Captcha Breaking using Segmentation and Morphological Operations

Bandu Madar
Assistant Professor
Anurag Group Of Institutions

G. Kiran Kumar
Assistant Professor
Anurag Group Of Institutions

C. Ramakrishna
Assistant Professor
Anurag Group Of Institutions

ABSTRACT
Segmentation subdivides a CAPTCHA image into its constituent regions or objects. The point to which the subdivision is carried depends on the problem being solved. That is, segmentation should end when the objects of interest inan application have been isolated. Without a good segmentation algorithm, an object may never be identifiable. Image segmentation continues to be an vital and active research area in image analysis. Many techniques have been proposed to deal with the image segmentation problem. They can be broadly grouped into the following categories. Histogram-Based Techniques, Edge-Based Techniques, Region-Based Techniques, Hybrid Techniques. The accuracy of segmentation is highly dependent on the success or failure of each computerized analysis procedure. After the segmentation process is over, we should be familiar with, which pixel belongs to which object, the discontinuities where abrupt changes lie, tell us the locations of boundaries of regions. The connectedness of any two pixels is identified when there exists a connected path wholly within the set, where a connected path is a path that always moves between neighboring pixels. Therefore, region is a set of adjacent connected pixels. Extensive researches have been made in designing and creating different segmentation algorithms, however, still no algorithm is found from the researches results that can be accepted and appropriate for all kinds of images, obviously, all segmentation algorithms cannot be equally applicable to a certain application.

Keywords
CAPTHA, histogram, edge-based, accuracy

1. INTRODUCTION
Since its first appearance in 2000, a safety mechanism based on Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA) has been subjected to multiple attacks that seek to compromise their efficiency [1–5]. Therefore various security verification methods have been proposed covering a broad spectrum of options for generation of robust CAPTCHAs able to resist attacks by malicious programs [6–8]. Recently used methods are based on solutions of CAPTCHA imagerecognition tasks; text- or voice-processing; logical and mathematical puzzles, that besides offering a recognition challenge make the user apply some additional knowledge; even more, complex approaches that analyze patterns of clicks or face recognition [2,5,6,9]. However, text-based systems appear to be the most populardue to their easy implementation and usability. For this reason a set of design rules has been proposed to increase CAPTCHA security without compromising the user experience [2,9–11]. This has significantly reduced the number of CAPTCHA APIs allowing only the most mature remain in preference of Web security providers. Those systems, which properly follow the CAPTCHA design guidelines, are currently used as safety mechanisms in high-traffic sites such as Facebook, Ticket Master, Gmail, Livenation, Uploading, CNN, YouTube and others [4,11–13].

Recently, two basic concepts are assumed to address automatic recognition of CAPTCHA: to break anti-segmentation techniques used to protect regions corresponding to characters and to over-come anti-recognition techniques for each character. While anti-recognition mechanisms alter individual letter features such as font size, type and count, distortion, blurring and independent rotation of each character, the main secure mechanism to avoid breaking CAPTCHAs relies on anti-segmentation techniques that guarantee their robustness [9,11,13]. Some of the principal anti-segmentation techniques used in new versions of CAPTCHA and reCAPTCHA are the variable orientation of characters in word, the collapse between letters in a word, the addition of random dots and lines of different sizes, cluttered backgrounds and similar foreground/background colors [2,3,9,10]. For example, reCAPTCHA test proposes to recognize two out-of-context words, where waviness and horizontal stroke were added to increase the difficulty of breaking the CAPTCHA by a computer program. According to Bursztein, the unpredictable collapse in CAPTCHA is the best option to avoid segmentation of characters now widely used by various sites like Google, Facebook, Twitter and others [9,13]. Developed techniques for breaking CAPTCHAs are also used in pattern recognition applications particularly, for handwritten text interpretation or for optical character recognition (OCR) during automatic degraded text scanning. For example, the proposed approach provides simple and fast character recognition mechanism for scanning books in large scale such as Google Books and News Archive Search and their conversion to plain text [14,15].

Thus, the main purpose of this project is to reduce vulnerability of CAPTCHAs from frauds and to protect users against cyber-criminal activities as well as to introduce a novel approach for recognizing either handwritten or damaged texts in ancient books, manuscripts and newspapers. Currently, there are numerous techniques breaking CAPTCHAs. The most complete analysis of CAPTCHA beating mechanisms provided by Bursztein presents various systems able to recognize CAPTCHAs of some popular Internet sites, which include Wikipedia, eBay, CNN, Baidu, Mega upload and others with accuracy rates ranging from 40% to 95% [9,11,13]. However, these systems do not recognize CAPTCHAs provided by sites like Google or reCAPTCHA of new versions [10,13]. In Ref. [16] Yan presented attack on previous 2010 version of Google CAPTCHA, where character segmentation is based on analysis of patterns grouped in following categories: (1) point-shaped patterns (letter such as i or j); (2) cycle shaped patterns (letters a, b, d, etc.); (3) cross-shaped patterns (letters t and f) and (4) pattern, which juxta-poses three vertical lines to form the character (m or w). Although Google recently revamped its reCAPTCHA system nevertheless, Google's

reCAPTCHA now is vulnerable once again after newly launched reCAPTCHA-solving/breaking service [1].Several well-known approaches have broken CAPTCHAs such as Yahoo early CAPTCHAs [17], the CAPTCHAs used by PayPal site [18], Windows Hotmail and Gmail free e-mail providers [19], Live Journal, php BB, e-banking CAPTCHAs used by a lot of financial institutions and other services [20,21]. After that, when the Newcastle University research team has broken segmentation of Microsoft CAPTCHA with 90% of success rate, Microsoft uses improved CAPTCHA Control ASP.NET 2.0 [22]. This approach creates histogram of black pixels found in column assuming that characters are not overlapped and subsequently, defines letter separation point when no pixels are found in column. This segmentation approach fails, when characters are connected at least by a single pixel.

Recently reported in scientific literature approaches apply different algorithms to obtain CAPTCHA image skeleton for easy manipulation of characters overcoming in this way anti-segmentation mechanisms [9, 11, 23]. The precision of the segmentation step reported by newCAPTCHA beating systems lies between 40% [10,13,16] and 95% [9,10,24]. Interesting approach is presented by Liu[25], which exploitsa set of morphological filters that break satisfactorily security mechanism based on asymmetric-ellipses sometimes presented in reCAPTCHA. Another approach presented by Indian research group [26] considers that the pre-processing stage is not necessarily must generate complete letter blobs. It may be used only for fast global feature extraction however the correct segmentation task must be handled by the recognition module, which looks along the CAPTCHA image to define the character boundaries. The proposed approach achieves recognition accuracy about 72% with response time less than 14.5s per 400 CAPTCHAs.Although these results could give an idea that the problem is already solved, unfortunately, these reports frequently present theoreticalproposal and have not formalevaluation of whole CAPTCHA breaking process. The main security mechanisms implemented in reCAPTCHA are focused on exploiting different font sizes, which suffer from a particular pattern of waving rotation and random collapse overlapping characters in words. That represents a challenge for binarization and correct segmentation of characters. Additionally, some extra security features such as length and text-size randomization, character tilting and waving are used, which may guarantee that CAPTCHA scheme is secure against attacks [9,27].

Another requirement of systems for automatic CAPTCHA beating is providing high-speed recognition useful for real-time applications that not always are reported in well-known approaches. As usually, these CAPTCHA beating schemes apply the following stages: preprocessing for removal of background clutter and noise, segmentation for sub-division of CAPTCHA image into single regions and recognition of characters. The most difficult task is segmentation step although the development of fast and robust classifier is also a challenging task.

In this research we propose to subdivide the CAPTCHA breaking process into the following stages: CAPTCHA image acquisition, preprocessing, segmentation and recognition.

1.1 Effect of Noise in CAPTCHA images

In CAPTCHA image processing, CAPTCHA images are corrupted by different types of noises. It is very important to obtain precise images without noise to facilitate accurate observations for the given application. Removing of noise from CAPTCHA images is now a very challenging issue in the field of CAPTCHA image processing and many researchers are working in this area. Most well known noise reduction methods, which are usually based on the local statistics of a CAPTCHA image, are not efficient for CAPTCHA image noise reduction.

Low image quality and even a fractional noise images are obstacles for effective feature extraction, analysis, recognition and quantitative measurements especially in CAPTCHA image processing. Even though sometimes the quality is good, a small noise leads to an improper diagnosis by CAPTCHA expert thus leads to a false treatment, which may affect seriously the patient. Therefore, there is a fundamental need of noise reduction on CAPTCHA images.

1.2 CAPTCHA Image Segmentation

CAPTCHA Image processing can be defined as the manipulation of anCAPTCHA image for the purpose of either extracting accurate information from the image or producing an alternative representation of the image. There are numerous specific motivations for CAPTCHA image processing but many fall into the following categories: (i) to remove unwanted signal components that are corrupting the image and (ii) to extract information by rendering it in a more obvious or more useful form.

CAPTCHA image segmentation is one of the most critical tasks of image analysis because the segmentation results will affect all the subsequent processes of image analysis, such as representation and description, feature measurement and even the following higher level tasks such as classification and interpretation. CAPTCHA image segmentation can be done using color, gray level, depth, CAPTCHA or any other feature of interest according to the specific application.

The segmentation algorithm depends on the envisioned application and on the imaging modality employed. For instance, segmentation of gray and white matter in a cerebral MRI induces vastly different constraints from that of a vertebrae in an X-ray of the vertebral column, in terms of target topology, prior knowledge, choice of target representation, signal to noise ratio, and dimensionality of the input data. The selection of an adequate segmentation paradigm is therefore pivotal as it affects how efficiently the segmentation system can deal with the target organ or structure and conditions its accuracy and robustness. A deeper understanding of both the anatomical characteristics of the tissues and organs of the human body (or, more precisely, of the sub-structures we distinguish within them) and of their inter-relationships is crucial in diagnostic and interventional medicine which is possible only by the effective and accurate segmentation process.

1.3 CAPTCHA Segmentation

The CAPTCHA image segmentation also plays a vital role in various pattern recognition applications such as cartography, remote sensing, robot vision, military surveillance, inspection of textile products and CAPTCHA imaging. CAPTCHA segmentation divides an image into a set of distinct regions based on CAPTCHA properties, so that each region is uniform with respect to certain CAPTCHA characteristics. Results of segmentation can be further used to image processing and analysis, for example, to object recognition. Similar to classification, segmentation of CAPTCHA also involves extracting features and deriving metrics to separate CAPTCHA s. However, segmentation is generally more complex than classification, since boundaries that separate

different CAPTCHA regions have to be detected in addition to recognizing CAPTCHA in each region. CAPTCHA segmentation could also be supervised or unsupervised based on if prior knowledge about the image or CAPTCHA class is available. Supervised CAPTCHA segmentation identifies and splits one or more regions that match CAPTCHA properties shown in the training CAPTCHA s. For CAPTCHA images unsupervised segmentation is preferred as the information to be segmented is not known in advance.Unsupervised segmentation has to first recover dissimilar CAPTCHA classes from an image before dividing them into regions. Compared to the supervised case, the unsupervised segmentation is more flexible for real world applications despite that it is generally more computationally expensive. Segmenting an image into uniform regions is very useful in various applications of proper identification of abnormality, shape and volume in CAPTCHA image processing, pattern recognition and machine learning.

2. THE PROPOSED APPROACH FOR CHARACTER SEGMENTATION AND RECOGNITION

Based on analysis of character morphology in CAPTCHA alphabet, we grouped them by some characteristics in the following categories: Some characters with circular regions such as a, b, d, e, g, o, p and q. These letters generate regions of 20 pixels wide. Characters with occurrence of more than one pixel per column for letters like c, e, f, k, s, t, z. There usually exist at least two pixels for each column.

U-shape pattern characters such as u, n, h are normallypresented as pattern with two narrow sections with more than one pixel in column separated by a wide section of columns with only one pixel (the part of letter that connects vertical segments). Characters of one pixel per column with slope representing letters like v, x, y with a slope about 4517101. Thin characters are letters such as i, j, l; they consist of a small vertical block of approximately 5 pixels wide. R-shape pattern characteris formed by narrow stripe with somepixels in column followed by a much larger section of columns with only one pixel.

Double Characters are letters m and w, which can be commonly confused with letter n or v only, when they are separated by column without black pixels. The obtained three-color bar may be enhanced using some proposed rules. They must be applied one at a time and in the following order to avoid interference between them:

1. Noise reduction: if a bar in generated three-color bar code is black and it is only of one pixel wide, then the bar is replaced by a white bar.
2. Slope calculation: calculate the slope of white segments
3. Mand w pattern matching: the m-type pattern can be wronglyinterpreted as two consecutive n characters. This character is detected by a segment of pixels represented by two wide white bars separated by black bar. To find m-type pattern in the three-color bar code we run a template matching algorithm using template image. For each found m-type pattern, the corresponding region is segmented in stringent CAPTCHA image.
4. Upattern matching: this pattern represented by two black barsseparated by a white bar is found in letters n, u, v, y and h. To find u-type pattern in three color bar code we run atemplate matching algorithm using the template. Then for each found patterns the correspondingregion is segmented in stringent CAPTCHA image.

2.1 Proposed Algorithm:

Step1: Read color CAPTCH image Step

2: Convert color CAPTCH into gray scale image Step

3: By using threshold value extract the CAPTCH from the back ground of image Step

4: Apply spatial filter for further cleaning Step

5: Segment the cleaned CAPTCHA into individual characters Step

6: Recognize the characters by using Template matching Step

7: Decode the characters print recognized characters and display confidence

3. RESULT AND DISCUSSIONS

The proposed method considered Google CAPTHAs and applied proposed method. The proposed method performed well in segmenting, and identifying the characters in CAPTHS and recognizing the CAPTHA. It is also computed confidence value which decides the accuracy of proposed method.

(a) Original CAPTHA

(b) Segmented result

(c) Character wise classification 7XUVV

Fig.1: CAPTHA breaking system results.

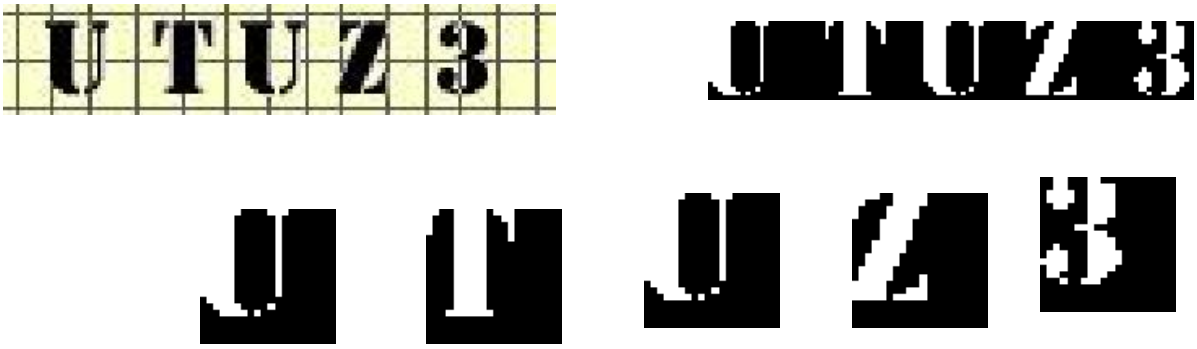

Fig.2: Different font of CAPTCHA results.

Table 1: Confidence value of proposed and existing methods.

Methods	Confidence
Preetiika et. al[28]	85.69
Mohammad[29]	92.10
Proposed Method	99.95

The Table 1 shows the confidence value of proposed and existing methods. From this table, it is clearly that the proposed method shows its dominance compared with exiting methods. And corresponding result is shown in Fig.3.

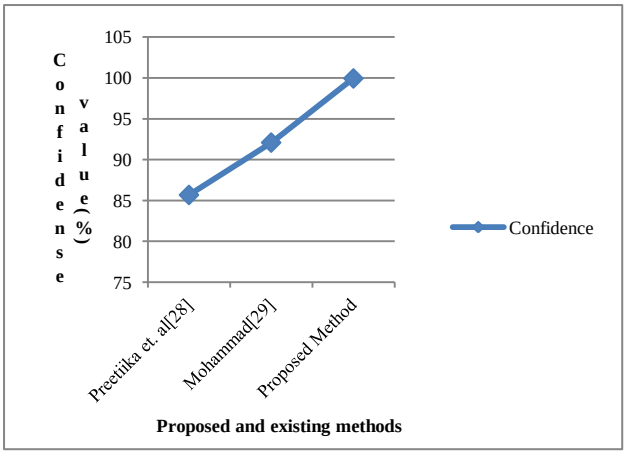

Fig.3: Comparison graph of proposed and existing methods.

4. CONCLUSIONS

In summary, three designs of text based CAPTCHA areproposed in this PROJECT. This CAPTCHA breaking design follows the principle "hard toseparate text from background using segmentation techniques". The CAPTCHAs are designed considering the techniques and concepts involvedin cracking various existing CAPTCHAs. The proposeddesigns of CAPTCHA are thus too strong to get cracked using template matching and at the same time very user friendly with high confidence rate. The proposed method classification performance as follows

- Pixel Counting: 8% Break Rate
- Vertical Projections: 97% Break Rate
- Horizontal Projections: 100% Break Rate
- Template Correlations: 100% Break Rate

5. REFERENCES

[1] D. Danchev, Google's reCAPTCHA under automatic fire from a newly launched reCAPTCHA-solving/breaking service, Internet Security Threat Updates & Insights, http://www.webroot.com/blog/2014/01/21/googles-recaptcha-automatic-fire-newly-launched-recaptcha-solving-breaking-service/⟩, 2014.

[2] C. Obimbo, A. Halligan, P. De Freitas, CaptchAll: an improvement on the modern textbased CAPTCHA, J. ProcediaComput. Sci. 20 (2013) 496–501.

[3] G. Baxter Bell, Strengthening CAPTCHA-based Web security, First Monday J. 17 (2012) 2 http://firstmonday.org/ojs/index.php/fm/article/view/3630⟩.

[4] S. Kulkarni, H.S. Fadewar, CAPTCHA based web security: an overview, Int. J. Adv. Res. Comput. Sci. Softw. Eng. 3 (11) (2013) 154–158 ⟨http://www.ijarcsse.com/docs/papers/Volume_3/11_November2013/V3I10-0379.pdf⟩.

[5] M. Serrao, S. Salunke, A. Mathur, Cracking CAPTCHAS for cash: a review of CAPTCHA crackers, Int. J. Eng. Res. Technol. 2 (1) (2013) 1–5.

[6] G. Goswami, B.M. Powell, M. Vatsa, R. Singh, A. Noore, FaceDCAPTCHA: face detection based color image CAPTCHA, Futur. Gener. Comput. Syst. 31 (2014) 59–68.

[7] L.D. Priya, S Karthik, Secure captcha input based spam prevention, Int. J. Emerg. Sci. Eng. 1 (7) (2013) 9–12.

[8] S. Azad, K. Jain, CAPTCHA: attacks and weaknesses against OCR technology, Global J. Comput. Sci. Technol. Neural Artif. Intell. 13 (3) (2013) 14–18.

[9] E. Bursztein, A.Moscicki, C. Fabry, S. Bethard, J.C. Mitchell, D. Jurafsky, Easy does it: more usable CAPTCHAs, in: Proceedings of the 32nd ACM Conference on Human Factors in Computing Systems, Canada, 2014, pp. 2637–2646, http://dx.doi.org/10.1145/2556288.2557322.

[10] C. Cruz-Perez, O. Starostenko, F. Uceda-Ponga, V. Alarcon-Aquino, L. Reyes-Cabrera, Breaking reCAPTCHAs with unpredictable collapse: heuristic character segmentation and recognition, in: J.A. Carrasco-Ochoa, J.F. Martinez-Trinidad, J.A. Overa Lopez, K. Boyer (Eds.), LNCS: Pattern Recognition, 7329, Springer-Verlag, Berlin Heidelberg, 2012, pp. 155–165.

[11] E. Bursztein, M. Matthieu, Text-based CAPTCHA strengths and weaknesses, in: Proceedings of the 18th ACM Conference on Computer and Communications Security, IL, USA, 2011, pp. 125–138, ⟨http://ly.tl/p22⟩.

[12] K. Fang, Z. Bu, Z.Y. Xia, Segmentation of CAPTCHAs based on complex networks, in: J. Lei, F. Lee Wang, H. Deng, D. Miao (Eds.), LNCS: Artificial Intelligence and Computational Intelligence, 7530, 2012, pp. 735–743.

[13] E. Burzstein, H..Paskov, et. al., Science of CAPTCHAs: solvability by humans and machines, AFOSR MURI Project, 2013, pp. 1–59.

[14] Committee on Institutional Cooperation. Google Book Search Project, 2014 ⟨http:// www.cic.net/projects/library/book-search/introduction⟩.

[15] C. Lim Tan, X. Zhang, L. Li, Image based retrieval and keyword spotting in documents, in: D. Doermann, K. Tombre (Eds.), Handbook of Document Image Processing and Recognition, Springer-Verlag, London, 2014, pp. 805–842.

[16] J. Yan, A. Salah, E. Ahmad, The robustness of a new CAPTCHA, in: 3rd Workshop on System Security, NY, USA, 2010, pp. 36–41, ⟨http://doi.acm.org/ 10.1145/1752046.1752052⟩.

[17] M. Wehner, Internet advertisers kill text-based CAPTCHA, ⟨http://news.yahoo. com/internet-advertisers-kill-text-based-captcha-205416291.html⟩, 2013.

[18] K. Kluever, R. Zanibbi, Breaking the PayPal CAPTCHA, ⟨http://www.kloover. com/2008/05/12/breaking-the-paypalcom-captcha/⟩, 2014 (retrieved on 25th of May).

[19] K. Dawson, Windows Live Hotmail CAPTCHA Cracked, Exploited, ⟨http://tech. slashdot.org/article.pl?sid=08/04/15/1941236&from=rss⟩ and Gmail CAPTCHA Cracked, ⟨http://it.slashdot.org/article.pl?sid=08/02/27/0045242⟩, 2014.

[20] S. Li, A. Syed, et. al., breaking e-Baking CAPTCHAs, in: Proceedings of the 26th Computer Security Applications Conference, NY, USA, 2010, pp. 171–180, ⟨http://www.acsac.org/2010/openconf/modules/request.php?module=oc_program&action=summary.php&id=53⟩.

[21] S. Kruglov, Defeating of weak CAPTCHAs, ⟨http://www.captcha.ru/en/break ings/⟩, 2013.

[22] Microsoft ASP.NET Team, Using a CAPTCHA to Prevent Bots from Using our ASP.NET Web Razor) Site, ⟨http://www.asp.net/web-pages/tutorials/security/using-a-catpcha-to-prevent-automated-programs-(bots)-from-using-your-aspnet-web-site⟩, 2012.

[23] S.E. Ahmad, J. Yan, M. Tayara, The Robustness of Google CAPTCHAS, Newcastle University Print, England, 2011 (Technical report).

[24] A. Baluni, S. Gole, Two-step CAPTCHA: using a simple two step turing test to differentiate between humans and bots, Int. J. Comput. Appl. 81 (16) (2013) 48–51.

[25] P. Liu, J. Shi, L. Wang, L. Guo, An efficient ellipse-shaped blobs detection algorithm for breaking facebook CAPTCHA, in: Y. Yuan, X. Wu, Y. Lu (Eds.), CCIS: Trustworthy Computing and Services, 320, Springer-Verlag, Berlin Heidelberg, 2013, pp. 420–428.

[26] D. Kapoor, H. Bangar, A. Chaurasia, A. Sethi, An ingenious technique for symbol identification from high noise CAPTCHA images, in: Proceedings of the Annual IEEE India Conference, 2012, pp. 98–103.

[27] M. Takaya, H. Kato, T. Komatsubara, Y. Watanabe, A. Yamamura, Recognition of one-stroke symbols by humans and computers, J. Procedia – Soc. Behav. Sci. 97 (6) (2013) 666–674.

[28] Preethi, Vikas,Security online Authentication using Captcha, International Journal of Advanced Research in Computer Science and Software Engineering, Volume , 5, , Issue ,6 ,June-2015.

[29] Mohammad Jabed Morshed Chowdhury, Narayan Ranjan Chakraborty, CAPTHA Based on Human Cognitive Factor, IJACSA,Vol. 4, No.11,2013

5. AUTHOR PROFILE

Mr.BANDU MADAR working as Assistant Professor in Anurag Group of Institutions , Hyderabad I have 9 years of Experience in teaching. I received M.Tech from Jawaharlal Nehru Technological University. Hyderabad, Telangana.

Mr. GARA KIRAN KUMAR working as Assistant Professor in Anurag Group of Institutions , Hyderabad I have 6 years of Experience in teaching. I received M.Tech from Jawaharlal Nehru Technological University. Hyderabad, Telangana.

Mr.CHITHARI RAMAKRISHNA working as Assistant Professor in Anurag Group of Institutions , Hyderabad I have 6 years of Experience in teaching. I received M.Tech from Nagarjuna University. Guntur, Andra Pradesh.