Semantic Web -1 (2024) 1–13
DOI 10.3233/SW-243660
IOS Press
CORRECTED PROOF

Ontology supported semantic based image retrieval

Akif Gaşi [a,*], Tolga Ensari [b] and Mustafa Dağtekin [c]

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

Editor: Cogan Shimizu, The Knowledge and Semantic Technologies (KASTLE) Laboratory, Wright State University, USA
Solicited reviews: Md Kamruzzaman Sarker, Department of Computer Science, Bowie State University, Maryland, USA; Jean Vincent Fonou-Dombeu, Department of Computer Science, University of KwaZulu-Natal, South Africa; Three anonymous reviewers

Abstract. In this study, a two-stage approach for developing a Semantic-Based Image Retrieval system supported by Ontology is proposed. In the initial stage, the Object Detection process is employed to identify objects within the image. Subsequently, a predicate describing the relationship between these two objects is determined using the developed Bi-directional Recurrent Neural Network (Bi-RNN) model. In the second stage, relations defined in the form of <subject-predicate-object> are transformed into Ontologies and utilized to search for images that are semantically similar. In addressing the primary challenge of Semantic Gap within the Semantic-Based Image Retrieval approach, the proposed solution involves measuring the number of similar relationships between two images through the utilization of entropy. The Semantic Gap between two images was computed using the Joint Entropy method, leveraging the number of relationships (X) identified in the query image and the total number of relationships (Y) in the image with similar relationships obtained as a query result. The proposed approach exhibits characteristics of a novel method within this field, distinct from other similar methods employed in Semantic-Based Image Retrieval through the utilization of Ontologies. In the performance measurement of the developed model, 91% accuracy was obtained according to the Recall@100 (Top-5 accuracy) result.

Keywords: Semantic based image retrieval, ontology, predicate detection, object detection

1. Introduction

In recent years, both the size and quantity of digital images have significantly increased. This density has made manual classification of image content nearly impossible. As a result, the automatic classification of image content has become essential. In today's rapidly evolving landscape, where automatic monitoring and recommendation systems are gaining importance, the computerized detection of image content and the generation of various information and alert scenarios are becoming increasingly significant. Image Retrieval (IR) is the process of querying a

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

database for images that exhibit similarity based on a specific keyword or certain attributes (color, shape, texture) related to an image. Fundamental research in the field of Image Retrieval has been concentrated on Text-Based Image Retrieval (TBIR) and Content-Based Image Retrieval (CBIR) approaches. Search operations in both areas are performed based on the used keyword or image features. In search operations with TBIR, a keyword is utilized, whereas in search operations with CBIR, features extracted from the image, such as color, shape, and texture, are employed. This situation can lead to the retrieval of results unrelated to the keyword or selected features (although the color, shape, and texture features of two images may be the same, their contents can differ). As a result of the search process, images that do not match with the keywords or extracted features from the image are also presented as results.The fundamental issue encountered in TBIR and CBIR processes is referred to as the Semantic Gap. The Semantic Gap is defined as the discrepancy between the textual or visual features used for query purposes and the semantic meaning associated with the image [25]. The suggested solution to address the Semantic Gap problem is the Semantic-Based Image Retrieval (SBIR) approach. Semantic-Based Image Retrieval involves searching a database for images with similar relationships between the perceived objects in an image and the relationships defined among those perceived objects.

The first stage of the Semantic-Based Image Retrieval approach is the Visual Relationship Detection (VRD) process. In Visual Relationship Detection, the image undergoes Object Detection followed by determining an expression (Predicate Detection/Recognition) that describes the relationship between objects. In a study conducted in this field [10], the initial step involves detecting objects in the image (such as "person" and "motorcycle"), followed by determining an expression ("on") that defines the relationship between the identified objects. Consequently, the image's meaning is inferred with the relation represented in the form of <person-on-motorcycle>, expressed as <subject-predicate-object>. The semantic inference obtained through the Visual Relationship Detection process can be utilized in the search for other images containing the same meaning. The Semantic Web initiative represents a transition from the current "Web of Documents" approach to the "Web of Data" approach, aiming to create a network of interconnected information rather than just a network of documents [1]. In this new approach introduced by the Semantic Web, ontologies are used for modeling concepts. The inferred meaning obtained through the Visual Relationship Detection process is transformed into a structure that computers can process using ontologies. This structured representation is then employed in the search for other images carrying the same meaning.

In this study, the Visual Genome dataset was utilized for training the Visual Relationship Detection model and constructing ontologies. [8]. The dataset contains object information for images, bounding box coordinates, class labels, and expressions (predicates) defining relationships between objects. Using the information from the Visual Genome dataset, a Bi-RNN model was trained for the Visual Relationship Detection process, and an ontology was created to model the relationships between objects. The generated ontology was then employed in the Semantic-Based Image Retrieval process.

In this study, to measure the Semantic Gap, which is the fundamental issue encountered in the Semantic-Based Image Retrieval approach, it is proposed to use entropy to calculate the number of similar relationships between two images.

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

- A novel approach was proposed for solving the problem of Semantic Gap in Image Retrieval using Ontologies.
- The ontology, crafted from data within the Visual Genome dataset, models relationships, establishing a structure that computers can process and infer meaning in the form of <subject-predicate-object>.
- The relationship represented by ontologies was used in the process of searching for other images that convey the same meaning.
- In the measurement of the Semantic Gap, it is suggested to use the Joint Entropy method to calculate the number of similar relationships between two images.
- The proposed method can be considered as a more general approach to the Semantic-Based Image Retrieval process.

2. Related work

2.1. Image retrieval

In the field of Image Retrieval, fundamental studies have focused on Text-based Image Retrieval and Content-based Image Retrieval approaches. In a study conducted in this field [26], a comprehensive research was conducted on feature extraction (color, shape, texture) from images, system architecture, and proposed solutions to the encountered problems. In the Text-Based Image Retrieval approach, the desired results cannot be obtained in search operations due to the mismatch caused by the semantic difference between the content of images and keywords. The main reason for this is the inability to fully express the meaning of the image with the label information (Tag) used to describe the image. The fundamental method used in the text-based document retrieval approach is expressed as Latent Semantic Indexing (LSI). In this method, clustering is performed to identify semantically similar words. In the search process, not only the words matching the keywords but also the words in the same cluster are searched. In a study [28], the LSI method was used in accessing web documents along with the image attributes found in the document. In the Content-Based Image Retrieval approach, the fundamental method used in search operations is defined as Query-by-visual-example (QBVE). In this method, low-level features (such as color, shape, texture) are extracted from the image used for querying, and as a result of the search process, images matching these features are retrieved from the database. In a study in this field [16], a method referred to as Query-by-semantic-example (QBSE) is employed, where in addition to image features, concept vectors describing the image are also utilized. With the development of web applications, there has been an increased necessity to identify image content on the web. Tag structures are used for describing image content. In many cases, the label information used to describe image content may lead to the use of inappropriate labels due to people's different perceptions of images. In a study [11], an image similarity graph was used in addition to image features. In a research study [4] aimed at tackling the issue of Semantic Gap, which emerges between low-level features (such as color, shape, and texture) and the high-level semantic concepts perceived by humans, an ontology-based model was developed. This model was designed to define objects and their relationships identified from satellite images. Subsequently, the created model was applied in the process of searching for similar satellite images. In our study, we propose a workflow that can be regarded as a more comprehensive approach to the process of Semantic-Based Image Retrieval.

2.2. Object detection

The initial phase of image inference involves the Object Detection process, where Computer Vision systems employ Deep Learning approaches. In a conducted study [29], the examination of performance results centered on Object Detection architectures (Region Proposal, Feature Pyramids, CNN), application areas (Salient object detection, Face detection, Pedestrian detection), and various datasets (PascalVOC, Microsoft COCO). This analysis was carried out by employing Deep Learning approaches. In another study [6], comprehensive details are provided on the performance of Object Detection utilizing the Fast Region-based Convolutional Network (Fast-RCNN) architecture across different datasets (Pascal VOC, Microsoft COCO). YOLO (You Only Look Once) is one the preferred object detection system in object detection processes. With YOLO, object detection can be performed over 9000 object categories [17]. In this study, Object Detection processes were performed using the Detectron2[1] framework developed by Facebook Research and based on the Faster-RCNN [20] method.

2.3. Visual relationship detection

In visual relationship detection, the goal is to determine an expression that describes the relationship between two objects detected in an image. The process of visual relationship detection involves extracting semantic meaning from the image. There are three distinct tasks in visual relationship detection. Performance results for each of these three tasks, namely Phrase Detection, Relationship Detection, and Predicate Detection/Recognition, have been examined in various studies [10,23,24,27]. The semantic inference of an image is performed using N detected objects in the

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

image and K predicates that describe the relationships between objects. In this case, the complexity of learning all possible relationships is expressed as $O(N^2K)$, and obtaining training data examples for all possible relationships is challenging. In the process of extracting meaning from an image, it is envisioned to only consider a useful number of relationships [2]. Defining the relationship between two objects is related to their interaction with each other (Spatial Relationships). This interaction is expressed with a predicate that defines the relationship between two objects, such as a spatial position (on, under) or an action (holds, moves). In a study in this field [15], the relationship between objects was represented in the form of subject-predicate-object using visual and spatial features of the objects. In the first stage of the Semantic-Based Image Retrieval process proposed in this study, Predicate Detection was performed referring to a conducted study [9].

2.4. Scene representation

To derive meaning from an image, it is essential to represent objects and their relationships using an appropriate method. In a conducted study [22], objects and their relationships were modeled using the Scene Graph method. In the Semantic Web approach, Ontologies are used to model concepts and establish relationships between them. In a general description, Ontologies are utilized to provide a clear and formal representation of concepts within a specific domain. In a study conducted on ontologies developed for multimedia applications [5], a new schema named Multimedia Web Ontology Language (MOWL) has been proposed for the representation of multimedia content. With the proposed schema, the term "concepts" is used to represent concepts related to real-world entities (Real World Entities), and "media objects" is used for the representation of multimedia concepts. In the second stage of this study, the representation of objects and their relationships was accomplished using Ontologies.

2.5. Semantic-based image retrieval

The Semantic-Based Image Retrieval approach is used to address the Semantic Gap problem. Low-level features (color, shape, texture) extracted from regions identified in the image through Object Detection are transformed into ontologies. This transformation aims to convert these features into a structure that computers can process and derive meaning from. In a study [12], an Object Ontology is created from low-level features, and keywords are defined for objects. The defined keywords are used in the representation of high-level concepts, and the search process is conducted based on these keywords. In the Ontology-Based Image Retrieval system created for nature images [18], the process involves searching for semantically similar images using query sentences created by users. In another study [14], an image retrieval system is proposed for the domain of the Asteroidae flower using an ontology that defines classes, subclasses, and properties. Whereas, in another study [7], the created Scene Graph was utilized in the process of searching for images that have a similar structure. In this study, a two-stage approach for Semantic-Based Image Retrieval using Ontologies has been proposed to address the Semantic Gap problem.

2.6. Similarity measurement

In this study, it is proposed to use the Entropy method to measure the Semantic Gap, which is the fundamental problem encountered in the Semantic-Based Image Retrieval approach. In information theory, the number of bits required to encode data that will be stored, transmitted, or compressed between computers is determined by the Entropy method developed by Claude E. Shannon [19]. In a study [3], a four-step process for measuring the Semantic Gap has been proposed. In the first step, image features were defined with ontologies. The color, size, texture, and location features of the image were divided into four parts, and an object ontology was created. In the second step, combinations of image features were defined, and using the Entropy method, the amount of information contained in the image was measured. In the third step, combinations of image features for query purposes were defined by users, and again using the Entropy method, the amount of information was measured. In the fourth step, the Semantic Gap measurement was made by examining the relationship between the measured amounts of information obtained from image features. In another study [21], the proposed Content-Based Image Retrieval (CBIR) approach involves measuring the similarity between two images using the Graph Matching method. Additionally, a language

model was developed to describe the similarities between two images in written form. In the measurement of the Semantic Gap, in thıs study suggested to use the Joint Entropy method to calculate the number of similar relationships between two images.

3. Materials and methods

Figures 1 and 2 illustrate the operational steps of the two-stage system proposed in this study. In the initial stage, objects were identified using the Object Detection algorithm, and dual combinations of these identified objects (Object Pairs) were generated. Utilizing the Word Embedding method, class labels associated with two objects were transformed into 100-dimensional vectors (Subject Vector, Object Vector). The representation of the relationship (including spatial information) between the two objects was accomplished following the equation provided in Eq. (2). A predicate, delineating the relationship between two objects, was ascertained through the Bi-RNN model. Consequently, the meaning of the image was inferred, expressed within the subject-predicate-object structure.

In the second stage, the relationship structured in the subject-predicate-object format was transformed into Ontologies (owl_Relation) and employed to search for images containing semantically similar relationships. The procedural steps of the second stage in the proposed system are depicted in Fig. 2.

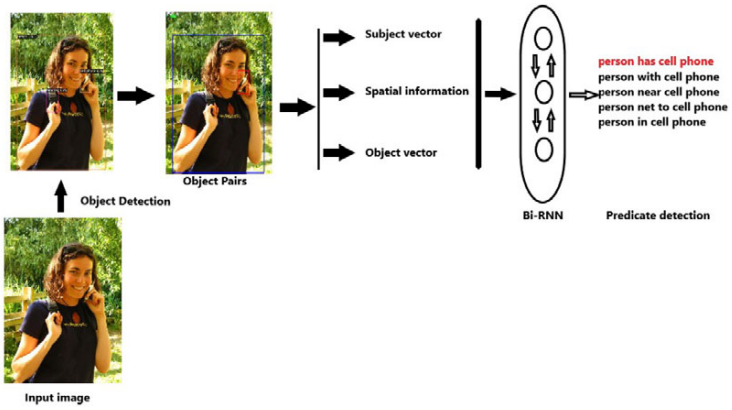

Fig. 1. System architecture.

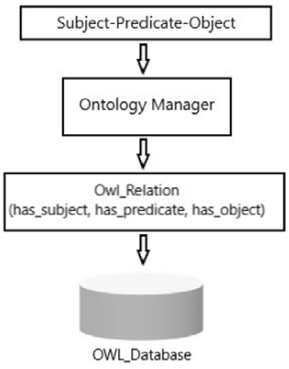

Fig. 2. Ontology creation and search process.

3.1. Dataset

To establish the relationship between two objects, Bi-RNN was employed for training the created model, and subsequently, it was used in the search operation. The Visual Genome dataset played a crucial role in generating the ontologies for this purpose. In this study, following essential data filtering and transformation processes, a total of 59,724 defined association samples were selected from the first 5,000 images within the Visual Genome dataset. The data selection was based on 100 object categories and 70 predicate categories, referencing a previous study [10]. For the training and testing phases of the model, the dataset was divided into 75% for training, 15% for validation, and 10% for testing data. The details regarding the selected data from the dataset, the training results of the model, and the processes of Ontology creation will be explicated in the relevant sections.

3.2. Word embedding

Word Embedding is employed in Natural Language Processing (NLP) applications to represent and demonstrate the semantic similarities between words. Each word is converted into a numerical vector using the word embedding method. Therefore, In our study, when delineating the relationship between two objects, the class labels of the objects (e.g., "person," "motorcycle") were converted into a 100-dimensional vector through the utilization of the word embedding method. Words with semantic relationships are positioned closely to each other in the vector space. In this study, the word embedding process was executed through the application of the Continuous Bag of Words (CBOW) method [13] within the Word2Vec algorithm. The Gensim[2] library was employed for creating word vectors in this study. 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 utilized. An example of the data selected from the dataset and employed in training the Word2Vec algorithm is as follows:

[['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, was used to describe the relationship between two objects. In the Bi-RNN architecture depicted in Fig. 3, the vector X_t represents the input sequence at time t, h_t denotes the hidden

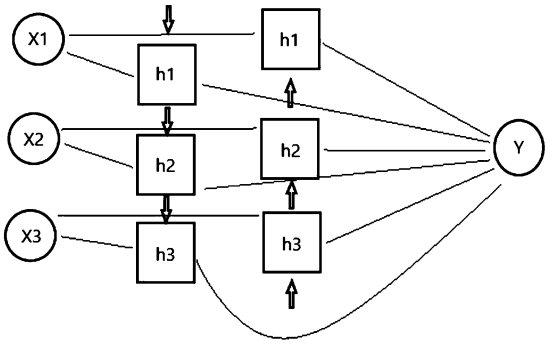

Fig. 3. Modified Bi-RNN model.

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

layer, and Y represents the output data corresponding to the input. When determining the outcome information at time t, Bi-RNN utilizes information from both the previous and next nodes in the sequence.

The Bi-RNN model performs the computation process in both the forward and backward directions in the hidden layer. In visual relationship detection processes, the arrangement of subject and object information in the expressed relationship <subject-predicate-object> affects the generated outcome (predicate). Therefore, the ordering of subject and object information in relationships such as <person-on-motorcycle> and <motorcycle-on-person> leads to different results. The reason for using the Bi-RNN model is its ability to learn the differences arising from the ordering of subject and object in relationships expressed in the <subject-predicate-object> form.

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.

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

A Bi-RNN computes the hidden states twice: a forward sequence $h_1{}^{\rightarrow}$ and a backward sequence $h_1{}^{\leftarrow}$ where $W_{h1y}^{\rightarrow}$ denotes the input-hidden weight matrix in forward direction.

During the training stage of the model, the selected data from the Visual Genome dataset was used. The data was arranged in a specific order and utilized as the input for the model. In this sequence, x1 represented the subject vector, x2 denoted spatial information, and x3 represented the object vector data. The representation of the relationship (including spatial information) between two objects was determined in accordance with Eq. (2).

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

In the Eq. (2), the x and y correspond to the center coordinates of the bounding box of an object, whereas w and h signify the width and height of the bounding box respectively, that encodes their relative spatial relationship which is important for relationship representation.

In this study, the predicate detection process was executed. In this process, the class labels associated with two objects and the location information in the image, referred to as Subject and Object, were provided. The aim was to determine the expression defining the relationship between them. This process can be represented using Eq. (3):

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

In the Eq. (3), the expression $P_i(O_1)$ gives the probability that the object detected from the image belongs to class i, whereas, the expression $P_j(O_2)$ gives the probability that the object belongs to class j in the Object Detection process. On the other hand, Bi-RNN (O_1, O_2, 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.

In this study, the Recall@K metric was utilized for the performance measurement of the proposed Visual Relationship Detection process in the first stage. The Recall@K performance is calculated according to the Eq. (4), providing the probability of the ground truth being among the first K predictions (Top-K Predicate).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

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

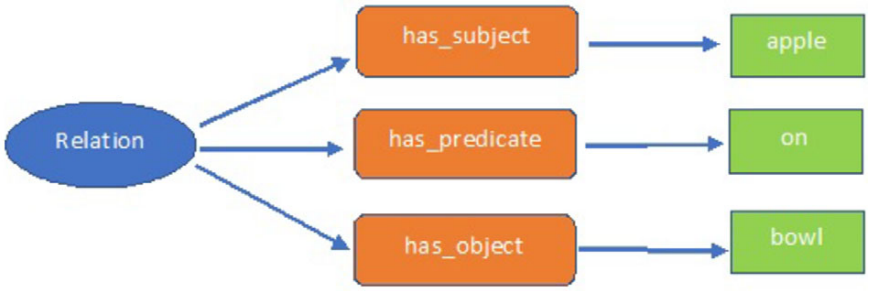

Fig. 4. Relation ontology.

Table 1
Experimental results of different methods in the VG dataset

Method	Recall@50	Recall@100
Ref. [6]	65.20	67.10
Ref. [17]	85.02	91.77
Proposed	90.00	91.00

- True positives: data points labeled as positive that are actually positive.
- False negatives: data points labeled as negative that are actually positive

The reason for choosing this metric is the selection of the Predicate Detection task in the proposed first stage of our study, where Visual Relationship Detection is employed. Predicate Detection aims to determine an expression (predicate) that defines the relationship between two objects. In this case, when evaluating the model's performance, we examined the cases where the ground truth is among the first five predictions as determined.

3.4. Ontology creation

Ontologies serve the purpose of describing entities by creating classes within a specific domain and establishing relations between these classes. In this study, an ontology was generated, as depicted in Fig. 4, utilizing the data structured in the subject-predicate-object format from the Visual Genome dataset. These ontologies were employed in the process of Semantic-Based Image Retrieval, and the outcomes of the search are detailed in the relevant section.

4. Experiments

In this study, a Bi-RNN model was trained using data chosen from the Visual Genome dataset to ascertain the expression defining the relationship between detected objects in images. For model training, the Subject and Object data were transformed into 100-dimensional vectors using Word2Vec. Additionally, a 4-dimensional vector was generated by utilizing Bounding Box coordinates to describe the relationship between two objects, following the formula in 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.

4.1. Meaning inference from the image

During the testing phase of the model on an image not present in the training dataset, objects within the image were initially detected using Detectron2. Subsequently, a predicate was established to define the relationship between the two objects by generating dual combinations of the identified objects. The results obtained from this process are illustrated in Fig. 5.

Fig. 5. Meaning inference from the image.

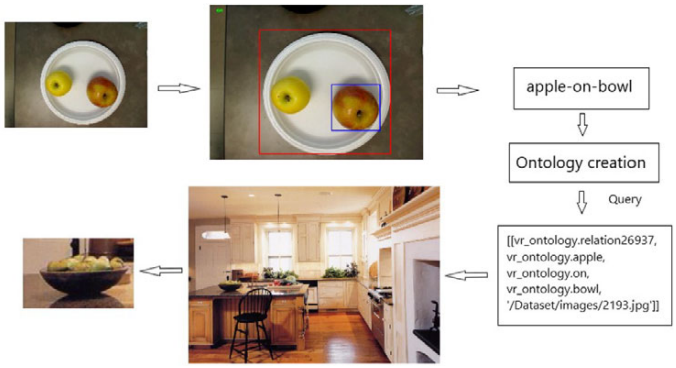

Fig. 6. Searching process using ontology.

4.2. *Searching process using ontologies*

Meaning inference was achieved by determining a predicate that describes the relationship between objects perceived from an image. The relationship structured in the subject-predicate-object format was then transformed into ontologies and employed in the search for semantically similar images. The outcome of the search for semantically similar images using ontologies is depicted in Fig. 6.

4.3. *Measuring the semantic gap*

The current study suggests measuring the Semantic Gap, a primary challenge in Semantic-Based Image Retrieval, by calculating the quantity of similar relationships between two images using entropy. The Entropy method, used to measure the amount of information in Information Theory, is utilized for measuring the Semantic Gap between two images. The entropy method generally expresses the disorder of a system. An increase in disorder within the system will lead to an increase in the entropy value. When the entropy of the difference between the numbers of relationships in two images is high, it indicates that there is a high semantic similarity between the two images. Conversely, when the entropy of the difference between the numbers of relationships is low, it can be understood that the semantic similarity is low.

Joint entropy is a concept that measures the uncertainty between two or more random variables. The joint entropy equation is used to calculate this uncertainty and is as described according to Eq. (5):

$$H(X, Y) = -\sum_{x,y} p(x, y) \log (x, y) \tag{5}$$

In the given Eq. (5)

- $H(X, Y)$: Represents the joint entropy of X and Y.
- $P(x, y)$: Indicates the probability of the joint occurrence of X and Y.

The Semantic Gap between two images was calculated using the Joint Entropy method. This involved utilizing the number of relationships (X) found in the image used for query purposes and the total number of relationships (Y) in the image with similar relationships identified as a result of the query.

According to Eq. (5), as a result of the calculation using the relationship numbers between the two images, three cases were examined:

$$H = \begin{cases} 1, & \text{“there is No semantic gap”} \\ 0 < H < 1, & \text{“there is a semantic gap”} \\ 0, & \text{“No similar images”} \end{cases} \tag{6}$$

In this case, according to the Eq. (6), if the entropy (H) is “1” it implies that the number of relationships between the two images is equal, and the two images are semantically similar. If the entropy value is between “0” and “1” it indicates that the number of relationships between the two images is different, and there is a semantic difference between them. When the entropy is “0” it signifies that there is no similar relationship between the two images, and the two images are different from each other.

In the search process utilizing the sample image illustrated in Fig. 7, for the objects “apple” and “bowl” detected from the image, a similar image was found as a result, conforming to the specified “apple-on-bowl” relationship.

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

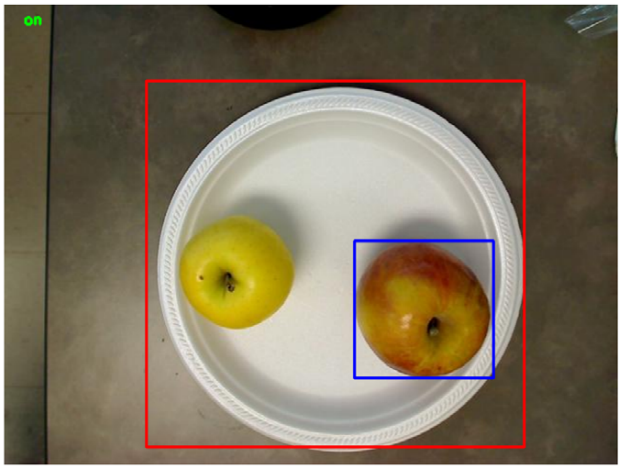

Fig. 7. Searching process using ontology.

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"

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

According to the entropy-based calculation, out of the six ground truth relationships defined in the image found in the query result, only one relationship matched. Based on the calculation result according to Eq. (5), it was determined that there exists a Semantic Gap.

5. Discussion

In the initial step of the two-stage Semantic-Based Image Retrieval approach proposed in this study, as illustrated in Fig. 1, the model was trained to determine the relationship between two objects and infer meaning. This training utilized 212 features organized as Object vector, Spatial Information, Subject vector. In other studies [9,10], 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, an ontology has been created for the relationship represented in the form of <subject-predicate-object> and utilized in the search for semantically similar images. In a previous study [7], the search process utilized the Scene Graph structure of the image. In contrast, in our study, an expression defining the objects within the image and the relationship between them is determined. This relationship in the subject-predicate-object structure is then transformed into Ontologies. Ontologies provide a more robust modeling of objects and their relations compared to the Scene Graph structure. They are employed to search for semantically similar images as part of the solution to address the problem referred to as the Semantic Gap. The complexity of semantic analysis, especially with large datasets, might lead to increased computational requirements, impacting the efficiency of the retrieval process. Addressing these computational challenges and optimizing algorithms for efficiency will be essential to mitigate potential drawbacks. Moreover, exploring strategies to balance

accuracy with computational resources and considering advancements in hardware capabilities can contribute to enhancing the overall performance of Semantic-Based Image Retrieval systems.

6. Conclusion

In this study, a novel approach was introduced to address the Semantic Gap problem in Image Retrieval by utilizing Ontologies. The relationships established between the detected objects in the image were converted into Ontologies and employed in the process of searching for semantically similar images. The proposed method can be considered as a more general approach to the Semantic Based Image Retrieval process. In comparison to the Scene Graph structure employed in other studies, more efficient results were achieved in conceptual modeling by utilizing Ontologies. This involves transforming information into a format that computers can process, facilitating the search for semantically similar images through the inference of 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).

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] G. Antoniou and F. van Harmelen, *A Semantic Web Primer*, 2nd edn, MIT Press, Cambridge, Massachusetts London, England, 2008.
[2] J. Cheng, L. Wang, J. Wu, G. Jeon, D. Tao and Z. Mengchu, Visual relationship detection: A survey, *IEEE Transactions on Cybernetics* **52**(8) (2022), 8453–8466. doi:10.1109/TCYB.2022.3142013.
[3] L. Chengjun and G. Song, A method of measuring the semantic gap in image retrieval: Using the information theory, in: *International Conference on Image Analysis and Signal Processing*, 21–23 October 2011, Wuhan, China, 2011, pp. 287–291. ISBN:978-1-61284-879-2.
[4] D.D. Dhobale, B.S. Patil, S.B. Patil and V.R. Ghorpade, Semantic understanding of image content, *International Journal of Computer Science Issues* **8**(3) (2011), 191–196.
[5] H. Ghosh, S. Chaudhury and A. Mallik, Ontology for multimedia applications, *IEEE Intelligent. Informatics Bulletin* **14**(1) (2013), 21–30.
[6] R. Girshick. R-CNN fast, in: *IEEE International Conference on Computer Vision*, 2015.
[7] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D.A. Shamma, M.S. Bernstein and L. Fei-Fei, Image retrieval using scene graphs, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3668–3678.
[8] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. Shamma, M. Bernstein and L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International Journal of Computer Vision* **123**(1) (2017), 32–73. doi:10.1007/s11263-016-0981-7.
[9] W. Liao, B. Rosenhahn, L. Shuai and M.Y. Yang, Natural language guided visual relationship detection, in: *IEEE Conference on Computer Vision and Pattern Recognition Worshops*, Long Beach, CA, USA, 2019, pp. 444–453.

[10] C. Lu, R. Krishna, R. Bernstein and L. Fei-Fei, *Visual Relationship Detection with Language Priors, European Conference on Computer Vision*, 2016.
[11] H. Ma, J. Zhu, M.R.T. Lyu and I. King, Bridging the semantic gap between image contents and tags, *IEEE Transactions on Multimedia* **12**(5) (2010), 462–473. doi:10.1109/TMM.2010.2051360.
[12] V. Mezaris, I. Kompatsiaris and M.G. Strintzis, An ontology approach to object-based image retrieval, in: *IEEE International Conference on Image Processing*, 14–17 September, 2003.
[13] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, in: *International Conference on Learning Representation*, 2013.
[14] R.I. Minu and K.K. Thyagharajan, Semantic rule based image visual feature ontology creation, *International Journal of Automation and Computing* **11**(5) (2014), 489–499. doi:10.1007/s11633-014-0832-3.
[15] J. Peyre, I. Laptev, C. Schmid and J. Sivic, Weakly-supervised learning of visual relations, in: *IEEE International Conference on Computer Vision and Pattern Recognition*, 22–29 October 2017, Venice, Italy, IEEE, 2017, pp. 5189–5198. ISBN:978-1-5386-1032-9.
[16] N. Rasiwasia, P.J. Moreno and N. Vasconcelos, Bridging the gap: Query by semantic example, *IEEE Transactions on Multimedia* **9**(5) (2007), 923–938. doi:10.1109/TMM.2007.900138.
[17] J. Redmon and A. Farhadi, YOLO9000: Better, faster, stronger, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6517–6525.
[18] S. Sarwar, Z. Qayyum and S. Majeed, Ontology based image retrieval framework using qualitative semantic image descriptions, *Procedia Computer Science* **22** (2013), 285–294. Barcelona, Spain, IEEE, ISBN:0-7803-7750-8, 511–514.
[19] C.E. Shannon, Communication theory of secrecy systems, *The Bell System Technical Journal* **28**(4) (1949), 656–715.
[20] R. Shaoqing, H. Kaiming, R. Girshick and J. Sun, R-CNN faster: Towards real-time object detection with region proposal networks, *IEEE Transactions and Pattern Analysis and Machine Intelligence* **39**(6) (2017), 1137–1149. doi:10.1109/TPAMI.2016.2577031.
[21] Z. Xinying and L. Linhu, Diverse image search with explanations, in: *Multimedia Tools and Applications*, 2023.
[22] D. Xu, Y. Zhu, C.B. Choy and L. Fei-Fei, Scene graph generation by iterative message passing, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3097–3106.
[23] P. Yaopeng, Z.C.H. Danny and L. Lanfen, Visual relationship detection with a deep convolutional relationship network, in: *IEEE International Confernce on Image Processing*, 25–28 October 2020, Abu Dhabi, UAE, IEEE, 2020, pp. 1461–1465. ISBN:978-1-7281-6395-6.
[24] Z. Yaouhi and J. Shuqiang, Deep structured learning for visual relationship detection, in: *Procedings of AAAI Conference on Artificial Intelligence*, 2018.
[25] L. Ying, Z. Dengsheng, L. Guojun and M. Wei-Ying, A survey of content-based image retrieval with high-level semantics, *Pattern Recognition* **40**(1) (2007), 262–282. doi:10.1016/j.patcog.2006.04.045.
[26] R. Yong and H. Thomas, Image retrieval: Current techniques, promising directions, and open issues, *Journal of Visual Communication and Image Representation* **62** (1999), 39–62.
[27] L. Zhang, S. Zhang, P. Shen, G. Zhu, S.A.A. Shah and M. Bennamoun, Relationship detection based on object semantic inference and attention mechanisms, in: *International Conference on Multimedia Retrieval*, 2019, pp. 68–72.
[28] R. Zhao and W.I. Grosky, Narrowing the semantic gap – Improved text-based web document retrieval using visual features, *IEEE Transactions on Multimedia* **4**(2) (2002), 189–200. doi:10.1109/TMM.2002.1017733.
[29] Z.Q. Zhao, P. Zheng, S.-T. Xu and X. Wu, Object detection with deep learning: A review, *IEEE Transactions on Neural Networks and Learning Sysems.* **30**(11) (2019), 3212–3232. doi:10.1109/TNNLS.2018.2876865.