Ontology-based Indexing of Annotated Images using Semantic DNA and Vector Space Model

Syed Abdullah Fadzli
Faculty of Informatics,
Universiti Sultan Zainal Abidin,
Kuala Terengganu, Malaysia.
scesae@cardiff.ac.uk,

Rossitza Setchi
Knowledge Engineering Systems Group,
Cardiff University,
Wales, United Kingdom.
setchi@cardiff.ac.uk

Abstract—**The study presented in this paper focuses on the pre-processing stage of image retrieval by proposing an ontology-based indexing approach which captures the meaning of image annotations by extracting the semantic importance of the words in them. The indexing algorithm is based on the classic vector-space model that is adapted by employing index weighting and a word sense disambiguation. It uses sets of Semantic DNA, extracted from a lexical ontology, to represent the images in a vector space. As discussed in the paper, the use of Semantic DNA in text-based image retrieval aims to overcome some of the major drawbacks of well known traditional approaches such as 'bags of words' and term frequency- (TF) based indexing. The proposed approach is evaluated by comparing the indexing achieved using the proposed semantic algorithm with results obtained using a traditional TF-based indexing in vector space model (VSM) with singular value decomposition (SVD) technique. The experimental results show that the proposed ontology-based approach generates a better-quality index which captures the conceptual meaning of the image annotations.**

Keywords – semantic image indexing; image annotation, vector space model

I. INTRODUCTION

With the rapid growth of the internet and the world-wide-web, the amount of digital image data available to users has grown enormously. Image databases are becoming larger and more widespread, and there is a growing need for more effective and efficient image retrieval (IR) systems.

Most IR systems adopt a two-stage approach to retrieve annotated images from image collections, which includes indexing and searching [1]. Indexing involves pre-processing of image annotations with the aim of creating a suitable and useful data representation, and storing it in an index base. The index terms used to represent each image may be obtained directly from the annotations or generated using complex algorithms. During search, the index terms used in the search query are compared with the index base to retrieve images with annotations most similar to the query.

Most indexing approaches use traditional 'bags of words' and term frequency (TF) methods to create sets of weighted index terms to represent images. The classic vector space model (VSM) is one of the most widely used models for representation [2, 3]. Using VSM, images are represented by their TF vectors by computing the number of times a term is repeated within a text associated with the image (e.g. an annotation, a document containing the image, etc.). As a result, the higher the term frequency, the higher its index weight. However, indexing based on 'bags of words' and TF values can lead to poor retrieval performance due to the following three main reasons [1, 2]. Firstly, the high dimensionality of the term vectors which is the size of the vocabulary across the entire dataset, leads to massive computational costs of indexing and searching [4, 5]. Secondly, the ambiguity of term senses may lead to many unrelated images being retrieved just because they match some of the query terms [6]. In other words, these approaches do not consider the context of the terms used, which leads to inaccuracies when calculating their weight. Finally, the variation in the terms used to describe a particular concept may result in relevant images not being retrieved because their annotations or the text associated with them do not contain any of the query terms [7].

To address these shortcomings, this paper proposes a novel indexing approach that transforms an image annotation into a combination of Semantic DNA strings (SDNA), represented in a SDNA vector space. An SDNA [8, 9] is a string of numbers which represents a sense of word according to the hierarchical structure of a lexical ontology. Previous work has shown that the extraction of Semantic DNA based on SDNA similarities has a number of drawbacks including the SDNA weighting scheme which is based on the co-occurrences of SDNAs. This scheme is biased towards long annotation where images with longer annotations tend to get higher weight compared to those with shorter annotations. An SDNA is assigned to each term (i.e. a word or a phrase) in an image annotation using an SDNA disambiguation method which weights the SDNA strings for each term to determine its sense and concept. However, previous work has shown that the use of concept likelihood in selecting the most relevant SDNA for each term leads to poor sense disambiguation, where a concept with rare global frequency will always have poor chance of been chosen. The use of index table to compare the queries and the index also results in poor retrieval performance as different images are represented by index tables with different size, depending on the size of the annotation.

This paper extends previous work by proposing a normalization technique designed to enhance the retrieval of documents with various length. The SDNA disambiguation is improved by applying the Hamming distance method and information theory in selecting the most relevant SDNA for each term. Furthermore, the SDNA indexing process is further

978-1-61284-353-7/11/$26.00 ©2011 IEEE

improved by adapting the classic vector space model (VSM) to provide better representation of the image annotations. Unlike the classical model, this paper suggests that each SDNA extracted from the annotations can be treated as an independent dimension in a high-dimensional vector space. Any image can then be represented in this space by a unique vector. For each image, the SDNA vector capturing the contextual meaning of the image annotation is calculated and stored in an SDNA index matrix.

Section II describes the traditional approach which employs 'bags of words' and the TF-IDF method. The lexical ontology used in the proposed approach is briefly introduced in section III. Section IV explains the proposed approach while section V describes the data collection used in the experiments. Section VI presents experimental results and discussion. Section VII concludes the paper.

II. Traditional Approach

In traditional indexing, the text of the image annotations is split into words, and some words (e.g. the stop words) are removed. Then the remaining words are transformed into their root or stem forms [1, 2, 3, 7]. The resulting stem words are used as terms for each image annotation. The two most popular stemmers are the Lovins stemmer [11] and the Porter stemmer [12]. The Lovins stemmer removes over 260 different suffixes using the longest-match algorithm. The Porter stemmer removes about 60 suffixes in a multiple-step approach, each step successively removing suffixes or transforming the stem.

Since its introduction, the VSM is the most popular model used in information retrieval. In this model, documents and queries are represented by vectors in a n-dimensional space, where n is the number of distinct terms used in a corpus. Each axis in this n-dimensional space represents a term. Given a query, documents most similar to the query vector are retrieved in a ranked list. Latent semantic indexing (LSI) is one of the proven successful indexing and retrieval method that uses VSM model based on singular value decomposition (SVD) technique [3]. SVD is used to identify patterns of term distribution across the document. LSI considers documents that have similar terms in common to be semantically closed.

The similarity between a query and a document is often defined as the cosine of the angle between their respective vectors. The ranking of the retrieved documents using similarity measures is undoubtedly the biggest challenge in retrieval systems as the goal is to obtain the most relevant results without spending a lot of time browsing through the retrieved documents.

In the traditional approach, a term weight is used to represent the relative importance of a term t in a document d. It is usually computed following the term-frequency - inverse document frequency (TF-IDF) weighting scheme:

$$tfidf_{t,d} = (n_{t,d} / |d|) \times log\ (|D| / |\{d: t \in d\}|) \qquad (1)$$

where $n_{t,d}$ is the number of occurrences of the term t in a document d while $|d|$ is the size of the document d. |D| denotes the number of documents in the collection and $|\{d: t \in d\}|$ is the number of documents containing the term t.

III. Lexical Ontology

This paper proposes a novel text-based image indexing approach which uses sense disambiguation to reveal the contextual meaning of the annotations. The use of SDNA in this approach is in the centre of the three main contributions of this research: (i) determining the semantic relationships between terms contained in image annotations, (ii) SDNA-based term sense disambiguation, and (iii) creating the SDNA index representing the images.

As defined in [8, 9], an SDNA is a string of numbers derived from a lexical ontology. This study uses a lexical ontology called OntoRo developed based on the digital version of the Roget's Thesaurus [13]. The effectiveness of OntoRo as a lexical ontology used in ontology tagging has been successfully exploited in the TRENDS project funded by the Framework Programme of the European Commission [10]. The SDNA is extracted from six OntoRo hierarchical levels, namely *class number, section number, subsection number, concept number, part-of-speech type* and *paragraph number*. For example, as Table I shows, the word *bond* has 16 senses, each of them having a distinct SDNA (see the online version of OntoRo [14]).

TABLE I. THE SDNA OF THE TERM 'POVERTY'

SDNA	Concept Sense	SDNA	Concept Sense
1-3-10-33-1-1	Smallness	5-27-62-627-1-2	Requirement
1-3-10-35-1-1	Inferiority	5-27-63-636-1-2	Insufficiency
1-3-10-37-1-1	Decrease	5-27-63-649-1-1	Uncleanness
1-3-11-53-1-5	Part	5-27-63-655-1-1	Deterioration
1-4-14-69-1-2	End	5-30-69-731-1-1	Adversity
4-25-58-572-1-1	Feebleness	5-34-73-774-1-1	Non-ownership
5-26-59-596-1-1	Necessity	5-34-76-801-1-1	Poverty
5-26-61-616-1-1	Evil	6-38-89-945-1-1	Asceticism

An SDNA represents a unique paragraph in Roget's Thesaurus consisting of words and phrases that can be used to explain a similar idea or concept. Given an idea or concept, OntoRo can help find words which can be used to express it. These words are not just synonyms but words that could clarify the thought. A term may have more than one possible sense which means that it may be contained in a number of paragraphs. Therefore, each term in an image annotation can produce a set of possible SDNA candidates that could represent different meaning.

IV. Proposed Approach

The goal of the proposed indexing approach is to select a set of SDNA that most adequately represents the underlying meaning of the image annotations by disambiguating the meaning of each term and assigning the most relevant SDNA to it. The approach is based on the assumption that the words contained in the image annotations are used to convey common meaning, which can be used to represent the image.

The approach is composed of several processes as illustrated in Fig. 1. First, keywords analysis analyzes the image annotations and produces a list of relevant terms. Next, all possible SDNA candidates are generated for each of these terms, and the most relevant among them are selected. Finally, a normalization process removes the advantage of long over short annotations by normalizing the weights of the selected SDNA, before storing them in an SDNA Index matrix.

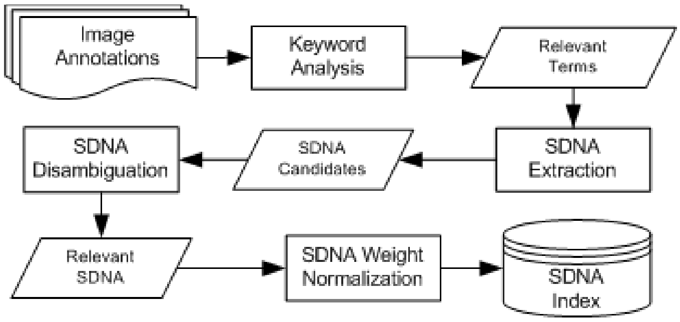

Figure 1. Semantic Image Indexing process flow.

A. Keyword Analysis

The keywords analysis process aims to identify the relevant terms among the existing words and phrases in each image annotation by directly comparing all words and phrases in the image annotations with the lexical ontology. A word or phrase is defined as a relevant term if it exists in the ontology. Stop words such as ‘the’ and ‘an’ are removed while stemming is only applied if a word or phrase cannot be matched directly with the ontology, in order to preserve the contextual meaning of the word or phrase. The Porter stemmer is used in this approach, which produces a less complex and higher speed algorithm.

If α is an image annotation and K is the set of all relevant terms w in it ($w \in K$), then $K_\alpha = \{ w_1, w_2, w_3, \ldots, w_n\}$, where n is the number of relevant terms in image annotation α or $|K_\alpha| = n$.

B. SDNA Extraction

The purpose of this process is to extract all possible SDNA candidates’ strings for each relevant term w_i in the annotation, where $w_i \in K_\alpha$. As explained earlier, each relevant term may have several possible senses (i.e. lexical ambiguity), where each sense is represented by an SDNA, therefore:

$$Senses(w_i) = \{w_i s_1, w_i s_2, \ldots, w_i s_n\} \qquad (2)$$

where $w_i s_j$ denotes an SDNA candidate s_j of the term w_i. Combining all SDNA candidates of each relevant term K for an image α will form a set of SDNA of the image, as defined by (3):

$$SetSDNA(\alpha) = \bigcup_{i=1}^{|K_\alpha|} Senses(w_i) \qquad (3)$$

C. SDNA Disambiguation

SDNA disambiguation is a process of determining, for each relevant term, which SDNA is the one which corresponds most closely to the context of the annotation. It can be formally described as the task of identifying a mapping A from terms to senses, such that $A(w_i) \in Senses(w_i)$. Only the most relevant sense is selected, i.e. $|A(w_i)|=1$, and every term w_i is supplied with a unique $A(w_i)$, that is $|A|=|K|$.

To help determine $A(i)$, information theory is applied as one of the factors in the SDNA weighting. Information Theory [15] suggests that the information contained in a statement is measured by the negative logarithm of the likelihood, meaning that, as likelihood increases, information decreases, therefore decreasing the chance of a sense to be chosen. In other words, an SDNA which appears in many annotations is less important than the one which occurs in a smaller number. This is the global information used in choosing the most relevant sense for each keyword.

Let C be an image collection, where $\alpha \in C$, and s is an SDNA from $SetSDNA(\alpha)$ or $s \in SetSDNA(\alpha)$; $freq(s)$ is the frequency of an SDNA s calculated from all available annotations in the image collection. Each SDNA occurring in an SDNA set is counted as an occurrence of that SDNA. Therefore the information content of an SDNA s is calculated as follows:

$$IC(s) = -log\ (freq(s)\ /\ N) \qquad (4)$$

where N is the cardinality of the image collection.

To create the relationship between an image and an SDNA, the SDNA weight $sw()$ is calculated to measure the importance of a particular SDNA for that image. In other words, an SDNA might be more important to an image than another SDNA if its weight is higher. The method proposed here is based on the following observations:

- Each word in the same OntoRo’s part of speech group (5th level of SDNA) and paragraph (6th level of SDNA) represents the same idea within the same context;
- The words in the same paragraph are semantically closer than the words in the same part of speech group;
- Different words that embody the same idea tend to have similar hierarchies, therefore share similar SDNA;
- Same words with different senses (i.e. lexical ambiguity) tend to belong to different hierarchies, thus producing different SDNA;
- The selection of the most significant SDNA can determine the exact sense of the word.

The proposed method identifies, measures and utilizes the information shared between the SDNA of the words in image annotations. The SDNA weight determines two important factors:

- The most relevant SDNA for each term, selected according to the highest weight, and
- the importance of the SDNA in relation to the image.

In this method, Hamming distance is used to measure the similarity between two SDNAs; it is a way to compare a single SDNA with all other SDNA in the same set of strings. The Hamming distance between two SDNA is the number of level(s) at which the corresponding numbers are different. Let s be an SDNA in *SetSDNA*, such that for any $s \in SetSDNA$, $hamdis(s_i,s_j)$ is the Hamming distance between SDNA s_i and SDNA s_j. It is an expression for the number of mismatched levels in any two SDNA strings. Formally, the similarity between any pair of SDNA is calculated as follows:

$$sim(s_i,s_j) = L - hamdis(s_i,s_j) \tag{5}$$

where L is the number of levels used in an SDNA. In the case of using OntoRo as a lexical ontology, the number of levels is 6, therefore $L=6$. For example, for s_i=*1-2-3-4-5-6* and s_j=*1-2-3-4-10-7*, $hamdis(s_i,s_j)=2$ and $sim(s_i,s_j)=6-2=4$.

A higher similarity score when two SDNA strings are compared indicates higher relevancy with the particular image. Thus, the total similarity value is calculated as the cumulative similarity, formally:

$$totalsim(s_i) = \left(\sum_{j=1:j \neq i}^{|SetSDNA|} sim(s_i, s_j) \right) \tag{6}$$

Table II shows the SDNA to SDNA similarity matrix for an example SDNA Set α, where $SetSDNA(\alpha) = \{w_1s_1, w_1s_2, w_2s_1\}$. $totalsim(w_1s_1)$ is calculated by the summation of all similarity measurements between w_1s_1 and all other SDNA strings in $SetSDNA(\alpha)$ (i.e. w_1s_2 and w_2s_1). The value in position (i,j) of the matrix represents the similarity between SDNA i and j in $SetSDNA(\alpha)$.

Equations (4) and (6) are then used to calculate the weight $sw(s)$ of each SDNA in *SetSDNA*. It is the product of $totalsim(s)$ and the information content of s, as follows:

$$sw(s) = totalsim(s) \cdot IC(s) \tag{7}$$

TABLE II. SDNA TO SDNA SIMILARITY MATRIX

SDNA	$\boldsymbol{w_1s_1}$	$\boldsymbol{w_1s_2}$	$\boldsymbol{w_2s_1}$	***totalsim()***
$\boldsymbol{w_1s_1}$	-	$sim(w_1s_1,w_1s_2)$	$sim(w_1s_1,w_2s_1)$	$totalsim(w_1s_1)$
$\boldsymbol{w_1s_2}$	$sim(w_1s_2,w_1s_1)$	-	$sim(w_1s_2,w_2s_1)$	$totalsim(w_1s_2)$
$\boldsymbol{w_2s_3}$	$sim(w_2s_3,w_1s_1)$	$sim(w_2s_3,w_1s_2)$	-	$totalsim(w_2s_3)$

Finally, for each relevant term, $w_i \in K$, $Senses(w_i)$ represents the SDNA candidates for w_i, where $s_j \in Senses(w_i)$. The SDNA with the highest weight $sw(s_j)$ for each w_i is chosen as the most relevant SDNA for the term $A(w_i)$, thus determining the most relevant sense of w_i in the context of K.

At this step of the indexing, images with longer annotations will most probably get higher average *sw()* compared to those with shorter annotations. The next step is designed to deal with the problem of having annotations of varying lengths in the collection.

D. SDNA Weight Normalization

An efficient normalization technique has to be integrated in order to balance the SDNA weight *sw()* when having SDNA sets with smaller and bigger size.

Different with free text or long documents, image annotations only tend to explain the content of the image (i.e. elements, objects, spatial information, context). Thus, all keywords in the annotation are equally important regardless of the size of the annotations. For example, the annotations of two different images, the first being an image of a tiger, and the second showing a lion, both contain the keyword '*wild'*. However, the first image has a short annotation with 10 words only while the annotation of the second image contains 30 words. In both cases the keyword '*wild'* needs to be treated equally although the size of the two annotations is different. Without normalization, the keyword '*wild'* would have different weight due to the different size of the two annotations.

The normalization method proposed is a probabilistic model based on Okapi BM25 which is one of the most important and widely used methods in information retrieval [16, 17]. It has been thoroughly studied and tested, widely used in industrial applications and proved to be effective. Applying Okapi BM25 method, the SDNA weight calculation is defined as:

$$SW(s_i) = IC(s_i) \cdot \frac{totalsim(s_i) \cdot (k_1 + 1)}{k_1 \cdot (1 - b + b \cdot p(s)) + totalsim(s_i)} \tag{8}$$

where k_1 and b are two tuning parameters which are adjustable according to the requirements of the specific application. k_1 is a positive parameter that calibrates the document frequency scaling. A k_1 value of 0 corresponds to a binary model with no term frequency, and a large value corresponds to using raw term frequency. b is another tuning parameter which determines the scaling by document length, where $b=(0, .., 1)$. $b=1$ yields to fully scaling the term weight by the document length, while $b=0$ yields to no document length normalization. The values of $k_1=9$ and $b=0.80$ have been selected using empirical observations.

E. SDNA Vector

Let S represents the number of all distinct SDNA strings 'existing' in the lexical ontology, then the images can be modeled in an S-dimensional space where each image is represented by an S-dimensional vector.

Fig. 2 shows an example of a three-dimensional SDNA index space, where each image is identified by three distinct SDNA strings. The three dimensional example may be extended to *S* dimensions when *S* distinct SDNA strings are used. Given a query vector, images are retrieved by computing all SDNA vectors and returning the images with the closest vectors; the distance is measured through the angle between them.

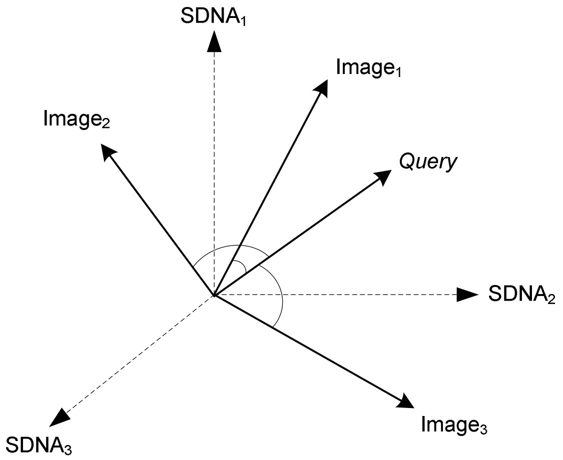

Figure 2. Vector representation of SDNA space

In a vector space, the image similarity between $Image_1$ and query *Q*, $isim(Image_1,Q)$ can be defined as the cosine of the angle between image vectors $Image_1$ and *Q*, formally:

$$isim(Image_1,Q) = Image_1 \cdot Q \,/\, |Image_1| \cdot |Q| \qquad (9)$$

V. Data Collection

Research collaboration with VisconPro Limited has provided this study with 157,639 digital images. VisconPro Limited is one of the UK leading online companies which hosts an image stock website called fotoLIBRA© [18].

They are currently hosting 392,728 high quality images covering a broad range of topics. The images, owned by more than 20,000 photographers, have already been manually annotated by them. Table III provides information about the image collection used.

The fotoLIBRA© collection is selected as an experimental test bed in this research for the following reasons:

- Large collection of high quality images.

While most image libraries include images of various quality, the fotoLIBRA's collection contains a large amount of high quality images. Only images with a certain quality standard are included. Most of the images are originally taken by cameras, though a reasonable manipulation is permissible. Their strict submission guidelines provide their collection with high standard.

- Valuable image content.

The images are included in the collection by the owners for one main reason, which is to sell them to potential buyers. All images uploaded are checked for content and anything which is pornographic, racist, sexist, defamatory, obscene or offensive is rejected.

- Accurate annotation.

All images have been properly annotated by their owners for making their photos findable by others; this is a method called social-organization. The situation is different with other large online image collections such as flickr© and facebook©, where the owners tag their images for self-organization as well as self-communication and social-communication, hence providing less contextual information of the image content [19]. The owners of the images in fotoLIBRA describe their images as accurately as possible in order to increase the chance of being retrieved by the right people who really appreciate their content. fotoLIBRA also provides and strictly follows guidelines on annotating keywords.

- Covering broad range of topics.

As a global picture library, fotoLIBRA offers images illustrating universal subjects from broad categories including animals, architecture, arts, events, health, heritage, leisure, lifestyle, nature, people, plants, science, society, sport, transport, travel and work.

TABLE III. FotoLibra Image Collection Information

Collection Details	Amount
Total Number of Images	157,639
Total Number of Owners	7,326
Total Number of Keywords in Annotation	3,354,388
Average Number of Keywords per Image Annotation	20.78

VI. Experimental Results

In this section, several examples from the experimental results are shown to illustrate the efficiency of the proposed approach. The experiments involve all 157,639 images contained in the data collection. The collection has being indexed using both the traditional approach (as discussed in section II) and the proposed approach, using OntoRo as the lexical ontology. To increase the size of the annotations, the image titles and image keywords are also used and combined with the image annotations. The next subsections describe the experiments conducted.

A. Vector dimension

The first experiment aims at assessing the dimension reduction (defined as the number of index terms used to represent an image) in a vector representation. The experiment shows that the traditional indexing approach has produced a massive 80,225 x 157,539 term to image matrix with 80,225 dimensions corresponding to the 80,225 distinct terms used in the collection.

The proposed semantic approach on the other hand has produced a smaller index matrix of 6,239 x 157,539, with only 6,239 SDNA dimensions. This indicates that compared to the traditional approach, the proposed approach reduces the index dimensions by about 92.2% which leads to better computational efficiency.

B. SDNA disambiguation results

The second experiment aims at assessing the capability of the SDNA disambiguation method to select the most relevant sense of each term in an image annotation. Another aim is to assess the capability of SDNA index to provide additional words that can be used to describe the same idea or concept. Fig. 3 shows an example chosen from the collection. The image is annotated in the following way:

'*Fatherly love, father, child, dad, daddy, son, boy, model, hands, love, bond, life, fingers, summer, sun, fashion, advertising*'.

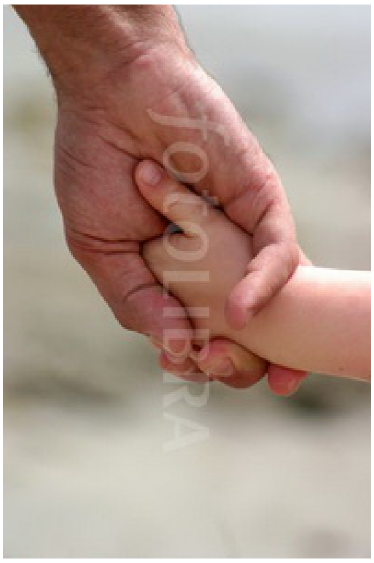

Figure 3. An image example (Image ID 107602).

Table IV lists top 10 terms used in the annotation of the image shown in Fig. 3 (Image ID 107602) and their most relevant SDNA indicating the correct sense of each term within context of the image. The table also lists the words that are semantically related to the particular SDNA. For example, the term '*bond*' has 15 senses including '*relation, union, bond, support, interjacency, hindrance, subjection, prison, promise, compact, security, retention, money, dueness, duty*'. From 15 different senses, the proposed SDNA disambiguation method had chosen the sense*'relation'* as the correct sense of the term '*bond*' in the context of the annotation. According to OntoRo, other words that can be used to describe the same sense include '*relatedness, connectedness, rapport, reference, respect, regard, bearing, direction, concern, concernment, interest, import, importance, involvement'*. These words are considered as semantically related words which could be used to describe the same sense.

SDNA disambiguation results in Table IV shows that the proposed method is capable of identifying the appropriate sense for each term, or at least the closest sense.

C. Indexing results

The aim of the third experiment is to assess the relevancy of the SDNA index used to represent the images. The term '*love*' has been chosen to illustrate the value of the indexing results. According to OntoRo, '*love*' is related to 29 different senses including '*zero, parentage, pleasure, concord, feeling, moral semibility, desire, wonder, repute, friendship, courtesy, love, relationship, endearment, darling, benevolence, jealousy, approbation, disinterestedness,* and *divineness'*. Two senses have been chosen for the purpose of this experiment, '*parentage*' with an SDNA of *1-8-28-169-1-4* and '*romance*' with *6-37-83-887-1-1.* Images related to each of these SDNA strings are retrieved to assess the relevancy of the SDNA index generated.

Fig. 4 and 5 show two sets, each containing 20 images indexed with SDNA *1-8-28-169-1-4* ('*parentage*') and SDNA *6-37-83-887-1-1* ('*romance*'). The image IDs used are those used in the fotoLibra online database [18] where all images can be viewed by searching for their ID numbers. Fig. 4 shows that the top 20 images having an *1-8-28-169-1-4* SDNA ('*parentage*') are about love and close relationship between parents and an offspring. The highest ranked image of holding hands between a father and a son, closely reflects the concept of '*love*' in the sense of '*parentage*'. On the other hand, the top 20 images in Fig. 5 mainly represent the feeling of affection between loving couples, which reflects the concept of '*love*' in the sense of '*romance*'.

These illustrative examples clearly demonstrate that the SDNA produced by the proposed approach is a useful representation of the indexed images. The results obtained using the proposed approach are compared with traditional approach, the term '*love*' is used again to retrieve images indexed by the traditional indexing approach.

The results shown in Fig. 6 indicate that the term '*love*' has been used as an index term in a variety of contexts including romance, name of flower (Love-in-a-mist), parentage and a movie (First Love). These results might be acceptable in many cases but are unsatisfactory if the expected result is images that illustrate a particular idea or an abstract concept.

TABLE IV. SDNA INDEX FOR IMAGE ID 107602

Term	Possible Senses	SDNA Index	SDNA Sense	Related words
daddy	*parentage*	1-8-28-169-1-3	Parentage	*paternity; fatherhood; father; dad; daddy; pop; papa; pater; governor; the old man; single dad;*
dad	*parentage*	1-8-28-169-1-3	Parentage	*paternity; fatherhood; father; dad; daddy; pop; papa; pater; governor; the old man; single dad;*
fatherly	*parentage, benevolence*	1-8-28-169-2-1	Parentage	*parental; paternal; maternal; matronly; fatherly; fatherlike; motherly; step motherly; family;*
father	*consanguinity, causation, propagation, parentage, male, clergy*	1-8-28-167-2-3	Propagation	*generate; evolve; produce; bring into being; bring into the world; usher into the world; give life to; bring into existence;*
son	*posterity, male*	1-8-28-170-1-2	Posterity	*descendant; son; daughter; chip off the old block; infant; child; scion; shoot; sprout; beneficiary; love child;*
love	*zero, parentage, pleasure, concord, feeling, moral semibility, desire, desire, wonder, repute, friendship, courtesy, love, romance, endearment, darling, benevolence, jealousy, approbation, disinterestedness, divineness*	3-15-48-376-3-1	Pleasure	*enjoy; relish; like; quite like; love; adore; feel pleasure; experience pleasure; take pleasure in; be pleased; thrill to; be excited; luxuriate in; revel in; riot in;*
boy	*child, male*	1-6-22-132-1-2	Child	*youngster; juvenile; young person; young adult; young hopeful; young'un; young people; yoof; youth; boy;*
bond	*relation, union, bond, support, interjacency, hindrance, subjection, prison, promise, compact, security, retention, money, dueness, duty*	1-2-5-9-1-1	Relation	*relatedness; connectedness; rapport; reference; respect; regard; bearing; direction; concern; concernment; interest; import; importance; involvement;*
hands	*safety, agent*	5-27-63-660-1-2	Safety	*protection; conservation; preservation; insurance; surety; caution; patronage; care; sponsorship; good offices; auspices; aegis; fatherly eye; aid; protectorate;*
child	*child, posterity, littleness, fool, innocence*	1-8-28-170-1-1	Posterity	*progeny; issue; offspring; young; little ones; child; breed; race; brood; seed; litter; farrow; spawn; young creature;*
model	*copy, prototype, superiority, composition, conformity, attribution, littleness, littleness, form, comparison, supposition, manifestation, representation, painting, sculpture, sculpture, plan, perfection, amusement, repute*	1-2-7-22-1-1	Copy	*copy; exact copy; clone; reproduction; replica; replication; facsimile; tracing; fair copy; transcript; transcription; counterpart; analogue; cast; death mask;*

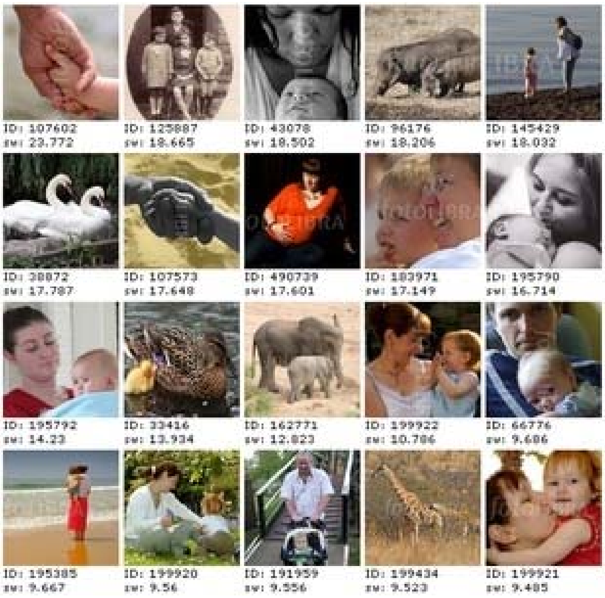

Figure 4. Images related to '*love'* in the sense of '*parentage'*.

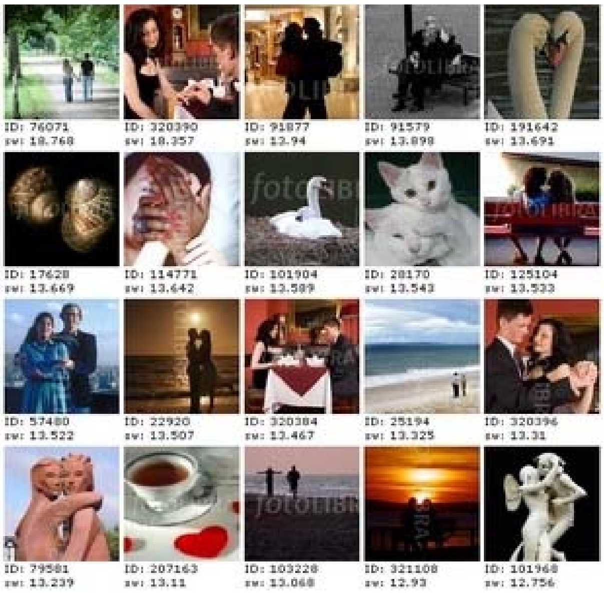

Figure 5. Images related to '*love'* in the sense of '*romance'*.

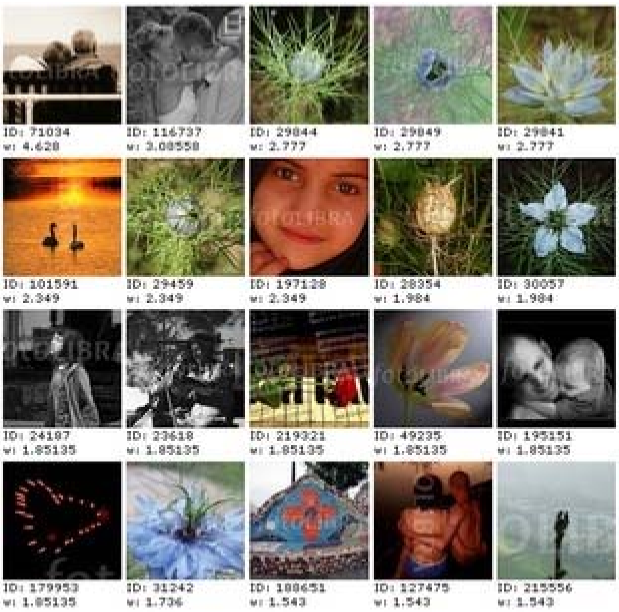

Figure 6. Images related the term *'love'* using traditional indexing approach.

VII. CONCLUSION

The paper presents a novel approach to image indexing that uses an SDNA vector space to model image annotations. The proposed approach aims at overcoming major drawbacks of traditional approaches which use 'bag of words' and the TF-IDF weighting scheme, namely the high dimensionality of the index vector generated, and the lack of term disambiguation and semantic expansion (links to words used to describe the same concept or idea). The comparison between the indexes created using both the proposed approach and the traditional approach indicates that the proposed approach offers 92.2% reduction in index dimensions which helps to improve the process performance.

The experimental results indicate that the SDNA disambiguation method is capable of identifying the relevant sense of each term according to the annotation context. Each SDNA chosen, which represents the sense of a term, is also related to several other semantically related words.

In summary, the proposed approach overcomes some of the major drawbacks of traditional indexing approach and could potentially be useful in information retrieval systems. Future work includes combining Semantic DNA indexing with content-based image retrieval to eliminate the full dependency of this method on image annotations.

ACKNOWLEDGMENT

Special thanks to all members of the Knowledge Engineering Systems group for their valuable help with this research, the Ministry of Higher Education of Malaysia and Universiti Sultan Zainal Abidin in Malaysia for their sponsorship.

REFERENCES

[1] Baeza-Yates, R and Ribeiro-Neto, B. "Modern Information Retrieval". ACM Press, ISBN: 020139829, 1999.

[2] Salton, G., Wang, A., Yang, C.S. "A vector space model for automatic indexing", Communication of the ACM 18 (11) (1975) 613–620.

[3] Salton, G., and McGill, M.: "Introduction to Modern Information Retrieval". McGraw-Hill, New York (1983)

[4] Kang, B.-Y. and S.-J. Lee. "Document indexing: a concept-based approach to term weight estimation.". In Proc. of Information Processing & Management 41(5): 1065-1080, (2005)

[5] Manning, C.D. & Hinrich, S. "Foundations of statistical natural language processing" (pp. 529–574). Cambridge, Massachusetts: MIT Press, 2001.

[6] Zhang, W., Yoshida, T., Tang, X. "A comparative study of TF*IDF, LSI and multi-words for text classification". Expert Systems with Application, vol. 38 (3), (2011)

[7] Zhang, W., Yoshida, T., Tang, X. "TF TFIDF, LSI and multi-word in information retrieval and text categorization", In: Proc. of IEEE International Conference Systems, Man and Cybernatics, Singapore, (2008)

[8] Fadzli, S.A., Setchi, R., "Semantic Approach to Text-based Image Retrieval Using a Lexical Ontology". In: Proc. of the Knowledge Engineering and Emotion Research Int. Conf., Paris, France, March 2-4, 2010, pp. 866-876

[9] Fadzli, S.A., Setchi, R., "Semantic Approach to Image Retrieval Using Statistical Models Based on a Lexical Ontology". In: Proc. of the 14th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Cardiff, Wales, 2010.

[10] Setchi R, Bouchard C, "In search of design inspiration: a semantic-based approach", ASME Journal of Computing and Information Science in Engineering, invited paper for a Special Issue on Knowledge-Based Design (23 pages) , 10 (3), 2010.

[11] J.B. Lovins, "Development of a stemming algorithm", Mechanical Translation and Computational Linguistics 11 (1–2) (1968) 22–31.

[12] M.F. Porter, "An algorithm for suffix stripping", Program 14 (3) (1980) 130–137.

[13] Roget, P. M., "Roget's thesaurus of English words and phrases". Longman Harlow, Essex, 2004.

[14] OntoRo Online, http://kes.engin.cf.ac.uk/sdna/ontoro, last accessed 30 January 2011.

[15] Ross, S. M., "A first course in probability". Prentice Hall, 2002.

[16] Beaulieu, M. M. et al. eds., "Okapi at TREC-5". Proceedings of the Fifth Text REtrieval Conference. Gaithersburg, USA, 1997.

[17] Robertson, S. E. et al. eds., "Okapi at TREC-7". Proceedings of the Seventh Text REtrieval Conference. Gaithersburg, USA, 1998.

[18] fotoLIBRA, http://www.fotolibra.com, last accessed 30 January 2011.

[19] Ames, M. and Naaman, M. eds., "Why we tag: motivations for annotation in mobile and online media." ACM, 2007.