Available online at www.sciencedirect.com

ScienceDirect

Materials Today: ProceedingsXX (2016) XXX–XXX

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

ELSEVIER

PMME 2016

INDEXING NEAR-DUPLICATE IMAGES IN WEB SEARCH USING MINHASH ALGORITHM*

#S.Thaiyalnayaki, *J.Sasikala*

[a]Assistant Professor,Dhanalakshmi srinivasan college of engineering and technology,ECR,Mamallapuram,Chennai-603104
[b]Assistant Professor, Annamalai University,Chidambaram-608002.

Abstract

Near-duplicate images cause problems of redundancy and copyright infringement in large image collections. The trouble is minor in the web, where appropriation of images without acknowledgment of source is prevalent. Near duplicates can be exact copies or else differ slightly in their visual content. Near-duplicate detection has received substantial attention over the past few years due to the applications in copyright enforcement, organizing large size image databases, increasing focus on image search, redundancy elimination of logos, managing storage space by removing duplication, etc. In the paper, a method has proposed for Indexing Near-Duplicate Images in the Web Search. First image enhancement is done in user query image then features are extracted based on SURF (Speeded up Robust Features) that is to extract the local invariant features of each image. After this similarity measure is calculated among the feature extracted images using min-hash algorithm. Finally, Locality Sensitive Hashing (LSH) is used for indexing near duplicate images based on user query. We demonstrate that our indexing approach is highly effective for collections of up to a few hundred thousand images.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Indexing,near-duplicates, near-duplicate detection, Image Enhancement

1. INTRODUCTION

Billions of photos and videos generates in the World Wide Web are growing every day. Users who are browsing the internet will rapidly encounter many duplicates of images in multiple locations. For instance, several news sources use the same photo the devastation of an earthquake, while the same funny image of the baby wearing

*SThaiyalnayaki,J.Sasikala. Tel.: +919566208899.
*E-mail address:*thaiyalvijayo@gmail.com

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

clothes like a tiger may be shared by hundreds of person on a social network. In general, being able to detect and trace duplicates is useful in various application areas, including: Large storage space: Only a single image needs to be stored, instead of keeping a backup copy of each duplicate which is especially useful for photo sharing websites. Understanding behaviour and interests: Tracing how images are shared and how they spread across the internet can give insights into the social behaviour of people and their interests. Duplicate images means exact duplicates, indicating the images are completely similar in appearance, or near-duplicates, indicating the images are not exact similar but differ slightly in content. Most of the systems perform feature extraction as a pre-processing step, obtaining global image features like colour histogram or local descriptors like shape and texture. In this paper, the near-duplicate images are detected based on user query image and retrieving the near duplicate image based on indexing. This process is achieved by three steps; initially features are extracted from the query image. Second is after extracting the features of each images, similarity is calculated. Finally, indexing of near duplicate images based on user query is done. For indexing we use Locality Sensitive Hashing (LSH).

2. RELATED WORKS

In the field of near-duplicate detection or image retrieval, there are many different methods according to different specifications or patterns. A common solution to retrieve images similar to each other in content properties like shape is based on Content-Based Image Retrieval (CBIR).[1] This paper proposes an indexing method based on stroke density code. Initially segment the document image to retrieve all the Chinese character images, then calculate its stroke density of each Chinese character image, and at last make the stroke density code of each character image. [4] Used to detect image spam effectively and also it is necessary to analyze the image content. Near-duplicate image spam is detected based on CE (cross entropy), in which the SURF is used to extract the local invariant features of each image (spam and ham); then the GMM (Gaussian Mixture Models) of local invariant features which are fit using CE as the distance measurement between Gaussian distributions, improve the K means to all the cluster the GMMs since our dataset is very large. [6]Proposed to estimate the degree of correlation between two given patterns, auto correlation and cross correlation are the common statistical techniques used in the areas of image processing and pattern recognition. [8] From the results of the experiment it is clear to see that the classification ant-colony algorithm is able to automatically identify different target categories in the process of image classification. In this algorithm all ants are divided into two groups: Stochastic and intellectual ants. The stochastic ants can remember the positions of the target pixels when they pass, and ascertain the initial clustering center. The stochastic ants provide search guidance for the intellectual ants. [9] In this paper a signature for content-based image copy detection is described. The extraction of the signature is fast and the detection speed is very efficient – millions of signatures can be compared in a second. Also the precision and recall rates are high for many commonly used modifications of images.

3. PROPOSED WORK

In proposed system, Indexing and identifying near duplicate images is done using SURF features. Identification of duplicate images consists of four steps. a) Image Enhancement b)Speeded Up Robust Features(SURF) c) Min-Hash Algorithm d)Locality Sensitive Hashing. Min hashing means convert large sets to short signatures, while preserving similarity. Locality-sensitive hashing means focus on pairs of signatures which are likely to be similar.

3.1 Dataset

In this paper, real images are used. Actually this dataset contains real camera photos taken directly from a real user’s photo gallery. This collection consists of different types of near duplicate and exact duplicate images.

3.2 Image Enhancement

Aim of image enhancement is to improve the query image quality. To improve the image quality, transformation function is needed which takes the pixel’s intensity value of the query image and generate the new intensity value for corresponding pixel to produce the enhanced image.

To estimate the quality of the enhanced image automatically, an evaluation function is needed which will tell us about the quality of our enhanced image.

3.3 Speeded Up Robust Features (Surf)

The work of Speeded Up Robust Features (SURF) is to detect the features of the image. In proposed work SURF is used for extracting image features and detects the interest points among the images.

SIFT was comparatively slow and people needed more speeded-up version. In 2006, SURF: Speeded Up Robust Features” was introduced. As name suggests, it is a speeded-up version of SIFT.SURF, sometimes referred to as the Fast-Hessian detector and is essentially based on the Hessian matrix with the Laplacian-based detectors such as Difference of Gaussian (DoG) [13].

The search for discrete image point correspondences can be divided into three major steps. Primarily, ‘interest points’ are selected at distinctive locations in the image, like blobs, corners and T-junctions. Next, the neighbourhood of each and every interest point is provided by a feature vector. Also the descriptor has to be distinctive and at the same time resistive to noise, detection displacements and geometric and the photometric deformations. Finally, the descriptor vectors are compared between different images. The matching is based on the distance between the vectors SURF (Speeded up Robust Features) is used for extracting the local invariant features of every labelled image [3]. This is a scale-invariant and the rotation-invariant interest point detector and descriptor which is achieved by relying on integral images for image convolutions.

Interest point detection

To achieve the fast robust features, the SURF algorithm employs the integral images approach which reduces the computation time.

Integral images

Occasionally, this approach is the summed area table [15] image from the summing of pixels’ intensity of the input image I within a rectangular region formed around location x as the following [13]

$$I\varepsilon(X) = \sum_{i=0}^{i\le x}\sum_{j=0}^{j\le y} I(i,j) \qquad (1)$$

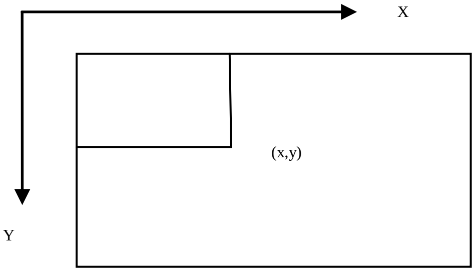

fig. 1 Summing of pixels

The integral image computes values at each pixels (x,y) that is sum of pixel values above and to left of (x,y) as shown in the figure with recursive definition shown below the integral image can be computed quickly one pass through as shown in below equation

Sum(x,y)=sum(x,y-1)+i(x,y) (2)

I(x,y)=I(x-1,y)+s(x,y) (3)

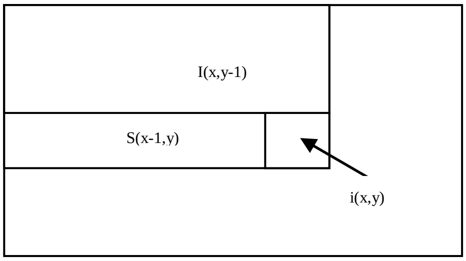

Fig. 2 Recursive definition of integral image

3.4 Min-Hash Algorithm

In this section, it is described how a method originally developed for text near-duplicate detection and it is adapted to near-duplicate detection of images. The goal is to retrieve all the image in the web images that are similar to a query images. This section looks at an efficient randomized hashing based procedure that retrieves near duplicate images in time proportional to the number of the near duplicate images. The outline of the algorithm is as follows: Initially a list of min-Hashes is extracted from each web images and query image. Usually a min-Hash is a lone number having the property that two sets w1 and w2 have the same value of min-Hash with their probability equal to their similarity Sims (w1, w2). For efficient retrieval the min-Hashes are grouped into n-tuples called as sketches. Identical sketches of images are then efficiently found using a hash table. Images with at least h identical sketches (i.e., sketch hits) are considered as possible near duplicate images and their similarity is then estimated using all the available min-Hashes**.**

The distance measure between the query image and web images is computed as the similarity of sets w1 and w2, which is defined as the ratio of the number of elements in the intersection over the union:

$$\mathrm{sim}(w1,w2)=\frac{w1\cap w2}{w1\cup w2} \qquad (4)$$

Then we discussed how to compare sets, specifically using the Jacquard similarity. For example if a query image is qi = {0, 1, 2, 5, 6} and web image is wi = {0, 2, 3, 5, 7, 9}.

The Jaccard similarity is defined

$$\mathrm{Sim}(qi,wi)=\frac{qi\cap wi}{qi\,U\,wi} \qquad (5)$$

$$=\frac{\{0,2,5\}}{\{0,1,2,3,5,6,7,9\}} = \frac{3}{8}$$

3.5 Locality Sensitive Hashing

Locality-sensitive hashing (LSH) algorithm by Indyk&Motwani[10], is an approximate similarity search technique that works capably even for high-dimensional images.The algorithm builds a set of l such hash functions and each of this selects k bits from the bit string and each function uses a different and randomly selected set of k bits). These k bits are hashed once more to index into the buckets in the hash table, and a 32-bit checksum hash value is also generated. The two parameters k and l will enable the designer to select an appropriate trade-off between accuracy and running time.

LSH and Min Hash are computed using following algorithm.

3.6 Algorithm For Min hash and LSH

Procedure for calculating Similarity.

Input: Features of query image.

Output: Indexing a near and Exact duplicate images

```
1. Features of Query image qi like $q_{i1},q_{i2},q_{i3},\ldots\ldots,q_{in}$, Features of Web image wi like
$w_{11},w_{12},w_{13},\ldots\ldots,w_{1n}$,
$w_{21},w_{22},w_{23},\ldots\ldots w_{2n},W_{31},w_{32},w_{33},\ldots\ldots w_{3n},\ldots\ldots,W_{i1},w_{i2},w_{i3},\ldots\ldots w_{in}$
2. for all images F=1,……………K do
          If($q_{iF}$==$w_{iF}$) then Increment the sim [i]
          Increment the Features
      End
End
3.If sim[i]==k then        EDI = $W_{if}$
  Else   NDI=$W_{if}$
End
4. Increment the image I;
```

3.7 Algorithm For Indexing A Near Duplicate And Exact Duplicate Detection

```
1.for all i=1,...................,n
If(Sim[i==sim[i+1])
          Pos[i]=sim[i];
          Pos[i+1]=sim[i+1];
          End
End
2.j=i
3.For all i=1,..................,n;
If(sim[i]!=sim[i+1])
Pos[j+1]=sim[i];
Pos[j+2]=sim[i+1];
End
End
4.Display of pos[i];
5.stop.
```

3.8 Overall Proposed System

The proposed system is used to detect all the near duplicate images and exact duplicate images corresponding to the query image; Overall Block Diagram of the Proposed System is

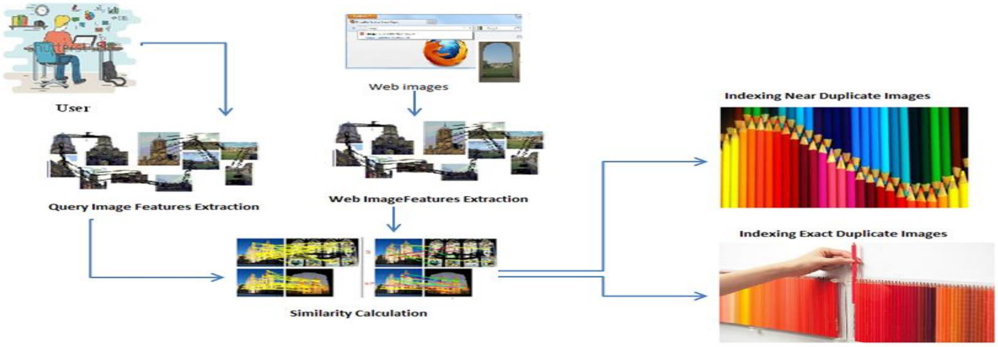
User
Web images
Query Image Features Extraction
Web ImageFeatures Extraction
Similarity Calculation
Indexing Near Duplicate Images
Indexing Exact Duplicate Images

Fig.3 Overall Block Diagram of Proposed System

The steps in Proposed Work can be depicted using the flow chart –

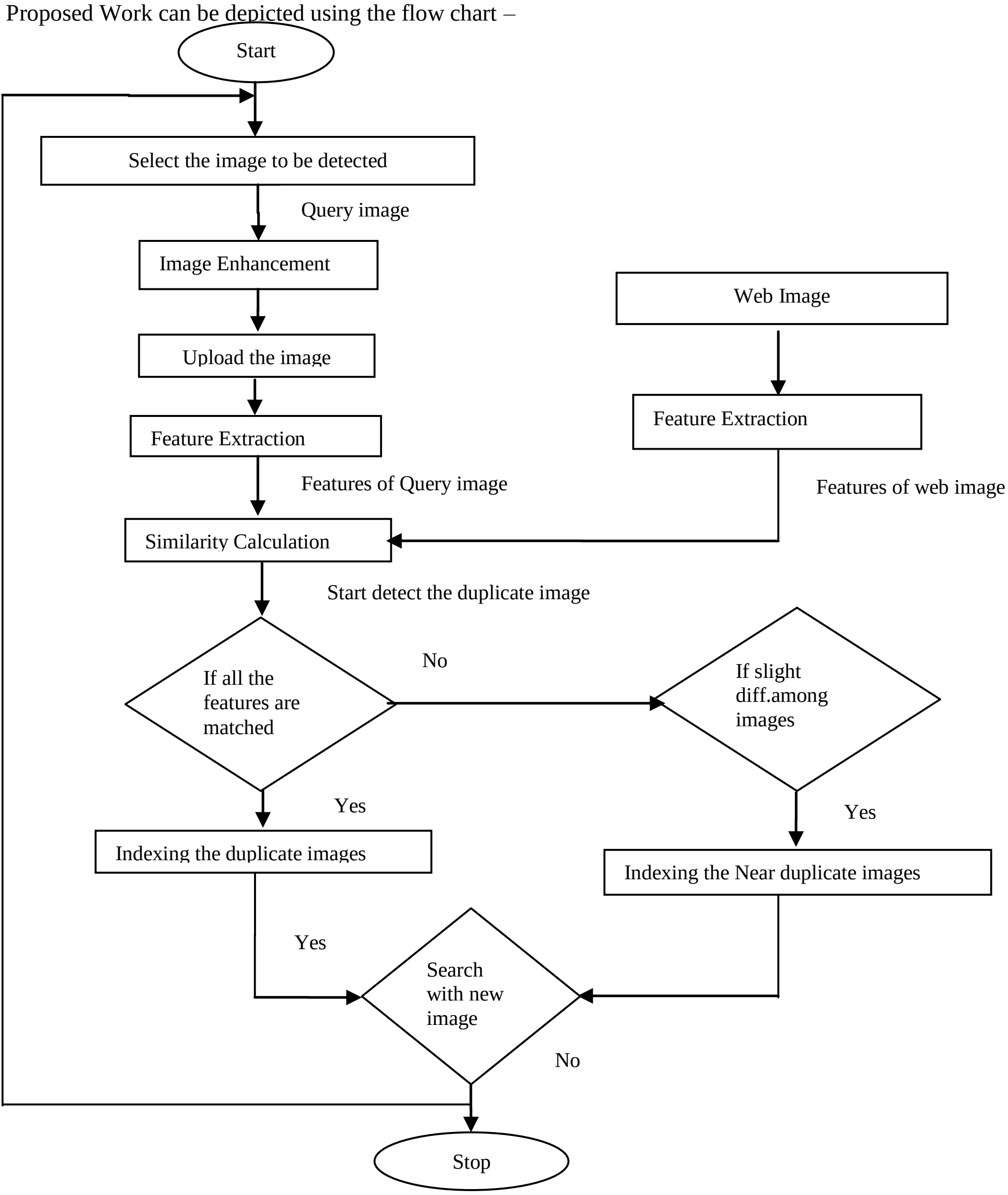

Fig.4 Flowchart of Proposed System

5. EXPERIMENTAL RESULTS

This paper is proposed mainly for detecting the near duplicate images by indexing the image using SURF and Min hash algorithm. Once the features are extracted, then similarity is calculated. Finally indexing is done. These indexing are used to detect the near duplicate images corresponding to the query image. This is done by computing the similarities between the query image and the web image. The small distance between the query image and the web image indicates that the web image is relevant to query image.

In this paper, Real images were used. Actually this dataset contains real camera photos taken directly from a real user's photo gallery. This collection consists of different types of near duplicates and exact duplicate images. Initially very little images were taken by the proposed system and later the entire dataset is used.

Figure 6 show uploading of the enhanced query image that is unfamiliar sample results for web search and figure 7 refers the identification of exact duplicate and near duplicate images

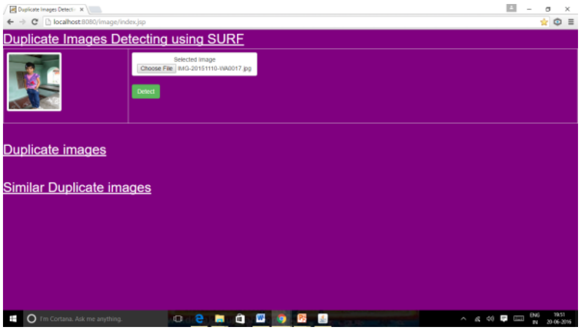
Duplicate Images Detecting using SURF
Selected Image
Choose File IMG-20151110-WA0017.jpg
Detect
Duplicate images
Similar Duplicate images

Fig.5 Upload the unfamiliar query image

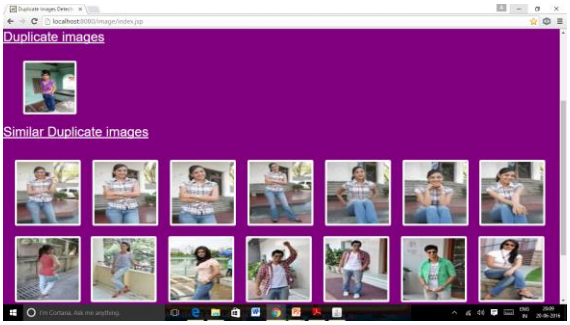

Fig. 6 Indexing the Near Duplicate images

The following figures show another familiar sample results for web search

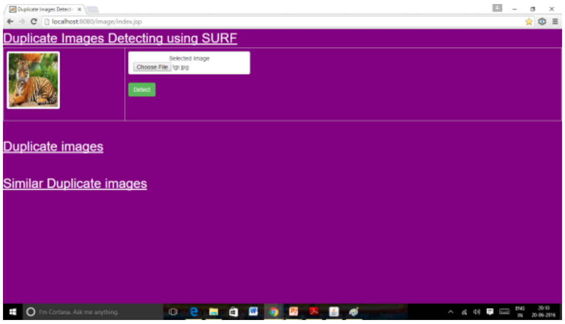

Fig. 7 Upload the familiar query image

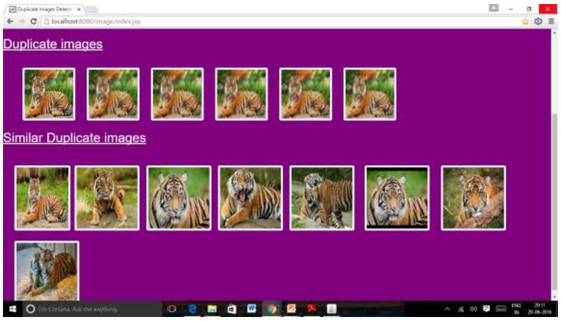

Fig.8 Indexing the Near duplicate and exact image detection

5. CONCLUSION

The overall work here is detecting near duplicate images and indexing those images from a collection of dataset. In this paper, methodology is presented to perform indexing of near-duplicate images. Initially, the query is passed by the user to the search engine and the search engine results set of query related images. These images contain duplicate as well as near-duplicate images. Here we concentrate in detecting near-duplicate images and index those images. This is done using following steps – initially enhance the user query image and then features are extracted. After features are extracted similarity is measured and finally indexing the near duplicate images. This results in indexing of images. We conclude that our indexing approach is highly effective for collections of up to a few hundred thousand images.

References

[1]Yaodong He, Zao Jiang, Bing Liu and Hong Zhao**,** content-Based Indexing and Retrieval Method of Chinese Document Images,Shenyang,China

[2] Bart Thomee, Mark J. Huiskes, Erwin M. Bakker, Michael S. LewAN EVALUATION OF CONTENT-BASED DUPLICATE IMAGE DETECTION METHODS FORWEB SEARCH

[3] DharmendraPatidar, Nitin Jain, AshishParikhPerformance Analysis of Artificial Neural Networkand K Nearest Neighbors Image ClassificationTechniques with Wavelet features, 2014 IEEE International Conference on Computer Communication and Systems(ICCCS '14),

Feb 20-21, 2014, Chennai, ThlDIA

[4] WANG MuNi, ZHANG WeiFeng,, ZHANG YingZhou, JI XiaoHua,**Detecting Image Spam Based on Cross Entropy,** 2011 Eighth Web Information Systems and Applications Conference

[5]Changjing Shang and Dave Barnes,Support Vector Machine-BasedClassification of Rock Texture ImagesAided by Efficient Feature Selection, WCCI 2012 IEEE World Congress on Computational IntelligenceJune, 10-15, 2012 - Brisbane, Australia

[6]Imran Ahmad,MuhammadTalalIbrahim,**Image Classification and Retrieval using Correlation,** Proceedings of the 3rd Canadian Conference on Computer and Robot Vision (CRV'06)0-7695-2542-3/06 $20.00 © 2006 IEEE

[7] Chi-Man Pun and Moon-ChuenLee,Extraction of Shift Invariant Wavelet Featuresfor Classification of Images with
Different Sizes, 2009 International Conference on Environmental Science and Information Application Technology

[8] Wei-jiu Zhang1, Li Mao2, Wen-bo Xu2,**Automatic Image Classification Using the Classification Ant-Colony Algorithm,** 2009 International Conference on Environmental Science and Information Application Technology

[9]Karol Wnukowicz 1 , GrzegorzGaliński 1, Ruben Tous 2,Still Image Copy Detection Algorithm Robust to Basic Image Modifications, 50th International Symposium ELMAR-2008, 10-12 September 2008, Zadar, Croatia

[10] OndrejChum,JamesPhilbin,MichaelIsard,AndrewZisserman,Scalable Near Identical Image and Shot Detection University of Oxford,Silicon Valley

[11] A. Joly, O. Buisson, and C. Frélicot. Content-based copy detection using distortion-based probabilistic similarity search. IEEE Transactions on Multimedia, to appear, 2007.

[12] Ryuji Funayama,Hiromichi,Yanagihara,Luc Van Gool,TinneTuytelaars,HerbertBay,"ROBUST INTEREST POINT DETECTOR AND DESCRIPTOR"",published 2009-09-24.

[13] H. Bay, *et al.*, "SURF: Speeded up robust features," in *Computer Vision - Eccv2006 ,Pt1, Proceedings*. vol. 3951, A. Leonardis, *et al.*, Eds., ed Berlin: Springer-Verlag Berlin,2006, pp. 404-417.

[14]X. Anqi and G. Dudek, "A vision-based boundary following framework for aerialvehicles," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ InternationalConference on*, 2010, pp. 81-86.

[15].M. Kruis, "Human pose recognition using neural networks, synthetic models, andmodern features," Master Of Science Elecrtical Engineering, Oklahoma StateUniversity, Stillwater, OK, 2010.