**I J C T A**, 9(2) 2016, pp. 1409-1420


# Automatic scraper of celebrity images from heterogeneous websites based on face recognition and sorting for profiling

**Sneha K.* and N. Lalithamani****

**ABSTRACT**

Now a days it has become trend to follow all the celebrities as we consider them as our role models. So instead of searching the images of various celebrities in different websites we can find them in a single website by sorting all the images. Reliable database of images is essential for any image recognition system. Through Internet we find all the required images. These images will serve as samples for automatic recognition system. With these images we do face detection, face recognition, face sorting using various techniques like local binary patterns, haar cascades. We make an overall analysis of the detector. Using opencv we detect and recognize images. Similarity matching is done to check how the images are related to each other. Collection of images is based on user defined templates, which are in web browser environment. With the help of this system one can give their requirement and the image of celebrity is displayed based on that.

***Keywords:*** Celebrity Images; Image recognition, Image sorting.

## 1. INTRODUCTION

The World Wide Web gives the information of anything we need these days. It acts as a collection of all data together placed in single unit. Anything you name is available on internet these days. It is can be said as net of piled data. Not only sharing information across the globe it is very useful for developing many product and services. Since everyone now a days are using many websites even for extracting the contents from the given websites. Even though the structure of the data present in the website is very similar, every page is different and the structure of one webpage differs from the other webpage. The systematical browser of the world wide web is termed as Web crawler. [9]. It is used for downloading all the links in the webpage and even all the contents present in the webpage. Extraction of the specific data from an web page is called Web Scrapping. In different kind of websites there are different types of templates with which they have been build. Example of this kind is amazon store, where the product details of the amazon page is always having same structure, and differs in the products that have been upadated consequently. Similarily alike amazon the structure of online shopping may be same but differs in the way in which they have been presented.

After the crawling of images these images are taken for face detection, face recognition and sorting them which is termed as similarity matching. The algorithms used for face detection, face recognition are being discussed in detail. After the face recognition these images are being sorted to check the similarity among images.

* PG Student Dept of Computer Science and Engineering, Amrita School of Engineering, Coimbatore Amrita Vishwa Vidhyapeetham Amrita University India, *Email: snehakandacharam@gmail.com*

** Assistant Professor (SG) Dept of Computer Science and Engineering, Amrita School of Engineering, Coimbatore Amrita Vishwa Vidhyapeetham Amrita University, India, *Email: n_lalitha@cb.amrita.edu*

## 2. RELATED WORK

There are huge amount of papers available for crawling of webpages[14]. The extraction rule of each and every webpage is different and follows a certain pattern which helps us to extract any content[15]. The content which is required to extract can be of HTML element[7]. Web wrapper is used for downloading webpages from source code of any webpage and extracts its content. The standard used for taking any content from the data can be of HTML or Xpath [1][8]. Many people use this in their works. CSS selectors are also used for extracting the data which we need.

This can also be done using beautiful soup which is an python library for pulling data from HTML and XML[11][12]. It works as an appropriate parser which is used in so many ways like navigating, searching, and modifying the parse tree. It saves us from a lot of coding problems.

Face recognition has its own beauty in attracting people working in image processing side and people who are doing research in this are not only engineers but also neuroscientists, since it has many practical implementions even in other fields like control systems.[19]. First and foremost face detection is an important part of face recognition as it can be taken as first step towards facial related task. As we know face detection is not a standard set of process since it has lot of variations of image appearance, such as pose variation (front, non-front), occlusion, image orientation, illuminating condition and facial expressions [20].

Digital images and video processing are arriving a lot these days and plays an immense role in world of multimedia Information [2]. The human face is considered as an essential object in an image or video. Detection of human faces in an image and extracting the facial features in it is necessary which is present in different set of applications, such as human face recognition, human computer interfacing, video-conferencing, etc.

A problem statement of machine recognition of faces can be briefed as follows: having video image of a scene, identification or verification of one or more persons in the scene using the faces available in the

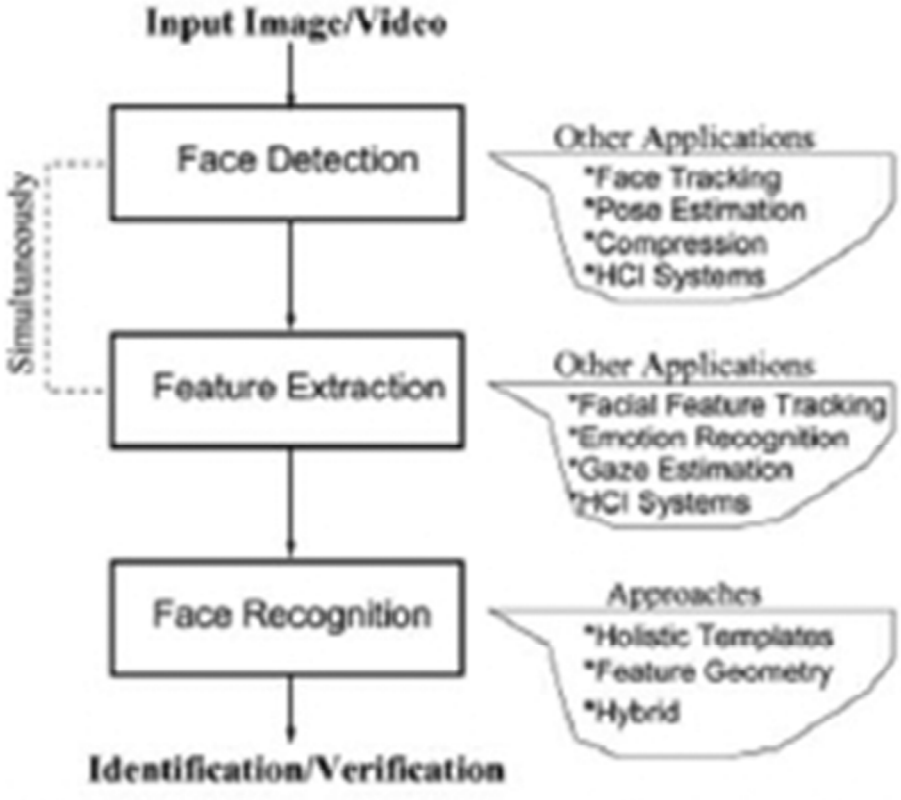



**Figure 1: Configuration of generic face recognition system**

databases. The other related information like race, age, gender, facial expression, or speech is used in reducing the burden of the searching process (enhancing recognition). The solution here is as follows: segmentation of faces from cluttered scenes, certain features extracted from face, recognition or verification (Fig 1)[16]. For identifying, the input to the system is an random face, and the system sends back the determined identity from a database of known faces of people , and in verification problems, the system needs to say whether the claimed face is correct or will that reject the face.

Figure 1 [16] Face detection is also helpful in challenges like patten classification and there is so much of learning as well as wide range of research oriented work . When an unknown or filtered image is taken as input to a pattern classifier, the dimensions of the feature that are being considered are very large (i.e., the number of pixels present in training images). The classification of face and non-face images is decided by multimodal distribution functions and effective decision boundaries which are nonlinear in the image space. To be a good classifier, either it should be able to roll back from a finest number of training samples or be correct while dealing with a very huge number of high-dimensional training samples.

Finding similarity between visual data is necessary in many computer vision tasks that includes object detection and recognition, action recognition, texture classification, data retrieval etc. Methods available for these kind of tasks are based on an image which represents some global or local image properties, and compares those properties with them using some similarity measure.

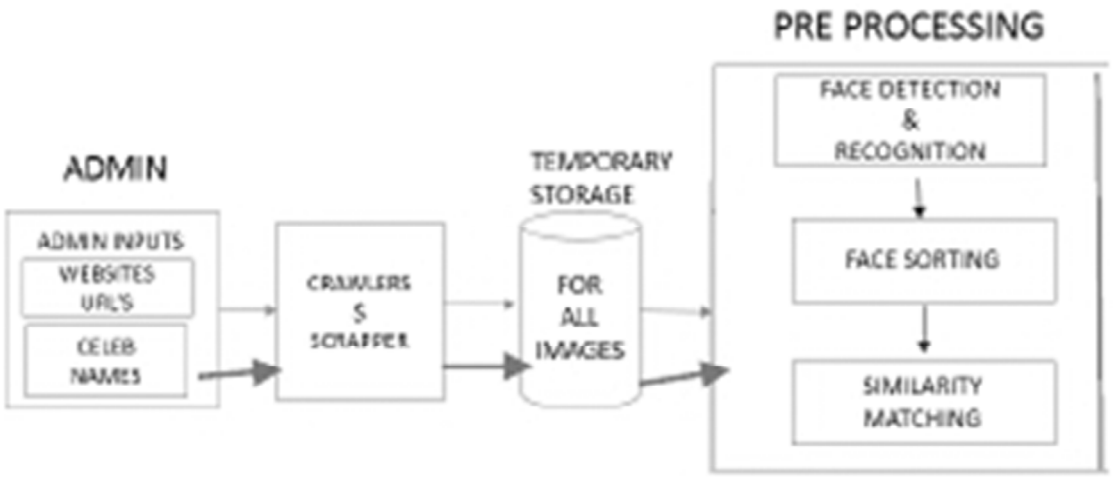
## 3. AUTOMATIC SCRAPPING AND PROFILING

The main motivation for this work is development of website which consists of all the celebrity images by doing face detection, face recognition and sorting them.

This system will be more useful for people who want to follow present trend as well as they can follow them as per their requirements. First of all there are two Modules. First Module is Admin module and another is User Module.

In Admin Module there are websites or URLs present. Along with them we give celebrity names as inputs. We develop a crawler which stores all the images of celebrities and stores it in a temporary database. Then the preprocessing of the images takes place where it consists of another three sub modules. Face detection and recognition, face sorting and similarity matching.

After the pre-processing the filtered images are taken and tested for image processing related tasks like lipstick and eye liner, eye wear information , dress information and gender information. After all this processing it is given to the database indexer. All the filtered images and database indexer combined together and form a Filtering and Image fetching module where all the images will be stored with the required



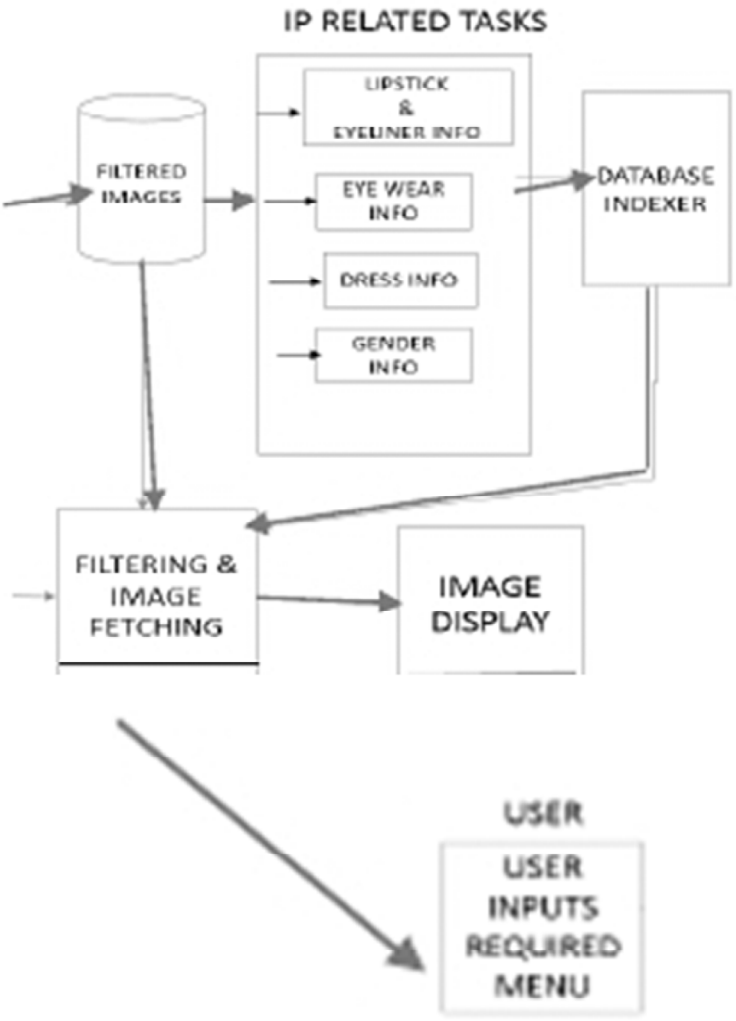



**Figure 2: Architecture of the Automatic scrapper and profiling system**

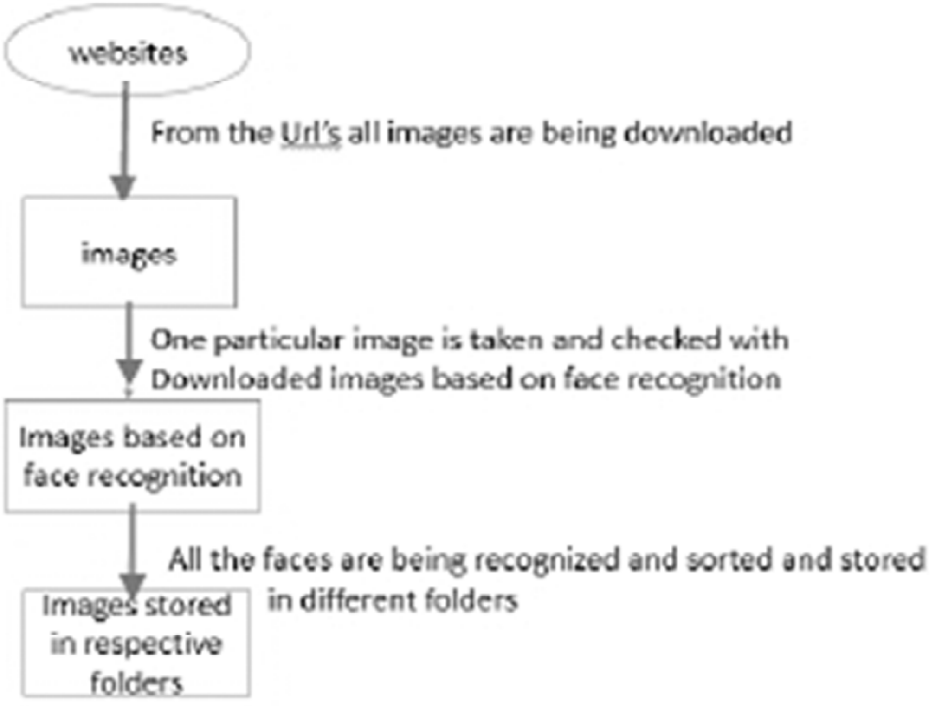



**Figure 3: Flow diagram of the Automatic scrapper and profiling.**

content. Then the user can request for any kind of image data and can find the image in image display. The work flow is as follows

From Fig 3 we get that Images are being downloaded using python library called Beautiful Soup. After downloading all the images from different websites one particular image is taken and check with all the other images. These images are sorted according to face recognition. This face recognition is done using opencv python. One particular image is saved in one folder and next checks for same image and saves into that folder. It removes all the junk images as well.

## 4. FACE DETECTION AND RECOGNITION

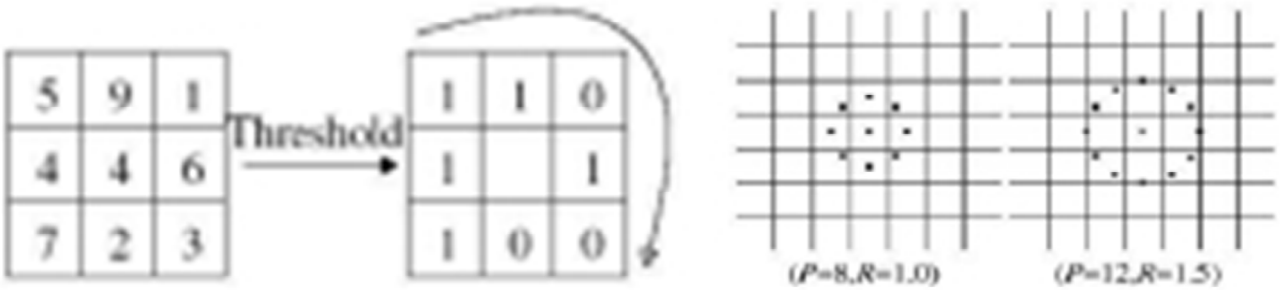
Face recognition has its own tale in attracting lot of people and its research has been tremendously increased by engineers, since it has many powerful applications in computer vision and automatic access control system. Especially, face detection is surely an necessity of face recognition as it is considered as the first step of automatic face recognition[26]. However, face detection is just not a single task but has lot of combinations like it has lots of variations of image appearance, such as pose variation (front, non-front), image orientation, illuminating condition and facial expression.

Automatic facial expression recognition involves two vital important things to be considered: facial feature representation and classifier design[27]. Facial feature representation is for deriving a set of features from original face images which actually minimizes within available class variations of expressions[25]. There are two main methods used for extracting facial features: geometric feature-based methods and appearance-based methods.

Local Binary Patterns (LBP) is a genuine low-cost discriminative feature used for human facial expression recognition[4]. It was originally developed for texture analysis. A facial image is divided into a set of small regions where LBP histograms are generated automatically and formed into a spatially enhanced feature histogram[4][21][23]. The simple LBP feature can be easily got from an raw image which helps in retaining complete facial information which is present in a compact representation[6].

The original LBP operator was introduced by Ojala et al . The operator labels the pixels of an image by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number (see left of Fig 4 for an illustration)[4]. The histogram of the labels are actually used as a texture descriptor.

From the above figure we can understand how an LBP operator working. Figure 4[4]. Left: Is The basic LBP operator and Right: Two examples of the extended LBP : a circular (8, 1) neighborhood, and a circular (12, 1.5) neighborhood. The only limitation of the basic LBP operator is its 3×3 neighborhood that cannot capture features with large scale structures [24][4]. Hence the operator was again used to check with different sizes. Using circular neighborhoods and interpolating the bilinearly pixel values is used for allowing any




**Binary – 11010011**
**Decimal - 211**

**(P = 8, R = 1)**
**(P = 12, R = 1.5)**

**Figure 4: Texture descriptor**

radius or any number of pixels in the neighborhood. Examples of the extended LBP are shown in right of Figure 4, where (P, R) denotes P sampling points on a circle of radius of R. Further extension of LBP is to use uniform patterns [4].

A Local Binary Pattern is uniform if it has at most two bitwise transitions from 0 to 1 or vice versa when the binary string is taken as circular. For example, 00000000, 001110000 and 11100001 are uniform patterns. It is observed that uniform patterns account for nearly 90% of all patterns in the (8, 1) neighborhood and for about 70% in the (16, 2) neighborhood in texture images [18][4].

Local binary patterns can also found by using opencv in pyhton. With help of this face detection has been done. It can detect more than one person in the image.

```
File  Edit  Format  Run  Options  Window  Help
from sklearn.svm import LinearSVC
from imutils import paths
import argparse
import cv2
import os

ap = argparse.ArgumentParser()
ap.add_argument("-t", "--training", required=True,
        help="path to the training images")
ap.add_argument("-e", "--testing", required=True,
        help="path to the tesitng images")
args = vars(ap.parse_args())

desc = LocalBinaryPatterns(24, 8)
data = []
labels = []

for imagePath in paths.list_images(args["training"]):
        image = cv2.imread(imagePath)
        gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
        hist = desc.describe(gray)

        labels.append(os.path.split(os.path.dirname(imagePath))[-1])
        data.append(hist)
```

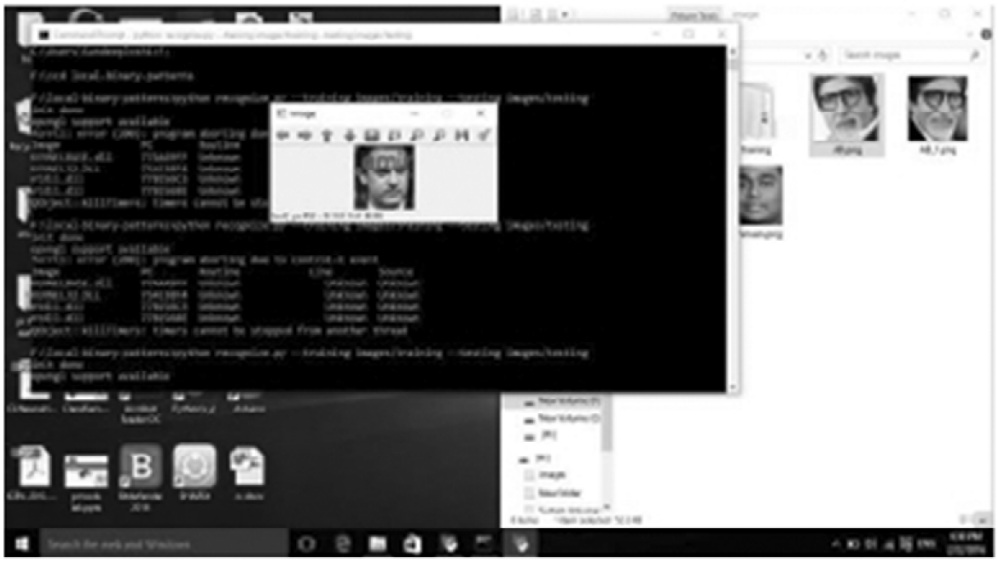


**Figure 5: Here face recognition is being done using opencv**

## 5. HAAR CASCADE DETECTION

Haar cascade detection here plays an important role for face detection. Here in the haar cascade detection it has so many complex classifiers which focuses on main regions of the images[3]. Usually with the help of this we can easily determine where an particular object might occur [17, 8, 1]. The complexity of these main parts of image is very high. The measure that has been taken in this whole process is the false negative nature which has to be relatively very low. Even if we consider for an object as an instance the main part of the image is need to be checked first.

The normal functions that was used by Papageorgiou et al is initially considered as important haar based function. Basically there are three features that are being taken into consideration. The two-rectangle feature value is the difference between the sum of the pixels added from the two rectangular regions. The regions are having the similar size, shape and which are horizontally or vertically adjacent (see Figure 6). When an three-rectangle feature therefore calculates the sum wof two outside rectangles which is again subtracted from the sum formed in the center of the rectangle. Hence a four-rectangle feature calculates the difference among diagonal pairs of rectangles that are taken into account by the new filter.

In the Figure 6 [13] is an example which helps in obtaining the detection features of the image. The sum of the pixels is calculated by subtracting it from region of grey rectangles to the region of white rectangle.. In (A) and (B) two rectangle feature is given. (C) is a three-rectangle feature, and (D) shows a four-rectangle feature. Haar cascade detection hence plays an vital role in face detection because of its immense behavior of detection of objects focused on important areas. Below we consider an integral image which is the alternate representation for any image and find out the haar cascade detection to it.

### 5.1. Integral Image

The integral image is an alternative representation of any image which helps us in considering the rectangle features of the image. The integral image at particular location s,t contains the addition of the pixels above and left of s,t , inclusive[4]:

$$ii(s,\ t) = \sum s' \leq\ s,\ t' \leq\ t\ i(s',t')$$

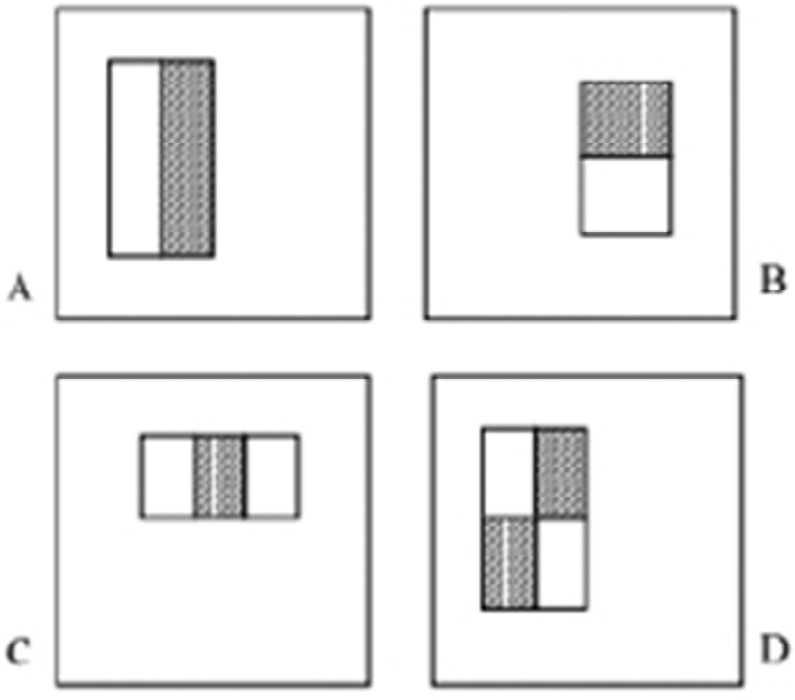



**Figure 6: An detection window**

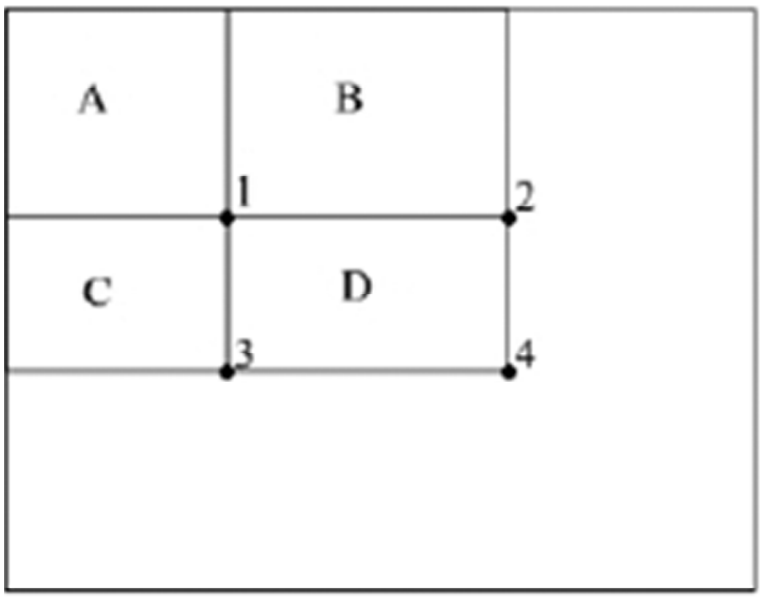



**Figure 7: To compute sum of pixels**

From Figure 7. [4]. The addition of the pixels can be calculated with the references of four arrays. The actual value of the integral image taken at location 1 is the addition of the pixels in rectangle A. The value of the location 2 is sum of A,B , at location 3 is sum of A,C and at location 4 is sum of A,B,C,D. The total sum within the image is calculated as 4 + 1 –(2 +3). where ii(s,t) is the integral image and i(s,t) is the original image.

The following pair of recurrences is used:

$$S(s, t) = S(s, t-1) + i(s, t) \quad (1)$$

$$ii(s, t) = ii(s-1, t) + S(s, t) \quad (2)$$

(where $S(s, t)$ is cumulative row sum, $S(s, -1) = 0$ and $ii(-1, t) = 0$) the integral image can be immediately formed with one single pass of the original image. With the help of integral image any calculation of the rectangular areas of image is done with four references of arrays (see Figure 7). The differential part of the two rectangles can be calculated with references of four arrays. The two-rectangle features given above has adjacent rectangular sums they are calculated in six array references, when it is three rectangle features eight references of arrays is taken, and nine in the case four-rectangle features. So intergral image is useful in obtaining an accurate result of detecting images.

## 6. SIMILARITY MATCHING

Comparison among the images is the first and foremost operation done for finding the similarity among the pictures or images . The similarity matching of two images can be present in different hierarchical levels from pixel-by-pixel level, feature space level, object level, and semantic level. In every system that we compare the pixel to pixel comparison alone is not sufficient: when we consider the similarity not only the correlation part but also it should differ in the semantic part of the images[5]. These matching techniques are created mainly for object recognition under several distortable conditions and when we take similarity measures, are used when applications like image databases are taken. Similarity matching and dissimilarity usually doesn't depend on same set of features that are taken for testing them and also differs in the approaches.

Similarity checking of images plays prominent role in theories of human knowledge representation, behavior, and problem solving and when it comes to human perception and cognition nothing else can beat

them. As Tversky describes the similarity concept as "an organizing principle by which individuals classify objects, form concepts, and make generalizations." Retrieval of images by similarity matching [10].

The following three main points are required for any similarity matching technique to be followed:

- Feature extraction or specific signatures from the images, and an unique representation for the images as well as the data which is precalculated has to be stored separately.
- When similarity measure is taken it should have meaningful information of the image taken and when it is matched with databases present it should have same perspective as the other images in that database.
- Usually the user when considers about finding the similarity it should be applied for retrieval of similar images or pictures, how the images are presented , this helps in the feedback of images of similar kind.

Similarity Matching can be found using open cv where we can find the similarity between images and an histogram which shows the similarity.

Figure 8. shows the similarity between two same images but from different website.

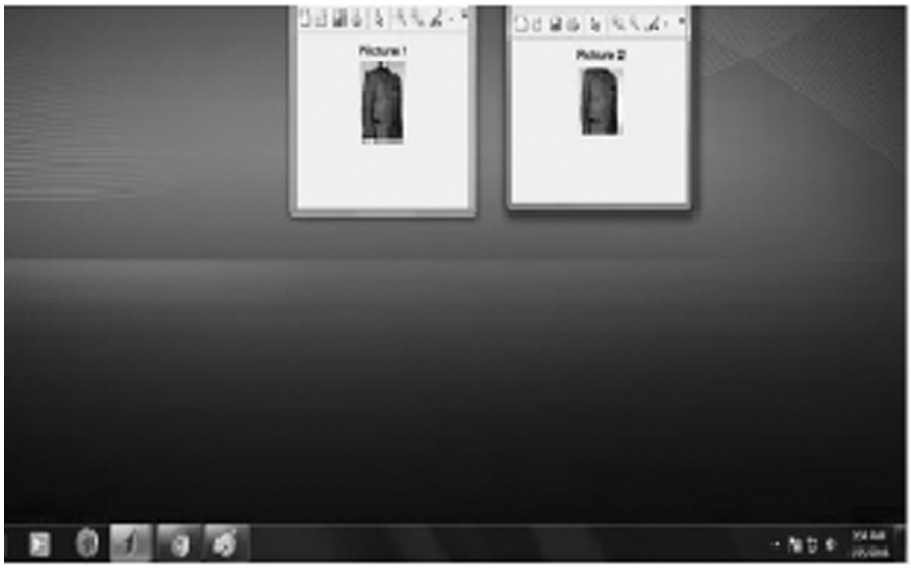


**Figure 8: Similarity Matching**

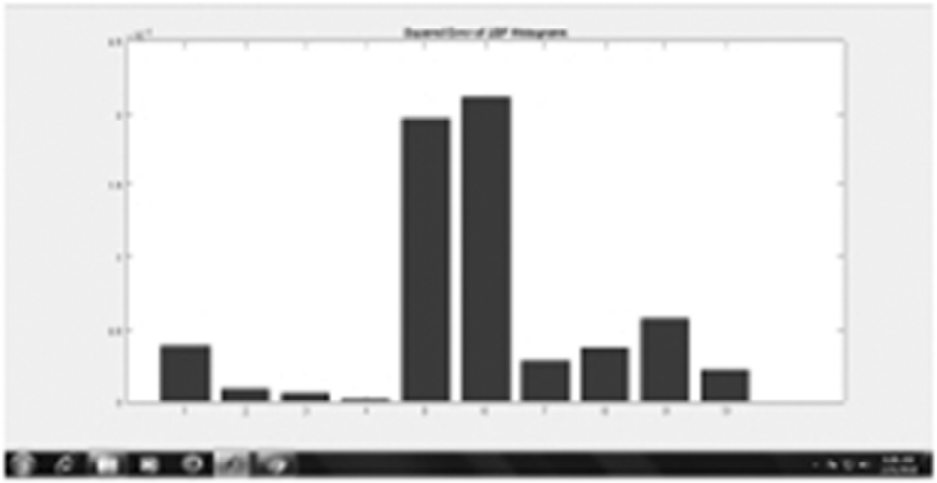


**Figure 9: Shows the histogram of the images which is having pixels range from 0 to 255 [22]**

## 7. DATABASES

For experimental purposes , we have taken the below websites for downloading the images and have performed face detection , face recognition and similarity matching and found significant results.

http://www.behindwoods.com/index.html

http://www.missmalini.com/

http://www.harpersbazaar.com

http://www.highheelconfidential.com/

http://fashionista.com/2015/02/most-influential-style-bloggers-2015

## 8. CONCLUSION AND FUTURE ENHANCEMENT

Face recognition has got its importance on a wide range in Image Processing. It is an part of biometrics which helps in recognizing human faces in digital images. Similarity matching also plays a vital role where we can know how similar those images are when they are matched.

The proposed system helps to crawl the images and perform face detection, recognition and similarity matching which helps to develop any facial recognition system using the algorithms Local Binary Patterns and Haar cascades. Using these we will be able to recognize any face in digital images and check similarity.

## ACKNOWLEDGMENT

I thank the great Almighty and my parents for showering their blessings on me and helping my efforts turn into this fruitful contribution. I express my sincere gratitude towards my Guide Ms. N. Lalithamani, Assistant Professor (SG), Department of CSE for giving me an opportunity to work on this project and guiding me through the right track which helped in obtaining the aspired goals of the project.

## REFERENCES

[1] Michal vagac , Miroslav Melichercik , Matus Marko (2015). Crawling images with web browser support : IEEE 13th International Scientific Conference on Informatics. Informatics’2015 . November 18-20 . poprad. Slovakia

[2] Junghoo Cho , Hector Garcia-Molina, Lawrence Page (2012). Reprint of: Efficient crawling through URL ordering : Elsevier Journal , Computer Networks 26( 2012 ) 3849-3858.

[3] Paul Voila , Michael Jones . Rapid Object Detection using a Boosted Cascade of Simple Features. Accepted Conference on Computer Vision and Pattern Recognition 2001.

[4] Caifeng Chan , Shogang Hong , Robust Facial Recognition using Local Binary Patterns . Image Processing, 2005. ICIP 2005. IEEE International Conference on (Volume:2 )

[5] Eli shechtman , Michal Irani. Matching Self local similarities across Image and Videos. http://www.wisdom.weizmann.ac.il/~vision/VideoAnalysis/Demos/Self Similarities

[6] V. Ferrari, T. Tuytelaars, and L. V. Gool. Object detection by contour segment networks. In ECCV, May 2006.

[7] R. Baumgartner , S. Flesca and G. Gottlob, Visual web information extraction with lixto , In Proceedings of the 27th International Conference on Very Large Data Bases, VLDB 01, pages 119-128, San Francisco, CA, USA: Morgan Kaufmann publishers Inc, 2001.

[8] V. Crescenzi , P. Merialdo , and D. Qui, Alfred : Crowd assisted data extraction, In Proceedings of the 22nd International Conference on World Wide Web Companion , WWW 13 Companion, pages 297-300, Republic and Canton of Geneva, Switzerland, 2013.

[9] T.Furche , G. Gottlob, G.Grasso , C. Schallhart and A. Sellers, Oxpath: A language for scalable data extraction , automation, and crawling on the deep web, The VDLB Journal , 22(1):47-72, Feb. 2013.

[10] R. Brooks, T. Arbel, D. Precup, Anytime similarity measures for faster alignment, Computer Vision and Image Understanding 110 (3) (2008) 378–389.

[11] T.Grigalis, Towards web-scale structured web data extraction , In proceedings of the Sixth ACM International Conference on Web Search and Data Mining , WSDM 13, pages 753-758, New York, NY, USA:ACM, 2013.

[12] K. Kanaoka , Y. Fujii , M. Toyama , Ducky : a data extraction system for various structured web documents , In Proceedings of the 18th International Database Engineering & Applications Symposium, IDEAS ‘ 14. Pages 342-347, New York , NY, USA: ACM, 2014

[13] N. Kushmerick , Wrapper induction : Efficiency and expressiveness, Artificial Intelligence , Vol 118, Issue 1-2 , pages 15-68. Essex, UK: Elsevier Science Publishers Ltd., 2000.

[14] M. Tlo and M. Suzuki, Design and implementation of a facility for wandering and manipulating the structure of on-line data, In Information Science and Applications (ICISA), 2011 International Conference on, pages 1-8, April 2011.

[15] M. Geel , T. Church and M.C . Norrie, Sift : An end –user tool for gathering web content on the go , In Proceedings of the 2012 ACM Symposium on Document Engineering , DocEng 12, pages 181-190, New York, NY, USA: ACM, 2012.

[16] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In International Conference on Computer Vision, 1998.

[17] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Patt. Anal. Mach. Intell., 20(11):1254–1259, November 1998

[18] H. Greenspan, S. Belongie, R. Gooodman, P. Perona, S. Rakshit, and C. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1994.

[19] Edgar Osuna, Robert Freund, and Federico Girosi. Training support vector machines: an application to face detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[20] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. to appear in PAMI, 2006.

[21] Di Huang , Caifeng Shan, Mohsen Ardabilian. Local Binary Patterns and Its Application to Facial Image Analysis: A Survey. IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 41, no. 6, november 2011 .

[22] Chi-Ho Chan, Josef Kittler, Kieron Messer, Multi scale local Binary Pattern Histograms for Face Recognition. Advances in Biometrics Volume 4642 of the series Lecture Notes in Computer Science pp 809-818.

[23] G. Zhang, X. Huang, S. Z. Li, Y. Wang, and X. Wu. Boosting Local Binary Pattern (LBP)-based face recognition, volume 3338. Springer Berlin / Heidelberg, 2004.

[24] S. Yan, S. Shan, X. Chen, and W. Gao. Locally assembled binary (lab) feature with feature-centric cascade for fast and accurate face detection. 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008.

[25] B. Fröba and A. Ernst. Face detection with the modified census transform. In Sixth IEEE Int. Conference on Automatic Face and Gesture Recognition, pages 91–96, 2004

[26] H. Jin, Q. Liu, H. Lu, and X. Tong. Face detection using improved lbp under bayesian framework. In Third Int. Conference on Image and Graphics, pages 306–309, 2004.

[27] M. Heikkilä, M. Pietikäinen, and C. Schmid. Description of interest regions with local binary patterns. Pattern Recognition, 42(3):425– 436, 2009.