Ontology supported semantic based image retrieval

Akif Gaşi [a,*], Mustafa Dağtekin [b] and Tolga Ensari [c]
[a] *Computer Engineering, Istanbul University-Cerrahpasa, Türkiye*
E-mail: akif.gasi@ogr.iuc.edu.tr
[b] *Computer Engineering, Istanbul University-Cerrahpasa, Türkiye*
E-mail: dagtekin@iuc.edu.tr
[c] *Department of Computer and Information Science, Arkansas Tech University, AR, USA*
E-mail: tensari@atu.edu

Editor: Pascal Hitzler, Wright State University, Bari
Solicited reviews: Claudia d'Amato, Università degli Studi di Bari, Italy; anonymous reviewer
Open review: Eva Blomqvist, Linköping University, Sweden

Abstract. In this study, a two-stage approach for developing a Semantic Based Image Retrieval system supported by Ontology is proposed. In the first stage, objects are detected with the Object Detection process from the image and a predicate describing the relationship between the two objects is determined with the developed Bi-directional Recurrent Neural Network (Bi-RNN) model. In the second stage, relations defined as <subject-predicate-object> are converted into Ontologies and used to search for semantically similar images. In the measurement of Semantic Gap, as the main problem encountered in the Semantic-Based Image Retrieval approach, it is proposed to calculate the number of similar relationships between two images by using entropy. By using the number of relationships (X) found in the image used for query purposes and the total number of relationships (Y) of the image with similar relationships that was found as a result of the query, the Semantic Gap between two images was calculated with the Joint Entropy method. The proposed approach has the characteristics of a new method used in this field and gives more effective results compared to other similar methods that are used in Semantic Based Image Retrieval by using Ontologies.

Keywords: Semantic based image retrieval, Ontology, Predicate detection, Object detection

1. Introduction

Image Retrieval (IR) is the process of searching in a database for images that have similarity according to a keyword or certain features (color, shape, texture) of an image used for query purposes. In search operations with Text Based Image Retrieval-TBIR and Content Based Image Retrieval-CBIR approaches that are used for Image Retrieval purposes, results that are not related to the desired result are also obtained. A keyword is used in searches made with TBIR, whereas, features such as color, shape, texture extracted from the image are used in searches performed with CBIR. As a result of the search process, images that do not match with keywords or features extracted from the image are also brought as a result. TBIR (Text-Based Image Retrieval) and CBIR (Content-Based Image Retrieval) processes face the fundamental problem known as Semantic Gap. The Semantic Gap is

*Corresponding author. E-mail: akif.gasi@ogr.iuc.edu.tr.

defined as the discrepancy between the textual or visual features used for query purposes and the semantic meaning associated with the image [1]. As the proposed solution for addressing the Semantic Gap problem is the Semantic-Based Image Retrieval (SBIR) approach. With Semantic-Based Image Retrieval, objects detected from an image and images with similar relationships defined between detected objects, are searched in a database

The first stage of the Semantic-Based Image Retrieval approach is Visual Relationship Detection (VRD) process. With Visual Relationship Detection, Object Detection from the image and an expression that defines the relationship between objects (Predicate Detection/Recognition) are determined respectively. In a study conducted in this field [2], objects are first detected in the image ("person," "motorcycle"), and an expression ("on") defining the relationship between the two objects is determined. As a result, a meaning of the image is inferred with the relation in the form of <person-on-motorcycle>, which is shown as <subject-predicate-object>.

The meaning inference made as the result the Visual Relationship Detection process can be utilized in the searching operations for other images containing the same meaning. The Semantic Web initiative represents the transition from the "Web of Documents" approach to the "Web of Data" approach, as it is used nowadays [3]. In this new approach brought by the Semantic Web, Ontologies are used for modeling concepts. The meaning derived from the Visual Relationship Detection process is transformed into a structure that computers can process using Ontologies and is utilized in the search process for other images that carry the same meaning.

In our study, the Visual Genome dataset was used for training the Visual Relationship Detection model and creating Ontologies [4]. The dataset contains object information, bounding box coordinates, class labels, and predicate defining the relationships between objects in the images. By using the information in the Visual Genome dataset, the Bi-RNN model used for Visual Relationship Detection was trained and an Ontology was created to model the relationship between objects. The generated Ontology was then utilized in the Semantic-Based Image Retrieval process.

The main contributions of our study can be summarized under the following headings:

- A novel ontology-supported approach is proposed for Semantic-Based Image Retrieval.
- By using the information in the Visual Genome dataset, the relationships with the created Ontology are modeled and a structure that computers could process and infer meaning was created (subject-predicate-object).
- The relations displayed by ontologies are utilized in the search process for searching other images with the same meaning.
- The use of ontologies resulted in more effective results in the retrieval of semantically similar images compared to current studies.

2. Related Work

Image Retrieval: In the field of Image Retrieval, fundamental studies have focused on Text-based Image Retrieval and Content-based Image Retrieval approaches. Search operations in both areas are performed based on the utilized keywords or the features associated with the image. This situation leads to the retrieval of unrelated results that may not have any relevance to the keywords or selected feature (even if two images have the same color, shape, and texture features, their contents can still be different). In a study conducted in this field [5], a comprehensive research was conducted on feature extraction (color, shape, texture) from images, system architecture, and proposed solutions to the encountered problems. In a research study [6] addressing the problem of Semantic Gap, which arises between low-level features (color, shape, texture) and high-level semantic concepts perceived by humans, an ontology-based model was created to define objects and their relationships detected from satellite images, and the created model was then used in the process of searching similar satellite images.

Object Detection: The first stage of image inference is the Object Detection process. Computer Vision systems utilize Deep Learning approaches in Object Detection processes. In a conducted study [7] by utilizing Deep Learning approaches, performance results were examined on Object Detection architectures (Region Proposal, Feature Pyramids, CNN), application areas (Salient object detection, Face detection, Pedestrian detection) and various datasets (PascalVOC, Microsoft COCO). Whereas, in another study [8], detailed information are given on Object Detection performance using Fast Region-based Convolutional Network (Fast-RCNN) architecture and different datasets

(Pascal VOC, Microsoft COCO). In this study, Object Detection processes were performed using the Detectron2[1] framework developed by Facebook Research and based on the Faster-RCNN [9] method.

Visual Relationship Detection: There are three different tasks (Task) in Visual Relationship detection operations. In the **Phrase detection** process, an expression describing the relationship between the objects is determined using the Bounding Box coordinates of the objects belonging to the image, and the area containing the relationship between the two objects (provided that IoU is > 0.5) is displayed with a general bonding box. In the **Relationship detection** process, besides specifying the class labels of the objects and a predicate that defines the relationship between the two objects, it should also correspond to IoU > 0.5 ratio with the exact reference data (Ground Truth) of the bounding boxes, which describes the positions of objects in the image. Whilst in the **Predicate detection** process, the aim is to find a predicate ("on", "with") that describes the relationship between objects using the bounding box coordinates (x, y, w, h) and class labels ("person", "motorcycle"). The accuracy of the Predicate Detection process is measured without the influence of the Object Detection algorithm (or without the influence of all other possible factors). In the studies conducted in the field of Visual Relationship Detection [2, 10, 11], performance measurement is being made for all three tasks, and the results obtained are being compared in different datasets. (Visual Relationship Dataset, Visual Genome) In the first stage of the Semantic-Based Image Retrieval process proposed in this study, Predicate Detection was performed referring to a conducted study [12].

Scene representation: In order to extract meaning from an image, it is necessary to represent objects and their relationships using an appropriate method. In a conducted study [13], objects and their relationships were modeled using the Scene Graph method. Whereas, in another study [14], the created Scene Graph was utilized in the process of searching for images that have a similar structure. In this study, the representation of objects and the relationship between them in the proposed second stage was made with Ontologies.

3. Materials and Methods

Figure 1 and Figure 2 show the working steps of the two-stage system proposed in this study. In the first stage, objects were determined with the Object Detection algorithm and dual combinations of the determined objects (Object Pairs) were created. With the Word embedding method, class labels belonging to two objects were converted to 100-dimensional vectors (Subject Vector, Object Vector) and the relationship representation (spatial information) between two objects was done according to the equation given in Eq. (2). A predicate defining the relationship between two objects was determined using Bi-RNN model and as the result the meaning of the image was inferred with the relation in the subject-predicate-object structure.

Whereas, in the second stage, the relationship in the subject-predicate-object structure was converted to Ontologies (owl_Relation) and was used to search for images containing semantically similar relationships. The second stage steps of the proposed system are shown in Fig. 2.

3.1. Dataset

In order to determine the relationship between two objects by using Bi-RNN for training of the created model and using it in searching operation, the Visual Genome dataset was utilized to generate the ontologies. In this study, after performing necessary data filtering and transformation processes, for the first 5000 images included in the Visual Genome dataset, a total of 59724 defined association samples were selected. Selection of the data was performed based on 100 object categories and 70 predicate categories, taking as a reference a conducted study [2]. The dataset used in the training and testing stage of the model was divided as 75% train, 15% validation, 10% test data. The data selected from the dataset, the test results of training the Model and the Ontology creation processes will be explained in the relevant sections.

[1] https://detectron2.readthedocs.io/en/latest/

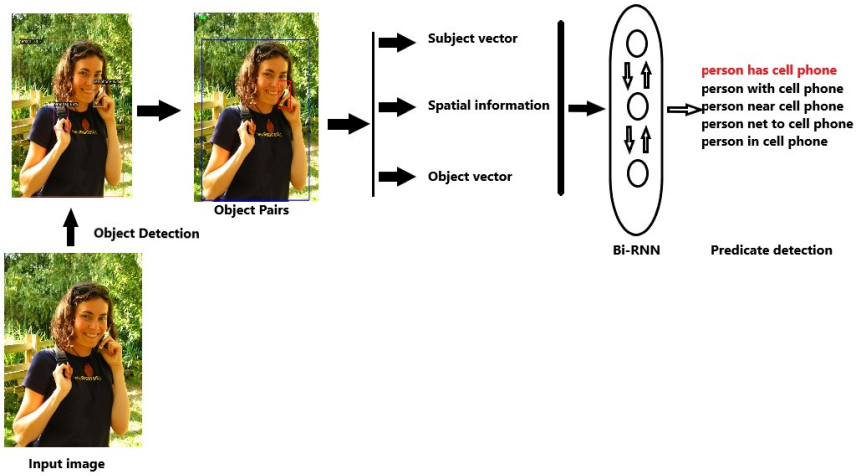

Fig. 1. System architecture.

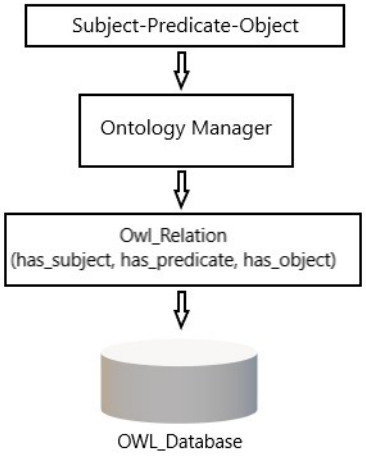

Fig. 2. Ontology creation and search process.

3.2. *Word Embedding*

Word Embedding is used in Natural Language Processing (NLP) applications to display the semantic similarities between words. Each word is transformed into a numerical vector. Therefore, in our study, when describing the relationship between two objects, the class labels of the objects (e.g., person, motorcycle) were transformed into a 100-dimensional vector using the word embedding method. Words that have semantic relationships are located close to each other in the vector space. In this study, the word embedding process was carried out by using the Continuous Bag of Words (CBOW) method [15] of the Word2Vec algorithm. The Gensim[2] library was used for the creating word vectors. During the training stage of the Word2Vec algorithm the data in the subject-predicate-object structure selected from the Visual Genome dataset specific to our problem were used. Example of data selected from the dataset and used in training the Word2Vec algorithm:

[2]https://radimrehurek.com/gensim/index.html

[['shade', 'on', 'street'], ['car', 'has', 'headlight'], ['sign', 'on', 'building'], ['tree trunk', 'on', 'sidewalk'], ['man', 'has', 'shirt'], ['sidewalk', 'next to', 'street'], ['car', 'has', 'back'], ['man', 'has', 'glasses'], ['parking meter', 'on', 'sidewalk'], ['man', 'has', 'shoes']]

The word vectors generated by the Word2Vec algorithm were utilized as the input data for training the Bi-RNN model.

3.3. Model

In this study, a redesigned Bi-RNN model, which is shown in Fig. 3 and specific to our problem, was used to describe the relationship between two objects.

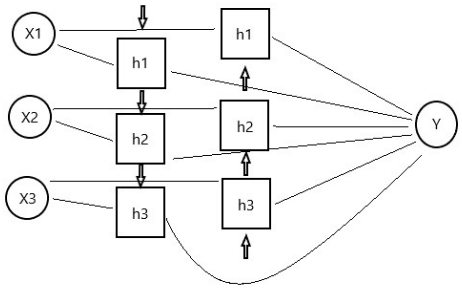

Fig. 3. Modified Bi-RNN Model.

The model was designed according to Eq. (1). The first term of the equation represents the forward sequence of information, while the second term represents the backward sequence. The model takes the subject, spatial information, and object data (x1, x2, x3) as input. The output, displayed by $Y = R^K$ equation, provides the distribution generated for a total of K predicates selected for the first 5000 images.

$$Y = \left(W_{h1y}^{\rightarrow} h_1^{\rightarrow} + W_{h2y}^{\rightarrow} h_2{}^{\rightarrow} + W_{h3y}^{\rightarrow} h_3{}^{\rightarrow}\right) + \left(W_{h1y}^{\leftarrow} h_1{}^{\leftarrow} + W_{h2y}^{\leftarrow} h_2{}^{\leftarrow} + W_{h3y}^{\leftarrow} h_3{}^{\leftarrow}\right) + by \tag{1}$$

During the training stage of the model, the selected data from the Visual Genome dataset was used. The data was organized in a specific order and used as the input for the model. In this order, x1-subject vector, x2-spatial information, x3-object represent vector data. The relationship representation (spatial information) between two objects was made according to Eq. (2). The x and y used in this equation represent the center coordinates of the bounding box of an object, while w and h represent the width and height of the bounding box, respectively.

$$Z = \left[\frac{x1 - x2}{w1}, \frac{y1 - y2}{h1}, \frac{w2}{w1}, \frac{h2}{h1}\right] \tag{2}$$

In this study, the process of predicate detection was performed. In the process, the class labels belonging to two objects and the location information in the image were given, referred to as Subject and Object, and the expression that would define the relationship between them was tried to be determined. This process can be represented using Eq. (3):

$$Rel\,(s, p, o) = P_i\,(O_1)\,(BiRNN\,(O_1, O_2, Z))\,P_j\,(O_2) \tag{3}$$

In the Eq. (3), the expression Pi(O1) gives the probability that the object detected from the image belongs to class i, whereas, the expression Pj(O2) gives the probability that the object belongs to class j in the Object Detection

process. On the other hand, Bi-RNN (O1,O2,Z) shows the expression that determines the predicate and defines the relationship between two objects.

The developed model was trained on the Turkish National e-Science e-Infrastructure (TRUBA)[3] provided by TÜBİTAK ULAKBİM High Performance and Grid Computing Center. The Recall@X metric was used as the performance measure. Recall@X gives the exact ratio of ground truth within the Top-K result produced by the model. The performance results obtained on the test dataset of the model will be given in the relevant section.

3.4. Ontology creation

Ontologies are used to describe entities. In order to describe entities, classes belonging to a specific Domain are created and relations between classes are determined. In this study, an Ontology was created as shown in Fig. 4 from the data in the subject-predicate-object structure defined in the Visual Genome dataset which will be used in the Semantic Based Image Retrieval process. The created ontologies were used in the search for semantically similar images. Search results are explained in the relevant section.

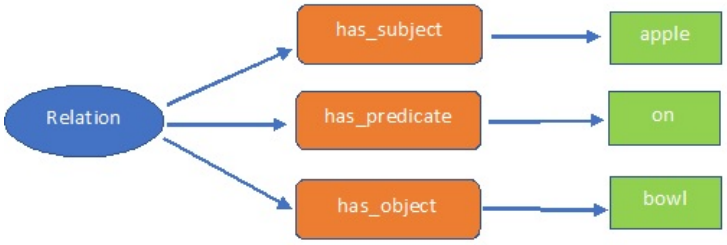

Fig. 4. Relation ontology.

4. Experiments

In this study, a Bi-RNN model was trained using data selected from the Visual Genome dataset to determine the expression that describes the relationship between objects detected in images. For training purposes of the model, the Subject and Object data were converted into 100-dimensional vectors using Word2Vec. By using Bounding Box coordinates that describe the relationship between two objects, a 4-dimensional vector was created according to Eq. (2). The data representing the relationship between the two objects, was combined into a 12-dimensional vector [x1, y1, w1, h1; x2, y2, w2, h2; Z], The model was trained using a total of 212 features in the form of subject_vector-spatial_information-object_vector. The performance results of the model trained with the Visual Genome dataset compared to other studies on the same dataset are given in Table 1.

Table 1
Zero-Shot Predicate Prediction Model results.

Method	Recall@50	Recall@100
Ref [11]	65.20	67.10
Ref [12]	85.02	91.77
Proposed	90.00	91.00

[3]https://www.truba.gov.tr/index.php/en/main-page

4.1. Meaning inference from the image

In the process of testing the model on an image that is not in the training dataset, firstly objects from within the image were detected with Detectron2, and then a predicate was determined that defines the relationship between the two objects by creating dual combinations of the objects. The obtained results are shown in Fig. 4.

Fig. 5. Meaning inference from the image.

4.2. Searching process using ontologies

Meaning inference was made by determining a predicate describing the relationship between objects perceived from an image. The relationship in the subject-predicate-object structure was converted into ontologies and used in the search for semantically similar images. The result of searching for semantically similar images using ontology is shown in Fig. 5.

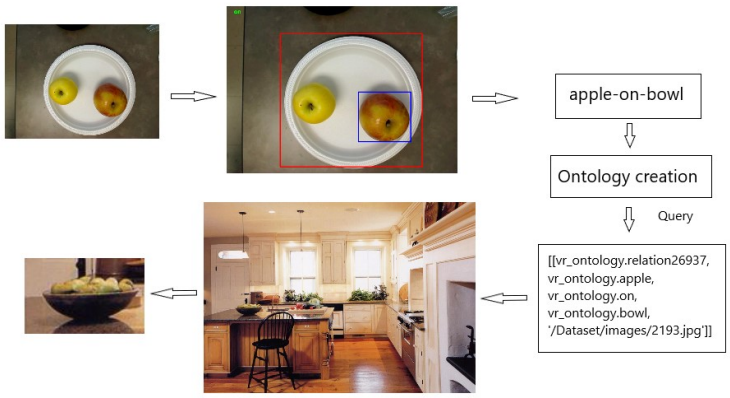

Fig. 6. Searching process using ontology.

4.3. Measuring the Semantic Gap

In the present study, in the measurement of Semantic Gap, as the main problem encountered in the Semantic-Based Image Retrieval approach, it is proposed to calculate the number of similar relationships between two images by using entropy. The Entropy method, that is used to measure the amount of information in Information Theory, is utilized for measuring the Semantic Gap between two images. By using the number of relationships (X) found in the image used for query purposes and the total number of relationships (Y) of the image with similar relationships that was found as a result of the query, the Semantic Gap between two images was calculated with the Joint Entropy method.

$$H(X,Y) = -\sum_{x,y} p(x,y)\log(x,y) \tag{4}$$

According to Eq. (4), as a result of the calculation using the relationship numbers between the two images, 3 cases were examined:

$$H = \begin{cases} 1, & \text{"there is No semantic gap"} \\ 0 < H < 1, & \text{"there is a semantic gap"} \\ 0, & \text{"No similar images"} \end{cases} \tag{5}$$

In the search process using the sample image shown in Fig. 7, for “apple” and “bowl” objects detected from the image, an image similar to the defined “apple-on-bowl” relationship was found as a result.

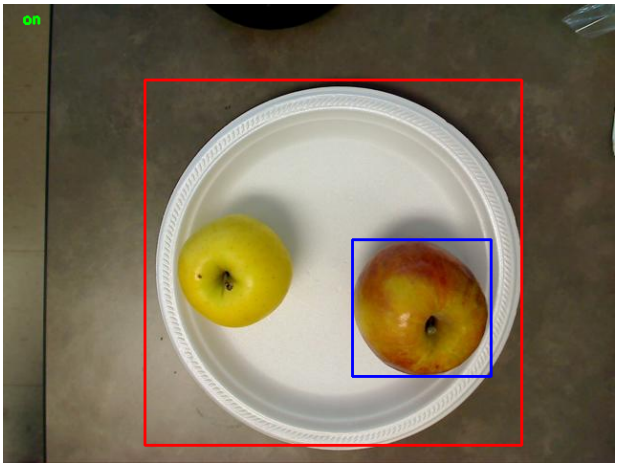

Fig. 7. Searching process using ontology.

As a result of the query, information about the image with a similar relationship is given in Table 2:

As a result of the search process, a relationship (apple-on-bowl) that was found in the image used for query purposes was also found in another image. The result of the Semantic Gap measurement between the two images is given in Table 3.

According to the calculation by using entropy, from six ground truth relationships defined in the image found in the query result only one relationship matched and based on the calculation result according to Eq. (5) it was determined that there is a Semantic Gap.

Table 2
Query results.

Query:	apple-on-bowl
Number of ground truth relations for 2193.jpg:	6
Images with similar relations:	['2193']
Number of images:	1
Number of relations:	1
Result image:	Fig. 8
Ontology of relation for image:	[[vr_ontology.relation26937, vr_ontology.apple, vr_ontology.on, vr_ontology.bowl, '/Dataset/images/2193.jpg']]

Fig. 8. Result image.

Table 3
Semantic gap measuring results.

Number of relations for 2193.jpg:	1
Number of ground truth relations for 2193.jpg:	6
Calculated Entropy for 2193.jpg:	0.59
Result:	"there is Semantic Gap"

5. Discussion

In the first step of the two-stage Semantic-Based Image Retrieval approach proposed in this study, as shown in Fig. 1, with the aim to determine the relationship between two objects and to infer meaning, the model in use was trained with 212 features arranged as Object vector, Spatial Information, Subject vector. In other studies [2, 12], more features extracted from the objects in the image were used in training the model. In this study, by using fewer features compared to other studies, better results were obtained, according to Recal@50 and also closer results were obtained, according to Recal@100.

In the second stage as shown in Fig. 2, by using the proposed ontologies more effective results were obtained in the search process compared to the Scene Graph structure used in other studies. In a previously conducted study [14], the Scene Graph structure of the image selected for the search process was used in the search process, in our study, an expression defining the objects belonging to the image and the relationship between them is determined

and the relationship in the subject-predicate-object structure is converted into Ontologies. With Ontologies objects, and their relations are better modeled compared to the Scene Graph structure and it is used to search for semantically similar images in the solution of the problem expressed as Semantic Gap.

6. Conclusion

In this study, a new approach was proposed for solving the problem of Semantic Gap in Image Retrieval using Ontologies. The relationship created between the objects detected in the image was converted into Ontologies and used to search for semantically similar images. The proposed method can be considered as a more general approach to the Semantic Based Image Retrieval process. Compared to the Scene Graph structure used in other studies, more efficient results were obtained in modeling concepts by using Ontologies, converting information into a form that computers can process, and searching for semantically similar images by inferring meaning from this information.

In our future studies, we aim to extend the proposed approach to a more general solution for use in Semantic Based Image Retrieval operations with simultaneous execution of more efficient feature extraction and Ontology generation to identify relationships between objects detected from images.

Acknowledgements

The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Declarations

Conflict of interest: The authors declare that they have no conflict of interest.

Data availability: The data that support the findings of this study are available on [https://homes.cs.washington.edu/~ranjay/visualgenome/api.html]

References

[1] Ying, L., Dengsheng, Z., Guojun, L., Wei-Ying, M., A survey of content-based image retrieval with high-level semantics, Pattern Recognition, 2007, 40 (1), 262–282.

[2] Lu, C., Krishna, R., Bernstein, R., Fei-Fei, L., Visual Relationship Detection with Language Priors, European Conference on Computer Vision, 2016.

[3] Antoniou, G., v.Harmelen, F., A Semantic Web Primer, 2nd, MIT Press, 2008, Cambridge, Massachusetts London, England.

[4] R. Krishna., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L-J., Shamma, D., Bernstein, M., Fei-Fei, L., Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, International. Journal of Computer Vision, 2017, 123 (1), 32–73.

[5] Yong, R., Thomas, H., Image Retrieval: Current Techniques, Promising Directions, and Open Issues, Journal of Visual Communication and Image Representation, 1999, 62, 39–62.

[6] Dhobale, D.D., Patil, B.S., Patil, S.B., Ghorpade, V.R., Semantic understanding of Image content, International Journal of Computer Science Issues, 2011, 8 (3) 191–196.

[7] Zhao, Z.Q., Zheng, P., Xu, S-T., Wu, X., Object Detection with Deep Learning: A Review, IEEE Transactions on Neural Networks and Learning Sysems., 2019, 30 (11), 3212–3232.

[8] Girshick, R., Fast R-CNN, IEEE International. Conference on Computer Vision, 2015.

[9] Shaoqing, R., Kaiming, H., Girshick, R., Sun, J., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions and Pattern Analysis and Machine Intelligence, 2017, 39 (6), 1137–1149.

[10] Yaouhi, Z., Shuqiang, J., Deep structured learning for visual relationship detection, Procedings of AAAI Conference on Artificial Intelligence, 2018.

[11] Zhang, L., Zhang, S., Shen, P., Zhu, G., Shah, S.A.A., Bennamoun, M., Relationship detection based on object semantic inference and attention mechanisms, International Conference on Multimedia Retrieval, 2019, 68–72.

[12] Liao, W., Rosenhahn, B., Shuai, L., Yang, M.Y., Natural language guided visual relationship detection, IEEE Conference on Computer Vision and Pattern Recognition Worshops, 2019, Long Beach, CA, USA, 444–453.
[13] Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L., Scene graph generation by iterative message passing, IEEE Conference on Computer Vision and Pattern Recognition, 2017, 3097–3106.
[14] Johnson, J., Krishna, R., Stark, M., Li, L-J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L., Image retrieval using scene graphs, IEEE Conference on Computer Vision and Pattern Recognition, 2015, 3668–3678.
[15] Mikolov, T., Chen, K., Corrado, G., Dean, J., Efficient estimation of word representations in vector space, International Conferance on Learning Representation, 2013.