Journal of Computer Science and Cybernetics, V.35, N.1 (2019), 39–56
DOI 10.15625/1813-9663/35/1/13097

SIR-DL: AN ARCHITECTURE OF SEMANTIC-BASED IMAGE RETRIEVAL USING DEEP LEARNING TECHNIQUE AND RDF TRIPLE LANGUAGE*

VAN THE THANH[1,a], DO QUANG KHOI[2], LE HUU HA[1], LE MANH THANH[3]

[1]*Faculty of Information Technology, HCMC University of Food Industry*
[2]*Center for Training and Fostering, Quang Nam University*
[3]*Faculty of Information Technology, University of Science Hue University*
[a]*vanthethanh@gmail.com*

Abstract. The problem of finding and identifying semantics of images is applied in multimedia applications of many different fields such as hospital information system, geographic information system, digital library system, etc. In this paper, we propose the Semantic-Based Image Retrieval (SBIR) system based on the deep learning technique; this system is called as SIR-DL that generates visual semantics based on classifying image contents. Firstly, the color and spatial features of segmented images are extracted and these visual feature vectors are trained on the deep neural network to obtain visual words vectors. Then, we retrieve it on ontology to provide the identities and the semantics of similar images corresponds to a similarity measure. In order to carry out SIR-DL, the algorithms and diagram of this image retrieval system are proposed and after that we implement them on ImageCLEF@IAPR, which has 20,000 images. Based on experimental results, the effectiveness of our method is evaluated; these results are compared with some of the works recently published on the same image dataset. It shows that SIR-DL effectively solves the problem of SBIR and can be used to build multimedia systems in many different fields.

Keywords. Bag of visual word; Deep learning; Ontology; SBIR; Similarity measure; Similar images.

1. INTRODUCTION

Global digital data has been increasing rapidly and reaching enormous amounts. This leads to the need for a good method to solve the problem of data mining and information retrieval. According to International Data Corporation (IDC), global data in 2012, 2013 reached 2.8 zettabytes and 4.4 zettabytes. It is estimated, at the end of 2020, global data is 300 times more than that in 2005, which is an increase from 130 exabytes to 40,000 exabytes (40 trillion gigabytes = 40 zettabytes), of which data generated by mobile devices accounted for 27%. By 2025, global data will reach about 163 zettabytes, which is a tenfold increase compared with 2017 [15]. In addition, digital photos have become familiar with people. They are used in many multimedia information retrieval systems

*This paper is selected from the reports presented at the 11th National Conference on Fundamental and Applied Information Technology Research (FAIR'11), Thang Long University, 09 - 10/08/2018.

[22, 27] such as hospital information system, geographic information system, digital library system, biomedicine, education and training, entertainment, etc. In 2015, the total number of images across the globe reached 3.2 trillion photos; in 2016, there were 3.5 million photos shared shared and stored online. In 2017, the world created 1.2 trillion photos so that the total number of photos on global in 2017 was 4.7 trillion photos, of which the images generated from smart phones and mobile devices are 90% [7]. Therefore, the problem of data mining and information retrieval related to digital images need to be solved as well as the finding of similar images is one of the important problems of many multimedia systems [17, 25].

There were many systems of semantic-based image retrieval, which have been published and applied in a variety of fields such as a semantic framework image retrieval based on high-level semantics and image annotations applied on CT images [5], a semantic-based medical imaging retrieval using Convolutional Neural Network (CNN) for brain MRI image [30], a semantic-based application in the distributed information systems [9], a medical case-based image retrieval based on textual and image information in RadLex ontology [2], etc. In each of the different areas, multimedia systems need to be extracted the semantic of objects to describe content. So, SBIR extracts features to identify meaning of images; then, it retrieves the related images in visual features and extracts semantics of contents of these images [8, 26, 29]. The first challenge of SBIR is to extract visual features after that map it into semantics to describe content of image. The second challenge is to describe semantics and search for related images [27]. In this paper, SBIR based on Deep Neural Network (DNN) and RDF triple language (SIR-DL) is built. The experiment of SIR-DL is executed on ImageCLEF dataset [4, 10, 13]. We identify the semantics of similar images on ontology, which describes semantics of visual features of images. The process of image retrieval is executed based on semantic classification of SIR-DL according to the visual feature vector of the query image from which it produces a visual word vector. SIR-DL shows the semantics of input image as well as queries by semantics to find out similar images based on RDF triple language.

The proposed model using DNN is based on visual content images, from which we automatically generate SPARQL queries and execute on ontology using RDF triple. We build the semantic-based image retrieval system based on content of the image using DNN, BoW, RDF, ontology and SPARQL. We combine these tools to create the new model. From there, the algorithms are proposed based on this model; at the same time, we prove the theoretical and empirical correctness. In the experimental results, our suggestions are effective. The contributions of the paper include: (1) using Bag-of-Visual-Word (BoVW) and deep learning techniques to classify images into visual semantic vectors based on color and spatial features; (2) building ontology for image dataset and creating RDF triple language; (3) creating a SPARQL query to retrieve similar images based on visual word vector and ontology; (4) proposing model and algorithms of SIR-DL to retrieve similar images by semantics; (5) constructing the experimental application based on SIR-DL model and proposed algorithms.

The rest of paper is as follows. In Section 2, we survey and analyze related works. Section 3, the general architecture of SIR-DL is described to construct an SBIR. Section 4 & 5, we present the components and the proposed algorithms in SIR-DL. Then, we build the experiment and evaluate the effectiveness of proposed method. Conclusions and future works are presented in Section 6.

2. RELATED WORKS

There were many techniques of multimedia retrieval by semantics that have been widely applied in many different fields such as query techniques on Ontology-based for the purpose of exact meaning interpretation of user query [19], visual encoding model based on convolutional neural network [31], semantic-based natural image retrieval using bag of visual word model and distribution of local semantic concepts [3], an efficient video retrieval based on semantic graph queries [12], an adaptive image search engine for deep knowledge and meaning of the image applied in Ontology-based to produce a new level of image meaning [18], content based semantics and image retrieval system for hierarchical databases [24], etc.

In 2018, M. Tzelepi and A. Tefas proposed a CNN training method for content-based image retrieval based on Caffe Deep Learning framework. In this paper, the authors classified images from low-level features based on relevance feedback and applied to the problem of similar image retrieval [21]. Xiao Xie et al. proposed a method of classifying the visual features of images based on CNN and rendering semantic keywords to find similar images. These authors did not perform a query on Ontology to determine semantic of images [27]. Safia Jabeen et al. built a model of image retrieval based on BoVW by clustering the visual features associated with the semantics of the categories of images [23]. However, clustering low-level visual features can create clusters of images with different semantics that lead to the searching semantic of query image is inaccurate. Therefore, the method of semantic classification from low-level features needs to be applied to map these features into semantics of the images.

In 2014, Yalong Bai et al. used DNN to classify feature vectors of image to map into bag-of-word (BoW). The phase of image retrieval is executed based on BoW from which a set of images is given corresponding to this BoW [28]. This model has not converted visual features into semantics and has not yet retrieved directly from a given image. Thus, a method of classification for mapping from low-level visual features to semantics of images must be constructed to create input of the semantic search problem. J. Wan et al., surveyed deep learning technique to solve the image retrieval problem. The results of paper showed that effectiveness of applying this method to classify images by semantics [16].

In 2016, Yue Cao et al. used CNN to classify images to generate binary feature vectors. On the base of this, the authors proposed Deep Visual-Semantic Hashing (DVSH) model to identify a set of similar images by semantics [29]. However, this method must perform two classification processes of visual and semantic features. If a image lacks one of these two features, the retrieved similar images are inaccurate. Furthermore, the method has not yet mapped from visual features to semantics of images. Vijayarajan et al. performed image retrieval based on analyzing natural language to create a SPARQL query to find similar images based on RDF image description [26]. The process of image retrieval depends on analyzing grammar of language to form keywords describing the content of image. This method has not yet implemented classification of image content from the color and spatial features to obtain keywords to perform retrieval; therefore, the search process does not proceed from a given query image.

In 2017, Hakan Cevikalp et al. executed image retrieval based on graph-cut structure and binary hierarchy tree. Training was implemented using Support Vector Machines (SVM) based on low-level image features [14]. This method tested on ImageCLEF dataset and after that it compared effectiveness with other methods, but it did not classify the semantic of

images. M. Jiu and H. Sahbi used a multilayered neural network based on different nonlinear activation functions on each layer. The SVM technique was used to classify images at the output layer to determine meaning of similar images based on BoW [20]. In this method, neural network is fixed the number of layers, so the classification of deep learning technique is limited. B. B. Z. Yao et al. (2010) introduced the Image to Text (I2T) tool to generate RDF that describes image semantic from which users can query through this semantics. The And-or Graph (AoG) was used to transform relationships of components of image into natural semantics to describe the image [27]. This is a method of semantic image retrieval and it makes the problem of image retrieval according to semantics is more complete.

On the basis of inheriting and overcoming limitations of related works, we propose SIR-DL model by classifying the features of images into visual semantics using deep learning technique and transform it into a SPARQL command from which to execute the query on RDF triple language according to Ontology of the given image dataset.

3. THE ARCHITECTURE OF SIR-DL SYSTEM

3.1. The model of SIR-DL

The general architecture of SIR-DL is described in Figure 1 and it is implemented by classifying images into visual word vectors based on deep learning network and performing image retrieval on RDF triple language. This model is built based on combination of components including deep learning network [16, 20, 21], BoVW technique [23, 28, 29], and semantic query on ontology in SPARQL language [3, 8, 18, 26]. Based on deep learning, the classification model of semantic images is trained on dataset to create inputs for the problem of image retrieval on Ontology. The query is performed by automatically creating the SPARQL command and searching images on the Ontology described in RDF triple language. The SIR-DL consists of two phases including: (1) extracting feature vectors of image datasets to generate inputs for training DNN based on classifying using BoVW; (2) for each image, its features are classified based on SIR-DL to generate the visual word vector. Then, the SPARQL query is generated at the same time performing the query on Ontology which was described as RDF triple language.

3.2. Pre-processing phase of SIR-DL

The result of pre-processing stage of SIR-DL is a classification model. Each image in the dataset is extracted regions to create a set of feature vectors. Then, the DNN of SIR-DL is trained based on the method of reducing gradient in the direction of error function to find the optimal value of weights. The process of pre-processing phase consists of the following steps:

Step 1: Extract a sample (x, y) of each region corresponding to each image in dataset, where x is the feature vector, y is the semantic classification;

Step 2: Train DNN of SIR-DL according to each epoch based on Gradient reduction method combined with momentum value;

Step 3: Build Ontology as RDF triple language to describe semantics for image dataset.

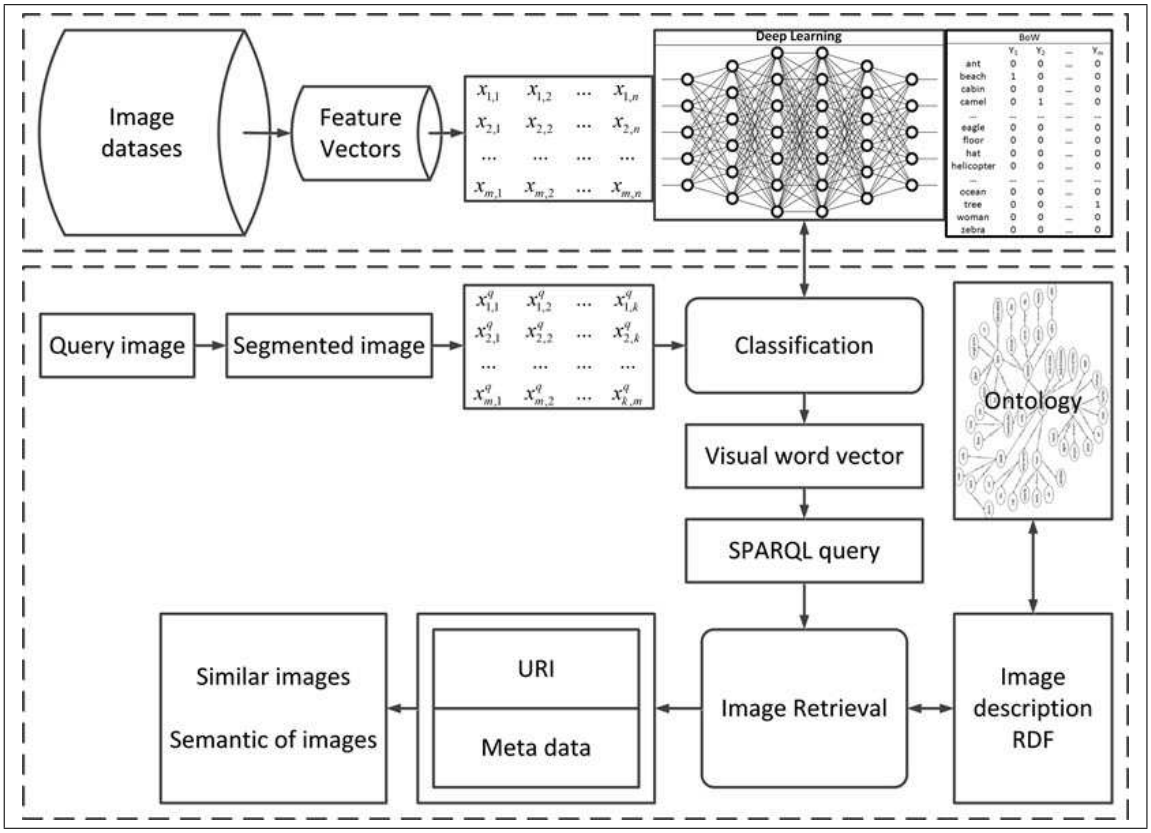

Figure 1. Model of semantic-based image retrieval SIR-DL

3.3. Image retrieval phase of SIR-DL

The searching of similar images is performed with input as a visual word vector at the same time it generates SPARQL command. Then, SIR-DL performs this query on Ontology to get results as a set of URIs and metadata of similar images. The process of image retrieval is performed as follows:

Step 1: Each query image, the feature vectors of region of image are extracted and classified to form the visual word vector based on the trained DNN of SIR-DL;

Step 2: Create SPARQL query based on visual word vector and perform image retrieval on Ontology to get result as a set of URIs and metadata of images;

Step 3: Give similar images from URIs and arrange them by similarity measure according to the query image.

4. CREATING THE COMPONENTS OF SIR-DL SYSTEM

4.1. Extracting visual features of images

Each image in dataset is segmented into different objects according to Hugo Jair Escalantes method [10]. Figure 2 shows an original image and five regions belonging to the classes

including cloud, hill, ruin-archeological, road, group-of-persons. Each region is extracted a feature vector including characteristics: Region area, width and height; Features of locations including mean and standard deviation in the x and y-axis; Features of shape including boundary/area, convexity; Features of colors in RGB and CIE-Lab space including average, standard deviation and skewness [4, 13].

Figure 2. Original image and segmented images

4.2. Creating similarity measure between images

The similarity measure is created based on feature vectors to evaluate the similarity between two images. Because each image has a different number of feature vectors, the Earth Mover's Distance (EMD) distance is applied to evaluate the similarity between two images by distributing among regions of images [1]. Given two set of features of images I and J as $F_I = \{f_I^i | i = 1, ..., n\}$ and $F_J = \{f_J^j | j = 1, ..., m\}$, respectively. The similarity of feature vector f_I^i of image I with image J is evaluated by the following formula

$$\text{dis}_{I,J}^i = \text{dis}(f_I^i, J) = \frac{1}{m} \sum_{j=1}^{m} ||f_I^i - f_J^j|| \tag{1}$$

with $||f_I^i - f_J^j|| = \sqrt{\left(f_J^{j1} - f_I^{i1}\right)^2 + ... + \left(f_J^{jk} - f_I^{ik}\right)^2}$.

On the base of formula (1), the similarity vectors of two images I and J are $D_{I,J} = (\text{dis}_{I,J}^1, ..., \text{dis}_{I,J}^n)$ and $D_{J,I} = (\text{dis}_{J,I}^1, ..., \text{dis}_{J,I}^m)$, respectively. The feature distance from image I to image J is defined as follows

$$DF(I, J) = \frac{1}{n} \sum_{i=1}^{n} \text{dis}_{I,J}^i. \tag{2}$$

Proposition 1. *The feature distance $DF(I, J)$ in formula (2) is a metric.*

Proof. This is easy to prove because $DF(I, J)$ is a metric. ∎

Let $E = (e_{ij})$ be a distance matrix between two images, with $e_{ij} = ||f_I^i - f_J^j||$, let $F = (f_{ij})$ be a distribution matrix between $D_{I,J} = (\text{dis}_{I,J}^1, ..., \text{dis}_{I,J}^n)$ and $D_{J,I} = (\text{dis}_{J,I}^1, ..., \text{dis}_{J,I}^m)$, with f_{ij} as a distribution value between $\text{dis}_{I,J}^i$ and $\text{dis}_{J,I}^j$, then, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} = \min\{\sum_{i=1}^{n} \text{dis}_{I,J}^i, \sum_{j=1}^{m} \text{dis}_{J,I}^j\}.$$

On the base of transport problem, the similarity measure between two images I and J is defined by the following formula

$$EMD(I,J) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} e_{ij}f_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}}. \quad (3)$$

Proposition 2. *The similarity measure EMD in this case is a metric.*

Proof. This is easy to prove because $\text{dis}^{i}_{I,J}$ and $DF(I,J)$ are metrics. ∎

4.3. Training deep neural network

Deep Neural Network (DNN) of SIR-DL is designed including an input layer, an output layer, and multi-hidden layers; each node of next layer is fully connected to nodes in previous layer. At each layer, the bias element is connected to all nodes of that layer to assist in the implementation of classification process [16, 21, 28]. In SIR-DL model, DNN has input layer is a feature vector of region of image as $f_i = (f_i^1, ..., f_i^t)$, output layer is a vector $y_k = (y_k^1, y_k^2, ..., y_k^s)$; the values of vector y_k are mapped into a unit vector, then a label class as $l_k \in \{l_1, l_2, ..., l_m\}$ is created. Therefore, the training set of DNN is $T = \{(f_i, y_k)|i = 1, ..., n; k = 1, ..., m\}$. The result of training process is a set of weights at each layer $W = \{W_k, Wb_k|k = 1, ..., K\}$, with W_k as a weight matrix of connections between two layers, Wb_k as a weight vector of connections corresponding to bias of each layer. The softmax and tanh function are used to active functions of output layer and hidden layers, respectively. In order to train DNN, with each input value f_i, the output values y_k are calculated based on the propagation process from input layer to output layer. The propagation algorithm to calculate output values y_k are done as follows:

Theorem 1. *Let* f_i^1, f_i^2, f_i^3 *be feature vectors and* $y_k^1 = \text{Out}(W, f_i^1)$, $y_k^2 = \text{Out}(W, f_i^2)$, $y_k^3 = \text{Out}(W, f_i^3)$. *Then, if* $|f_i^1 - f_i^2|| \leq ||f_i^1 - f_i^3||$ *there holds* $||y_k^1 - y_k^2|| \leq ||y_k^1 - y_k^3||$.

Proof. Because of $\text{Out}(W, f_i)$ function using a weight matrix for three input values f_i^1, f_i^2, f_i^3 to calculate the output values of each node. In addition, Tanh and softmax are continuous, single-valued, and monotonic functions.

So, we have

$$\text{If } |f_i^1 - f_i^2|| \leq ||f_i^1 - f_i^3|| \text{ then } |y_k^1 - y_k^2|| \leq ||y_k^1 - y_k^3||.$$

∎

From Theorem 1, we have a conclusion that if two regions of image have the same features then they are classified in the same class.

Proposition 3. *The complexity of DLO algorithm is* $O(m \times n)$.

Proof. DLO algorithm carries out on the connection weight matrices, so the complexity is $O(m \times n)$.

∎

The training algorithm of deep learning network is done using back-propagation method, which updates weights from input layer to output layer according to values of Gradient vector at each layer. The DNN training algorithm is described as follows:

Algorithm 1 DLO

Input: $f_i = (f_i^1, ..., f_i^t)$, $W = \{W_k, Wb_k | k = 1, ..., K\}$;
Output: $y_k = (y_k^1, y_k^2, ..., y_k^s)$;
Function: DLOut(W, f_i);

Begin
Initializing values of input layer as $f_i = (f_i^1, ..., f_i^t)$;
for $(W_k, Wb_k) \in W$ **do**
  for $w_{ij} \in W_k$ **do**
    $h_{kj} = \text{Tanh}(bias_{kj} + \sum_{i=1}^{a} h_{kj} \times w_{ij})$;
  end for
end for
for $i = 1 : s$ **do**
  $y_k^i = \text{softmax}(bias_{Ki} + \sum_{j=1}^{b} o_i \times w_{Kj})$;
end for
Return y_k;
End.

Theorem 2. *Let (f_i, y_k) be an example of training set. Then we have $||y_k - y_o(t+1)|| \leq ||y_k - y_o(t)||$.*

Proof. Because Tanh and softmax are continuous functions, single-valued, and monotonic. In addition, the values of weights are updated by Gradient vector. So that, each (f_i, y_k), we have $||y_k - y_o(t+1)|| \leq ||y_k - y_o(t)||$. ∎

From Theorem 2, we have a conclusion that on the same example, the training error must be lower than the previous one.

Proposition 4. *The complexity of DLT algorithm is $O(N \times m \times n)$, with N as the number of epochs in training set.*

Proof. Because DLT algorithm trains weight matrix for each epoch, the complexity is $O(N \times m \times n)$.

∎

4.4. Creating ontology of image dataset

In order to query by SPARQL, an ontology domain is created, which describes semantics of image dataset [8, 25, 26]. In this paper, each region of image is designed an individual belonging to a class that links to meaningful image. In order to describe meaningful images, the ontology is built on RDF triple language as Turtle using semantics on ImageCLEF dataset and is described in Figure 3. The diagram of ontology is extracted from Protg using the set of triples and is described in Figure 4. The descriptions of RDF/XML ontology are presented in Figure 5.

Algorithm 2 DLT

Input: $T = \{(f_i, y_k)|i = 1, ..., n; k = 1, ..., m\}$, learning rate α, momentum η, hidden layers H;
Output: $y_k = (y_k^1, y_k^2, ..., y_k^s)$;
Function: DLTraining(T, α, η, H);

Begin
Initializing the set of weights W;
for *epoch* in T **do**
  for (f_i, y_k) in *epoch* **do**
    y_o = DLOut$(W,\ f_i)$;
    $e_i = ||y_k - y_o||$;
    $\nabla E_i = (\frac{\partial e_i}{\partial w_1}, \frac{\partial e_i}{\partial w_2}, ..., \frac{\partial e_i}{\partial w_K})$;
  end for
  for h in H **do**
    for (f_i, y_k) in *epoch* **do**
      $\nabla E_{ih} = (\frac{\partial e_{ih}}{\partial w_{i1}}, \frac{\partial e_{ih}}{\partial w_{i2}}, ..., \frac{\partial e_{ih}}{\partial w_{in}})$;
    end for
  end for
  for $(W_k, Wb_k) \in W$ **do**
    for $w_{ij} \in W_k$ **do**
      $w(t)_{ij} = w(t-1)_{ij} - \alpha * \frac{\partial E_{ik}}{\partial w_{ij}} - \eta w(t-1)_{ij}$;
    end for
    for $w_{ij} \in Wb_k$ **do**
      $w(t)_{ij} = w(t-1)_{ij} - \alpha * \frac{\partial E_{ik}}{\partial w_{ij}} - \eta w(t-1)_{ij}$;
    end for
  end for
end for
Return W;
End.

4.5. Image retrieval

On the base of trained DNN, each query image is extracted feature vector and is classified to create a visual word vector. The classification algorithm of image is done as follows.

Proposition 5. *The complexity of DLR algorithm is* $O(r \times m \times n)$.

Proof. DLR algorithm executes r times to calculate DLOut(W, f_I^i), so the complexity of DLR algorithm is $O(r \times m \times n)$. ∎

On the base of visual word vector, SPARQL command is created to query on Ontology. The result is a set of URIs and metadata of similar images. Figure 6 shows a SPARQL command which is generated from a visual word vector.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix cla: <http://www.ittc.edu.vn/itcsites/imageclef/class#>.
@prefix cat: <http://www.ittc.edu.vn/itcsites/imageclef/catetories#>.
@prefix img: <http://www.ittc.edu.vn/itcsites/imageclef/imagename#>.
@prefix des: <http://www.ittc.edu.vn/itcsites/imageclef/description#>.

cla:group-of-persons rdf:type cat:humans;
    cat:catetories cat:humans;
        img:segimage "112_1.jpg";
    img:imageID "112.jpg";
    des:description "112.eng" .

cla:rock rdf:type cat:landscape-nature;
    cat:catetories cat:landscape-nature;
        img:segimage "112_2.jpg", "112_3.jpg", "116_4.jpg";
    img:imageID "112.jpg", "116.jpg";
    des:description "112.eng", "116.eng" .

cla:branch rdf:type cat:trees;
        cat:catetories cat:trees, cat:vegetation, cat:landscape-nature;
    img:segimage "112_4.jpg", "112_5.jpg", "25_1.jpg";
    img:imageID "112.jpg", "25.jpg";
    des:description "112.eng", "25.eng" .
```

Figure 3. An example of ontology on ImageCLEF by Turtle

Algorithm 3 DLR

Input: $F_I = \{f_I^i | i = 1, .., r\}$, $W = \{W_k, Wb_k | k = 1, ..., K\}$;
Output: visual word vector V ;
Function: DLRetrieval(F_I, W);

Begin
Initializing the visual word vector V;
for $f_I^i \in F_I$ **do**
  $y = \text{DLOut}(W, f_I^i)$;
  $v = \text{DLClassification}(y)$;
  $V = V \cup v$;
end for
Return V;
End.

5. EXPERIMENTS

The experiment of SIR-DL is built including two stages: (1) pre-processing stage is done based on training the model of DNN in SIR-DL to classify semantics of image features; (2) image retrieval stage is executed semantic retrieval of query image.

SIR-DL is built in dotNET Framework 4.5, and C# programing language. It is shown in Figure 7. Pre-processing stage of SIR-DL is done on server which has CPU Intel(R) Xeon(R) 20 Core x 2 CPU ES-2680 v2 @ 2.80GHz (2 processors), OS Windows Server 2012 64-bit,

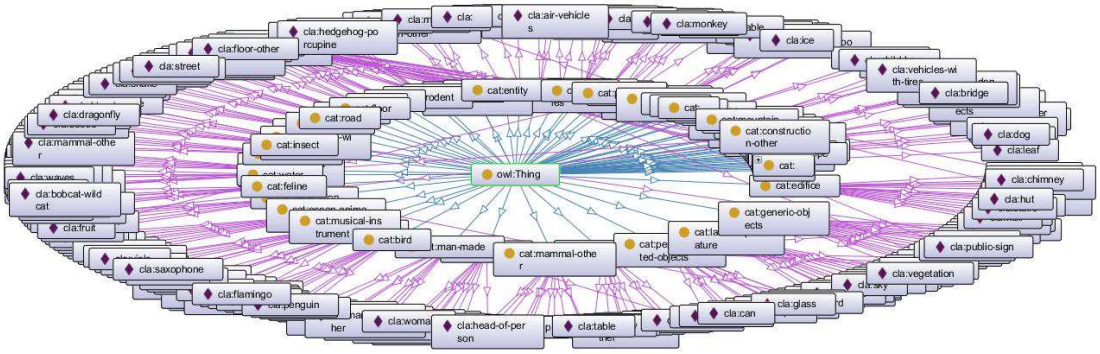

Figure 4. Ontology of ImageCLEF dataset on Protege

Algorithm 4 DLC

Input: vector y;
Output: an unit vetor v;
Function: DLClassification(y);

1: **Begin**
2: $v = (v_1, v_2, ..., v_n)$, so that $c_i = 0$;
3: $k = argMax(y_i)$;
4: $v_k = 1$;
5: **Return** v;
6: **End.**

RAM 128 GB. Image retrieval stage is carried out on computer, which has CPU Intel(R) CoreTM i7-2620M, CPU 2,70GHz, RAM 4GB, and OS Windows 7 Professional.

The results of experiment are evaluated on ImageCLEF dataset, which has 20,000 images including 276 classes and stores in 41 folders (from 0-th folder to 40-th folder); the volume size of this dataset is 1.64 GB. In order to assess effectiveness of proposed method, the experiment is shown values including precision, recall, and F-measure. These values are described by the recall-precision and ROC curves. The formulas of these values are as follows [1]

$$\text{precision} = \frac{|\text{relevant images} \cap \text{retrieved images}|}{|\text{retrieved images}|}, \tag{4}$$

$$\text{recall} = \frac{|\text{relevantimages} \cap \text{retrievedimages}|}{|\text{relevantimages}|}, \tag{5}$$

$$\text{F-measure} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}. \tag{6}$$

Our empirical data set is divided into two sections, one for training data and one for test data. Number of photos is taken randomly. The results of experiment of SIR-DL are shown in Figure 8, Figure 9, Figure 10, and Figure 11. Performance of SIR-DL is given

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://www.w3.org/2002/07/owl#"
     xml:base="http://www.w3.org/2002/07/owl"
     xmlns:img="http://www.ittc.edu.vn/itcsites/imageclef/imagename#"
     xmlns:des="http://www.ittc.edu.vn/itcsites/imageclef/description#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     xmlns:cat="http://www.ittc.edu.vn/itcsites/imageclef/catetories#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:cla="http://www.ittc.edu.vn/itcsites/imageclef/class#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
    <Ontology/>
    <AnnotationProperty rdf:about="http://www.ittc.edu.vn/itcsites/imageclef/catetories#catetories"/>
    <AnnotationProperty rdf:about="http://www.ittc.edu.vn/itcsites/imageclef/description#description"/>
    <AnnotationProperty rdf:about="http://www.ittc.edu.vn/itcsites/imageclef/imagename#imageID"/>
    <AnnotationProperty rdf:about="http://www.ittc.edu.vn/itcsites/imageclef/imagename#segimage"/>
    <Class rdf:about="http://www.ittc.edu.vn/itcsites/imageclef/catetories#aerostatic-balloon"/>
    <NamedIndividual rdf:about="http://www.ittc.edu.vn/itcsites/imageclef/class#aerostatic-balloon">
        <rdf:type rdf:resource="http://www.ittc.edu.vn/itcsites/imageclef/catetories#air-vehicles"/>
        <cat:catetories rdf:resource="http://www.ittc.edu.vn/itcsites/imageclef/catetories#air-vehicles"/>
        <cat:catetories rdf:resource="http://www.ittc.edu.vn/itcsites/imageclef/catetories#entity"/>
        <cat:catetories rdf:resource="http://www.ittc.edu.vn/itcsites/imageclef/catetories#man-made"/>
        <cat:catetories rdf:resource="http://www.ittc.edu.vn/itcsites/imageclef/catetories#vehicle"/>
        <des:description>9854.eng</des:description>
        <img:imageID>9854.jpg</img:imageID>
        <img:segimage>9854_2.jpg</img:segimage>
    </NamedIndividual>
</rdf:RDF>
```

Figure 5. An example of ontology on ImageCLEF dataset by RDF/XML

```
PREFIX cla: <http://www.ittc.edu.vn/itcsites/imageclef/class#>
PREFIX cat: <http://www.ittc.edu.vn/itcsites/imageclef/catetories#>
PREFIX img: <http://www.ittc.edu.vn/itcsites/imageclef/imagename#>
PREFIX des: <http://www.ittc.edu.vn/itcsites/imageclef/description#>
SELECT DISTINCT *
WHERE{
{select * where { cla:column img:imageID ?img.}}
 UNION
 {select * where { cla:rodent img:imageID ?img.}}
 UNION
 {select * where { cla:guitar img:imageID ?img.}}
 UNION
 {select * where { cla:wolf img:imageID ?img.}}
 UNION
 {select * where { cla:ground-vehicles img:imageID ?img.}}
 UNION
 {select * where { cla:vehicle img:imageID ?img.}}
 UNION
 {select * where { cla:hedgehog-porcupine img:imageID ?img.}}
 UNION
 {select * where { cla:bull img:imageID ?img.}}
 UNION
 {select * where { cla:plant-pot img:imageID ?img.}}
}
```

Figure 6. A SPARQL command

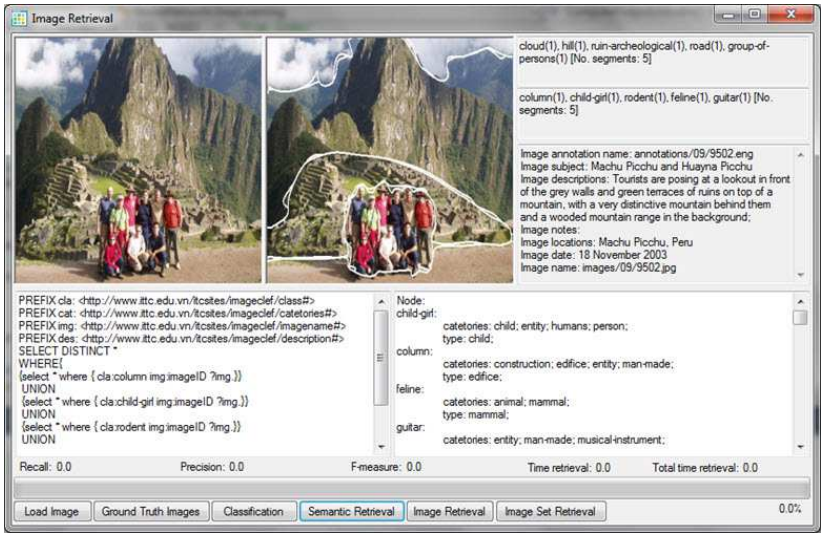

Figure 7. The application of SIR-DL for semantic-based image retrieval

Table 1. Performance of image retrieval of proposed method on ImageCLEF dataset

ID	No. images	Ave. recall	Ave. precision	Ave. F-measure	Ave. query time (ms)
00-10	2460	0.401259	0.609260	0.431001	875.1342
11-20	1797	0.410326	0.589953	0.430598	829.8472
21-30	1239	0.418620	0.607360	0.440907	828.1287
31-40	1431	0.437902	0.640513	0.470151	674.1342

in Table 1, which has 6927 query images; the averages of performance are 0.4123; 0.6054; 0.4381; 834.1439. Accuracies and errors in the process training of deep neural network are shown in Figure 9. The values of accuracy increase and errors decrease show that DLT training algorithm is exact in experiment. Figure 10 shows the curves of Precision-Recall and ROC, each curve describes a set of query images, which are retrieved. The areas under these curves show that the accuracy of image retrieval is not high; however, it has many curves above the average line.

Figure 11 shows the average of precision, recall, and F-measure of 39 subjects on ImageCLEF dataset. The values of Mean Average Precision (MAP) of proposed method are compared with other methods on the same dataset. They are described in Table 2, which shows that the accuracy of SIR-DL is higher than that of other methods.

In Y.Cao's work [29], the author performs image retrieval rely on CNN using AlexNet. In this method, two vectors are created including the image vector and the sentence vector. Then the authors search similar images but it does not create semantic of image content as well as does not query on Ontology. In this way, the authors only find similar images and can not find the semantic of each image, so this method only performs the first stage of the semantic image retrieval. So that, the accurate of this method more than the one of proposed

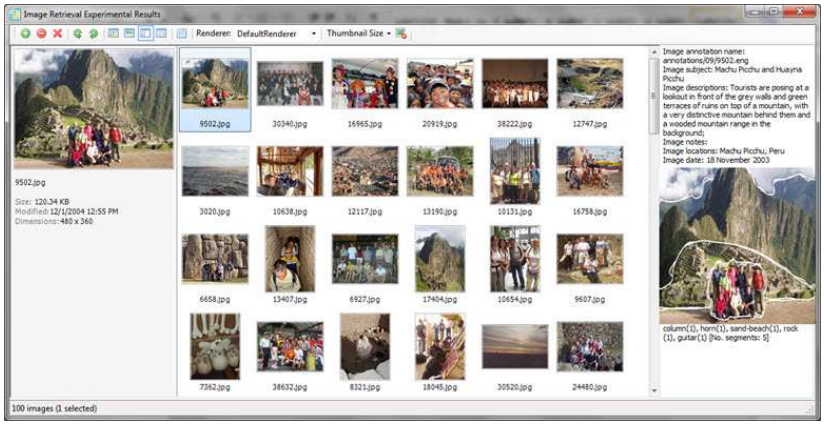

Figure 8. The result of semantic image retrieval using SIR-DL architecture

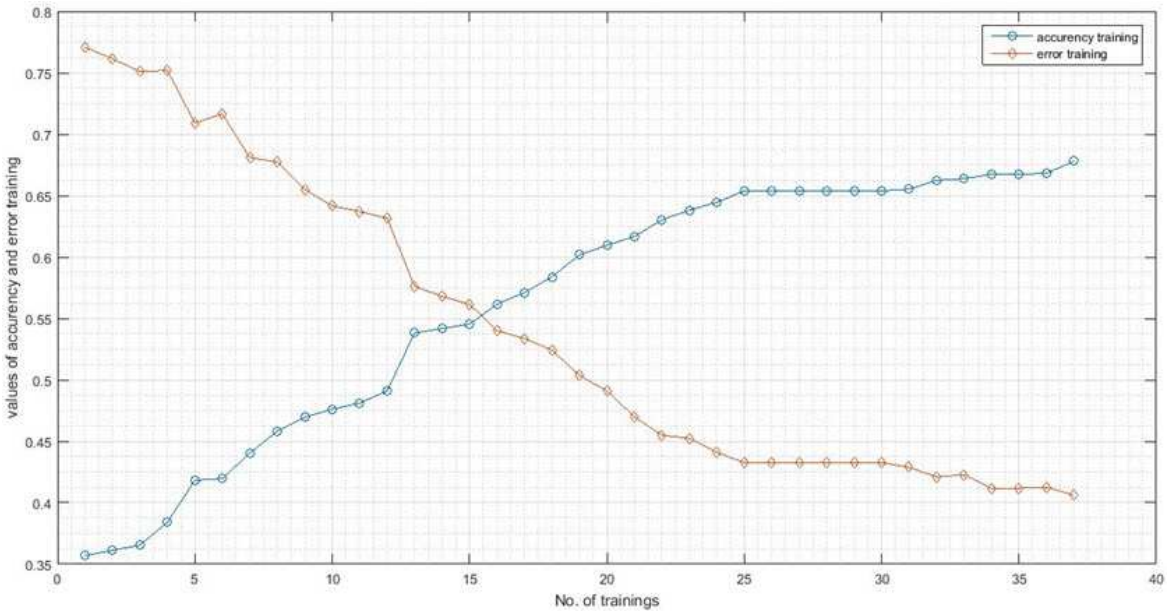

Figure 9. The accuracies and errors training of DNN in SIR-DL

method of this paper. We compared this work to show the difference between two problems, including the image retrieval based on semantic and the semantic-based image retrieval.

In the proposed method, we extracted semantics of image from content based on DNN and query on Ontology. Therefore, each query image, we generate semantics from image content and then automatic create a query based on SPARQL language. This shows that we can interpret the content of each image and easily apply in multimedia systems such as Hospital Information System, Geographic Information System, Digital Library System, etc. In addition, if our proposed method compared with the last four years, our results are more effective than the results of the other works. This shows that the effectiveness of our work.

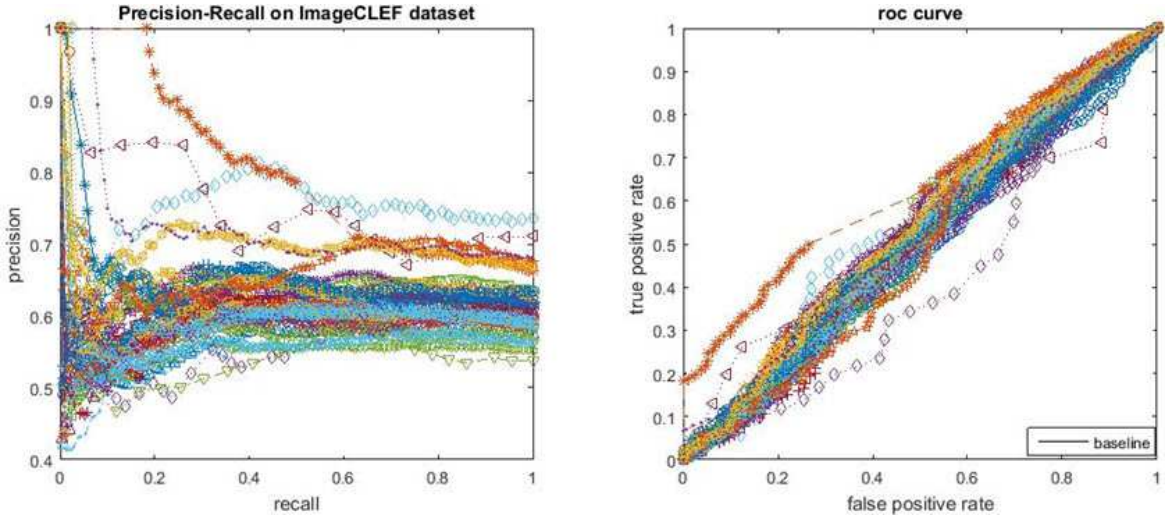

Figure 10. The Precision-Recall and ROC curves of SIR-DL on ImageCLEF dataset

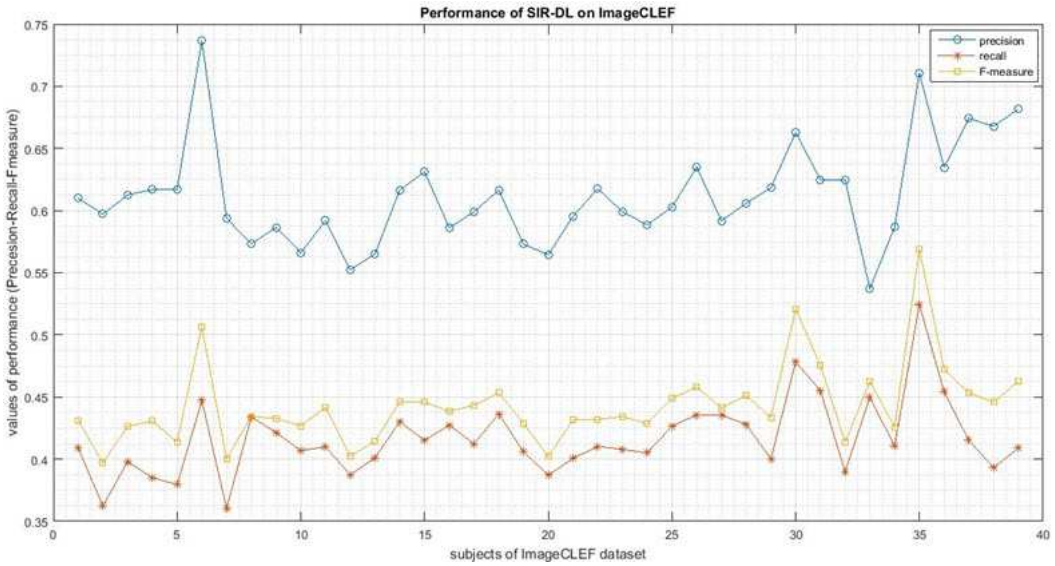

Figure 11. The average of Precision, Recall, F-measure of SIR-DL on ImageCLEF dataset

6. CONCLUSIONS AND FUTURE WORKS

In this paper, the model of SIR-DL was built to retrieve similar images based on semantics. The process of image retrieval is done by semantic classification using image content from which create a visual word vector to generate a SPARQL query. The results of image retrieval were accessed from the Ontology, which describes image meaning of ImageCLEF dataset. On the base of SIR-DL model, the algorithms were proposed and after that they were assessed performance based on the values of recall, precision, F-measure, and query

Table 2. Comparison mean average precision (MAP) of methods on ImageCLEF dataset

Methods	MAP
M. Jiu, 2017 [20]	0.5970
H. Cevikalp 2017 [14]	0.4678
Y. Cao, 2016 [29]	0.7236
V. Vijayarajan, 2016 [26]	0.4618
S. Fakhfakh, 2015 [11]	0.5400
C.A. Hernndez-Gracidas, 2013 [6]	0.5826
our proposed method (SIR-DL)	0.6054

time (milli-seconds). The experimental results of SIR-DL were compared with the result of the other methods on the same dataset from which show that the proposed method is relatively effective. The experiments have shown the correctness of the proposed model and algorithm, so SIR-DL can be improved for semantic image retrieval systems. The future works of SIR-DL are creating the process of online extraction. Then, the training and finding of images can be extracted for online data from WWW based on URIs from which creates image retrieval systems such as HIS, GIS, etc.

ACKNOWLEDGMENT

The authors wish to thank the Faculty of Information Technology, HCMC University of Food Industry, the Faculty of Information Technology, University of Sciences/Hue University, Vietnam, and the Center for Training and Fostering, Quang Nam University, Vietnam. We would also like to thank the anonymous reviewers for their helpful comments and valuable suggestions.

REFERENCES

[1] N. R. A. Alzubi, A. Amira, "Semantic content-based image retrieval: A comprehensive study," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 20–54, 2017.

[2] L. J. A.B. Spanier, D. Cohen, "A new method for the automatic retrieval of medical cases based on the radlex ontology," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 3, pp. 471–484, 2017.

[3] Y. Alqasrawi, "Bridging the gap between local semantic concepts and bag of visual words for natural scene image retrieval," *International Journal of Sensors Wireless Communications and Control*, vol. 6, no. 3, pp. 174–191, 2016.

[4] L. S. C. Hernndez-Gracidas, "Markov random fields and spatial information to improve automatic image annotation," in *Advances in Image and Video Technology. PSIVT 2007. Lecture Notes in Computer Science.* Springer, Berlin, Heidelberg, 2007, pp. 879–892. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-77129-6_74

[5] C. B. C. Kurtz, A. Depeursinge, "A semantic framework for the retrieval of similar radiological images based on medical annotations," in *IEEE International Conference*

on Image Processing, TCIP. Paris, France: IEEE, 2004. [Online]. Available: https://ieeexplore.ieee.org/document/7025454/

[6] M. M.-y.-G. C.A. Hernndez-Gracidas, L.E. Sucar, "Improving image retrieval by using spatial relations," *Multimedia Tools and Applications*, vol. 62, no. 2, pp. 479–505, 2013.

[7] Deloitte, "Photo sharing: trillions and rising," Deloitte Touche Tohmatsu Limited, Deloitte Global, Tech. Rep., 2016.

[8] B. Z. Y. et al., "I2t: Image parsing to text description," in *Proceedings of the IEEE*, vol. 98, no. 8. IEEE, 2010, pp. 1485–1508. [Online]. Available: https://ieeexplore.ieee.org/document/5487377/

[9] C. B. et al., *Emerging Semantic-Based Applications*. Springer, 2016.

[10] H. E. et al., "The segmented and annotated iapr tc-12 benchmark," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 419–428, 2010.

[11] S. F. et al., "Image retrieval based on using hamming distance," *Procedia Comp. Sci.*, vol. 73, pp. 320–327, 2015.

[12] Z. Z. V. S. G. Castanon, Y. Chen, "Efficient activity retrieval through semantic graph queries," in *International conference on Multimedia*. Brisbane, Australia: ACM, 2015, pp. 391–400. [Online]. Available: https://dl.acm.org/citation.cfm?id=2806229

[13] M. Grubinger, "Analysis and evaluation of visual information systems performance," School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, Tech. Rep., 2007.

[14] S. O. H. Cevikalp, M. Elmas, "Large-scale image retrieval using transductive support vector machines," *Computer Vision and Image Understanding*, vol. 173, pp. 2–12, 2018.

[15] D. R. J. Gantz, "The digital universe in 2020: Big data, bigger digi tal shadows, and biggest growth in the far east." IDC iView, EMC Corporation, Tech. Rep., 2012 and 2014.

[16] S. H. P. W.-J. Z. Y. Z. J. L. J. Wan, D. Wang, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the 22nd ACM International conference on Multimedia*. Orlando, Florida, USA: ACM, 2014, pp. 157–166. [Online]. Available: https://dl.acm.org/citation.cfm?id=2654948

[17] H. A. L. Deligiannidis, *Emerging Trends in Image Processing: Computer Vision, and Pattern Recognition*, ser. Graduate Texts in Mathematics. Elsevier, USA: Morgan Kaufmann, Waltham, MA 02451, 2015.

[18] Y. Li, "Semantic image similarity based on deep knowledge for effective image retrieval," Department of Computer Science., Hong Kong Baptist University, Tech. Rep., 2014.

[19] E.-I. C. M.-H. Lee, S. Rho, "Ontology based user query interpretation for semantic multimedia contents retrieval," *Multimedia Tools and Applications*, vol. 73, no. 2, pp. 901–915, 2014.

[20] H. S. M. Jiu, "Nonlinear deep kernel learning for image annotation," *IEEE Trans. on Image Processing*, vol. 26, no. 4, pp. 1820–1832, 2017.

[21] A. T. M. Tzelepi, "Deep convolutional learning for content based image retrieval," *Neurocomputing*, vol. 275, pp. 2467–2478, 2018.

[22] L. G. P. Muneesawang, N. Zhang, *Multimedia Database Retrieval: Technology and Applications*, ser. Graduate Texts in Mathematics. Springer, New York Dordrecht London, 2014.

[23] T. M.-T. S.-A. R. M. M. S. Jabeen, Z. Mehmood, "An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-wordsmodel," *PLoS ONE*, vol. 13, no. 4, pp. 1–24, 2018.

[24] H. S. Pandey, PriteeKhanna, "A semantics and image retrieval system for hierarchical image databases," *Information Processing & Management*, vol. 52, no. 4, pp. 571–591, 2016.

[25] S. M. Sohail Sarwar, Zia Ul Qayyum, "Ontology based image retrieval framework using qualitative semantic image descriptions," *Procedia Computer Science*, vol. 22, no. open access, pp. 285–294, 2013.

[26] P. T. M. L. V. Vijayarajan, M. Dinakaran, "A generic framework for ontology based information retrieval and image retrieval in web data," *Human-centric Computing and Information Sciences*, vol. 6, no. 18, pp. 1–30, 2016.

[27] J. Z. N. C.-Y. W. X. Xie, X. Cai, "A semantic-based method for visualizing large image collections," *IEEE Transactions on Visualization and Computer Graphics, IEEE Computer Society*, pp. 1–15, 2018. [Online]. Available: https://doi.org/10.1109/TVCG.2018.2835485

[28] T. X. C. X.-K. Y. W.-Y. M. T. Z. Y. Bai, W. Yu, "Bag-of-words based deep neural network for image retrieval," in *International Conference on Multimedia*. Orlando, Florida, USA: ACM, 2014, pp. 229–232. [Online]. Available: https://dl.acm.org/citation.cfm?id=2656402

[29] J. W. Q. Y.-P. Y. Y. Cao, M. Long, "Deep visual-semantic hashing for cross-modal retrieval," in *Inter. Conf. on Knowl. Discovery and Data Mining, SIGKDD*. California, USA: ACM, 2016, pp. 1445–1454. [Online]. Available: https://dl.acm.org/citation.cfm?id=2939812

[30] D. Z. Y. Chou, D.J. Lee, "Semantic-based brain mri image segmentation using convolutional neural network," in *Advances in Visual Computing. ISVC 2016. Lecture Notes in Computer Science*, vol. 10072. Springer, Cham, 2016, pp. 628–638. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-50835-1_56

[31] W. H. Z. Zheng, H. Bu, "An approach to classify visual semantic based on visual encoding with the convolutional neural network," in *Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery*. Zhangjiajie, China: IEEE, 2015, pp. 854–858. [Online]. Available: https://ieeexplore.ieee.org/document/7382054/

Received on September 13, 2018
Revised on December 13, 2018