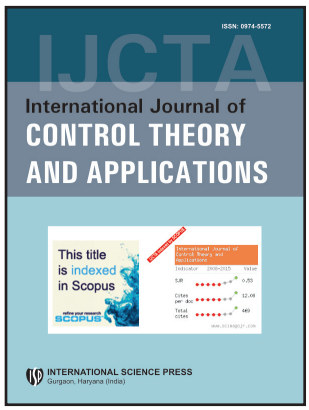


International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 12 • 2017

Web Image Based Auto Clustering of Cartoons using Contour Filter and Refine Technique

C. Menaka[a] and N. Nagadeepa[b]

[a]*M.C.A., M.Phil. Assistant Professor, Research Scholar, Bharathiar University, Coimbatore*
[b]*Principal, Karur Velalar College of Arts and Science for Women, Karur*

Abstract: Downloading accurate images using internet is a difficult task. The classification of images is a challenging task in web mining research. Number of techniques is available to classify the images in the process of web image classification. In this work, technique considers two HTML tags namely alt and src for extracting images. In a group of web pages these two tags are taken into account to download the images. Mainly this approach considers the cartoon image category for example the character like Dora, Pokeman, Disney and cartoon web link for the extraction and storing. Three different modules are used here. LTP (Lexical Tag Parsing) technique is applied here to parse the given tags. Images are clustered and stored in their respective folders as per the category after clustering process. CFR (Contour Filter and Refine) algorithm is used here to refine the images for storing. MIA (Multilevel Image Annotation) technique is applied here to give annotation for all images which is in the cluster for best retrieval. Finally based upon the given input as image resultant image can be searched from various available clusters and return to the user along with its detailed description.

Keywords: Image clustering, LTP, MIA, CFR, Image annotation, SIC.

1. INTRODUCTION

Web mining is the processing of information in web for any application. Text extraction is not so difficult but Image processing finds difficult task in web mining since it consists of various file formats and because of its storage capacity. Web image classification is a most challenging task in image processing. Most of the web mining algorithms use decision making technique. Here in this work, analyzing the web page of any cartoon websites for the classification of cartoon images.

Initially, cartoon website is considered for downloading. All the pages in the website will be downloaded. Then Lexical Tag Parsing is implemented to parse all the tags in the website. On parsing each tag can analyze the image and its text which will be used for classification of image. This approach is taken because: (1) there are two possibly conflicting sources of signals for a web image: the visual signal and the text signal and there are no simple way to combine or reconcile them; and (2) text signals are more reliable and can be better captured by

our novel techniques. In this framework, extract two types of context of a given image: the HTML <IMG> tag context which consist of the image and ALT tag which consists of annotation. Most of the web page includes the images using <IMG> tag only. So if we are parsing [4] this tag, it will be easy to download the images from the web page. After downloading it analyzing the annotation of the image using <ALT> tag. A note on image will be definitely stored in the ALT tag. So on probing this tag, then easily find the annotation of the image which says the category of image.

There are many efforts attempted to make image classification by analyzing the properties of the online image. It helps user to group the image of same family. Here image family means any properties like color, label, image source etc. In this example, we can download the images in number of cartoon websites and also can classify and store for later references. So if it need to make an offline image search to make use of this application. We can make a classification depending upon the cartoon's characters like Tom as one classification and Jerry as another like wise.

So this will enable the users to make an image search more easily as they have much classification [6] of cartoon images in hand. This kind of image classification can be made for any context depending on the application.

After parsing the images, all the images can be downloaded and its relevant text. So after downloading this, it will find the similarity between the images using their relevant texts.

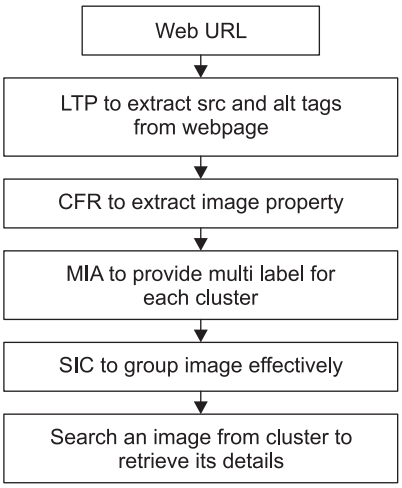


Figure 1: Flow of filter and refine technique

While classification of images is made, CFR algorithm is used to create a more accurate classification. This will make the image processing on each image to find the similarity of images. Sometimes more than one image will have same ALT tag text and some alt tag content may be repeated. In order to handle this problem MIA algorithm is used. Like that all images will be downloaded with its ALT text and compared with each other for its similarity [2]. If it's similar, then classified as same family or else it will be alienated. This continues until all the images are clustered effectively. Finally Synergic Image Clustering (SIC) is performed which to make more accurate classification of each image in their appropriate folder. There are number of existing systems for image retrieval and analysis from web.

2. STATEMENT OF PROBLEM

Web based image classification technique to classify cartoon characters from famous cartoon websites. Image classification itself a sensitive job in image processing whereas online image processing double the complication.

Moreover while downloading images the web page may give irrelevant data [3], Noisy data, Redundant data. If downloading capacity increases bandwidth also increases. While downloading images size of bytes may acquire more. It consumes more data .If the web page consists of 30% percent of irrelevant data it also considers the amount of time which is spent for downloading that. Moreover in downloading image can also be displayed along with the other objects or it may embossing on the object. For example, the name of the image is given as DORA it may consider the cartoon character, name of the person whose name is Dora, Dora image on other object like in *t*-shirt.

3. EXISTING WORKS

For the above mentioned problem few strategies have been used. D.S. Xia [1] defined clustering web image by their low-level visual features. Here, the low-level visual features denote the features wrench pointblank from the image data as Sift features, Giber wavelet features and so on. For example, Gordon and Yang extracted feature vectors to represent the images and then perform the clustering algorithms on the feature representations. Because there is semantic gap between the low-level visual features [1][2] of images and high-level intellect of users, those methods can't obtain the expected cluster results.

Since the web image is frequently encircled by some semantic related texts and researchers began to extract the textual features from the encircled texts. The textual feature is always critical words or text tags. It has been some excellent work proposed in the last few years. In particular, Cai proposed a graph model to co cluster the represented features and their web images. Their algorithm separates the visual features and textual features [1] [2] was executed in a two-step process. All the images were firstly clustered into different semantic groups by employing the link features and textual. This step was after pursued by visual feature based clustering of images in every symbolism group. Then identifies the problems in the results of the first step may be augmented in the pursuing process, Gao et al proposed another clustering framework to simultaneously integrate both the textual features and visual for clustering. In their work, spectral clustering [3] was applied and iteratively using half certain programming to cluster. Similarly, Rege proposed another clustering algorithm persistent Isoperimetric High Order Co clustering (CIHC) [3] with using isoperimetric concept to concatenate both textual features and visual at same time.

Since the graph models are widely used in the Web image clustering, hyper graph [8][9] models are also introduced to cluster the Web images. For example, Zhou and his team proposed a hyper graph partitioning model. Based on the hyper graph models, Wu proposed a clustering method based on the random walk and that received great performance especially on large dataset.

4. PROPOSED METHODOLOGY

Three different techniques have been proposed.

Lexical tag parsing technique which parses the web page for image source. After detecting the image source we are extracting the annotation of each image using ALT tag. After downloading these two information classification of the cartoon images by use of contour filter and refine technique.

Multilevel image annotation technique is used to classify the images more accurately by detailed annotation of each image.

Finally the synergic image clustering is made to combine the output into a complete package of fulfilled classification with noise eliminations.

Shukui Bo [2] defined as Mean shift algorithm is a nonparametric technique for estimation of the density gradient. Let $\{xi\} i = 1, \ldots, n$ be an arbitrary set of n points in the d-dimensional Euclidean space Rd. Assume

that the probability density function $p(x)$ of the d-dimensional feature space vectors x is unimodal. A sphere Sx of radius r, centered on x contains the feature vectors y such that $||y - x||_r$. The expected value of the vector $z = y - x$, given x and Sx is

$$\mu = E[z \mid Sx] = \int Sx(y - x)\, p(y \mid Sx)\, dy = \int Sx(y - x)\, p(y)/p(y \in Sx) \times dy.$$

If Sx is sufficiently small we can approximate

$$P(y \in Sx) = p(x)\ .\ V_{Sx}$$

Where, VS *c rx* is the volume of the sphere. The first order approximation of

$$P(y) = p(x) + (y - x)^{T}\, \nabla p(x)$$

Where, $\nabla p(x)$ is the gradient of the probability density function in x.

$$\mu = \int_{Sx}(y - x)\ (y - x)T/V_{Sx} \times \nabla p(x)/p(x) \times dy$$

Since the first term vanishes,

The integral is $$\mu = r^2/d + 2 \times \nabla p(x).$$

R^2 Image clustering [3] is to segment the image by clustering pixels into clusters based on the spectral similarity of each pixel. Mean shift algorithm is used in our image clustering. In mean shift procedure, the scale parameter '*r*' controls the result of clustering. The smaller the scale parameter is, the more clusters will be obtained in the data set. For image clustering, we are likely to know how many clusters are there are in the image. However, the number of clusters [3] is not specified a priori in the mean shift procedure. A general method for the choice of scale parameter '*r*' is heuristic in mean shift based image clustering. Firstly, an arbitrary choice of *r* is used and *m* clusters are defined by the *m* convergent points in mean shift [2] procedure. Then, if the number of clusters *m* is larger than the number of clusters in the image, a larger *r* is adopted to cluster the original image. On the contrary, a smaller *r* is adopted if *m* is smaller than the number of clusters in the image. In this way, mean shift algorithm [2] results in the correct number of clusters in the image. This heuristic method usually gains the correct number of clusters with a very large scale parameter '*r*'. Mean shift based clustering with a large window radius may produce increased probability of error. This section we proposed the so-called re-clustering technique based on mean shift that overcomes the drawback of the general image clustering [5] [8]. The re-clustering is a two-step method. In the next step, the mean shift procedure with a small window radius is performed in order to cluster the feature space of an image at a fine level of detail. It means that the mean shift procedure produces more clusters than that in the image.

A. Lexical Tag Parsing

Generally lexical analysis is the process of making the sequence of characters into tokens so that the meaning of each character would be analyzed. A technique called parser can be used with the lexical analyzer. A parser is basically used to check whether a natural language obeys the rules of formal language. It will be combined with parser to analysis the syntax of the programming language by comparing each token with the predefined format to check whether the program is in the format or not.

In this project by making use of this lexical analyzer to detect the image tags in the web page.

The image extraction from website needs parsing of webpage to detect the image element. It can be done using our LTP Technique which involves making a bag of words which audit the tags in the web page. So every page script consists of images and query strings which produce the page and the image. To analyze the context of a web pages, mainly there are two process namely image tag analysis and plain text analysis.

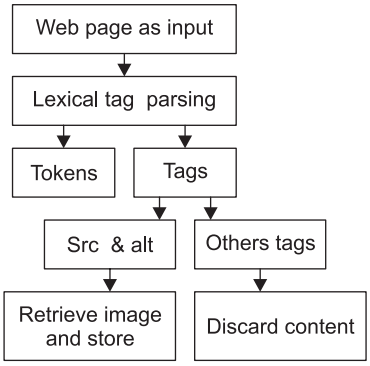


Figure 2: Flow of Lexical tag Parsing

Algorithm 1: Lexical Tag Parsing

Image tag analysis is to parse the src tag and the plain text analysis detects the alt tag to detect the labels of the image. It can be defined by

LTP(x) where x is the webpage

T – number of Tokens in webpage

k – number of tags in webpage

s – <src> tag

a – <alt> tag

C[n] – Cluster array

Img_{src} – cluster named folders

$\text{img}_k = \{\}$, a set to contain image in it.

Parse each word in x and to a Bag of words set W_b initially $W_b = \{\}$

for $n \leftarrow 1$ to T

do for $i \leftarrow 1$ to k

if ($i = s$)

$\text{Img}_{k <-}$ add(s) // add image into the set

If ($i = a$)

C[n] $\leftarrow 1$ // keep track of active cluster

Create folder Img_{src} named with Img_k

for each image in Img_k

for j 1 to N

do $\text{Img}_{\text{src}} \leftarrow \text{SIM}(\text{Img}_{\text{src}}, \text{Img}_k)$

$\text{Img}_k \leftarrow \text{add}(k)$

C[n] $\leftarrow 0$ (deactivate cluster)

Here the SIM() function is used to find the similarity of the folder name and the image name. If both are similar then the image will be stored in that folder.

It is practically impossible for the web page to be analyzed for images and textual contents by analyzing all types of tags. Because there are number of image representing tags in different types of scripting languages. But most of the scripting languages like Html, Java script and PHP uses SRC tag to include image in a web page. So by probing this tag we can easily detect the image file being included and can be downloaded. While downloading, all types of images will be downloaded, from relevant to irrelevant. Relevant images will be sent to CFR the next level of classification and the irrelevant images which have no proper labels and their annotation will be sent to separate folder call "Irrelevant".

The LTP process the given web page by considering the text for search once it opens the link it identifies the name of the image as its name in that one draw back is 90% of the image is identified with its name the remaining cannot be identified since it has different name for images.

For Example:

Input: www.cartoonnetwork.com, the given text is parsed by using LTP method here.

After parsing the necessary text "cartoon network: is taken for searching. Once the page is opened tokens and tags are separated.

Tokens include literals, keywords, identifiers, and user defined names. This web page consists of the following keywords: lang, meta, property, content, class, sizes, type, vocab, target, typeof, data-item.

In the same tags are divided into two ways to separate the images.src and src are the tags used to identify the images and the images can be into cluster.

If this process traverse into the tags other than src and alt the content can be discarded. It means it may not taken into account. Other tags included in this web page are as follows. , <meta>, , <script>, <H>, , <nav>.

Example:

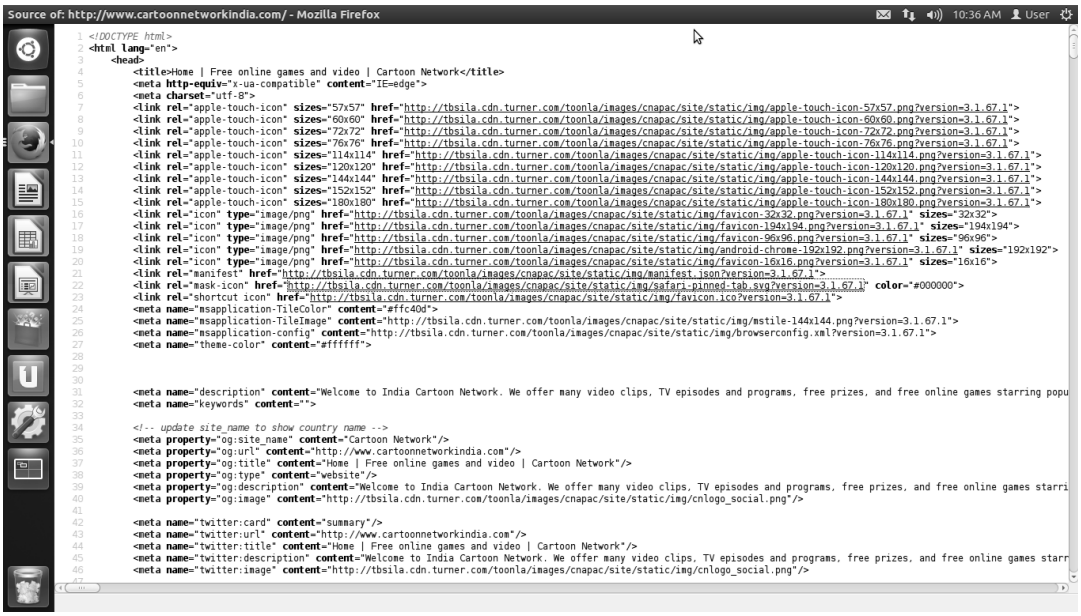


Figure 3: HTML page with tags

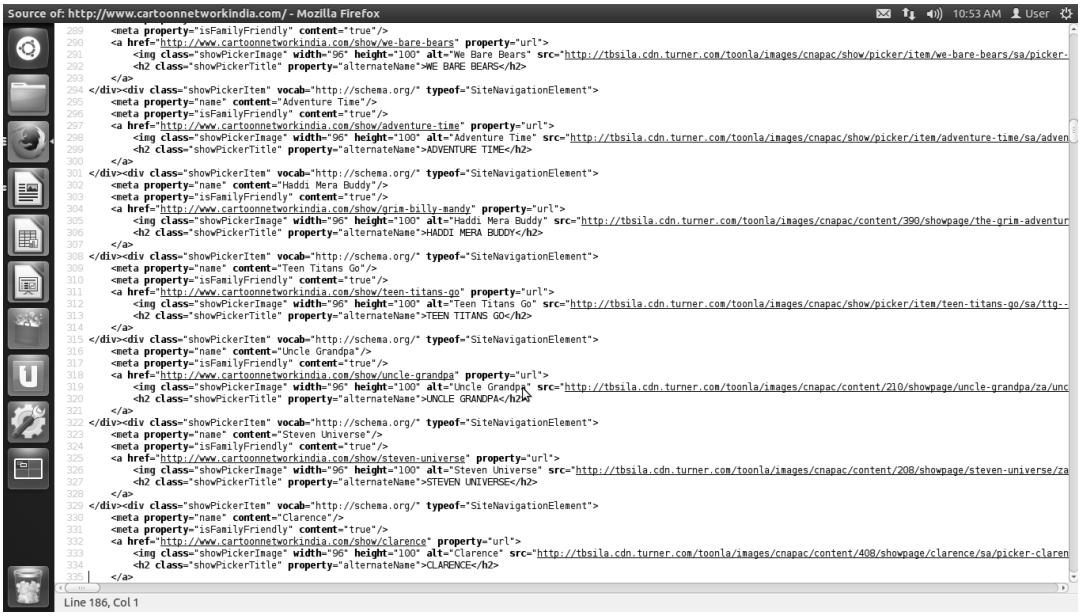


Figure 4: HTML page with tags

B. Contour Filter and Refine Technique (CFR)

Contour tracing is a preprocessing technique performed on digital images to extract image properties like shape and structure. Once the properties are extracted, it's different characteristics can be obviously examined and used as features which can be later used in pattern classification techniques. So accurate tracing of the contour will produce more accurate features which will increase the chances of classifying a given pattern more accurately. So to analyze the edge histogram, color layout, Texture properties of the images in order to find the similarity between two images. These standard properties of images can be used to cluster images by performing step by step approach. First the filtering of images on color properties of a particular cluster will be made so that irrelevant images to that cluster will be pruned out. After that refine method will be done so that the most matching images will be grouped with that particular cluster. Traditional image based search and clustering techniques mainly focused on content based as a whole. So it will be little time consuming as individual image will be compared with all the image to find its similar image.

Initially pruning of image will be made using

Algorithm 2. Filter by Refine

Input: $Img_{src,,}$ Img_k

Output: Img_{src}^* (containing clusters of similar images)

Description:

1. $Img_{src}^* = 0$
2. for each image in Img_k do
3. M_{MS} = Maximum Similar (Img_k);
4. if $|M_{MS}| > \beta$ then
5. add $Img_{k\ to}$ Img_{src}^*

6. else
7. discard Img_k
8. end if
9. end for
10. return Img_{src}^{*};

C. Multilevel Image Annotation (MIA)

The main objective of this technique is to divide the query images into clusters and assign few labels to each cluster. That is, each cluster is given with a few representative labels and make the labels unique across clusters. This is done to formulate a non-negative matrix factorization with sparsity and orthogonality constraints, which is well-suited for clustering images with our label feature representations more accurately. By employing this matrix factorization for classification or labeling to make it sure that the benefit of the low rank structure will also be highlighted when the label feature representation is combined with more than one constraint. This organize the clustering and labeling [7] techniques in a proper matrix factorization framework.

There are many possibilities for one type of images to be grouped into two different clusters. In order to address that issue by making this factorization so that detailed information can be acquired by giving more than one label description for a cluster. So on comparing the labels defined in each clusters and the label of the image that is being downloaded by getting an idea that the image is well suited to be placed in this cluster itself.

D. Synergic Image Clustering (SIC)

Synergic image clustering is the synergistic final delivery of i mage [8], its cluster and the image's description. This technique is used to make an effective and efficient image search result clustering. SIC as an efficient technique to organize Web image search results into very accurate semantic clusters. Different from all the existing web algorithms that can only cluster the top images using either visual or link features, proposed technique first identifies several semantic clusters [11] [12] related to the given image, then assigns all the resulting images to its corresponding clusters. This algorithm has three advantages over existing ISRC (Image Search Result Clustering) algorithms. First the most important image groups can be found more accurately. Second, entire group of images will be taken into consideration in the clustering process instead of only a smaller part. And finally, this algorithm is efficient enough to be implemented in practical systems.

Given the cluster names, merging and pruning technique is utilized to obtain the final cluster names. First, merged the same or very similar candidates from different set of sources [3] [8]. Second, the description of the images is utilized to prune out the candidate cluster names of possibly unhelpful clusters. Finally, the resulting cluster names are utilized as queries to search a description of that cluster.

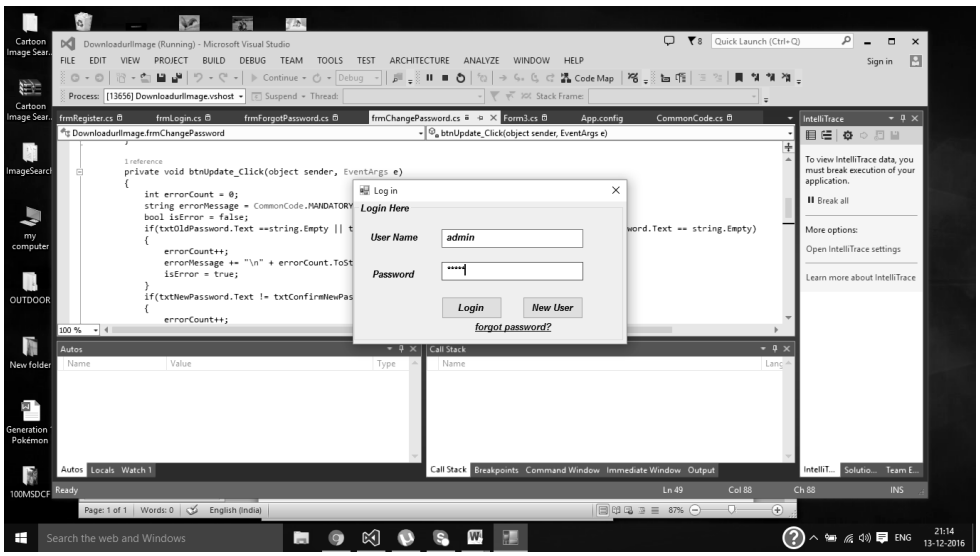
The cluster names with many or too few resulting images are first considered for analysis. The reduced thumbnails of top ranked images are used as representation images of the clusters. The by products, two problems of the existing web image search engines are solved to some extent by this algorithm. One is that with the existing image search engines, one kind of images tends to dominant the search results. For example, for query images on "Tom", this character can be in any color and structure and some other characters in cartoon can also be similar with these each other like color, layout *r* texture. So here it needs to synergize all the properties of the images to cluster them into exact classification. While with SIC these are considering our MIA and CFR techniques, other cartoon characters, e.g. Jerry or Ben 10 could be easily rationalized and discarded by our application. The other limitation addressed is that for some general queries, especially those game related queries, e.g. Power

puff, Sally and so on that are similar to cartoon will also be ranked very high. With SIC, the most related key phrases using the technique multilevel annotation could be found for each image could be filtered out and more relevant images could be grouped with the appropriate classification.

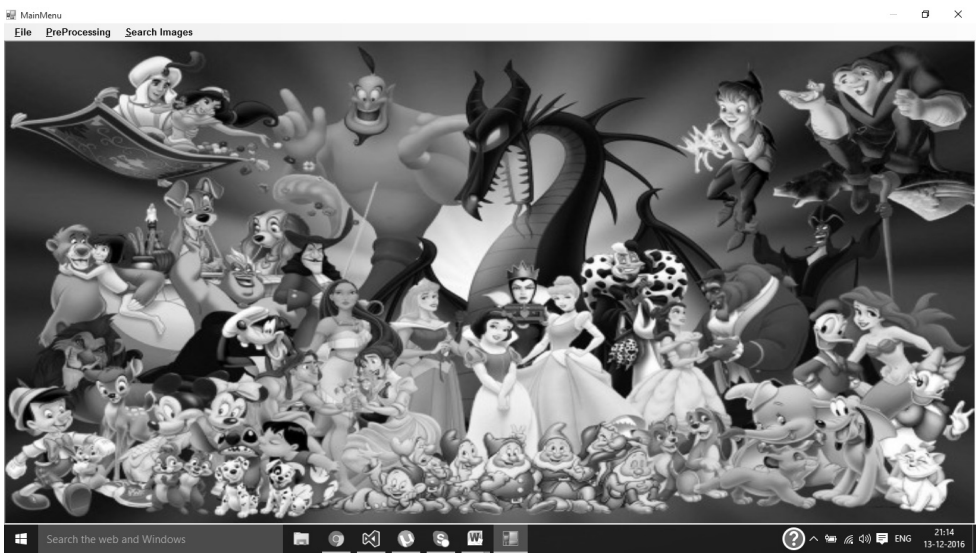
E. Noise Elimination

The output images after downloading definitely will have some irrelevant set of images which cannot be classified with any category. This includes some banners of ads in the website, small size files which are not clear for usage, some irrelevant gif files, etc. So this types of files will be eliminated by the application and it will be stored in a separate file called "Irrelevant" and it will not be considered for classification. After noise elimination, finally we can obtain the classification of files, their labels and their annotations.

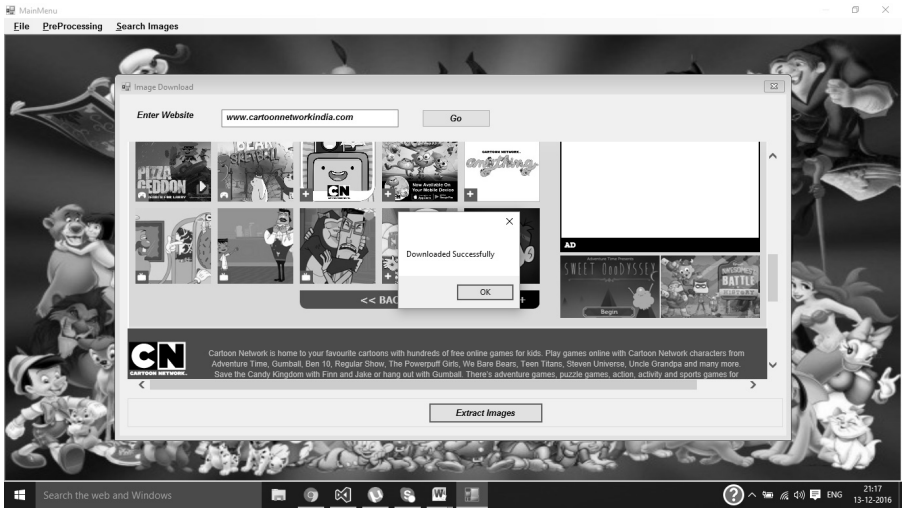
F. Sample Screenshots



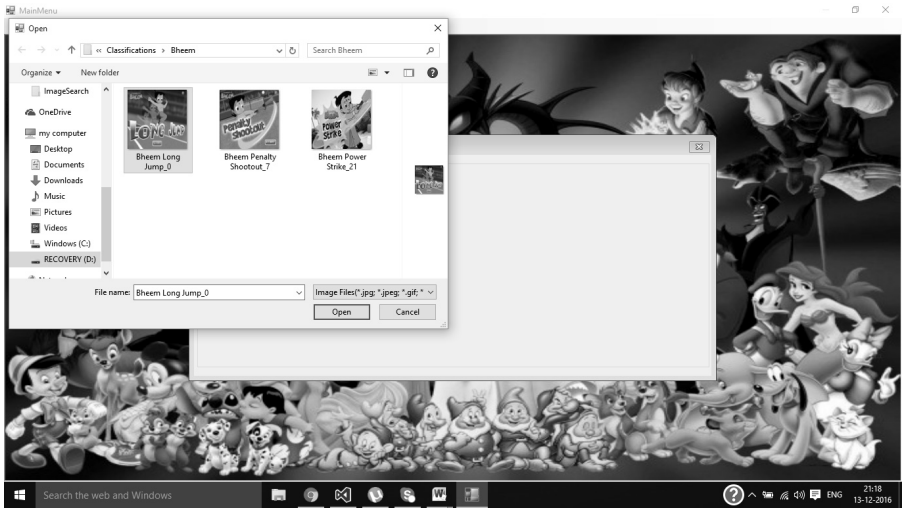
A. Login form



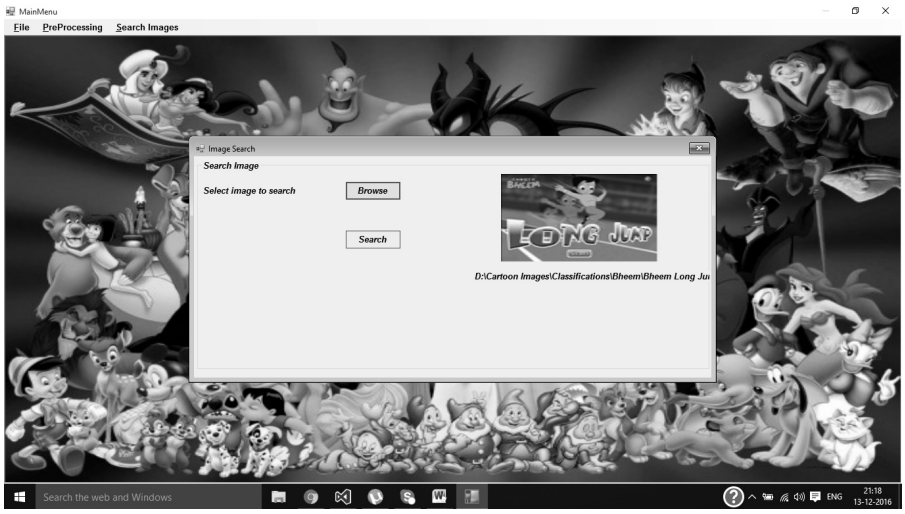
B. Initial Home page



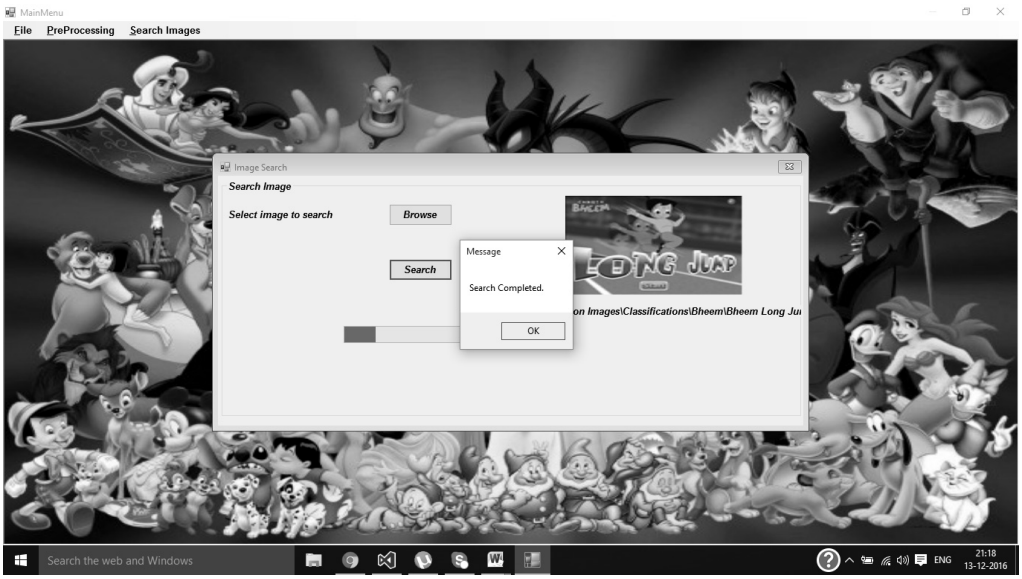
C. Displaying web page by getting input as web link



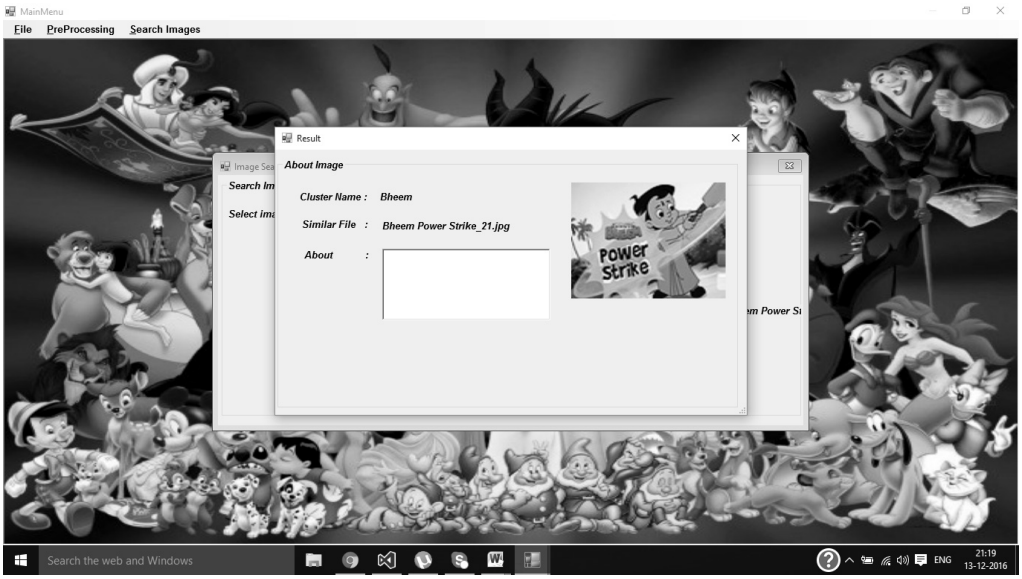
D. Clustering folder contains images



E. Search screen to select image as input



F. Searched image result



G. Displaying image with description

5. CONCLUSION AND FUTURE WORK

In this work, a novel framework for classification of web images using its textual annotation has been proposed. The context information including the encircled text and the image properties has been extracted using CFR. Proposed system has two level of classification. Initially it compares labels of each image in the website to find the similar annotation. After annotating, the color properties will be compared to detect more relevant images. The experimental results and user studies justify that the proposed MIA based image search scheme is very helpful and effective. The prototype system is practical as it is able to classify the images very quickly as well as give the annotation for the cluster image.

Also, training set data considered here from the following websites: www.cartoonnetwork.com, www.chottabheem.com. The HTML tags considered alt and src since most of the images comes under this type of

tags. In future this can also be applicable to go ahead with other related tags than alt and src by extending this algorithm.

6. RESULT AND DISCUSSION

The Lexical Tag parsing (LTP) is used for parsing all the available tags in the web site. This helps to analyze the images in the given web url.

Then Contour filter and refine technique (CFR) is used for better understanding of the shape and structure of images.. Here query images can be divided into cluster using Multilevel Image Annotation (MIA) for giving annotation for images. It provides more accuracy by the use of non-negative matrix factorization. Next, synergic Image clustering (SIC) gives the effective and efficient image search results.

Finally, proposed technique has three advantages over ISRC one is accuracy next it considers the entire Systems.

7. GRAPHS AND TABLES

Table 1
Comparison with their parameter

S. No	*Parameter*	*Description*	*Scale of efficiency*	
			Positive	*Negative*
1	Html Parsing	No of images detecting by parsing	GATE Algorithm	55/120 images
2	Image Annotation Creation	Number of annotations made per cluster	Image Web Graph	<4 annotations per cluster

Table 2
Feature comparison

S. No	*Feature*	*Positive*	*Negative*
1	Html Parsing	70	30
2	Image Annotation Creation	85	15

Table 3
List of websites

S. No	*List of Websites Used to download*
1	http://www.chotabheem.org/
2	http://www.cartoonnetworkindia.com/
3	http://www.cartoonindia.com/
4	www.pokemon.com
5	www.johnnybravo.com
6	www.powerrangers.com
7	www.Doraemon.com
8	www.cartoonnetworkasia.com
9	www.dangermouse.com
10	http://www.cartoonnetwork.co.uk/

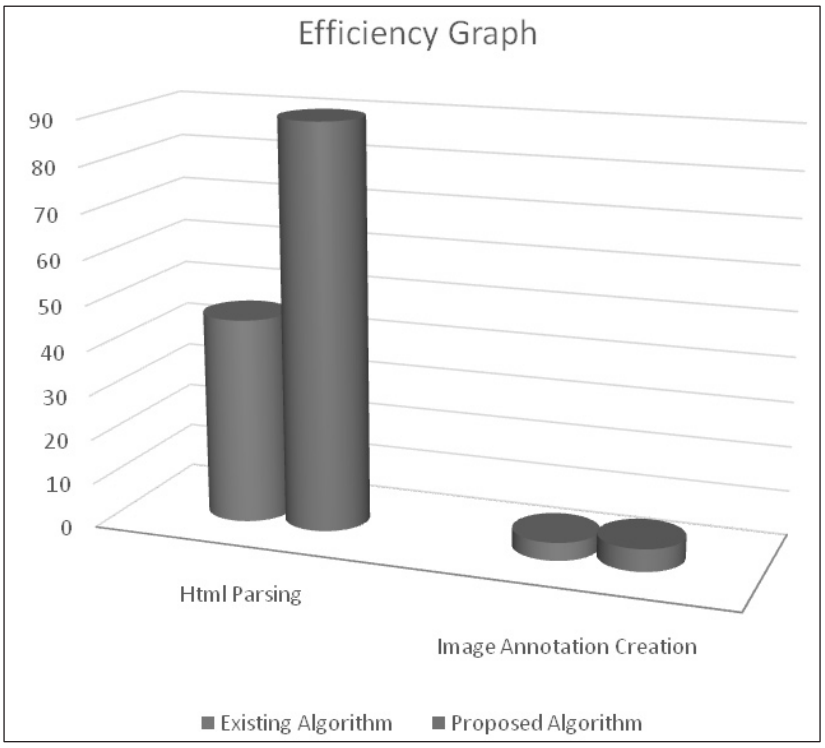


Figure 5: Efficiency Comparison of two algorithms

8. REFERENCES

[1] D.S. Xia, Z. Q. Xiang, Y.X. Zou* ADSPLAB/EL LIP, "Integrating Visual and Textual Features for Web Image Clustering", IEEE International Conference on Multimedia Big Data.

[2] Shukui Bo, Yongju Jing, "Image Clustering Using Mean Shift Algorithm", Fourth International Conference on Computational Intelligence and Communication Networks.

[3] A. Kannan, Dr. V. Mohan, Dr. Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques", IEEE International Conference on Computational Intelligence and Computing Research.

[4] D. Guillamet, & J. Vitri`a, Evaluation of distance metrics for recognition based on non-negative matrix factorization, Pattern Recognition Letters, 24(9-10), 2003, 1599-1605.

[5] "Public Control Algorithm for a Multi Access Scenario comparsing GPRS and UMTS", at Department of Computer Science and Engineering, National Conference on "Intelligent computing With IoT on April 16 2016 in Dhirajlal Gandhi College of Technology.

[6] "Teleimersion" Research Journal of Pharmaceutical, Biological and Chemical Sciences on March – April 2016 issue. (Impact Factor 0.35Indexed in Scopus). http://www.rjpbcs.com/pdf/2016_7(2)/%5b131%5d.pdf

[7] B. Xu, J. Lu, & G. Huang, A constrained non-negative matrix factorization in information retrieval, Proc. 2003 IEEE Int. Conf. on Information Reuse and Integration, Las Vegas, NV, 2003, 273-277.

[8] Y. Wang, Y. Jia, C. Hu, & M. Turk, Fisher nonnegative matrix factorization for learning local features, Proc. 6th Asian Conf. on Computer Vision, Jeju Island, Korea, 2004.

[9] V. Vapnik, The nature of statistical learning theory (Berlin: Springer-Verlag, 1995).

[10] K. Yu, L. Ji, & X. Zhang, Kernel nearest-neighbor algorithm, Neural Processing Letters, 15(2), 2002, 147156.

[11] P. Vincent, & Y. Bengio, K-local hyperplane and convex distance nearest neighbor algorithms, in T.G. Dietterich, S. Becker, & Z. Ghahramani (Ed.) Advances in Neural Information Processing Systems, 14, (Cambridge, MA: MIT Press, 2002) 985-992.

[12] E. Pe¸kalska, D. de Ridder, R.P.W. Duin, & M.A. Kraaijveld, A new method of generalizing Sammon mapping with application to algorithm speedup, Proc. 5th Annual Conf. of the Advanced School for Computing and Imaging, Heijen, the Netherlands, 1999, 221-228.