-

posed by (Trieu and Lee, 2016) called bag of words, that uses SURF for the identification of Korean machine printed and handwritten texts that co-exist in the same document image and achieved better results compared with methods employed based on structural features of Korean documents.

3 MATERIAL AND METHODS

Most document image analysis and pattern recognition systems start from dataset preparation, prepossessing, model training and then followed by model performance evaluation. The proposed system processes an Amharic document image based on three main stages. The first stage is the preprocessing module where, for a given document image, a set of text areas are localized, segmented and resulting a series of text-lines images. The second stage is the feature extraction module where a vector of automatic characteristics that represents the character and or text-line image properties are assigned to each text-line and finally the classification module for separating the handwritten from the machine printed Amharic text-lines images. An overview of the system is shown in Fig. 1.

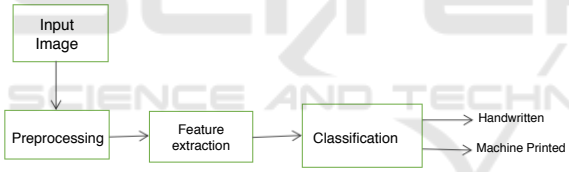


Figure 1: An overview of the proposed system.

3.1 Dataset

This study is concerned to develop a model that can separate a handwritten from machine printed Amharic text-line image. Since there is no any ready-made dataset for our experiment, we have also prepared our own dataset which contains about 1150 pages of Amharic document images. In each page of the document, both the handwritten and machine printed text-lines are co-exist with a random distribution. A sample scanned page of Amharic document from our dataset is shown in Fig. 2.

During dataset preparation, we have collected Amharic texts from different books, newspapers, magazines and websites. We compile the collected text documents by leaving some free space in between random lines of texts per each page. Then we gave each page of the printed document for different groups of university students to rewrite each

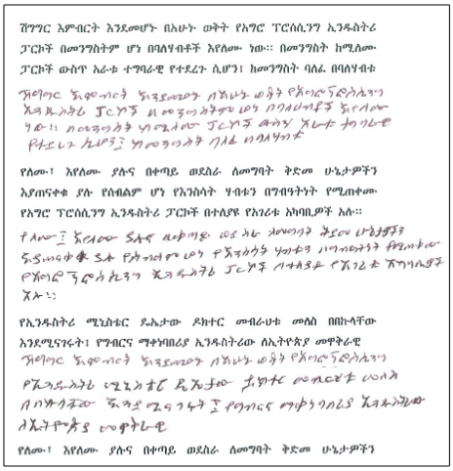


Figure 2: Sample scanned page of Amharic document, that contains a mixed handwritten and machine printed documents, from our data set.

printed text-line on the free space under it. Since the documents are distributed randomly, the students used pens with different level of color degradation and type.

Finally, we collect the documents and scanned each page of the document that contains both the handwritten and machine printed texts together. These scanned page documents are segmented into lines using, OCRopus, the open source OCR framework (Thomas, 2008).

Once line segmentation process is completed, we labeled each segmented text-line image manually. Since the images have not exactly the same size, considering the state-of-the-art works and the nature of image, each segmented text-line images were resized with the width of 128 and height of 64 pixel. A typical diagram of OCRopus text-line segmentation engine is shown in Fig. 3

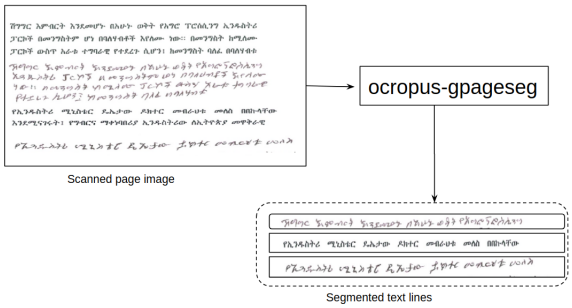
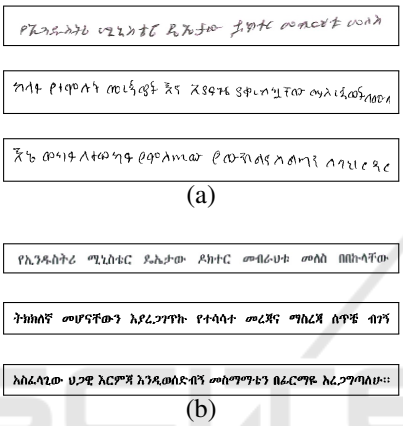


Figure 3: A Typical diagram of OCRopus for text-line segmentation. The scanned page of a mixed handwritten and machine printed Amharic document image is feed to the ocropus-gpageseg engine.The Engine is responsible for segmenting each page of the scanned document in to text-line images.

To develop and evaluate the effectiveness of the model, we have prepared the training and testing database which contains a total of 41,718 Amharic text-line images. The training set included 33374 text-line images while the remaining text-line images are the test set. The distribution of handwritten and machine printed texts, in the whole data, are distributed about equally while the training, validation and test datasets are randomly selected using Scikit-learn library (Garreta and Moncecchi, 2013).

A sample segmented handwritten and machine printed Amharic text-line images shown in Fig 4. All these images are normalized and in binary format.



(a)

(b)

Figure 4: Sample segmented and normalized Amharic text-line images. (a) handwritten and (b) machine printed.

3.2 The Proposed Method

After the preprocessing stage, already discussed in section "3.1", completed; the feature extraction and classification stage were followed to extract structural characteristics of text lines and to distinguish the handwritten from the machine printed text lines of Amharic document image respectively. An overview of the proposed system architecture is shown in (see Fig. 5).

The proposed architecture has two classifiers. The first classifier is Fully connected network (FCN) with the soft-max and motivated with its success on binary classification we employee a binary Support Vector Machine (SVM) as second classifier. Both classifiers use automatic feature extracted using CNN.

In the case of fully connected network based classification, we use seven convolutional layers with a 2 x 2 max-pooling at each two blocks of convolution with rectifier linear unit as an activation function. On top of the seventh single convolutional layer, after applying a 2×2 max-pooling, we stacked two fully connected layers followed by the soft-max function (Zhao et al., 2017) and then we train the network end-to-end.

During an end-to-end training, the dimension of output features in each convolutional layer can be computed by equation "(1)".

$$M = \left(\frac{N - K + 2 * P}{S}\right) + 1 \qquad (1)$$

where N is the dimension of input size, K is the kernel size S is the stride in the convolution, P is padding, and M refers the dimension of output size.

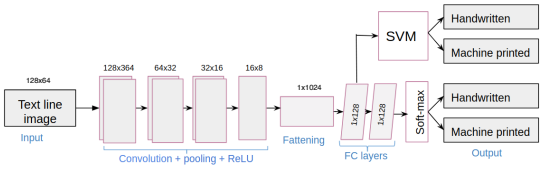


Figure 5: A proposed model architecture. This method has two classifiers. The first classifier is fully connected layer with soft-max which is trained with an end-to-end scenario while the second classifier is a binary SVM classifier which is stacked on top by removing the last fully connected and soft-max layer. The binary SVM classifier uses the features extracted by the CNN layers, once reshaped and passed through the first fully connected layer, as an input. Therefore, classification has been done using fully connected network with a soft-max function first and then SVM latter.

The number of learnable parameters is very important criteria to measure the complexity of network architecture and also it can be used to make comparison between different network architectures. Parameter at each subsuming layer is zero while the number of parameters for a single convolutional layer is computed using equation "(2)".

$$P_i = (Kh \times Kw \times F_{\text{m-1}} + 1) \times F_{\text{m}}) \qquad (2)$$

where P_i is the total number of parameters in the i^{th} layer , Kh is the size of height kernel, Kw the size of width kernel at i^{th} convolutional layer, $F_{\text{m-1}}$ and F_{m} are the number of feature maps at the previous layer and current layer respectively. While the number of parameters for each fully connected layer can is calculated as $((n_{\text{i-1}} + 1) \times n_{\text{i}})$ where $\text{n}_{\text{i-1}}$ and n_{i} are the number of neurons at the previous layer and current layer respectively. The added value 1, here, is the bias term for each filter.

Once, an end-to-end training completed, we just remove the last fully connected network with the soft-max layer and then the binary SVM becomes in place to discriminate the features extracted by the Convolutional layers into handwritten and machine printed text-line.

The final output of the network, in an end-to-end training, is determined by a Soft-max function, which

tells the probability that either of the classes are true, is computed using equation "(3)".

$$f(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad (3)$$

Where z is input vector for output layer, j is the indexes which runs from 1 to k and k is the number of outputs.

At training time, the fully connected network tried to minimize the cross entropy loss while the support vector machine minimizes margin-based loss.

Unlike the soft-max, which predicts the probability of the class, the SVM classifier will directly predicts classes by finding the best hyperplane, which leaves the maximum margin between classes, to separate the two classes (Hastie et al., 2001). For a set of training point x_i along with the class y_i and some dimension d, in which $x_i \in R^d$ and y_i=$\pm$1 then the hyperplane can be defined by equation "(4)"

$$f(x) = x^{'}\beta + b = 0 \quad (4)$$

where $\beta \in R^d$ and b is real number. Then the best separating hyperplane is a boundary with the value of β and b that minimize $||\beta||$ such that all data points (x_i, y_i), $y_i f(x_i) \geq 1$

4 EXPERIMENTAL RESULTS

Experiments are done following the network architectures and classification settings introduced in section "3.2" and figure "5" using an Amharic text-line image database which are, our own dataset, annotated manually. We develop a model using Keras application program interface, an open source deep neural network library, on a tensorflow backend with python programming. To select a suitable parameters for both the FCN and SVM model, different values of the parameters were considered and tuned during experimentation.

We conduct two experiments. The first experiment was conducted using convolutional neural network as feature extractor and fully connected network with soft-max as a classifier. Adam optimizer were used during training and the network converges after 15 epochs. The second experiment was used a binary SVM as classifier. This experiment uses the output of the late CNN layers, from the first experimental setup, as a feature and train a separate SVM. In case of binary SVM, the best result is recorded with the parameter; radial basis function kernel, 0.0001 gamma value and penalty weight (C) value of 10 which were selected using Grid search algorithm explained in (Gaspar et al., 2012).

In both experimental setups, all text-line images are resized to a width of 128 and height of 64 pixels. A CNN based features were used for model development in both classifiers. For training and testing the proposed model, we created our own database of 41,718 text line images from 1150 page of mixed handwritten and machine printed Amharic document image.

The performance of the proposed model is measured using accuracy which is computed as the ratio of correctly classified Amharic text-line images and total number of Amharic text-line images. According to experimental result, 93.45% and 95.35% discrimination performance were obtained using fully connected with soft-max and support vector machine respectively. The result shows that, our proposed system can discriminate handwritten and machine printed text-lines with an overall recognition accuracy of 95.35% using binary SVM classifier which outperforms the fully connected network.

In addition, the model requires enough mixed handwritten and machine printed text-line image data so as to perform better on a varieties of test data. In general, the performance of the proposed method obtained a promising results and this work will used as benchmark for future works on Amharic document image analysis.

5 CONCLUSION

Amharic is an official and working language of Ethiopia that has it own indigenous script and rich in a bulk of mixed historically printed and handwritten documents dated back 12^{th} century. However, it is underprivileged group of scripts in Natural language Processing due to lack of extensive research in the area and lack of annotated dataset. Therefore, in this paper, we present a method to discriminate handwritten from machine printed text-line images in Amharic documents by employing a FCN with soft-max and binary SVM. We evaluated our model with unseen mixed, hand written and machine printed, Amharic text-line images and achieved state-of-the-art results. A better classification accuracy were recorded with the Binary SVM classifier.

The proposed model can be used during handwritten dataset preparation from a mixed Amharic document image and also used as preprocessing stage for OCR application in the area of signature verification, application form and bill processing. As future work, we plan to enhance the performance of the model and integrate it with a system that can transcribe mixed handwritten and machine printed document images.

REFERENCES

Assabie, Y. and Bigun, J. (2008). Writer-independent offline recognition of handwritten ethiopic characters. *Proc. 11th ICFHR*, pages 652–656.

Banerjee, P. and Chaudhuri, B. B. (2012). A system for handwritten and machine-printed text separation in bangla document images. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 758–762. IEEE.

Belay, B., Habtegebrial, T., Liwicki, M., Belay, G., and Stricker, D. (2019a). Amharic text image recognition: Database, algorithm, and analysis. In *International Conference on Document Analysis and Recognition (ICDAR)*. IEEE.

Belay, B., Habtegebrial, T., Liwicki, M., Belay, G., and Stricker, D. (2019b). Factored convolutional neural network for amharic character image recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 2906–2910. IEEE.

Echi, A. K., Saidani, A., and Belaid, A. (2014). How to separate between machine-printed/handwritten and arabic/latin words? *ELCVIA: electronic letters on computer vision and image analysis*, 13(1):1–16.

Fan, K.-C., Wang, L.-S., and Tu, Y.-T. (1998). Classification of machine-printed and handwritten texts using character block layout variance. *Pattern Recognition*, 31(9):1275–1284.

Garreta, R. and Moncecchi, G. (2013). *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd.

Gaspar, P., Carbonell, J., and Oliveira, J. L. (2012). On the parameter optimization of support vector machines for binary classification. *Journal of integrative bioinformatics*, 9(3):33–43.

Guo, J. K. and Ma, M. Y. (2001). Separating handwritten material from machine printed text using hidden markov models. In 6th *International Conference on Document Analysis and Recognition*, pages 439–443. IEEE.

Hastie, T., Tibshirani, r., and Friedman, J. (2001). The elements of statistical learning. data mining, inference, and prediction.

Kassa, D. and Hagras, H. (2018). An adaptive segmentation technique for the ancient ethiopian geez language digital manuscripts. In *10th Computer Science and Electronic Engineering (CEEC)*, pages 83–88. IEEE.

Kavallieratou, E., Stamatatos, S., and Antonopoulou, H. (2004). Machine-printed from handwritten text discrimination. In *9th International Workshop on Frontiers in Handwriting Recognition, ICFHR-9*, pages 312–316. IEEE.

Medhat and et al. (2018). Tmixt: A process flow for transcribing mixed handwritten and machine-printed text. In *IEEE International Conference on Big Data*. Newcastle University.

Meshesha, M. (2008). *Recognition and Retrieval from Document Image Collections*. PhD thesis, IIT, Hyderabad, India.

Mozaffari, S. and Bahar, P. (2012). Farsi/arabic handwritten from machine-printed words discrimination. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 698–703. IEEE.

Pal, U. and Chaudhuri, B. B. (2001). Machine-printed and hand-written text lines identification. *Pattern Recognition Letters*, 22(3-4):431–441.

Pardeshi, R., Hangarge, M., Doddamani, S., and Santosh, K. (2016). Handwritten and machine printed text separation from kannada document images. In 10th *International Conference on Intelligent Systems and Control (ISCO)*, pages 1–4. IEEE.

Pathak, R. and Tewari, R. K. (2015). Distinction between machine printed text and handwritten text in a document. *International Journal of Scientific Engineering and Research (IJSER)*, 3(7):13–17.

Peng, X., Setlur, S., Govindaraju, V., and Sitaram, R. (2013). Handwritten text separation from annotated machine printed documents using markov random fields. *International Journal on Document Analysis and Recognition (IJDAR)*, 16(1):1–16.

Peng, X., Setlur, S., Govindaraju, V., Sitaram, R., and Bhuvanagiri, K. (2009). Markov random field based text identification from annotated machine printed documents. In 10th *International Conference on Document Analysis and Recognition.*, pages 431–435. IEEE.

Sahare, P, D. S. (2018). Separation of machine-printed and handwritten texts in noisy documents using wavelet transform. *IETE Technical Review*, pages 1–21.

Shalini Puri, S. P. S. (2019). An efficient devanagari character classification in printed and handwritten documents using svm. In *International Conference on Pervasive Computing Advances and Applications*.

Srivastava, R., Tewari, R. K., and Kant, S. (2015). Separation of machine printed and handwritten text for hindi documents. *International Research Journal of Engineering and Technology (IRJET)*, 2(2):704–708.

Teferi, D. (1999). Optical character recognition of typewritten amharic text. Master's thesis, Addis Ababa University, Addis Ababa.

Thomas, B. (2008). The ocropus open source ocr system. In *Document Recognition and Retrieval XV*, volume 6815, page 68150. International Society for Optics and Photonics.

Trieu, S. T. and Lee, G. S. (2016). Machine printed and handwritten text discrimination in korean document images. *Smart Media Journal*, 5:1–5.

Xiao-Hui Li, F. Y. and Liu, C. (2018). Printed/handwritten texts and graphics separation in complex documents using conditional random fields. In *13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 145–150. IEEE.

Zhao, H., Hu, Y., and Zhang, J. (2017). Character recognition via a compact convolutional neural network. In *International Conference on Digital Image Computing: Techniques and Applications*, pages 1–6. IEEE.

Zheng, Y., Li, H., and Doermann, D. (2004). Machine printed text and handwriting identification in noisy document images. *IEEE transactions on pattern analysis and machine intelligence*, 26(3):337–353.