Image Retrievals based on Statistical Modeling and Cosine Similarity

Anuradha Padala, Yarramalle Srinivas, M. H. M. Krishna Prasad

***Abstract:** The technological evolution encompassed drastic updates in the field of technology and communication. This technological updates have facilitated to pool variety of data ranging from text to videos. Therefore a repository of data is built indirectly and mining from these heterogeneous groups relevant information is a challenging task. This article addresses the concept of retrieving the images using statistical approaches together with cosine similarity.*

***Index Terms:** Image Retrieval, Unstructured Data, technological evolution, cosine similarity, and statistical model.*

I. INTRODUCTION

In today's global scenario, most of the data is made available for extracting meaningful information. These information can be with regard to medical data, textual data and any data for that matter with the only imitation that most of the data in the data pool available is of unstructuredness. Therefore retrieving the data based on the context is extremely challenging. Along with this challenge, as the size of the data is more, accurate retrieval and timely retrievals are another challenging issues. Therefore effective mechanisms are to be designed and developed for effecting retrieval from the globally pooled data. In this article, an insight is thrown with the objective of the retrieval of images based on the users' interest. Image retrieval is considered as a field of search based on the context and the information available in the Meta data. This information that is available can be accessed from different levels ranging from cell phone monitor to digital computer. Many methodologies are available in the literature which focuses on different methods addressing the retrieval of the images based on content. The technological innovations also helped to formulate a pool of people integrated together in the name of content sharing, chatting and discussing of relevant information by means of online video chatting. These groups are considered to be the twitter, Facebook, Orkut etc. Among the various foresaid interaction groups, each group as its own advantages and limitations. Among the various social networking groups that were formulated in the history of communications into classmates.com 1995 which gives to be initial site aimed at primary chatting. The main disadvantage is that the direct friend group cannot be established and every connection is based on through the schooling data and it is basically a site that is developed for school mates. Sixdegree.com 1997 has tried to overcome some of the limitations of the above social group and it is considered as first online networking group and suffers with the limitations that it lags in user profile customization. MySpace 2003 is considered to be a first online website by which sharing and transferring of multimedia data is possible. However this site is not user site. Later flicker, Youtube have dominated among the social networking group. Currently social networking groups such as whatsapp, instagram are also in use serving the above problems. Inspite of these networking groups, the main objective beyond the networking groups of these kind are very much useful for the student community where relevant and useful information can be shared across that can help during the settlement. However most of the present day technologies could not able to meet the above requirements. Many researchers [1][2][3][4][5] have proposed profounding methods for image retrievals based on content, size, shape, color, texture etc. works are also presented in the literature based on features [6][7][8][9][10]. However the authors have considered either the low-level features or high level features. These features are not capable enough to extract meaningful information to the requirement of the user. Since they consider only the local features and mostly discard the global features. To overcome this disadvantages high level features are considered, where the retrieval is based on the visual perception of the users. Since the visual content cannot be represented exactly the retrieval accuracy has failed to meet the maximal desired level. To overcome these disadvantages, in the present article proposed a novel statistical modeling approach based technique by considering the semantic features and the cosine similarities. The rest of the article is articulated as follows. Section 2 of the article deals with a brief introduction of content based retrievals. Section 3 of the article highlights about the Generalized Gaussian Mixture Model (GGMM). Section 4 of the article present an overview regarding the dataset considered, the methodology is presented in section 5. The experimentation together with the results are proposed in section 6, the performance evaluation is highlighted in section 7 and the summarization of the paper is presented in the concluding section 8.

II. CONTENT BASED RETRIEVAL

Content based retrieval is a specialized area of image retrieval where the relevant data is extracted based on the content. In order to achieve this, the content is given as the query and this query is to map with the bold dataset to identify the relevancy of the data and thereby the relevant data is retrieved.

Manuscript published on 30 June 2019.
* Correspondence Author (s)

Padala Anuradha*, Department of Computer Science Engineering, GITAM (Deemed to be University), Visakhapatnam, India.
Yarramalle Srinivas, Department of Information Technology, GITAM (Deemed to be University), Visakhapatnam, India.
M.H.M Krishna Prasad, Department of Computer Science Engineering, JNT University, Kakinada, India.

The main challenges associated here is that, the complete success of the retrieval is merely based on the query. However, if the query is not processed exactly because of the semantic gap between the visual interpretation and information, thus retrieval of the relevant image is at stake, if the query image is incomplete or size of the database is large this query based image retrieval systems are deemed to be ineffective. Therefore every images are to be considered and the semantic gap between the low-level features and high level features are to be estimated and this semantic estimation is to be bundled while retrieving the content.

III. GENERALIZED GAUSSIAN MIXTURE MODEL

In order to present the present article we have considered the statistical model namely Generalized Gaussian Mixture Model. The main advantage beyond the consideration of this model is that it can help to retrieve the images either based on the dimension of the image or based on frequency count i.e., the number of times a particular image has been retrieved and considering the best frequency, that image is considered as reference image or query image and the whole retrieval process is subjected on this retrieval. These frequency based approach and size based approaches can even be assumed as time-line based approaches and breadth based approaches. The probability density function of the Generalized Gaussian Mixture Model is given by

$$f(Z\mid\mu,\sigma,\rho)=\frac{1}{2\Gamma(1+\frac{1}{\rho})A(\rho,\sigma)}e^{-\left|\frac{(Z_i-\mu_i)}{A(\rho,\sigma)}\right|^{\rho}}$$

$$\sigma>0,\ A(\rho,\sigma)=\left[\frac{\sigma^{2}\Gamma(1/\rho)}{\Gamma(3/\rho)}\right]^{\frac{1}{2}}$$

Where

IV. DATASET CONSIDERED

In order to present the model, we have considered a benchmark dataset namely Flicker dataset. It is an online dataset comprising of 14062 images with different frequencies, size and shapes. Each of the images are of grey scale and color, for the experimentation purpose among these images we considered 1000 images for training and testing phase. Each of the images that are considered for the experimentation are presented in the following table along with their frequencies.

Table 4.1 presenting the frequent contents-based tags

Image Tag	Frequency
Sky	845
Water	641
Portrait	623
Night	621
Nature	596
Sunset	585
Clouds	558
flower/flowers	510/351
Beach	407
Landscape	385
Street	383
Dog	372
Architecture	354
graffiti/streetart	335/184
tree/trees	331/245
People	330
city/urban	308/247
Sea	301
Sun	290
Girl	262
Snow	256
Food	225
Bird	218
Sign	214
Car	212
Lake	199
Building	188
River	175
Baby	167
Animal	164

The Sample- Dataset considered from Flicker, is presented, in figure 1

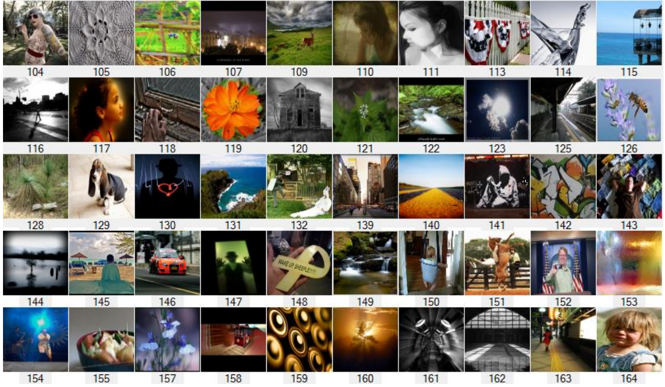


Fig. 1: MIRFlickr Dataset

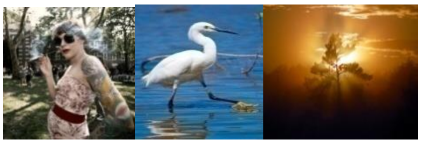
V. METHODOLOGY

In order to experiment the data, we have considered both the time-line based approach and breadth based approach into consideration. The time-spent on each image is considered and against each of the image based on the time, the relevant images that are in similar are presented along with the timestamp i.e., the time-spent in the following table.

Table 5.1 Showing the frequency of occurrence of each of the images

Image number	Duration in a _day	Time spent	Image number	Duration in a _day	Time spent
1.	2	6.97	2.	26	4.69
3.	4	6.27	4.	24	4.67
5.	6	6.24	6.	2	4.65
7.	10	5.98	8.	6	4.62
9.	16	5.97	10.	75	4.6
11.	22	5.62	12.	1	4.58
13.	64	5.6	14.	3	4.57
15.	46	5.52	16.	1	4.54
17.	2	5.5	18.	6	4.54
19.	14	5.49	20.	6	4.52
21.	9	5.49	22.	5	4.5
23.	5	5.47	24.	16	4.49
25.	20	5.34	26.	10	4.46
27.	4	5.3	28.	36	4.46
29.	37	5.27	30.	47	4.45
31.	8	5.25	32.	10	4.43
33.	52	5.18	34.	4	4.41
35.	4	5.13	36.	27	4.4
37.	2	5.11	38.	1	4.35
39.	3	5.09	40.	4	4.35
41.	72	5.09	42.	1	4.34
43.	4	5.07	44.	64	4.32
45.	28	5.06	46.	2	4.24
47.	2	5	48.	1	4.23
49.	3	5	50.	1	4.16
51.	9	4.98	52.	1	4.11
53.	5	4.96	54.	1	4.04
55.	15	4.94	56.	4	3.97
57.	6	4.89	58.	81	3.89
59.	6	4.88	60.	71	3.86
61.	7	4.84	62.	2	3.86
63.	4	4.83	64.	3	3.75
65.	7	4.81	66.	4	3.73
67.	7	4.75	68.	4	3.69
69.	5	4.71	70.	4	3.66
71.	28	4.7	72.	4	3.14
73.	51	4.69			

From the images basing on the highest frequency-stamp the query are selected based on these query images using the features of color and texture the retrieval images are extracted i.e., each of the images are considered as query images are taken into account and the probability density function of these images are identified using the model based on generalized Gaussian mixture model presented in section 3 of the article. Based on the relevance of the probability density functions, the retrieved images against the query images are presented in the following figure.



Fig, 2: Query images

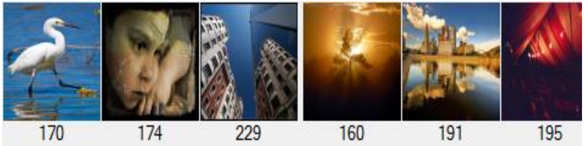


Figure 3: showing the relevant images retrieved

The results derived are tested for accuracy and are presented in section 6 of the article.

VI. EXPERIMENTATION

In order to evaluate the model, each image against the query image are considered and relevant probability density functions are extracted. Against each of the image, basing on the frequency of the retrievals, the possible query images in tune with the query images using the model the relevant probability density functions are extracted and are presented in table 6.1 of the article.

Table 6.1: Images retrieved based on PDF values

Input	Grey scale images	PDF Values	
		Min	Max
		0	16
		0	10
		0	37
		0	38
		0	30
		0	31

In order to identify the most similar items the cosine similarity measure is considered and the formula for evaluating the cosine similarity is given as follows

$$\mathrm{Cos}(d_1 d_2) = \frac{\mathrm{dot}(d_1, d_2)}{\|d_1\|.\|d_2\|}$$

And $\mathrm{dot}(d_1, d_2)$ is given by

$$d_{1[0]} * d_{2[0]} + d_{1[1]} * d_{2[1]}$$

Where, $\|d_1\| = \sqrt{d_1(0)^2} + \sqrt{d_1(1)^2} + \ldots\ldots\ldots..$

The images retrieved based on the Cosine Similarity, is presented in the following Figure-4

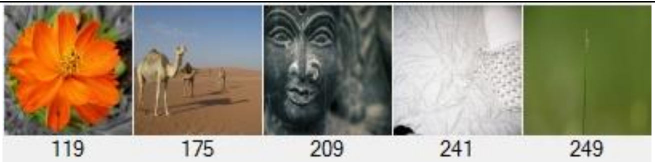
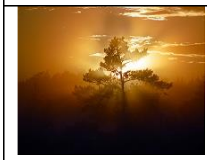
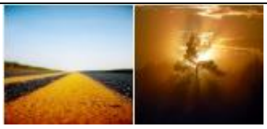
Query Image	Relevant Retrieved Images based on cosine similarity
	
	

Fig. 4: Relevant Retrieved Images based on cosine Similarity

In order to test the accuracy of the methodology, performance evaluations is carried out using benchmark metrics and are presented in section 7 of the article.

VII. PERFORMANCE METRICS

In order to identify the efficiency of the model benchmark metrics like precision and recall are considered and performance evaluation is carried out using these metrics. The formula for calculating the above metrics is given as follows.

a) Precision

The tradeoff between the images that are retrieved against a query can be measured using precision. It is generally given by actual number of appropriate images vs. total number of irrelevant images retrieved. The formula may be expressed as

Precision = (A / (A + C))* 100

A: Number of relevant images obtained against a query image.

C: Number of irrelevant images retrieved against a query image.

A + C: Total number of irrelevant + relevant images retrieved.

b) Recall

It is another measure which is generally used for identifying the relevancy. It is defined as the ratio of relevant images retrieved to the total number of relevant images and is expressed as

Recall = (A / (A + B)) * 100

A: Number of relevant images retrieved

B: Number of relevant images not retrieved

A + B: The total number of relevant images

Using the above metrics the retrieval images are tested for accuracy and the results derived are presented in the following tables and graphs.

The results obtained are presented in Table 7.1.

Query image	Content-based approach	
	Precision	Recall
	0.5	0.33
	0.66	0.4
	0.66	0.4
	0.33	0.25

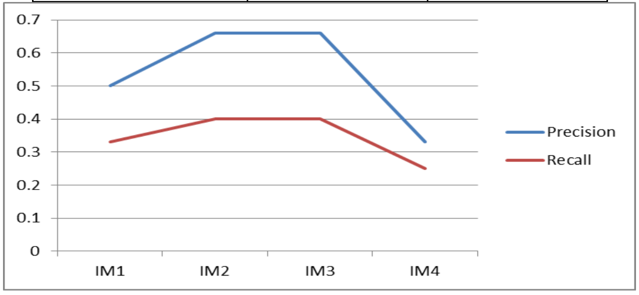
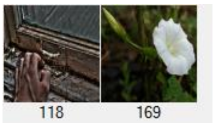


Fig. 5: Graph representing Precision and Recall

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Table 7.2: Images Retrieved based on PDF Values

S.No	Input	Images retrieved based on PDF Values	No of relevant images	Precision	recall
IM1			2	0.5	0.33
IM2			2	0.66	0.4
IM3			2	0.66	0.4
IM4			1	0.33	0.25
IM5			2	1	0.5
IM6			1	0.5	0.33

The relevancy of the retrievals based on the PDF is evaluated using the metrics precision and recall and the results obtained are presented in the Table 7.2

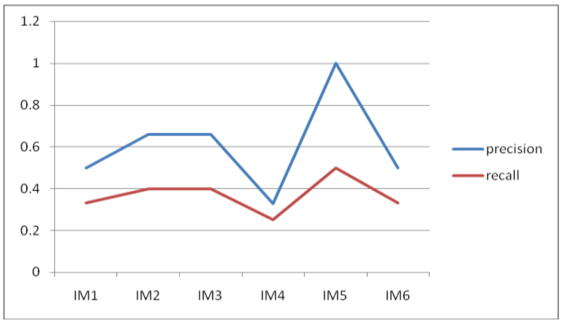


Fig. 6: Graph Representing precision and recall

VIII. CONCLUSION

In this article a novel methodology of retrieving a most retrieval information based on the content is presented. In order to present the article we have considered a benchmark dataset and the evaluation is carried out using this dataset and the proposed model based on generalized Gaussian distribution. In order to retrieve the similar images, approaches based on cosine similarity, frequency based are considered and presented. The experimentation is carried out with 1000 images and the result derived as case studies are tested using performance metrics like precision and recall and the results derived showcase the proposed model is giving a good retrieval accuracy of above 94% in most of the cases. This framework can be very much useful for students in particular to surf their interested topic based on relevancy.

REFERENCES

1. Y.Jhansi, Dr. E. Sreenivasa Reddy " Sketch Based Image Retrieval with Cosine Similarity" International Journal of Advanced Research in Computer Science, Volume 8, No. 3, March – April 2017, 691-695
2. Sukhdeep Kaur, Deepak Aggarwal "Image Content Based Retrieval System using Cosine Similarity for Skin Disease Images" ACSIJ Advances in Computer Science: an International Journal, Vol. 2, Issue 4, No.5 , September 2013, ISSN : 2322-5157
3. Hyun-chong Cho, Lubomir Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Mark Helvie, Chintana Paramagul, Alexis V. Nees "Similarity evaluation in a content-based image retrieval (CBIR) CADx system for characterization of breast masses on ultrasound images" Med Phys. 2011 Apr; 38(4): 1820-1831.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

4. Varish, Naushad, Kumar, Sumit, Pal, Aruo kumar “A Novel Similarity Measure for Content Based Image Retrieval in Discrete Cosine Transform Domain” Fundamenta Informaticae, vol. 156, no. 2, pp. 209-235, 201
5. Keneilwe Zuva,Tranos Zuva “Effectiveness of Image (dis)similarity Algorithms on Content-Based Image Retrieval” International Journal of Engineering and Science,ISSN: 2278-4721, Vol. 1, Issue 1 (August2012), PP 31-35
6. S.-H. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," International Journal of Mathematical Models and Methods in Applied Sciences, vol. 1, pp. 300-307, 2007.
7. Bei-ji Zou, Marie Providence Umugwaneza “Shape-based Trademark Retrieval using Cosine Distance Method” IEEE Eighth International Conference on Intelligent Systems Design and Applications, 978-0-7695-3382-7/08.
8. Niharika Rastogi, Anuj Bhargava, Prashant Badal “An Improve Method of Content Based Image Retireval by Using Different Distance etricon Color image” nternational Journal of Advance Engineering and Research Development,Volume 4, Issue 11,November-2017.
9. Gang Qian,Shamik Sural,Yuelong Gu,Sakti Pramanik “Similarity between Euclidean and cosine angle distance for nearest neighbor queries” 2004 ACM Symposium on Applied Computing, SAC’04, March 14-17, 2004, Nicosia, Cyprus