

# A New Algorithm for Skew Detection of Telugu Language Document based on Principle-axis Farthest Pairs Quadrilateral (PFPQ)

**MSLB. Subrahmanyam**

Research Scholar, JNTU Kakinada, Kakinada, 533001, India  
Email: subrahmanyam\_mtech@rediffmail.com

**V. Vijaya Kumar**

Director for CACR, Dean Dept. of CSE &IT, Anurag Group of Institutions (Autonomous), Hyderabad, India  
Email: vakula\_vijay@yahoo.com

**B. Eswara Reddy**

Professor of CSE Dept. & Principal of JNTU-A College of Engineering, Kalikiri, India  
Email: eswarcsejntu@gmail.com

Received: 29 September 2017; Accepted: 16 October 2017; Published: 08 March 2018

**Abstract**—Skew detection and correction is one of the major preprocessing steps in the document analysis and understanding. In this paper we are proposing a new method called “Principle-axis farthest pairs Quadrilateral (PFPQ)” mainly for detecting skew in the Telugu language document and also in other Indian languages. One of the popular and classical languages of India is Telugu language. The Telugu language is spoken by more than 80 million people. The Telugu language consists of simple and complex characters attached with some extra marks known as “maatras” and “vatthulu”. This makes the process of skewing of Telugu document is more complex when compared to other languages. The PFPQ, initially performs pre-processing and divides the text in to connected components and estimates principle axis furthest pair quadrilateral then removes the small and large portions of quadrilaterals of connected components. Then by using painting and directional smearing algorithms the PFPQ estimates the skew angle and performs the de-skew. We tested extensively the proposed algorithm with five different kinds of documents collected from various categories i.e., Newspapers, Magazines, Textbooks, handwritten documents, Social media and documents of other Indian languages. The images of these documents also contain complex categories like scientific formulas, statistical tables, trigonometric functions, images, etc. and encouraging results are obtained.

**Index Terms**—Indian languages, compound characters, complex categories, painting and directional smearing.

## I. INTRODUCTION

The document image processing (DIP) is a sub field of image processing. The main objective of DIP is to digitize the printed or handwritten or any type of document, to reduce physical storage of paper document and for quick access of the documents. Optical character recognition (OCR) is one way for converting document images from scanned or photo into editable format which computer can understand. Editable text format means, characters are in ASCII or Unicode. The accuracy of OCR depends upon the scanned quality of the image. If scanned image quality is not good then OCR gives noisy output. So it is very important to preprocess the document for improving OCR accuracy. One of the most important preprocessing steps for improving OCR accuracy is skew detection and its correction. The skew of document image specifies its text line deviation from the horizontal or vertical axis. The process of straightening an image that has been scanned or photographed is called de-skew. A lot of research on de-skew is available for European languages, Japanese, Chinese, Latin based languages, etc. Relatively, less work has been reported for Indian scripts. The recognition of Indian languages is a challenging task due to the complexity of scripts. The scanned documents may contains images or tables and with different font titles. So calculating skew angle for these kinds of scanned images is one of the challenging tasks. In the literature [1 – 23] a lot of skew detection techniques were presented which can be broadly divided into four categories based on projection profile [1-6], Hough

transform [7-15], neighborhood approach [16-18] and cross correlation approach [19-23]. In the project profile [1-6] the skew angle of the document is estimated from the maximum value of the projection profiles which are computed at various angles of the given document. The initial method is based on Horizontal projection profile [1], and it is computationally expensive. Several papers [2, 3, 4, 5, 6] has been published to reduce computational cost and improve accuracy in this category.

Hough Transform (HT) [7,8] is one of the popular methods for identifying arbitrary shapes within the image. In this method [7] the skew angle can be obtained from the peak value in the Hough space using voting procedure. Many research papers were published using this procedure for detecting skew of the document image [9 – 15]. To reduce computational complexity, Hinds [10] introduced new concept called “Burst image” for skew detection of document images. Manjunath[11] considered only centroids of connected components for HT instead of using all image pixels. Wang et al. [13] estimated document skew, using local peak in Hough space by using bottom pixels of the candidate objects within the selected region. Yu and Jain in [12] considered centroid of each connected component for applying HT using two angular resolutions. Block adjacency graph (BAG) was introduced by Singh [14] in preprocessing step for improving processing time using HT for skew detection. Abdelhak Boukharouba [15] proposed new algorithm based on randomized Hough transform for skew correction and baseline detection of Arabic documents.

In nearest neighbor approach skew angle of the document is calculated from the peak of the angle histogram of each connected component. In 1986 Hashizume et al. [16] used this method for estimating skew angle. O’Gorman [17] improved this using k-neighbor of each connected component. Lu and Tan [18] proposed high accuracy and language independent approach by using nearest neighbor chain. In interline cross correlation method skew is estimated based on deviation among test pixels in the vertical and horizontal direction of the document image. Yan [19] used this approach for estimating skew angle by cross-correlation between lines at a fixed distance. Gatos et al. [20] proposed cross-correlation between the pixels of vertical lines using this approach. Chou et al. [21] proposed piece-wise covering by Parallelograms (PCP). In [22], Deya and Nousath have proposed enhanced PCP (e-PCP) which can robustly estimate skew for both horizontal – flow and vertical flow documents with reduced number of computations. Alireza Alaei et. al. [23], proposed an iterative approach for detecting skew of the document image by applying Piece-wise Paining algorithm. The nearest neighborhood is treated as local based approach in the image processing. The local based feature extraction methods play a vital role in many image processing applications [24-33]. This paper is based on the nearest neighbor approach.

This paper is organized as follows Section II gives details about Telugu script characteristics, section III gives related work, section IV contains the description of

our proposed method, Section V contains experimental results our proposed method and conclusions are given in Section VI.

II. TELUGU SCRIPT CHARACTERISTICS

India is a multilingual country of more than 1.34 billion population with 18 constitutional languages and 10 different scripts. Telugu is one of the most popular languages of India that is spoken by more than 80 million people especially in South India [34]. Telugu Language is known to exist since the Time period 400 BC. Telugu is an official language in Andhra Pradesh, Telangana and union territory of Puducherry. It is also spoken by significant minorities in the Andaman and Nicobar Islands, Chhattisgarh, Karnataka, Maharashtra, Odisha, Tamil Nadu, and by the Sri Lankan Gypsy people. It is one of the six languages with classical status in India. Telugu ranks third by the number of native speakers in India (more than 74 million people) and is the widely spoken Dravidian language. A striking feature of Telugu character is that it has more than twice as many characters as Devanagari and nearly 100 times as many as English [35]. Telugu contains both simple and compound characters. There are 52 simple characters, out of which 16 represents basic vowel sounds, called achchlu and 36 represents simple consonants, called halluu. In addition there are two groups of symbols, semi-vowel or vowel modifiers called as ‘maatras’ and half-consonants called as ‘voththulu’. These groups i.e.

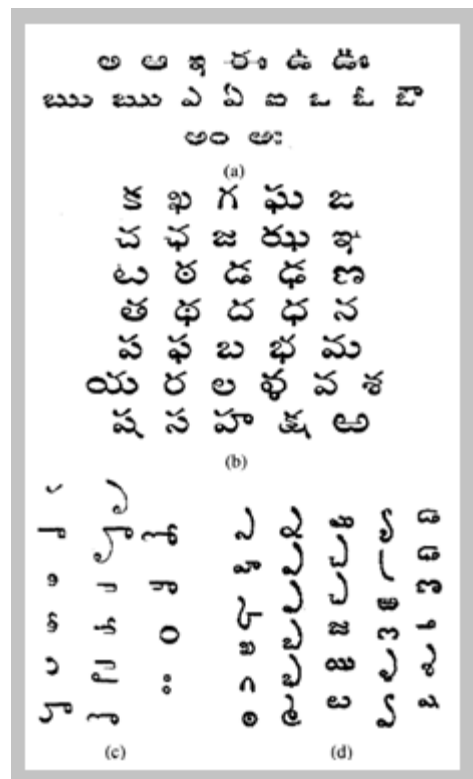


Fig.1. Telugu Alphabet ( a) Achchlu (b) Halulu (c) Maatras ( d) Voththulu.

maatras and voththulu are extra marks on the Telugu characters. The Telugu character with extra marks is known as compound character. The Telugu alphabet set is shown in Fig. 1. The compound characters of Telugu language is obtained by combination of simple consonants (halulu) with vowel modifiers (maatras) or with half consonants (voththulu) or with both maatras and voththulu[36]. The number of computed characters of Telugu language is estimated between 500s and 10000 [36].

### III. RELATED WORK

Skew detection for Indian languages in general is more complex than those of European languages due to the linguistic complexity of large number of vowels, consonants and various different combinations of vowels and consonants. Relatively less work has been reported for the skew detection and correction for Indian scripts. In 1997 B.B. Chaudhuri and U. Pal [37] proposed skew detection for Devanagari and Bangle scripts. In [38] M. Ramanan et al. proposed skew detection and correction of Tamil printed documents using Wiener filter, smearing technique and Radon transform. Lalita kumara et al. [39] proposed skew detection of Hindi, Bengali and Punjabi language scripts by partitioning whole document into sub blocks of fixed size and used histogram analysis on each sub block. A robust skew detection and correction of isolated words of machine printed Gurmukhi script is proposed using headline of the word [40]. Shamita Ghosh and Bidyut B. Chaudhuri[41] proposed skew detection method based on Hough transform.

### IV. METHODOLOGY

Till now no specific paper was published for skew detection and correction of Telugu documents. Skew detection of Telugu document is more challenging task because Unlike Latin script, Chinese script, Devanagari script or Bangla script, Telugu characters are rarely containing horizontal, vertical or diagonal lines. It is observed that majority of Telugu characters are compound characters with maatras and voththulu. This paper derives the skew angle of the entire Telugu document based on orientation of second order spatial moment. The compound Telugu characters make the process of detecting skew angle more complicated when compared to other languages. The matthras and voththulu are the bottle necks for estimating the skew angle of the document. At the time of skew angle detection, the matthras and voththulu (extra marks) associated with Telugu characters derives a completely wrong skew direction thus the skew estimation of the entire document will go wrong. To overcome this problem this paper removes these extra marks (maatras and voththulu) of the Telugu document using connected component (CC) approach. The Telugu characters with extra marks or compound characters are shown in Fig.2.

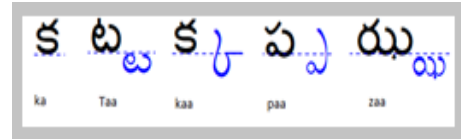


Fig.2. The extra marks of Telugu characters. The equivalent pronouncement/ representation in English are given below.

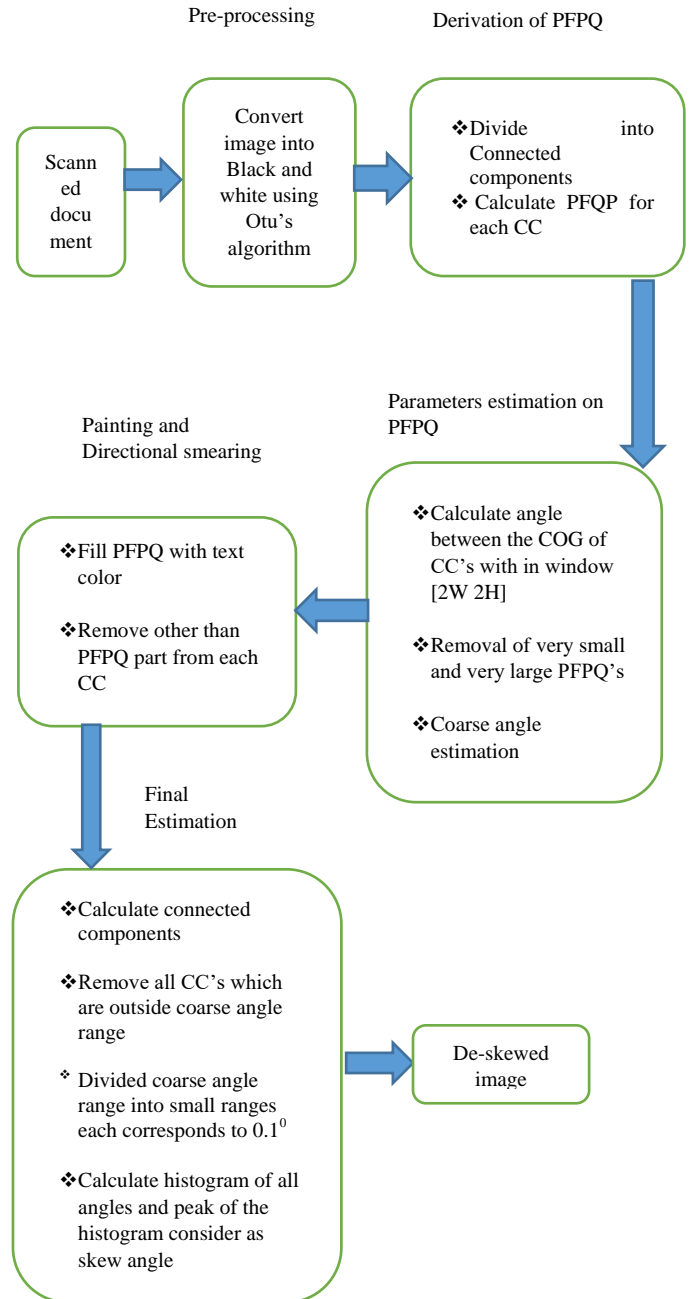


Fig.3. Block diagram of proposed PF PQ

After an in-depth study on the orientation of Telugu character we observed that the center of gravity (COG) and its orientation towards the principle axis is an important and crucial factor for the skew estimation. Based on this observation we found that only part around the principle axis of the character is very important to evaluate the skew of the character. Based on this

observation this paper proposes a new method for skew detection of Telugu document called “Principle-axis farthest pairs Quadrilateral (PFPQ)” algorithm.

The block diagram of the proposed model is shown in Fig. 3.

The proposed skew estimation process on Telugu document is given below.

*Step 1:* The proposed method initially performs a preprocessing operation based on Otu’s algorithm to convert given color or gray scale Telugu document into black and white image.

*Step 2:* To derive PFPQ this paper initially derived a connected component (CC) on Telugu document using 8-connectivity neighborhood algorithm. The center of gravity (COG) of all CC’s of Telugu document may not be in the same orientation due to the extra marks associated with Telugu characters. Thus to detect the accurate skew orientation, the proposed PFPQ algorithm derives quadrilateral for each CC. For this the proposed PFPQ initially computes the orientation  $\varphi$  of CC using the points inside or within the CC based on second order spatial moment as given below:

$$\varphi = \begin{cases} -\frac{1}{2} \arctan \left[ \frac{2\mu_{xy}}{\mu_{xx} - \mu_{yy}} \right] & \text{if } \mu_{xx} \neq \mu_{yy} \\ -\frac{\pi}{4} & \text{if } \mu_{xx} = \mu_{yy} \text{ and } \mu_{xy} \geq 0 \\ \frac{\pi}{4} & \text{if } \mu_{xx} = \mu_{yy} \text{ and } \mu_{xy} < 0 \end{cases} \quad (1)$$

Where

$$\begin{aligned} \mu_x &= \sum_{i=1}^n x_i/n \\ \mu_y &= \sum_{i=1}^n y_i/n \\ \mu_{xx} &= \sum_{i=1}^n (x_i - \mu_x)^2/n \\ \mu_{yy} &= \sum_{i=1}^n (y_i - \mu_y)^2/n \\ \mu_{xy} &= \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)/n \end{aligned} \quad (2)$$

Where  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  are  $n$  points within the CC and arc-tan is the inverse of tangent function. The range of arc-tan is between  $-90^\circ$  and  $+90^\circ$ . We have two solutions for  $\varphi$  which are  $\varphi$  and  $\varphi-90$ . The final value from these two are derived in this paper by using the least square method which minimizes the following sum of squares as given in equation 3

$$e^2(\varphi, \rho) = \sum_{i=1}^n (x_i \sin \varphi + y_i \cos \varphi - \rho)^2 \quad (3)$$

where  $\rho$  is the distance of the fitted line from the origin of the coordinate system to CC and  $\varphi$  is either  $\varphi$  or  $\varphi-90$  as shown in the Fig. 4.

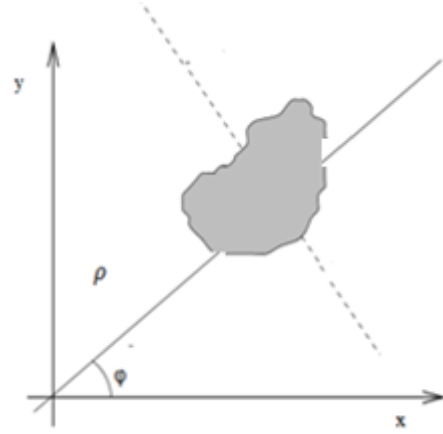


Fig.4.The estimation of  $\varphi$ .

After detecting  $\varphi$  the present paper identifies quadrilateral from the principle axes using center of gravity and orientation  $\varphi$  of CC in the following way.

Let  $\varphi$  be the angle of orientation of the CC. Let  $L_1$  and  $L_2$  be the major principle axis and minor principle axis respectively. The principle axis which passes through COG and makes an angle equal to orientation of CC is called major principle axis ( $L_1$ ). The principle axis which is perpendicular to major and passes through COG is called minor principle axis ( $L_2$ ). And (A,C) and (B,D) be the two pairs of opposite vertices of PFPQ as shown in Figure 5.

The equation of  $L_1$  and  $L_2$  are given by the following equations 4 and 5.

$$m^* x - y - m^*x_c - y_c = 0 \text{ where } m = \tan(\varphi) \quad (4)$$

$$m_1^* x - y - m_1^*x_c - y_c = 0 \text{ where } m_1 = \frac{-1}{m} \quad (5)$$



Fig.5. Representation of major ( $L_1$ ), minor ( $L_2$ ) and Quadrilateral vertices A,B,C and D of a given Telugu character ‘Ku’.

The first pair of PFPQ vertices A and C are derived in this paper based on the maximum perpendicular distance from the major principle axis  $L_1$  and first pair of PFPQ vertices A and C should satisfy the following inequalities 6 and 7 respectively and also shown in the Fig. 5.

$$m^* x - y - m^*x_c - y_c > 0 \quad (6)$$

$$m^* x - y - m^*x_c - y_c < 0 \quad (7)$$

Where  $m$  is the slope of  $L_1$ ,  $x_c$  and  $y_c$  are x-coordinate and y-coordinate of COG.

Similarly the second pair of PF PQ vertices B and D is calculated in the same way however they are in opposite sides of the minor axis  $L_2$  and the vertices B and D should satisfy the following inequalities 8 and 9 respectively.

$$m_1 * x - y - m_1 * x_c - y_c > 0 \quad (8)$$

$$m_1 * x - y - m_1 * x_c - y_c < 0 \quad (9)$$

Where  $m_1$  is the slope of  $L_2$ .

Finally the PF PQ is derived by joining the vertices A, B, C and D and it is shown in Fig 6.



Fig.6. Formation of PF PQ.

**Step 3:** The accuracy of skew estimation of document effects if it contains images, tables, sceneries or PF PQ with large size and also PF PQ with noise or small size. To identify them the present paper derived thresholds which are based on average width ( $aw$ ) and average height ( $ah$ ) of all PF PQ's of the document.

The average width ( $aw$ ) and average height ( $ah$ ) of all PF PQ's of the entire document are derived using equations 10 and 11 in the following way

$$aw = \frac{1}{N} \sum_{k=0}^N w_k \quad (10)$$

$$ah = \frac{1}{N} \sum_{k=0}^N h_k \quad (11)$$

Where  $N$  is the total number of PF PQ's and  $w_k$  and  $h_k$  be the width and height  $k^{th}$  PF PQ.

This paper estimated two thresholds  $t_1$  and  $t_2$  to identify images, tables, sceneries or PF PQ with large size as follows

$$t_1 = 3 * aw$$

$$t_2 = 3 * ah$$

skew estimation of the entire document. The minor disadvantage of the above coarse angle estimation is it gives the incorrect results when skew angle of document is close to the border between two adjacent ranges. To overcome this problem present paper extended the above coarse angle range by two degrees in both sides.

**Step 5:** To fill the gaps between two PF PQ's the present paper initially paints the each PF PQ with text color and then applies directional smearing algorithm using following formulas 12 and 13 to shape text lines

$$p_i = x_i + 2 * aw * \cos(\theta) \quad (12)$$

$$q_i = y_i + 2 * aw * \sin(\theta) \quad (13)$$

where  $(x_i, y_i)$  are co-ordinates of the text character pixels,  $\theta$  coarse angle and  $aw$  is the average width.

Some example of formation of PF PQ based on the proposed method are given in following Fig. 7. In Fig. 7 the original Telugu character (Nu) is rotated in three angles 7(a). The formation of principle axis by step 2b is shown in Fig. 7 (b). The formation of PF PQ is shown in Fig. 7 (c). The Fig. 7 (d) shows the filled quadrilateral with color of text and the remaining part of the character, which is out of the PF PQ is shown in Fig. 7 (e) in black color and PF PQ is represented with white borders.

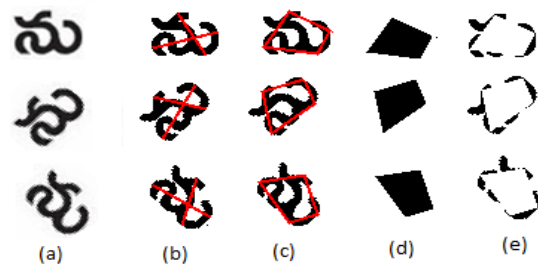


Fig.7. PF PQ algorithm steps (a) original Telugu character with tick mark (b) Formation of principle-axis (c) PF PQ generation (d) after painting (step 5) (e) The outside part of the PF PQ is shown in black color.

**Step 6:** To estimate the final skew angle of the entire document we calculate CC's and orientation of all connected components in the same way between two CC's using equations 1 and 2. Next we remove all CC whose orientation is out of the range of the coarse document skew angle. This paper divides the extended range of coarse angle into small ranges each corresponding to  $0.1^\circ$ . The peak of this histogram is considered as final skew of the document.

The Fig. 8 displays the three original Telugu documents, which contains text, logos, figures etc. and this makes the documents as complex documents. The stepwise results on the three complex of Fig.8 are shown in figures from 9 to 12.

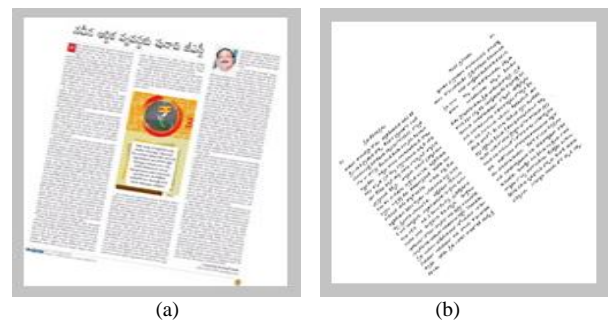


Fig.8. (a) and (b) are original Telugu documents.



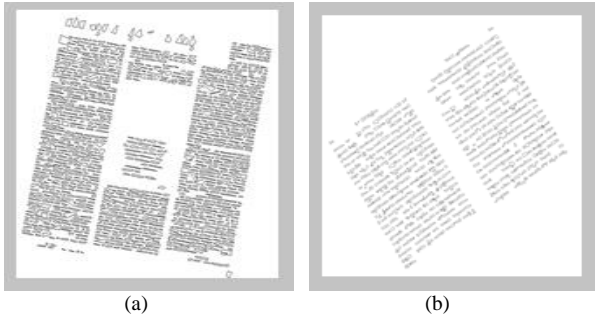


Fig.9. (a) and (b) are images obtained after applying PF PQ on Fig.8(a) and 8(b) respectively.

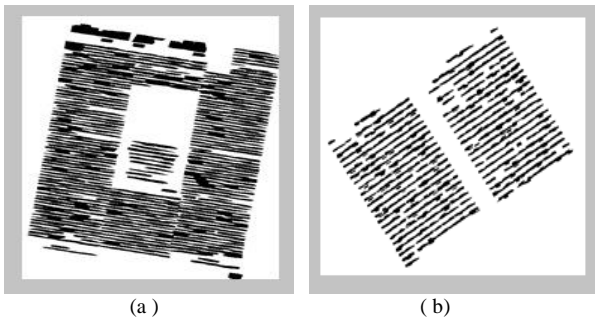


Fig.10. (a) and (b) are images obtained after painting and directional smearing on PF PQ of Fig. 9.

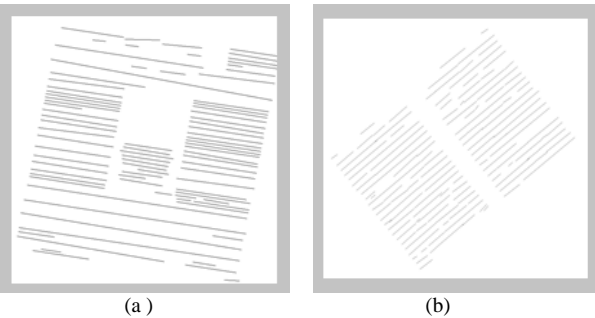


Fig.11. (a) and (b) are text document obtained after final angle estimation ( step 5) for images of Fig. 9.

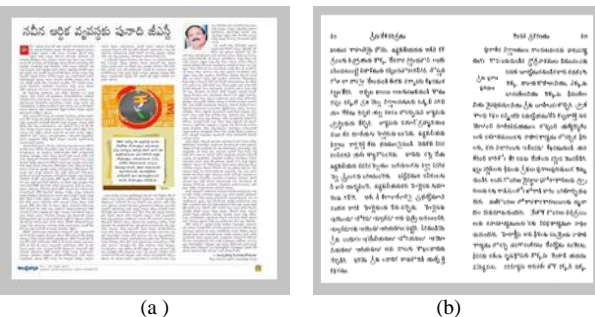
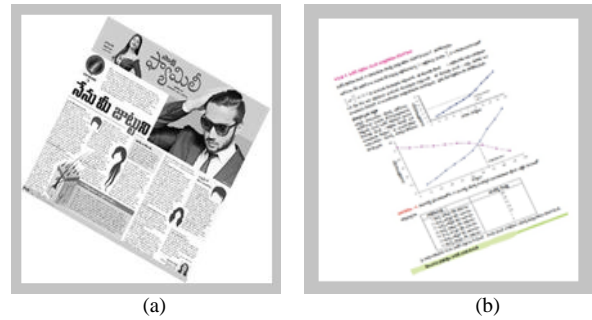


Fig.12. (a) and (b) are final text document De-skew.

### V. RESULTS AND DISCUSSION

This paper considered a total of 1000 text documents representing five categories. Out of five categories, the first four categories represent the different complex types of Telugu text documents. The considered documents are

in complex nature and they contain Images, tables, mathematical formulas etc. This paper collected 200 Telugu text documents from Newspapers, magazines and Text books and 25 documents of Telugu wedding cards and 25 Telugu official government documents (category 1). The samples of these shown in Fig 13. In category 2 the present paper considered 150 handwritten Telugu documents collected from housewives, students, government officers and others. Samples of these documents are shown in Fig. 15. In the third category the present paper considered 150 Telugu documents form social media. These Telugu documents may contain few English words also and sample images are shown in Fig.17. In category 4 the present paper considered single sentence and single word documents of 100 each and sample images are shown in Fig.19 and 21. In the fifth category, this paper considered 250 text documents collected from Newspapers of different Indian languages (other than Telugu) and sample images are shown in Fig.23. The corresponding de-skewed images of the above five categories obtained by the proposed method are displayed in Fig. 14,16,18,20,22 and 24.The results indicate the efficiency of the proposed PF PQ method.

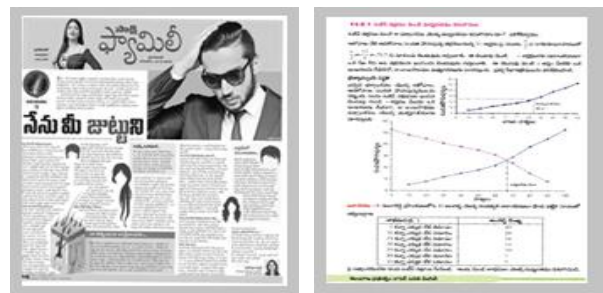


(a) (b)



(c)

Fig.13. Telugu document images with diagrams and tables etc.

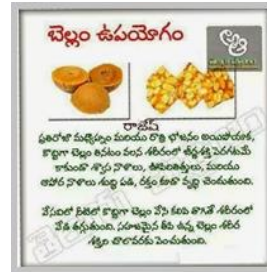


(a) (b)



(c)

Fig.14. De-skewed images of Fig 13.



(a)



(b)

Fig.18. De-skewed images of Fig. 17.

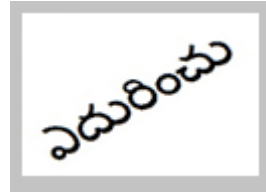


(a)

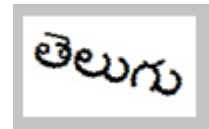


(b)

Fig.15. Handwritten documents.

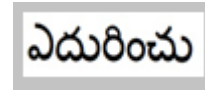


(a)

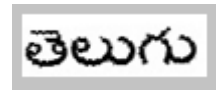


(b)

Fig.19. Single word images.



(a)

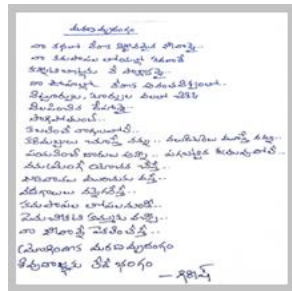


(b)

Fig.20. De-skewed images of Fig. 19.

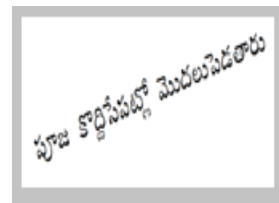


(a)

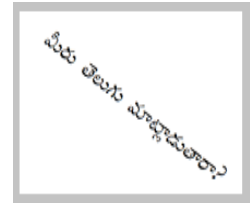


(b)

Fig.16. De-skewed images of Fig 15.



(a)



(b)

Fig.21. Single sentence images.

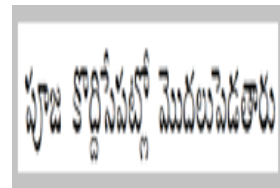


(a)

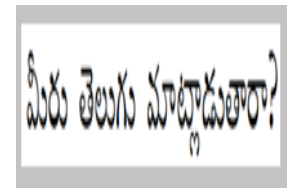


(b)

Fig.17. Telugu Images from social media.



(a)



(b)

Fig.22. De-skewed images of Fig. 19.





(a) Tamil

(b) Kannada

(c) Malayalam

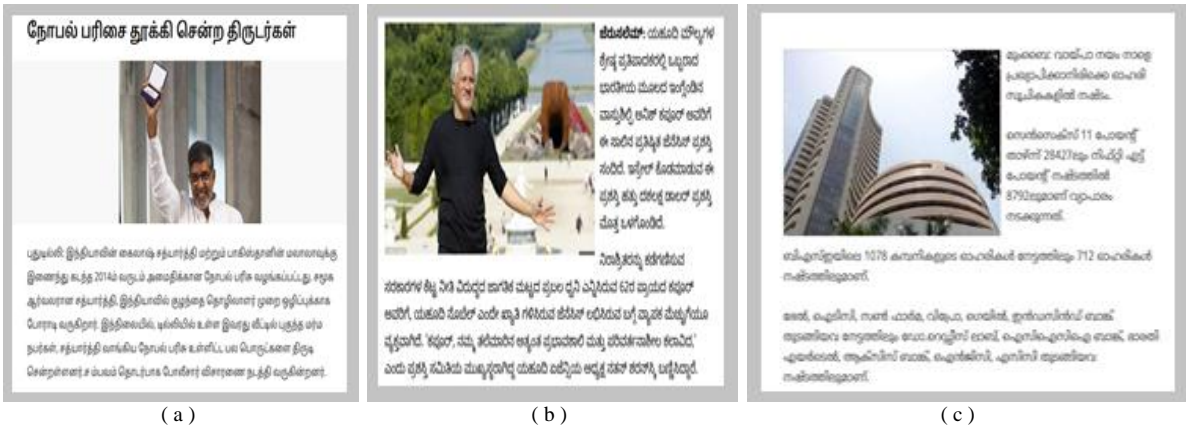


(d) Bangla

(e) Gujarati

(f) Hindi

Fig.23. Document images other than Telugu.



(a)

(b)

(c)



(d)

(e)

(f)

Fig.24. De-skewed images of Fig. 23.



The performance of present paper is evaluated using average error and variance as given in equations 14 and 15 respectively. The above average rates are considered for each category of document using following formulas 14 and 15.

$$\text{Average Error} = \mu = \frac{1}{N} \sum_{i=0}^N |\beta_i^* - \beta_i| \quad (14)$$

$$\text{Variance} = \sigma^2 = \frac{1}{N} \sum_{i=0}^N (\beta_i - \mu)^2 \quad (15)$$

Where  $\beta_i^*$  is the ground truth skew angle,  $\beta_i$  is the estimated skew angle by the proposed method, N is the number of documents in the dataset,  $\mu$  is average error and  $\sigma^2$  variance.

The low values of Average error and variance indicates the supremacy of the de-skew method. This paper computed average and variance on each of the document and average error rate and variance of each category is listed in Table 1. Further this paper evaluated the performance measured on the top 90% of the documents of each category that exhibited low error and variance rate and listed out in Table 1. The last row of Table 1 gives the average error rate and variance of all categories of document i.e. 1000 documents.

Table 1. Average error rate and variance of the proposed PFPQ method on five category of considered data.

Dataset Category	All text documents		Top 90% documents	
	Avg error rate	variance	Avg error rate	variance
Category 1	0.036	0.054	0.021	0.023
Category 2	0.052	0.067	0.036	0.035
Category 3	0.047	0.057	0.025	0.019
Category 4	0.007	0.005	0.004	0.0002
Category 5	0.063	0.061	0.029	0.020
Average of all Categories	0.041	0.048	0.023	0.019

A. Discussions:

From the experimental result, we observed that the proposed PFPQ achieved highest accuracy i.e. 99.3% for document of category 4 which contains either single word or single sentence. The reason for this is category 4 text documents don't contain any labels, images, logos, graphs or any other complex structure in the document. Next, we achieved good results in case of category 1 where the text document are collected from Newspapers, wedding cards, government official documents. These documents contain more of the Telugu text( on average of 90%) and below 10% of complex data, that's way they achieved low error rate and variance. The category 3 contains the social media Telugu documents which are similar to category 1 documents. In fact the category 3 and category 1 achieved almost the same accuracy rate however on average category 1 achieved higher accuracy rate. In category 2 the present paper collected or considered handwritten documents from different persons.

Each handwritten Telugu text document exhibits different kind of pattern because different persons write the same character with different styles and strokes that's way the handwritten Telugu documents are more complex in nature. The proposed PFPQ method achieved a higher accuracy rate of 94.8% on handwritten Telugu documents. This clearly indicates that the proposed PFPQ method is capable of de-skewing any type of Telugu document either it may be printed with simple or complex nature or handwritten. The present paper also applied the proposed PFPQ method on other languages text documents and achieved an accuracy rate of 93.7%. This accuracy rate is litter bit low when compared with type 1, 2, 3 Telugu documents. This indicates that the proposed method also suitable for de-skewing of other Indian languages.

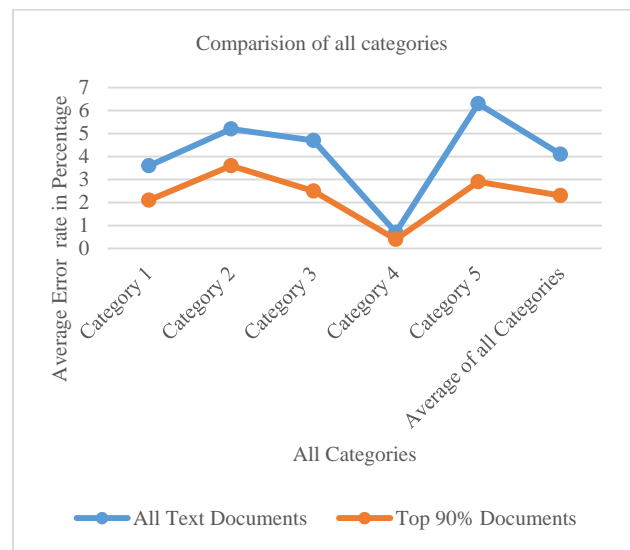


Fig.25. Comparison of all category data sets average error rate.

After analyzing the statistical results we observed that the accuracy of the proposed method depends on the accurate calculation of the coarse angle estimation. If the document contains text lines properly aliened the prosed method achieved more than 99% accuracy. The proposed method almost achieved 100% accuracy for fifth category documents i.e single word or single lines.

B. Comparative analysis of different methods:

Table 2. Comparison of the proposed PFPQ with the other methods

Method	All text documents		Top 90% documents	
	Avg error rate	variance	Avg error rate	variance
Project Profile	0.653	0.865	0.321	0.293
Hough Transform	0.477	0.582	0.258	0.267
Cross correlation	0.224	0.208	0.156	0.146
PPA	0.139	0.118	0.085	0.063
Proposed	0.041	0.048	0.023	0.019

The prosed PFPQ method is compared with other existing methods using classical project-profile [1], Hough transform [9], Yan [19] cross correlation and

Alireza et al.[23] PPA algorithms. The detailed comparison results are shown in the following table 2. From the table we can observe that the proposed method achieved very good results compared to all other methods. From these results, we conclude that the proposed method is very suitable for all Indian languages especially for Telugu documents. Comparisons with other methods are shown in Fig 26.

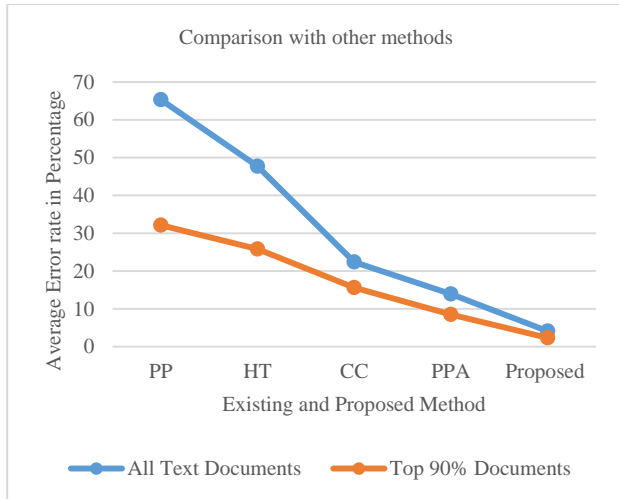


Fig.26. Comparison with other methods.

## VI. CONCLUSION

This Paper introduces a robust method for estimating skew angle from the different kinds of Telugu documents images. In this paper principle-axis-oriented quadrilateral painting and directional smearing algorithm is proposed to detected skew angle from the Telugu document. The proposed method is applicable for any document which may contain any textual, tabular or pictorial regions. This method works on any document without any angle restriction. The proposed method is independent on Font size, line spacing between Text lines etc. The robustness of our method was proved by conducting an experiment on five categories datasets with different contents. The proposed method doesn't require any user inputs and also not restricted to any angle. It can be applicable to any complex documents which may contain images, diagrams, tables, mathematical formulas etc. And proposed method can also be extended to other Our Indian languages.

## REFERENCES

- [1] Postl, W.: Detection of linear oblique structures and skew scan in digitized documents. In: Proc. 8th International Conference on Pattern Recognition, pp. 687–689 (1986)
- [2] Baird, H.S.: The skew angle of printed documents. In: Proc. SPSE 40th Symposium Hybrid Imaging, Rochester, NY, pp. 739–743 (1987)
- [3] Ciardiello, G., Scafuro, G., Degrandi, M.T., et al.: An experimental system for office document handling and text recognition. In: Proc. 9th International Conference on Pattern Recognition, pp. 739–743 (1988)
- [4] Ishitani, Y.: Document skew detection based on local region complexity. In: Proc. 2nd International on Document Analysis and Recognition, Japan, pp. 49–52 (1993)
- [5] Papandreou, A., Gatos, B.: A novel skew detection technique based on vertical projections. In: Proceedings of the 11th International Conference on Document Analysis and Recognition, ICDAR '11, pp. 384–388 (2011)
- [6] Papandreou, A., Gatos, B., Perantonis, S.J., Gerardis, I., 2014. Efficient skew detection of printed document images based on novel combination of enhanced profiles
- [7] Duda, R.O., Hart, P.E.: Use of Hough transformation to detect lines and curves in pictures. Commun. ACM 15, 11–15 (1972)
- [8] Hough, P.V.C.: Machine analysis of bubble chamber pictures. In: 2nd International Conference on High-Energy Accelerators, pp.554–558 (1959)
- [9] Srihari, N., Govindaraju, V.: Analysis of textual images using the Hough transform. Mach. Vis. A 2(3), 141–153 (1989)
- [10] Hinds, J., Fisher, L., D'Amato, D.P.: A document skew detection method using run-length encoding and the Hough transform. In: Proceedings of the 10th International Conference Pattern Recognition. IEEE CS Press, Los Alamitos, CA, pp. 464–468 (1990)
- [11] Manjunath, V.N., Kumar, G.H., Shivakumara, P.: Skew detection technique for binary document images based on Hough transform. Int. J. Technol. 13(3), 194–200 (2006)
- [12] Yu, B., Jain, A.K.: A robust and fast skew detection algorithm for generic documents. Pattern Recogn. 29(10), 1599–1629 (1996)
- [13] Wang, J., Leung, M.K.H., Hui, S.C.: Cursive word reference line detection. Pattern Recogn. 30(3), 503–511 (1997)
- [14] Singh, C., Bhatia, N., Kaur, A.: Hough transform based fast skew detection and accurate skew correction methods. Pattern Recogn. 41, 3528–3546 (2008)
- [15] Abdelhak Boukharouba A new algorithm for skew correction and baseline detection based on the randomized Hough Transform.
- [16] Hashizume, A., Yeh, P.-S., Rosenfeld, A.: A method of detecting the orientation of aligned components. Pattern Recognit. Lett. 4, 125–132 (1986)
- [17] Gorman, L.: The document spectrum for page layout analysis. IEEE Trans. Pattern Anal. Mach. Intell. 15(11), 1162–1173 (1993)
- [18] Lu, Y., Tan, C.L.: A nearest-neighbor chain based approach to skew estimation in document images. Pattern Recogn. Lett. 24, 2315–2323 (2003)
- [19] Yan, H.: Skew correction of document images using interline cross correlation. CVGIP: Graph. Models Image Process. 55(6), 538–543 (1993).
- [20] Gatos, B., Papamarkos, N., Chamzas, C.: Skew detection and text line position determination in digitized documents. Pattern Recogn. 30(9), 1505–1519 (1997)
- [21] Chou, C.-H., Chu, S.-Y., Chang, F.: Estimation of skew angles for scanned documents based on piecewise covering by parallelograms. Pattern Recogn. 40, 443–455 (2007)
- [22] Deya, P., Nousath, S.: e-PCP: A robust skew detection method for scanned document images. Pattern Recogn. 43, 937–948 (2010)
- [23] Alireza, A., Umamadham, P., Nagabhushan, P., Kimura, F.: An Efficient Skew Estimation Technique for Scanned Documents: An Application of Piece-wise Painting Algorithm 2016.

- [24] NVG Raju, V. Vijay Kumar, OS Rao, "Author based rank vector coordinates(ARVC) Model for Authorship Attribution", I.J. Image, Graphics and Signal Processing(IJIGSP), Vol. 5, 2016, pp:68-75, ISSN: 2074-9074
- [25] G. Srinivasa Rao, V.Vijaya Kumar, Penmesta Suresh Varma, "Cellular Automata Clustering Based on Morphological Reconstruction (CACMR)", I.J. Graphics, Vision and Image Processing (IJIGSP), Vol. 15, Iss.2, 2015, pp: 1-8, ISSN: 1687-398X,
- [26] K. Srinivasa Reddy, V.Vijaya Kumar, B.Eshwarareddy, "Face Recognition based on Texture Features using Local Ternary Patterns", I.J. Image, Graphics and Signal Processing (IJIGSP), Vol.10, 2015, pp: 37-46, ISSN: 2074-9082.
- [27] V.Vijaya Kumar, A. Srinivasa Rao, YK Sundara Krishna, "Dual Transition Uniform LBP Matrix for Efficient Image Retrieval", I.J. Image, Graphics and Signal Processing (IJIGSP), Vol. 8, 2015, pp: 50-57.
- [28] G S Murty ,J SasiKiran , V.Vijaya Kumar, "Facial expression recognition based on features derived from the distinct LBP and GLCM", International Journal of Image, Graphics And Signal Processing (IJIGSP), Vol.2, Iss.1, pp. 68-77,2014, ISSN: 2074-9082.
- [29] Lakhdar Belhallouche ,, Kamel Belloulata, Kidiyo Kpalma , A New Approach to Region Based Image Retrieval using Shape Adaptive Discrete Wavelet Transform, I.J. Image, Graphics and Signal Processing, 2016, 1, 1-14.
- [30] Prasanthi Jasmine1 ; P. Rajesh Kumar2 , Color and Rotated M-Band Dual Tree Complex Wavelet Transform Features for Image RetrievalM. I.J. Image, Graphics and Signal Processing, 2014, 9, 1-10.
- [31] V. Madhusudhana Rao, Dr. S.Pallam Setty, Dr. Y.Srinivas, An Efficient System for Medical Image Retrieval using Generalized Gamma Distribution, I.J. Image, Graphics and Signal Processing, 2015, 6, 52-58.
- [32] Abbas H. Hassin Alasadi, Saba Abdual Wahid, Effect of Reducing Colors Number on the Performance of CBIR System , I.J. Image, Graphics and Signal Processing (IJIGSP), 2016, 9, 10-16
- [33] Abdelhamid Abdesselam, Edge Information for Boosting Discriminating Power of Texture Retrieval Techniques s and techniques, I.J. Image, Graphics and Signal Processing (IJIGSP), 2016, 4, 16-28.
- [34] C. Vasantha Lakshmi, Ritu Jain, C. Patvardhan, "OCR of Printed Telugu Text with High Recognition Accuracies", ICVGIP 2006, pp. 786 – 795.
- [35] M.B. Sukhaswami, Seetharamulu, A.K Pujari, Recognition of Telugu characters using Neural Networks, Int. Journal of Neural Systems, September, 1995, 6(3), page 317 – 357.
- [36] Vijaya Kumar Koppula , Atul Negi Fringe Map Based Text Line Segmentation of Printed Telugu Document Images 2011 conference.
- [37] B.B. Chaudhuri and U. Pal Skew Angle Detection of Digitized Indian Script Documents IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 19, No. 2, February 1997.
- [38] M. Ramanan, A. Raman and and E. Y.A. Charles A Preprocessing Method for Printed Tamil Documents:Skew Correction and Textual Classification
- [39] Lalita Kumari1, Swapan Debbarma2, Radhey Shyam3 Text Orientation Detection from Document Image of Indian Scripts, *IJCCIS, Vol2. No1. ISSN: 0976-1349 July – Dec 2010*
- [40] Dharam Veer Sharma, Gurpreet Singh Lehal A Fast Skew Detection and Correction Algorithm for Machine Printed Words in Gurmukhi Script 2009 ACM 978-1-60558-698-4/09/07
- [41] Shamita Ghosh and Bidyut B. Chaudhuri **Composite** Script Identification and Orientation Detection for Indian Text Images 2011 International Conference on Document Analysis and Recognition

### Authors' Profiles



**MSLB Subrahmanyam** is working as a Head-imaging and algorithm practices at Srikari Impetus solutions pvt Ltd. He is a Research Scholar under Dr.V.Vijaya Kumar Director -Centre for Advanced Computational Research (CACR) and Dr B. Eswara Reddy professor of CSE Dept and principal of JNTU-A college of Engineering, Kalikiri,India from JNTU Kakinada. He received MTech(CS) from JNTU Hyderabad and MSc(Maths) from Osmania university, Hyderabad.



**Dr. V. Vijaya Kumar** is working as Dean in Dept. of CSE & IT, chairman Board of studies for CSE and IT and Director Centre for Advanced Computational Research (CACR) at Anurag Group of Institutions, (AGOI) (Autonomous), Hyderabad. He received integrated M.S.Engg, in CSE from USSR in 1989. He received his Ph.D. degree in Computer Science from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India in 1998 and guided 28 research scholars for Ph.D. He acted as principle investigator for various R&D projects. His research interests include big data, image analytics, image retrieval, image processing, pattern recognition, digital water marking and cloud computing. He is the life member of CSI, ISCA, ISTE, IE (I), IETE, ACCS, CRSI, IRS and REDCROSS. He published more than 100 research publications till now in various national, international journals and conferences.



**Dr. B. Eswara Reddy** Graduated in B.Tech.(CSE) from Sri Krishna Devaraya University in 1995. He received Masters Degree in M.Tech.(Software Engineering), from JNT University, Hyderabad, in 1999. He received Ph.D in Computer Science & Engineering from JNT University, Hyderabad, in 2008. Currently, he is working as Professor of CSE Dept. & Principal of JNTU-A College of Engineering, Kalikiri, India. He has published more than 100 publications in national and international conferences and journals. His research interests include Pattern Recognition & Image Analysis, Data Mining, and Cloud Computing. He received UGC Major Research Project titled 'Cloud Computing Framework for Rural Health Care in Indian Scenario'. He is co-author of the text books 'Programming with Java' (Pearson/Sanguine) and 'Data Mining' (Elsevier India). He is a life member of CSI, ISTE, ISCA, Fellow-IE (India) and member of IEEE.

**How to cite this paper:** MSLB. Subrahmanyam, V. Vijaya Kumar, B. Eswara Reddy," A New Algorithm for Skew Detection of Telugu Language Document based on Principle-axis Farthest Pairs Quadrilateral (PFPQ) ", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.10, No.3, pp. 47-58, 2018.DOI: 10.5815/ijigsp.2018.03.06