"Bodh", BPIT's International Journal of Technology & Management
ISSN: 2454-8421, Volume 7, Issue 2, July- Dec 2021, Page 1-10

Automatic Image Tagging Using Tensor and Gaussian Filter

Tanisha Madan[1], Tushar Patnaik[2], Deepali Virmani[3]

[1]*Department of Computer Science and Engineering, Bhagwan Parshuram Institute of Technology, Delhi, India,*
[2]*Education & Technology department, Centre for Development of Advanced Computing, Delhi, India*
[3] *Artificial Intelligence and Data Sciences Department,* VIPS-TC College of Engineering, *Delhi, India*
[1]tanishamadan@gmail.com, [2]tusharpatnaik@cdac.in, [3]deepalivirmani@gmail.com

Abstract- As automatic image tagging is important in commercial and research, there is a wide gap in visual representation and text generated by image. There is need to assign better quality tag to an image and remove the need of manual tagging. Gaussian filter is used to improve the quality of low-level visual features of the image and to minimize the semantic gap. All images are converted into tensors to form a group of similar features. 3-level tucker decomposition is then processed on tensors to find the better matching of the context group. Tensor formation and context estimation is used in the proposed approach to minimize the semantic gap. Sometimes irrelevant tags are assigned to the image and sometimes tags are not retrieved so this problem had been reduced through three level tensor decomposition and Gaussian blurring filter. The proposed algorithm is tested on the corel-10K dataset.

Keywords--- Tensor, Gaussian, Decomposition, Corel-10K

I. INTRODUCTION

In the last two decades a large amount of research has been carried out in automatic images tagging. Generally, the focused research part in this area is content based image retrieval system. While the current research shows there is a large semantic gap between image semantics understandable by human and content based image retrieval [1]. Distance between features that are visual in the image and the tag generated for the image is known as semantic gap [2]. Therefore, a large amount of research has been performed to minimize the semantic gap between tag generated and visual representation of image. In this paper the context of an image based algorithm is used to mend this gap. The information provided in the context can only complete the gap between the textual description and content of the image [2]. Context estimation is an essential task to incorporate the information for a better image annotation process. The context generated from the image itself with the help of visual features present in the image.

In this paper Gaussian filter based feature-independent and unsupervised context estimation is proposed for better classification and results are compared with the classical feature-independent and unsupervised context estimation method. Gaussian filters are used in this paper to improve the quality of visual representation of images and to minimize semantic gap. All the context images are converted into tensors to form a group of similar features; image.3 level tucker decompositions are then processed on tensors to find the better matching of the context group. The proposed algorithm is tested on the corel-10K dataset. Precision, accuracy, and recall are calculated for results and analysis purposes. The paper is organized as follows. Section1 deals with introduction. Section 2 deals with the related work. Problem formulation is described in the Section 3, Section 4 and its subsection deals with the estimation of the context information. Mathematical model for context estimation comes in Section 5. Section 6 deals with results and discussion. Section 7 deals with conclusion and future scope. Section 8 contains the reference part.

II. RELATED WORK

Automatic generation of image annotations has been studied for many years with the increasing popularity of social media. Several social media based approach [3,4, 5] have and been developed and proved its role in traditional applications as well as for personal needs. There are many methods in the past that are being developed for tagging but major drawback is that they are not giving appropriate tag to an image i.e. lacking the semantic gap. Machine translation of relevance models has been adapted for automatic image annotation [6-9]. The joint probability of the images with visual representation and textual description is modelled. These models are used for building a classifier that is used for tagging upcoming images. The tag refinement approach is proposed by which are based on employing random walk on a pair-wise graph, where the mined relations between tags are represented by the edge of the graph [10]. The authors of the research paper [11] have proposed metric learning based weight factor assignment to the neighbour images.

Auxiliary information can also be used with images to produce image tag. Usually, these models work with the images having news datasets. All the images available in the dataset must accompany with some news articles. In order to reduce the semantic gap between textual description and visual representation, auxiliary information based context estimation is used. The context estimation of all images must be performed for better accuracy. The video analysis tensors splitting based technique have been found suitable to recognize goals such as motion detection and action recognition. Automatic image tagging can be built using DenseNet, an advanced deep learning model [13].

AICRL model consist of one encoder and decoder. Encoder is built with ResNet 50 and decoder is built with LSTM [14]. CNN-LSTM is also used to recognize and generating the tag of images [15]. Neural network approaches are best in determining the tag of images automatically [16, 17]. Bengali tags can also be generated to tag the images [18]. AlexNet and GoogLeNet are also built for images tagging [19].Automatic annotation can also be done through mask RCNN and object can be detected through AWS [20]. The tensor based approach is used to better combine the similar context groups and 3-level tucker decomposition is used to evaluate the better correlated matrix. The variance matrix is used to better predict the tag for the testing images.

III. PROBLEM FORMULATION

The objective of this paper is to incorporate proper textual tag information for the process of automatic image annotation. For better accuracy estimation of the context information is used to incorporate the exact textual tag. Corel-10K[12] dataset is used for the training and testing purposes. First training folder contains set of training images with textual description. Let there are A training images having AxR vocabulary sets having the textual description of the image. All the training images will be filtered and form a context group. The context group must be form in such a way that all the images in one group must have some relation with the images in another group. The automatic image annotation will be performed on the testing images having M samples. The testing images will be annotated on the basis of trained samples. Estimation of the context information and context group formation are defined as in [2]. The flowchart of the proposed approach is as in fig1.

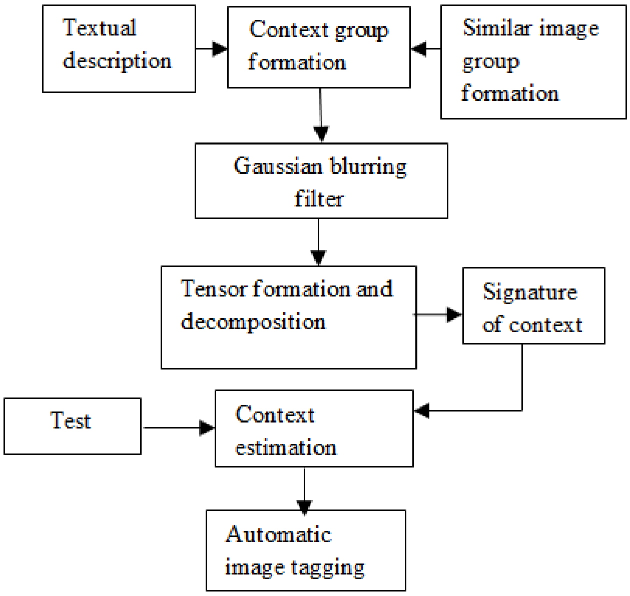


Fig 1: Proposed approach

The flowchart of proposed approach uses the major steps as:

1) Context group formation
2) Tensor Formation and Decomposition
3) Context Estimation
4) Context Based Automatic Image Annotation

IV. CONTEXT INFORMATION

Three step context estimation processes are proposed in this paper. In the first step groups having a limited number of training data are formed such that there must be some relation between images in different groups. The groups are then converted into tensors and then 3-level tucker decomposition is performed on the tensors to get better features. In the last step the textual description matrix called the label matrix is formed and assigned a number corresponding to the textual description. Example: Suppose textual description of an image are grass, water and animal then for ease of programming purposes grass will be assigned as 1, water as 2, and animal as 3.

A. Context Groups

In the first step context groups are formed that have high visual similarity between the images. The method being proposed requires two things:
1) The images should have high context similarity for better prediction for testing images.
2) Different groups must have some visual similarity for better learning and prediction purposes. The textual description of all the groups is available in the label matrix as discussed in the above section and must be highly correlated with the visual representation. Since each group contains N training images therefore tf- idf represents the X, denoted by t which is a vector having length A.
If A_{freq} is the frequency of ath word appearance in the textual description of the image X and A_{total} is the cumulative frequency of nth word in the dataset, then the value of nth sample in the vector t will be

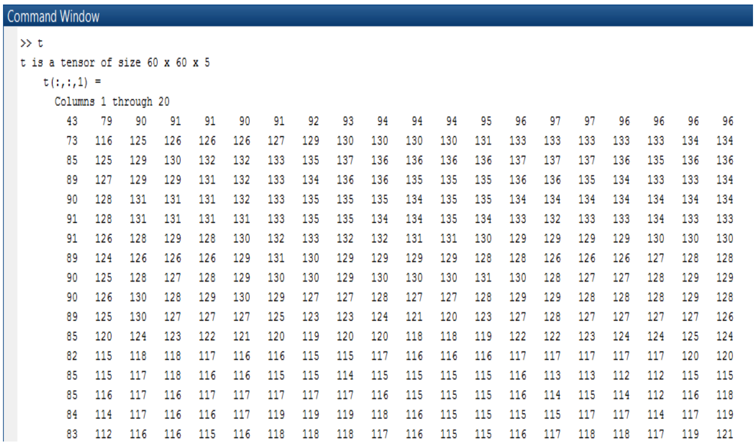
$$h_A = \frac{A_{total}}{A_{freq}} \quad (1)$$

The images are clustered on the basis of cosine similarity tf-idf vectors. This process ensures that visual similar image will come into the same context group. All the groups will be divided into two parts 1. General features based grouping and 2. Distinctive features based grouping. Each image will contain either general features or distinctive features. In the example, if the word 'water' occurs frequently in the textual description of all the images then it will be the 'general' vocabulary of the dataset. On the other hand, if 'animal' is textual description of few images then it will be 'distinct' vocabulary of the dataset.

Fig 2: Context group example based on visual similarity in the textual description

B. Tensor Formation and Decomposition

In the first step one context tensor $S\epsilon\, R^{X\times m\times n}$ is constructed for every groups of image as in the Fig. 2. The images are resize to a fixed dimension (height and width), gray-scale conversion, and passed through the Gaussian filter. Post filtering they combined to form a tensor where X, Y, and Z are width, height and number of images.



```
Command Window
>> t
t is a tensor of size 60 x 60 x 5
    t(:,:,1) =
      Columns 1 through 20
        43   79   90   91   91   90   91   92   93   94   94   94   95   96   97   97   96   96   96   96
        73  116  125  126  126  126  127  129  130  130  130  130  131  133  133  133  133  133  134  134
        85  125  129  130  132  132  133  135  137  136  136  136  136  137  137  137  136  135  136  136
        89  127  129  129  131  132  133  134  136  136  135  135  135  136  136  135  134  133  133  134
        90  128  131  131  131  132  133  135  135  135  134  135  135  134  134  134  134  134  134  134
        91  128  131  131  131  131  133  135  135  134  134  135  134  133  132  133  133  134  133  133
        91  126  128  129  128  130  132  133  132  132  131  131  130  129  129  129  129  130  130  130
        89  124  126  126  126  129  131  130  129  129  129  129  128  128  126  126  126  127  128  128
        90  125  128  127  128  129  130  130  129  130  130  130  131  130  128  127  127  128  129  129
        90  126  130  128  129  130  129  127  127  128  127  127  128  129  129  128  128  128  129  128
        89  125  130  127  127  127  125  123  123  124  121  120  123  127  128  127  127  127  127  126
        85  120  124  123  122  121  120  119  120  120  118  118  119  122  122  123  124  124  125  124
        82  115  118  118  117  116  116  115  115  117  116  116  116  117  117  117  117  117  120  120
        85  115  117  118  116  116  115  115  114  115  115  115  115  116  113  113  112  112  115  115
        85  116  117  116  117  117  117  117  117  116  115  115  115  116  114  115  114  112  116  118
        84  114  117  116  116  117  119  119  119  118  116  115  115  115  115  117  117  114  117  119
        83  112  116  116  115  116  118  118  118  117  116  115  115  116  117  118  118  117  119  121
```

Fig 3: Tensor of context group

The tensor $S\epsilon\, R^{X\times Y\times Z}$ are decomposed into smaller core tensor S and the matrices A, B and C such that

$$S \approx S \times_1 A \times_2 B \times_3 C = \sum_{i=1}^{X} \quad \sum_{j=1}^{m} \quad \sum_{k=1}^{n} \quad g_{ijk} a_i b_j c_k \tag{2}$$

where$A \in\ R^{X\times LEV}$, $B \in\ R^{m\times LEV}$, and $C \in\ R^{n\times LEV}$ are the orthogonal matrices,

$S \in\ R^{LEV\times LEV\times LEV}$ is the core tensor and $LEV \leq min(X, Y, Z)$.

The $\underline{\times_i}$ operator denotes the tensor namely,

$$\alpha = \underline{\beta \times_i \gamma} \lll=\ggg \alpha_{jk} = \sum_{i=1}^{N} \quad \beta_{ijk} \tag{3}$$

where$A$, B, and C are the matrix having dimensions$X \times LEV$, $M \times LEV$, and $N \times LEV$. The LEV is the rank of tucker decomposition. In this paper LEV is taken as 3. X, Y, and Z are the height, width, and the size of the context group. The matrices$A$, B and C are the similarity/dissimilarity of one image to its neighbour images. Since all the images belong to the same category, it will find a high similarity between all images. The matrix C is the compact signature of the context group.

C. Context Estimation

In this process the context for the image quantifies for their matching with different signature contexts. X_0 denotes the test image. Since there is no textual description available for the testing image, the variance in the elements of data in context signature matrix C will be very minimal because of the highly correlated visual appearance. If any foreign image insert into the context group, then the variance of that image in the context signature matrix$C$will be very high and will be easily rejected after the tucker decomposition process. The divergence in variance will be directly proportional to the dissimilarity in the image from the context group. The image which needs to be tagged is inserted at the location L in a tensor t by swapping the images kept at that location.

Now the matrix C' will be computed using tucker decomposition method. $|C - C'|$is used to measure the test image association with the context of context group into the tensor T.The distribution of conditional probability of the test image for every context group is calculated as:

$$W(H|X_0) = \frac{e^{(-(C'-C)^{t}\Gamma^{-1}(C'-C))}}{\sqrt{2\pi|\Gamma|}} \tag{4}$$

where ɼ is the covariance matrix, X_0 is the test image.

When the context groups are formed, the textual descriptions are given weight on the basis of their frequency. It is given less weight if it occurs very frequently and vice versa. Each test image X_0 is provided with the same conditional probability as the context group.

V. MATHEMATICAL MODEL FOR CONTEXT BASED AUTOMATIC IMAGE ANNOTATION

In Suppose there are n number of visual units such as V_n in an image and m number of textual description such as H_m. Let there are CC number of context categories where $T \in CC$ corresponds to one context group. Each training image will belong from the one of these T groups. By picking a context group with conditional probability over test image X_0 i.e. $W(H|X_0)$. By selecting a training image X_t within the training set TS with the probability $W(H|X_h)$
for i=1,2,…..,n
2.1 Pick a visual unit V_i having conditional probability $W_R(.|X_h)$

For j=1,2,…..,m

2.2 By selecting a word h_j from conditional probability $W_T(.|X_h)$

The main aim of the proposed approach is to enhance probability metric V and T over the training image X_h

$$P(X_h) = \sum_{H \in CC} \quad P(H|X_h) \sum_{X_h \in HS} \quad P(X_h|H) \prod_{j \in m} \quad w_T(H_j|X_h) \prod_{i \in n} \quad W_R(vI/X_t) \tag{5}$$

The $W_H(H_j|X_t)$ (Bernoulli distribution) is defined as:

$$W_H(H_j|X_h) = \frac{\mu\delta_{H_j} + N_{H_j}}{\mu + N_H} \tag{6}$$

where, A_{H_j} is the members of T with word T_j in their description

A_{H_j} is members of T_j

δ_{H_j} is set to be 1 if description of the image X_t has word T_j in it μ is empirically selected constant

$W_R(V_i|X_t)$ is the density estimate to generate the visual unit V_i for the training image X_t. Gaussian kernel is employed for this density estimate. Suppose if the visual units of the training image X_t are $\{VT_1, VT_2, \dots \dots, VT_n\}$ then

$$(V_i|X_h) = \frac{e^{(-(V_i - VH_n)^T (\Sigma \quad V_i - VH_n)^{-1})}}{\sqrt{2\pi|\Sigma|}} \tag{7}$$

where $\sum$ is the covariance matrix.

VI. RESULTS AND DISCUSSION

The proposed algorithm is tested on the Corel-10k dataset. In this paper results are compared without using filters and with using filters. In both the cases the tucker decomposition level is taken as 3. To check the effectiveness of the algorithm the comparison has been made between precision, recall, and accuracy. Since the dataset contains 10k images, the complete dataset are divided into small context groups for easy and fast analysis purposes. We have also checked the efficiency of the algorithm by taking different percentage combinations of training and testing data i.e. 70% and 30%, 50% and 50% etc.

Table I: Results analysis of different context groups

Group's name (Training images-Testing images)	No. of correct and retrieved tags in proposed approach	No. of correct and retrieved tags in base approach	Total tags that should be correct and retrieved
New 17(70-30)	59	45	60
New 19(70-30)	51	44	60
New 21(50-50)	101	83	113
New 12(70-30)	57	56	60
New 18(50-50)	92	86	100
New 32(50-50)	106	94	131
New 23(5-5)	11	11	12
New 24(5-5)	10	10	10

The results thus obtained are as in the table I-IV.

- Some tags are correct and retrieved
- Some tags are incorrect and retrieved
- Some tags are not retrieved giving some random value.

- **Precision:** Precision is calculated as a fraction of relevant tags among retrieved tags as in equation (8).
  Precision P=D/E (8)
  where D is the number of relevant images retrieved whereas E is the total number of images retrieved.

Table II: Precision table for both the cases

Group name(Training images-Testing images)	Proposed approach Precision in %	Base approach Precision in %
New 17(70-30)	98.3333	83.3333
New 19(70-30)	86.4407	81.4815
New 21(50-50)	90.9910	74.778
New 12(50-50)	98.2759	98.2456
New 18(50-50)	92.9293	88.6598
New 32(50-50)	79.6992	71.557

- **Recall:** Recall is calculated as a fraction of total relevant tags that are retrieved equation (9).
  Recall=D/F (9)
  where D is the number of relevant images retrieved. F is the number of images that are relevant in the dataset.

As it can be seen from Table I to IV that all the analysis parameters such as precision, recall, and accuracy have been improved many fold in the proposed approach as compared to the base approach.

Table III: Recall table for both the cases

Group 'name (Training images-testing images)	Proposed approach recall in %	Base approach Recall in %
New 17(70-30)	100	90
New 19(70-30)	98.0769	95.6522
New 21(50-50)	98.0583	90.2714
New 12(50-50)	96.6102	94.9153
New 18(50-50)	98.9247	97.7273
New 32(50-50)	85.4839	71.557

Accuracy is calculated as total no of correct observation divided by total no of observations.

Table IV: Accuracy table for both the cases

Group’s name(Training images-Testing images)	Proposed approach accuracy in%	Base approach accuracy in%
New 17(70-30)	88.2602	74.9217
New 19(70-30)	74.6538	70.4888
New 21(50-50)	78.1253	72.8023
New 12(50-50)	87.3810	84.7413
New 18(50-50)	83.9296	81.4908
New 32(50-50)	71.9161	64.4773

The model proposed by author Tariq et.al [2] has used single level tucker decomposition while in this paper three level tucker decomposition is used to model the base case. Even in the base case the results are better as compared to the results obtained by Tariq et. al [2]. Since all the images contain Gaussian noise by default therefore adopting the Gaussian filtering technique improves the results.

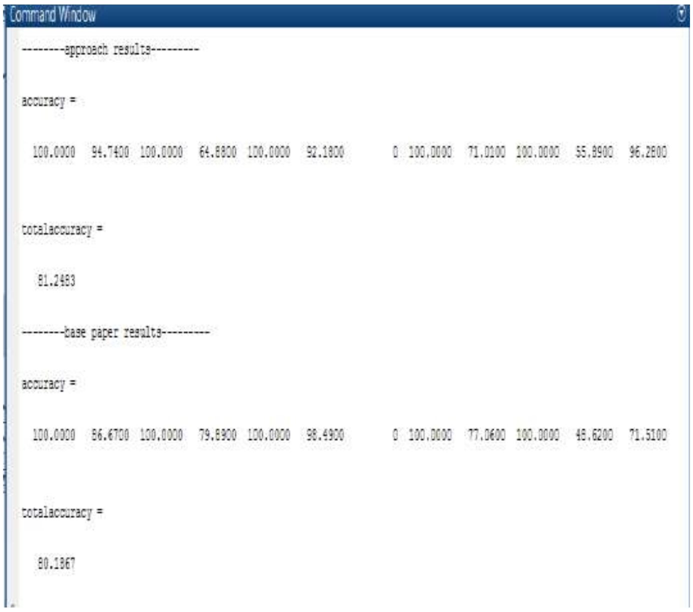


Fig 4: Snapshot of results obtained in MATLAB

The graphical representation of the above table can be seen in Fig. 5 to Fig. 7. MATLAB software is used for the simulation and analysis purposes. The results obtained during simulation are as in Fig. 4.

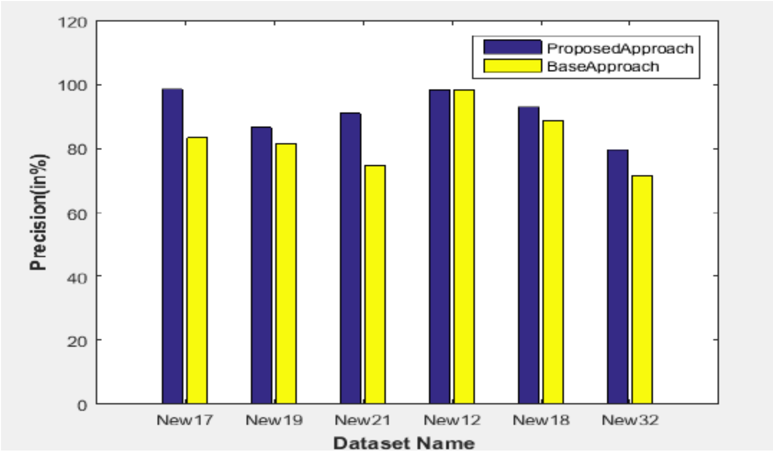


Fig 5: Precision graph for different context groups of dataset

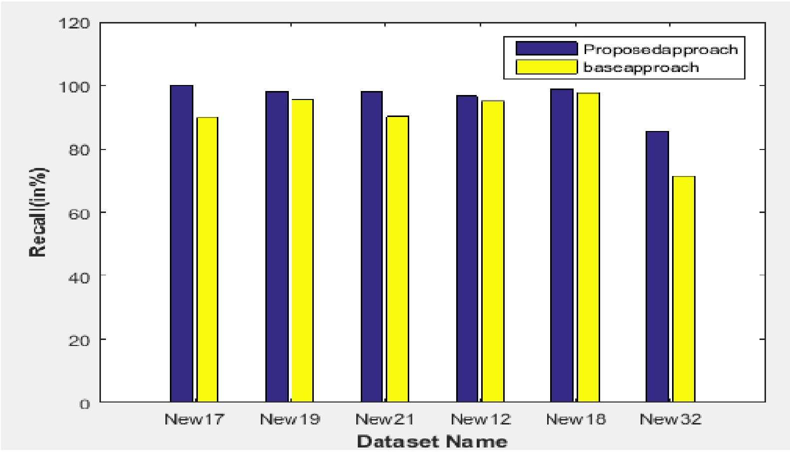


Fig 6: Recall graph for different context groups of dataset

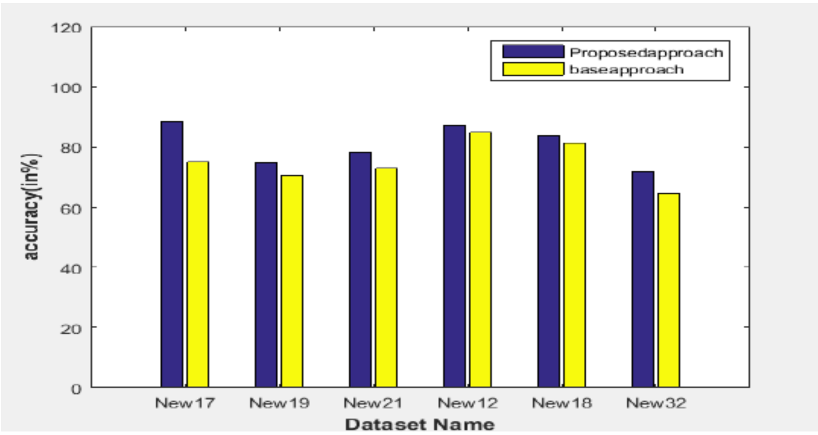


Fig 7: Accuracy graph for different context groups of dataset

VII. CONCLUSION AND FUTURE SCOPE

Gaussian filtering based novel context estimation and tensor decomposition system is proposed. Tensor formation and context estimation is used in this research paper to minimize the semantic gap while the tensor decomposition is used to find the best correlation between the context group images. Due to minimization of semantic gap the accuracy improves significantly.3-level tucker decomposition is adopted to model the framework for better correlation among the context groups. The results are compared between filtered context groups and unfiltered context groups. The evidence of the effectiveness of the proposed algorithm can be seen from the results and discussion section. In future, deep learning techniques can be used for better accuracy as well as to reduce the searching time.

REFERENCES

[1]. BahramiS, Abadeh M. S, Automatic Image Annotation Using an Evolutionary Algorithm (IAGA),7th International Symposium on Telecommunication, pp. 320–325, 2014.
[2]. Tariq A, Foroosh H, Feature-Independent Context Estimation for Automatic Image Annotation, IEEE,pp.1958-1965,2015
[3]. S. A. Zhu, C.-W. Ngo, and Y.-G. Jiang, “Sampling and ontologically pooling web images for visual concept learning,” IEEE Trans. on MM, vol. 14, no. 4, pp. 1068–1078, 2012.
[4]. X.-R. Li, C. G. M. Snoek, and M. Worring, “Learning social tag relevance by neighbor voting,” IEEE Trans. on MM, vol. 11, no. 7, pp. 1310–1322, 2009.
[5]. M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid, “TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation,” in ICCV, 2009.
[6]. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp.119-126,2003.
[7]. V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In Advances in neural information processing systems,2003.
[8]. S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1-8,2004.
[9]. S. Moran and V. Lavrenko. Optimal tag sets for automatic image annotation. In Proceedings of the British Machine Vision Conference,pp. 1-11, 2011.
[10]. D. Liu, X.-S. Hua, L.-J. Yang, M. Wang, and H.-J. Zhang, “Tag ranking,” in WWW,pp. 351-360, 2009.
[11]. A.Rae,B.Sigurbj¨ornsson,andR.-V.Zwol, “Improving tag recommendation using social networks,” in RIAO, 2010.
[12]. http://www.ci.gxnu.edu.cn/cbir/Dataset.aspx
[13]. Tran, T.-H.; Tran, X.-H.; Nguyen, V.-T.; Nguyen-An, K. Building an Automatic Image Tagger with DenseNet and Transfer Learning; IEEE: Piscataway, NJ, USA, pp. 34–41,2019
[14]. Chu, Y., et al.: Automatic image captioning based on ResNet50 and LSTM with soft attention. In: Wireless Communications and Mobile Computing, 2020
[15]. R. Subash November 2019 Journal of Physics Conference Series 1362:012096 : Automatic Image Captioning Using Convolution Neural Networks and LSTM.
[16]. Bai, S., An, S.: A Survey on automatic image caption generation. Neurocomputing 311, 291–304 (2018)
[17]. Tanti, M., Gatt, A., Camilleri, K.: What is the Role of Recurrent Neural Networks (rnns) in an Image Caption Generator? in: proceedings of the 10th International Conference on Natura l Language Generation, pp. 51–60 (2017).
[18]. Kamal, A.h., Jishan, M.a., Mansoor, N.: Textmage: The Automated Bangla Caption Generator Based on Deep Learning. in: 2020 International Conference on Decision aid Sciences and Application (DASA), pp. 822–826. IEEE (2020)
[19]. Teera Siriteerakul , Kunlabut Suriyakanon ,Sofia Sarideh , (2018) " Automatic Restaurant Image Tagging " , International Journal of Electrical, Electronics and Data Communication (IJEEDC) , pp. 1-4, volume-6,issue-4
[20]. Marielet Guillermo, Robert Kerwin Billones, Argel Bandala, Ryan Rhay Vicerra, Edwin Sybingco, Elmer P. Dadios, Alexis Fillone, "Implementation of Automated Annotation through Mask RCNN Object