Multimed Tools Appl (2012) 61:69–86
DOI 10.1007/s11042-010-0707-8

Multimodal concept fusion using semantic closeness for image concept disambiguation

Ahmad Adel Abu-Shareha · Rajeswari Mandava · Latifur Khan · Dhanesh Ramachandram

Published online: 11 January 2011

Abstract In this paper we show how to resolve the ambiguity of concepts that are extracted from visual stream with the help of identified concepts from associated textual stream. The disambiguation is performed at the concept-level based on semantic closeness over the domain ontology. The semantic closeness is a function of the distance between the concept to be disambiguated and selected associated concepts in the ontology. In this process, the image concepts will be disambiguated with any associated concept from the image and/or the text. The ability of the text concepts to resolve the ambiguity in the image concepts is varied. The best talent to resolve the ambiguity of an image concept occurs when the same concept(s) is stated clearly in both image and text, while, the worst case occurs when the image concept is an isolated concept that has no semantically close text concept. WordNet and the image labels with selected senses are used to construct the domain ontology used in the disambiguation process. The improved accuracy, as shown in the results, proves the ability of the proposed disambiguation process.

Keywords Disambiguation · Multi-modal data · Ontology · Path length · Semantic closeness

1 Introduction

The main focus of this work is to disambiguate the concepts that are extracted from visual stream with the help of identified concepts from associated textual stream using the semantically related information that may exist in the two related media. The semantically related information is identified based on exploiting ontologies (knowledge based). Recall

A. A. Abu-Shareha · R. Mandava (✉) · D. Ramachandram
School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia
e-mail: mandava@cs.usm.my

A. A. Abu-Shareha
e-mail: adel@cs.usm.my

D. Ramachandram
e-mail: dhaneshr@cs.usm.my

L. Khan
Department of Computer Science, University of Texas at Dallas, Richardson, TX 750830688, USA
e-mail: lkhan@utdallas.edu

that ontology is a collection of concepts and their interrelationships which can collectively provide an abstract view of the application domain [20].

As the example illustrated (Fig. 1), the Multi-modal data in the form of images and associated text in web-pages, image sharing, image annotation applications, etc. provide a rich source of information. Information derived from processing multi-modal data from diverse modalities are often semantically related. For example, in (Fig. 1), text stream includes "tourist", "terrace" and "wall" concepts; on the other hand, the image stream shows concepts of "people", "Green-Area" and "wall". From these two streams common concepts such as "wall" and "people" may be inferred. Recall that in the ontology, "tourist" is a sub-class of "people". Thus, it is obvious that the integration of such information is beneficial. However, in recent literature it may be observed that concept extraction from visual stream is more susceptible to errors as compared to concept extraction from textual stream [3]. This is because; images are cataloged in terms of features rather than words as in the text which may directly be related to concepts. In other words, concept extraction in textual stream is more robust than that of the visual stream. Therefore, in this paper we show how to resolve ambiguity of concepts in visual stream with the help of identified concepts from textual stream. For example, in (Fig. 1), visual stream receives relevant concepts, "wall" and "people" along with irrelevant concepts, "Tiger'" through an image object detector.

The concepts appearing in textual stream will help to discard irrelevant concepts, "tiger" and keep relevant ones. This is because "wall" and "people" from the image stream are semantically related to the concepts "tourists" and "wall" appearing in the textual stream.

The contributions of this paper are as follows: First, domain ontology is constructed based on selective senses from WordNet. Second, image and text streams are fused with the

<DOC>
<DOCNO>annotations/10/10002.eng</DOCNO>
<TITLE>Group photo in the citadel of Machu Picchu</TITLE>
<DESCRIPTION>**Tourists are posing on a green terrace with grey walls and more green terraces behind them;**
</DESCRIPTION>
<NOTES></NOTES>
<LOCATION>Machu Picchu, Peru</LOCATION>
<DATE>13 January 2004</DATE>

Fig. 1 Multimodal data of image and text

help of ontology and semantic unified representation is generated. Third, the image concepts with the help of concepts appearing in textual stream based on semantic closeness, are disambiguated. Finally, we have implemented a the proposed disambiguation mechanism using the benchmark dataset, ImageCLEF and have shown the effectiveness of our proposed work.

The rest of the paper is organized as follows: in Section 2 a brief review of the related work is given, the literature on disambiguation, multi-modal data and ontology mining is huge. Thus, a brief review is given as an introduction to the disambiguation task. Section 3 presents our proposed work. Section 4 describes the implementation. Section 5 presents the experimental results. Finally, Section 6 presents the conclusion.

2 Related work

It is observed that, the ambiguity in image extracted knowledge is characterized in the recently developed approaches for semantic image classifications. [16,28]. This ambiguity problem can be described as a sub-problem of the semantic gap problem and has been discussed widely in the literature related to semantic image classification and semantic image processing [2,28]. The semantic gap occurs in the process of transformation of low-level features into high-level semantics. Semantic gap characterizes the differences in machine extracted low-level features and the high-level semantics as humans observe. In the efforts to fill the semantic gap, the problem of ambiguity arises.

In the literature, the disambiguation task of image-related labels is carried out as a step that followed the multi-class detection or combined with the image classification scheme. It is assumed that the detection process outputs a candidate that certainly involves the true identity of the image objects. The disambiguation process then refines the candidate list using a trained context through some machine learning method. Xiaodong Fun [9] has utilized a pair-wise classifier for pair wise classes competition. In this work, each pair of classes are trained based on the frequency of their joint appearance, and the trained value was used for the contextualization process. Similarly, Galleguillos, et al., [11] have used a class co-occurrence side by side with location and appearance to train a context and maximize object label agreement using Conditional Random Field ‘CRF’. Kumar and Hebert [24] have used a two level trained context based on pixel and region observation dependent label interaction. Singhal, et al., [25] have used a belief propagation to construct context that refine the feature detector results. To bridge the semantic gap and facilitate the disambiguation process, Park and Lee [22] have used ontologies as knowledge modeling for semantic concept extraction. The developed approach includes two ontologies to transform the low-level features to high-level semantics. In biomedical application, Thies, et al., [26] have used a classifier model to transfer low-level features into a class. Zlatoff, et al., [29] have developed a knowledge source for image understanding framework based on scene modeling. Wu, et al. [27] have used a trained pair-wise concepts for misclassified concepts learning in video concepts.

Despite the differences in the disambiguation mechanism for the image labels, the ambiguity is not completely resolved. This observation suggests that image medium is not adequate for the extraction of semantic information. We can view the problem as a problem in multimedia understanding. Since image data, in some applications, may be composed of synchronized multi-modal information streams, comprising of visual and textual information, we propose a unified processing based on semantic representation by means of inter-modal collaboration. [1,13,17,18,21].

The proposed disambiguation process explores the visual and textual information jointly without a pair-wise restriction. Moreover, no learning is required to discover the association between the components of the modalities (image & text). Our work is similar to the work presented by Benitez and Chang, [5]. However, our work is different from Benitez and Chang, [5] in the sense that, the disambiguation is not based on direct string or concept matching, which might not exist, between the components of the streams. Instead, the proposed approach exploits the utilization of semantic closeness in a domain-specific context.

Semantic closeness is a function of semantic distance that has been used for schema matching and measurements of attributes' value similarity in the database. [12]. In such application, the distance is measured between components with similar semantic but different syntactic. A semantic distance measure in the knowledge mining has equivalent role as the string matching in the syntactic applications. This is supported by the recent evolving of the application in data mining to knowledge mining. [12]. In a different way, in this paper, the semantic distance is used to capture the relation between diverse modalities that refer to the same context.

3 Proposed work

This work proposes a mechanism that uses text data to disambiguate associated image. The disambiguation is achieved by measuring the semantic closeness/distance between the concepts extracted from the image and associated concepts in the text. The inputs to this process are multi-modal data: text, image and a reference to the WordNet. The multi-modal data is represented by a set of keywords relating to image objects and textual words. WordNet, a rich lexical database is used to construct the domain ontology.

Initially, the domain ontology is created using Web Ontology Language 'OWL' [23]. In this, each concept has two attributes: a label and its relationships with other concepts. The label is a word string where as the relationships create a position for the concept in the ontology hierarchy.

The domain ontology is created as a contextualization process based on WordNet which is a rich source of upper level knowledge containing multiple senses for each word, and each sense may be expressed in multiple words (synset). However, reflecting the domain knowledge using upper level knowledge does not reflect the relationships between domain-specific words in a correlated manner. For example, the words "fish" and "apple" have strong relationship in "food" domain. However, they have weak relationship in "Living-things" domain where in "fish" is an animal and "apple" is considered as a plant. It follows that specifying the domain, is necessary to reflect the true relationships between the concepts in that domain.

This ontology can be treated as "Topic model", representing the taxonomy of the domain concepts [6]. The ontology building process is a semi-automatic process that is implemented offline as shown in (Fig. 2). A set of keywords are selected from the domain images and used as input to build the domain-specific ontology through WordNet. As follows, each keyword is represented by abstract type or what is called a class 'owl:Class'. Then, the structure is initialized as a flat representation of the listed classes, with a single ancestor of 'owl:Thing' which is the top most class in any ontology. Then, the ontology is evolved by enriching each class with its ancestor classes with reference to the hyponyms relations in WordNet.

As such, each class in the flat representation is used as a query to WordNet which, in turn, responds with a chain of hyponyms of related words. However, WordNet has a diversity of senses for each word. This diversity is useful as each word might be used differently in diverse domains. The diversity of senses leads to an enormous set of

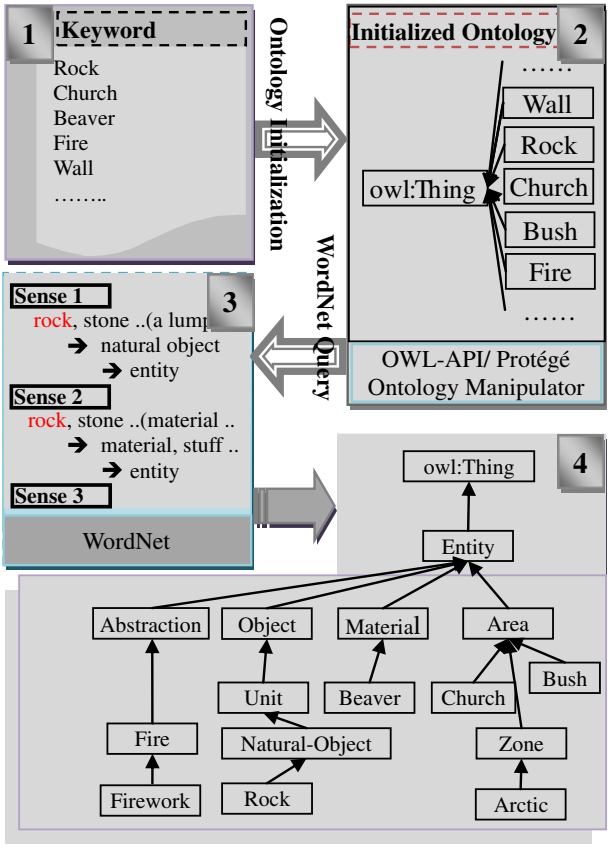


Fig. 2 Domain-ontology engineering

hyponyms chains for a single word which is not desirable for a domain-specific ontology. Thus, the most relevant sense out of the total retrieved senses and its correspondence chain is chosen to construct the domain ontology. The relevant sense is selected semi-automatically. The selected sense is the one that encloses in its hyponyms chain more domain-specific concepts. Domain specific concepts are those appear frequently in the chains of all the concepts in the domain.

The words in a selected chain are also represented as 'owl:Class' and added to the domain ontology, if it does not exist. The relationship between associated concepts in the selected chain are constructed as 'rdfs:subClassOf' relations, which is the backbone of the ontology's hierarchy structure. The keywords from domain text are linked to the domain ontology based on the most relevant sense extracted from the WordNet. The relevance of the senses for the domain text is measured comparing to the existing ontology created by the image domain keywords. At the end of this process, a comprehensive ontology for the domain is created.

The disambiguation process, illustrated in (Fig. 3), for a given multi-modal instance, starts by mapping all the extracted keywords from both image and text to the domain ontology. Here, straight mapping for the image concepts is achievable since all the possible keywords extracted from the image are used initially to construct the ontology. However, only some keywords from the text may be mapped to the ontology while the others will be discarded. This follows from the hypothesis that, text may involve many additional concepts that may not be relevant to the image under consideration. Thus, aligning the text

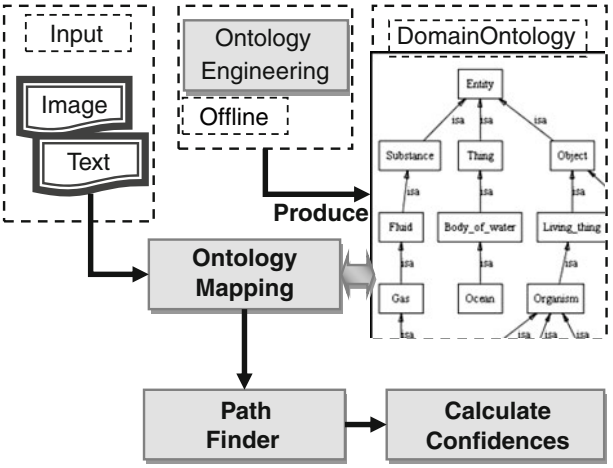


Fig. 3 Disambiguation framework

with the domain ontology shall produce, from the text, only those concepts that are relevant to the image while filtering out all others. Such mapping also facilitates word sense disambiguation using image [4].

After mapping both image and text keywords to the ontology concepts, a unified ontology structure 'G' is created as follows: All concepts are linked to their sub-classes concepts down to the leaves and to their super-classes concepts up to the root. The path finder finds the path between each concept and its sub-classes and super-classes concepts. These paths are then aggregated to produce 'G', a *hierarchy-based Minimum Spanning Tree*. (Fig. 4), illustrates an example of 'G'.

Finally, the confidence value of each image concept is calculated. The confidence value is a function of prior and contextualized confidence values. The prior confidence value is a

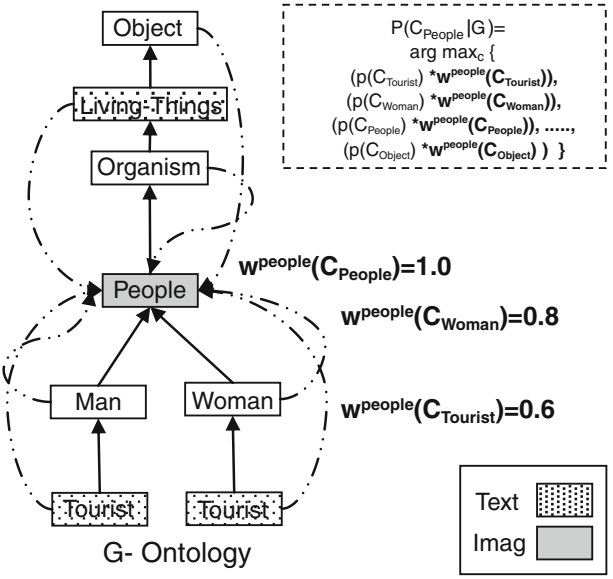


Fig. 4 Path of the concept (People)

weighted sum of its prior confidence from each of the modality. It is possible that a given concept may exist only in one of the modalities or both. The prior confidence is given as:

$$p(c_i) = \alpha p(c_i^t) + \beta p(c_i^g) \quad (1)$$

Where, $p(c_i^t)$ is the prior confidence value of the concept c_i from the text modality. While, $p(c_i^g)$ is the prior confidence value of the concept c_i from the image. A concept that appears in both the modalities is likely to be a non-ambiguous true concept, and thus, it will have high prior confidence. While, a concept that appears in only one of the modalities will have a low confidence value. The constants α and β are introduced to incorporate weightage to the modalities that may be available from prior knowledge of the application. These may be pre-assigned or determined empirically. The context-dependent confidence is a function of ontology-based distance which reflects the semantic closeness and it is calculated using 'G'. The context of each concept is computed through allocating a path to the root and some leaves as stated earlier, followed by computing the path length.

Algorithm 1, details the pseudo code for the path finder that is used to identify all the intermediate concepts between a concept and its root as well as its leaves. In lines 2 and 3 *path(c_i),* a set that will hold the concepts, and *distance(c_i)*, a set of corresponding distances are initialized. In line 5, the concept c_i itself is added as part of the concepts in *path(c_i)*. Lines 8–16 iteratively locate and identify the ancestors of the concept c_i. On the other hand, lines 18–26, iteratively locate the descendants (sub-classes) of the concept c_i. Given the path for each concept c_i, the final confidence value is calculated using Eq. 2.

$$p(c_i|G) = \arg \max_{c_j \in path(c_i)} \{p(c_j) * w^i(c_j)\} \quad (2)$$

Where, $p(c_j)$ is the prior confidence of the concept c_j along the path, $path(c_i)$. Note that $path(c_i)$ includes $Concept(c_i)$ itself. Each concept c_j is given a weight $w^i(c_j)$, inversely proportioned to its distance to c_i . The weight $w^i(c_j)$ is calculated using Eq. 3. The path, *path(people)* and the related final confidence calculation is illustrated in (Fig. 4).

$$w^i(c_j) = \left\{1 - \frac{\text{distance}(c_i, c_j)}{\text{distance}_{\max}}\right\} \quad (3)$$

The *distance(c_i , c_j)* in Eq. 3, is the number of intermediate concepts between the concepts c_i and c_j. This distance is normalized by the length of the maximum path between any two concepts $distance_{max}$.

Here, it may be noted that the weight $w^i(c_j)$ and the overall value of $p(c_i|G)$ are zero if the distance between a pair of concepts is equal to the maximum path length. The significant of such relation is low. However, the weight $w^i(c_j)$ for the concepts where in $c_i = c_j$, is equal to 1.

The final confidence value of each concept using this process, calculated using Eq. 2, is the maximum value obtained by mining all the concepts $c_j \in path(c_i)$. (Fig. 5), illustrates an example of such case. Based on the proposed method, from Eq. 2, it is obvious that $p(c_i|G) \geq p(c_i)$. i.e in $p(c_i|G) = \arg \max_{c_j \in path(c_i)} \{p(c_j) * w^i(c_j)\}$, when $c_j = c_i$, then $p(c_i) * w^i(c_i) = p(c_i)$, because, $w^i(c_i) = 1$.

In the unified ontology structure 'G', image concepts that have no link to the text concepts, have no increase in the prior confidence value. On the contrary, their confidence value, as computed in Eq. 1, is lower than that of their modality specific

confidence value. On the other hand, if an image concept is linked to associated concepts from the image or the text, its confidence will increase. The best augmentation for any image concept is realized if the same concept is stated clearly in the text. In such a case $p(c_i|G) \geq p(c_i)$, $p(c_i)$ will be high because $p(c_i)$ is an aggregated prior from both modalities.

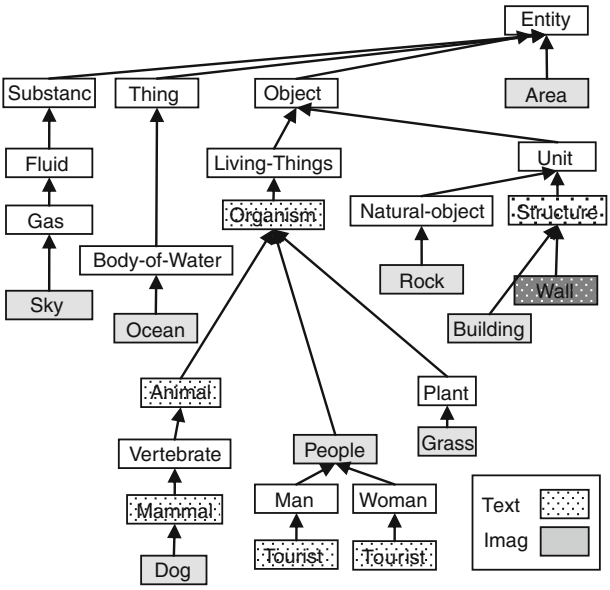
Algorithm 1: Path Finder

```
1.    INPUT: Concept c_i, Ontology O.
2.    Initialized path(c_i):{ø} .
3.    Initialized distance(c_i):{ø}.
4.    Initialized Concept c_j=null, p(c_j)=0,level=1.
5.    Begin:
6.      path(c_i) U {c_i}
7.      distance(c_i) = 0
8.      While Ancestor(c_i)≠null
9.          set c_j as Ancestor(c_i)
10.         p(c_j) = αP(c_j^t) + ß P(c_j^g).
11.         If(p(c_j) ≠0)
12.             path(c_i) U { c_j }
13.             distance(c_i) = level
14.         End If
15.         level = level+1
16.     End While
17.     Set level = 1
18.     While Descendant(c_i)≠null
19.         set c_j as Descendant (c_i)
20.         p(c_j) = αP(c_j^t) + ß P(c_j^g).
21.         If(p(c_j) ≠0)
22.             path(c_i) U { c_j }
23.             distance(c_i)= level
24.         End If
25.         level =level+1
26.     End While
27.   End
```

Many disambiguation scenarios can be described based on the proposed method. Here we state some of those examples with reference to (Fig. 5).

- **DM**: **D**irect **M**atching. The same concept is given by both modalities. For example, the concept 'Wall' is mentioned explicitly in both modalities. As stated earlier, the disambiguation process assigns a high confidence value to this case.
- **DP**: **D**irect **P**arent. A direct parent/ancestor of two or more siblings concepts. As an example, refer to the text concept 'Structure' and the image concepts 'Wall' and 'Building'. In this case, the concepts 'Wall' and 'Building' will be assigned same confidence value with reference to the concept 'Structure' as both have same distance to reference concept.
- **IDP: I**n-**D**irect **P**arent. A common ancestor of two or more concepts occurs at different levels of the hierarchy. Refer to the text concept 'Organism' and the image concepts 'Grass' and 'Dog' in (Fig. 5). Here, the concept 'Grass' will be assigned a higher confidence value as compared to 'Dog', owing to its semantic closeness to the text concept 'Organism'. On the other hand, with reference to the text concept 'Mammal,

Fig. 5 'G' Unified ontology



the image concept 'Dog' will receive a higher confidence value as compared to the concept 'Grass'.

- **IS**: **IS**olated image concept. As an example, in (Fig. 5), the concept 'Area' is an isolated image concept with no associated concept in the text. Thus, the disambiguation process assigns a degraded confidence value for this concept.

It may be argued that an alternative to the proposed mechanism may be a more straight forward mechanism that calculates the distance of each image concept c_i in the image to all other concepts in the text. However, the concepts in the image c_i cannot be disambiguated using another concept in the text that has a different path in abstraction domain. Example, in (Fig. 5), the image concept 'Grass' cannot be disambiguated using the text concept 'people' although the distance between them is short. Thus, the proposed path-based disambiguation is justified. As an experimental proof, a semantic similarity based on WordNet, WordNet Similarity (WNS), is implemented for the comparison purpose. In this method, the similarity measure proposed by Leacock and Chodorow [15] is used. Another method implemented for the comparison purpose that is the contextual disambiguation using Conditional Random Field (CCRF) similar to the work proposed by Xiaodong Fun [9] and Galleguillos, et al. [11]. It may be noted that the contextual disambiguation is purely on image domain alone.

4 Implementation & experimental results

The proposed disambiguation mechanism is implemented in NetBeans Java JDK 1.6. and using WordNet 2.1. [10] and Protégé-owl tools and API [14]. Protégé is an ontology engineering, manipulating and reasoning tool that is widely used.

The experiments are conducted over the IAPR TC-12 benchmark dataset provided by ImageCLEF [19]. This dataset contains more than 20000 images that are taken from MIR

Flicker and is complete with textual annotation in several languages. The group '00' of 254 images and the English annotation is chosen for the first set of experiments.

Please note that, image classification process is outside the scope of this research. Thus, the ambiguity that may be produced by a classification algorithm is mimicked. The true objects labels that are provided in the database for each image are ambiguiated by injecting randomly selected labels. Thus, the disambiguation process is implemented over an ambiguous set of labels and the results are compared to the true set. The text associated with each image is tokenized, the sentences are splattered and nouns and noun phrases are extracted using GATE, General Architecture for Text Engineering [8]. The nouns are matched directly with the ontology elements using a look-up mechanism.

The text concepts are given a prior confidence value $p(c_i^t)$ equal to 1.0. This is accordance with the earlier assumption that the text concepts are considered to be accurate. For the image input, the confidence value $p(c_i^g)$ of each label in the generated ambiguous list is set to be 0.5. As such, each label is given the same initialized confidence in order to be able to show the impact of the disambiguation process. Also, given the value of 0.5, each concept has an equal probability of being correct or incorrect. The ambiguities are mimicked in such a way that the true label set in each image is corrupted with equal number of labels. α and β values in Eq. 1 are chosen such that the text concepts are given more or equal strength. Thus, the pair values of (α=0.5, β=0.5), (α=0.6, β=0.4), (α=0.7, β=0.3) and (α=0.8, β=0.2) are tested. After running the disambiguation method as proposed, a list of concepts and their associated confidence values are generated as an output. The concepts with confidence value above some threshold are selected as final output. Three threshold values of 0.3, 0.4 and 0.5 are compared to show their ability in selecting true non-ambiguous concepts. This is followed by computing the accuracy values for the result. The accuracy is calculated based on Eq. 4.

$$Acc(I_x) = \frac{\sum_{c_i \in x} tp^x(c_i)}{\sum_{c_i \in x} c_i} \tag{4}$$

Where, $Acc(I_x)$ is the accuracy value for the image I_x. I_x is an image with a set of concepts, $tp(c_i)$ is the true positive value for the concept c_i, $tp^x(c_i)$ is the true positive value for the concept in the image I_x specifically. The results are presented in (Fig. 6) together with the results of the direct matching (DM) disambiguation process proposed by Benitez and Chang [5], contextual disambiguation using Conditional Random Field (CCRF) [9,11] and WordNet Similarity (WNS) [7] using the similarity measure of Leacock and Chodorow [15]. From this figure, it may be seen that the proposed disambiguation mechanism

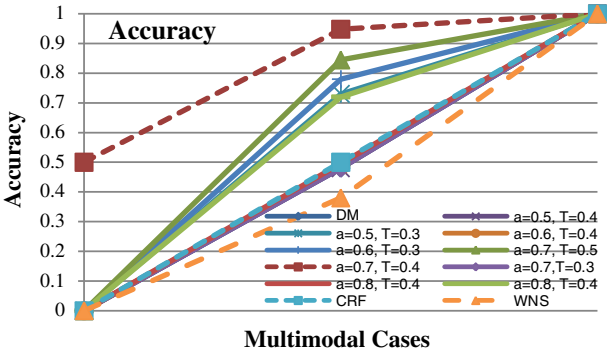


Fig. 6 Accuracy measure

performs far better than the disambiguation using direct matching CCRF and WNS. The pair values of (α=0.7, β=0.3) with threshold values of 0.4 and 0.5 give the best results. Thus, the rest of the experiments carried out using these values.

In the second set of experiments, the percentage ambiguity introduced is varied. In the first set of experiments 50% ambiguity is introduced by injecting equal number of corrupted labels to the correct labels. In this second set of experiments this percentage is varied from 33% to 66%. For each case, Precision, Recall and F-measured using Eqs. 5, 6 and 7 respectively are calculated.

$$\Pr(c_i) = \frac{tp(c_i)}{tp(c_i) + fp(c_i)} \tag{5}$$

$$Rcl(c_i) = \frac{tp(c_i)}{tp(c_i) + fn(c_i)} \tag{6}$$

$$F - M(c_i) = 2 * \frac{\Pr(c_i) * Rcl(c_i)}{\Pr(c_i) + Rcl(c_i)} \tag{7}$$

Where, $Pr(c_i)$, $Rcl(c_i)$ and $F\text{-}M(c_i)$ are the precision, recall and F-measure for the concept c_i. $tp(c_i)$ is the true positive value for the concept c_i. $fp(c_i)$ is false positive value and $fn(c_i)$ is the false negative value for the concept c_i

Unlike the previous experiments, these values are calculated for each label rather than for each image. The results of this experiment are, once again, compared with disambiguation using Direct Matching. For all the three parameters: Precision, Recall and F-Measure, the proposed disambiguation process has produced 20–25% improvement over Direct matching. The results are presented in Table 1. From the result, it is noted that (CCRF) and (WNS) have poor precision as compared to the other two methods. Even though Direct Matching (DM) performs consistently under varying ambiguity level, the proposed disambiguation mechanism performs better than all other method.

The final set of experiments are conducted over the complete dataset in a batch mode. A total number of 39 batches are tested. Here it may be noted that the batches are interrelated and have shared context since they are drawn from the same dataset. The average number of object classes in each batch is 127 and is 46% of the total number of object classes in the whole dataset.

Table 1 The average precision, recall and F-measure of the proposed mechanism compared to direct matching

	Method ambiguity	Direct match	Proposed disambiguation (T=0.4)	Proposed disambiguation (T=0.5)	Contextual (CRFF)	WordNet similarity (WNS)
Precision	33%	0.466	0.71	0.689	0.351	0.25
	50%	0.466	0.688	0.689	0.26	0.18
	66%	0.466	0.548	0.54	0.19	0.13
Recall	33%	0.256	0.49	0.465	0.48	0.36
	50%	0.256	0.49	0.465	0.46	0.34
	66%	0.256	0.4	0.38	0.46	0.343
F-Measure	33%	0.302	0.551	0.518	0.38	0.29
	50%	0.302	0.523	0.518	0.30	0.23
	66%	0.302	0.43	0.41	0.24	0.174

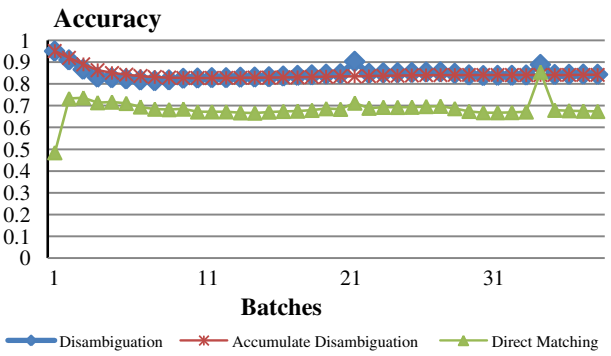


Fig. 7 Accuracy measurements of the whole dataset

Furthermore, the number of common object classes between any two the batches is high, with an average of 91. Finally, in a statistical approximation, the set of instances covering all the object classes in the dataset (276 distinct classes) can be obtained by combining an average of 8 batches.

The experimental parameters are: threshold value of 0.4; pair values of (α=0.7, β=0.3) and 50% ambiguity. An ontology is created for each batch, then the disambiguation process is carried out on the batch instances using the created ontology. Two different experiments categories are carried out: (i) batch specific and (ii) accumulated batches. In the batch specific experiments, each batch is tested with its associated ontology. For each batch, the averages of accuracy, precision, recall and F-measure are computed and presented in (Figs. 7, 8, 9 and 10) as *disambiguation series*. In the accumulated batches experiments, the ontologies with their associated batches are merged sequentially. First, a single batch with its associated ontology is tested, and then the second batch is added to the first batch and their associated ontologies are merged. The disambiguation process is carried out over the combined ontology for the combined batches. This process is continued until all the batches are added one by one. After adding each batch, the averages of accuracy, precision, recall and F-measure are computed. These results are presented in (Figs. 7, 8, 9 and 10) as *accumulate disambiguation series*.

The aim of the accumulated disambiguation is to test the proposed disambiguation process over a large number of instances and to examine the scalability of the ontology. The results illustrated in (Figs. 7, 8, 9 and 10) show the stability of the proposed disambiguation process with large number of instances. The trivial variation between the batch specific results and the accumulated disambiguation is due to the ontology enlargement.

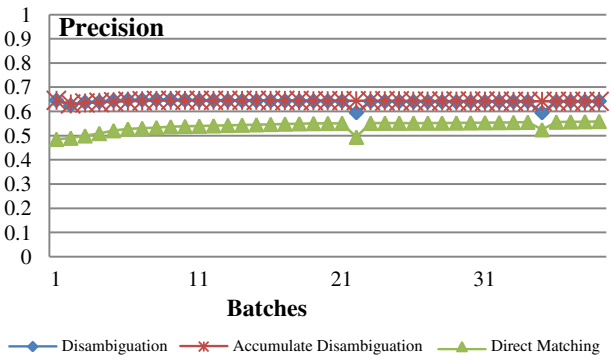


Fig. 8 Precision measurements of the whole dataset

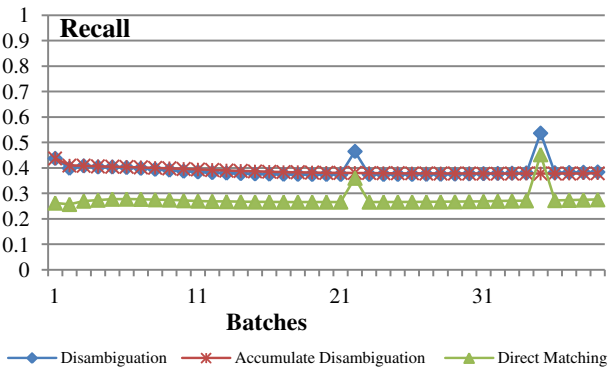


Fig. 9 Recall measurements of the whole dataset

Overall, the variation between the results of the two experiments is not significant and the process is applicable even with different portions of the ontology (ontology of each batch). Here, it is worth mentioning that the batches interrelation which is described earlier has aided in keeping the ontology consistent during the upscaling process. As may be seen from (Figs. 7, 8, 9 and 10) the results of these two experiments are better than those of the direct matching process. Additionally, the consistent results of the accumulate disambiguation series may be attributed to the accumulation in instances. With a total of 20,000 instances it is only natural that these results remain consistent unless the results of the continuously added bathes are exceptionally different.

5 Discussion

Figures 11 and 12 show comparative results of selected concepts using direct matching, proposed disambiguation with T=0.4 and T=0.5. It is to be noted that the text annotation may not include all the image labels or concepts of the image. Thus, the ability of direct matching may vary. While on the other hand, the proposed disambiguation process shows some stability. In (Fig. 11), the concept ‘Man’ and the concept ‘Woman’ are assigned high precision using direct matching as well as using the proposed disambiguation process. On the other hand, the concept ‘Child-girl’ is assigned a high precision under the proposed disambiguation process, while direct matching assigns zero precision to that concept. This

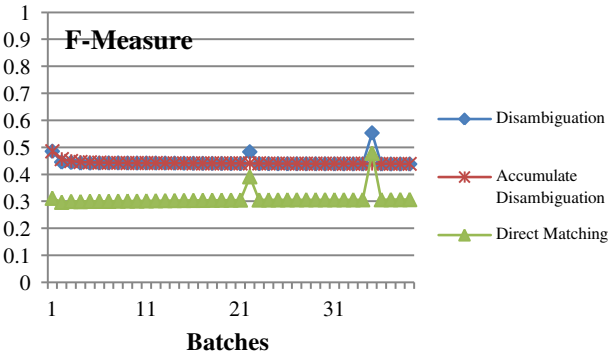


Fig. 10 F-Measure of the whole dataset

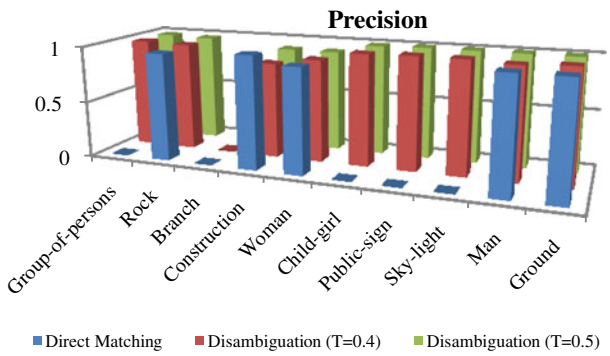


Fig. 11 Precision of Selected Concepts

is because the concept 'Child-girl' may be referred in the text using some other concepts that fall in the path, *path(Child-girl)* such as: kid, child and human. More concepts with comparative recall are illustrated in (Fig. 12). The concept 'Statue' is assigned high recall by both processes. On the other hand, the concept 'Person' is assigned a high recall by the proposed disambiguation process, while direct matching assigns zero recall for that concept.

In situations where the ambiguous concepts are located in the same portion of the domain ontology, with equal distance to the text concept, the disambiguation process is helpless. As an example, consider the concepts 'Man' and 'Woman' from the image and 'People' from text, in (Fig. 5). Thus, the effectiveness of the proposed disambiguation process varies from concept to concept depends upon the annotator's choice of keywords or their perceptual preference in noticing certain concepts over the others.

In addition, the performance of the disambiguation mechanism depends on the nature of the multi-modal data and the way modalities relate to each other. For example, the text may be describing, annotating, or clarifying the associated image. In these types, there are different probabilities to find related concepts in the text to be used in the disambiguation process. In all these cases, the proposed disambiguation may be used. On the other hand, the direct matching process has limited capabilities as the text and image may have different representations, granularity and convergence. It is possible that the degree of association between the text and the image may be varied. In such situation having a fixed threshold in the disambiguation process may not be a wise choice. Dynamic threshold that maybe determined as a function of the Euclidean distance between the concept vectors of the text and the image may be a better choice.

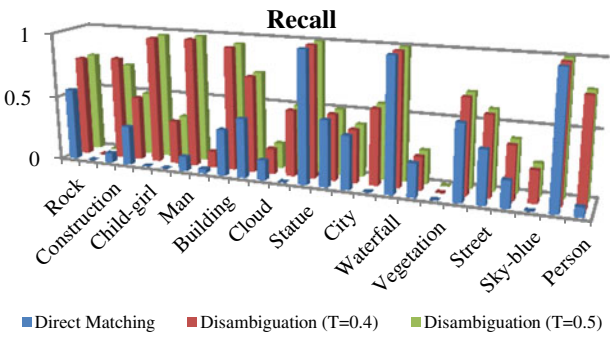


Fig. 12 Recall of selected concepts

6 Conclusion

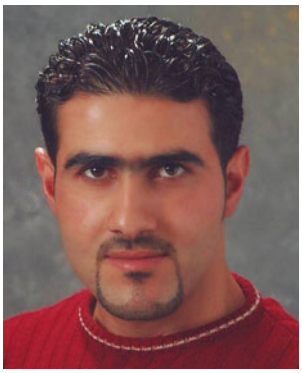
This paper proposed a method to disambiguate the semantic concepts extracted from image with identified semantic concepts from the associated text. First, domain ontology is constructed as the medium wherein the semantic closeness is measured. A compact created ontology has lead to enhance the precision of the disambiguation results. Second, the image and text are fused over the aforementioned ontology and semantic unified presentation is generated. Third, the disambiguation of the image concepts over the generated unified presentation is achieved based on the calculated semantic closeness with the textual concepts. As such, discovering the lexically unrelated but semantically related concepts lead to enhance the recall of the disambiguation results. Overall, concept-level disambiguation of image using associated text in multi-modal data is an efficient solution especially when the fusion at low or middle level is not possible. The improved accuracy, as shown in the results, proves the ability of the proposed disambiguation process. The disambiguation mainly depends on the incorporation of the semantically close concepts in the constructed domain ontology have been incorporated to solve the ambiguity that encapsulates the images.

Acknowledgments This work was supported by a Research University grant titled 'Multimodal Meaning Normalization through Ontologies' (No:1001/PKOMP/811021).

References

1. Angelo C, Vincenzo M, Antonio P, Antonio P (2008) Scene detection using visual and audio attention. Paper presented at the Proceedings of the 2008 Ambi-Sys workshop on Ambient media delivery and interactive television, Quebec City, Canada
2. Athanasiadis T, Mylonas P, Yannis A, Stefanos K (2008) Semantic image segmentation and object labeling. IEEE Trans Circuits Syst Video Technol 17(3):298–312
3. Barnard K, Forsyth D (2001) Learning the semantics of words and pictures. Paper presented at the International Conference on Computer Vision
4. Barnard K, Johnson M (2005) Word sense disambiguation with pictures. Artif Intell 167(1–2):13–30. doi:10.1016/j.artint.2005.04.009
5. Benitez AB, Chang S-F (2002) Semantic knowledge construction from annotated image collections. ICME Lausanne, Switzerland
6. Boyd-Graber J, Blei DM, Zhu X (2007) A topic model for word sense disambiguation. Paper presented at the Empirical Methods in Natural Language Processing, Prague, Czech Republic
7. Chin Y, Khan L, Wang L, Awad M (2005) "Image annotations by combining multiple evidence & WordNet" In Proc. of 13th Annual ACM International Conference on Multimedia (MM 2005), Singapore,November 2005, pp 706–715
8. Cunningham H, Maynard D, Bontcheva K, Tablan V (2002) GATE: a framework and graphical development environment for robust NLP tools and applications. Paper presented at the the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002
9. Fan X (2004) Contextual disambiguation for multi-class object detection. Paper presented at the International Conference on Image Processing
10. FELLBAUM Ce (1998) WordNet: an electronic lexical database. MIT Press
11. Galleguillos C, Rabinovich A, Belongie S (2008) Object categorization using co-occurrence, location and appearance. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. pp 1–8
12. Garcia ACB, Ferraz I, Santarosa Vivacqua A (2009) From data to knowledge mining. Artif Intell Eng Des Anal Manuf 23(4):427–441. doi:10.1017/S089006040900016X
13. Jie Y, Jiebo L (2008) Leveraging probabilistic season and location context models for scene understanding. Paper presented at the Proceedings of the 2008 international conference on Content-based image and video retrieval, Niagara Falls, Canada
14. Knublauch H, Fergerson R, Noy N, Musen M (2004) The Protege OWL Plugin: An Open Development Environment for Semantic Web Applications. In: The Semantic Web ISWC 2004, pp 229-243

15. Leacock C, Chodorow M (1998) Combining local context and WordNet similarity for word sense identification. In Fellbaum 1998, pp 265–283
16. Manjunath KN, Renuka A, Niranjan UC (2007) Linear models of cumulative distribution function for content-based medical image retrieval. J Med Syst 31(6):433–443. doi:10.1007/s10916-007-9075-y
17. Manolis D, Guillaume G, Patrick G (2008) Audiovisual integration with segment models for tennis video parsing. Comput Vis Image Underst 111(2):142–154. doi:10.1016/j.cviu.2007.09.002
18. Margarita K, Emmanouil B, Constantine K, Ioannis P (2007) A neural network approach to audio-assisted movie dialogue detection. Neurocomput 71(1–3):157–166. doi:10.1016/j.neucom.2007.08.006
19. Michael G, D. CP, Henning M, Thomas D (2006) The IAPR benchmark: a new evaluation resource for visual information systems. Paper presented at the International Conference on Language Resources and Evaluation, Genoa, Italy, 24/05/2006
20. Miller G (1995) WordNet: a lexical database for english. Commun ACM 38(11)
21. Ming-Fang W, Yung-Yu C (2008) Multi-cue fusion for semantic video indexing. Paper presented at the Proceeding of the 16th ACM international conference on Multimedia, Vancouver, British Columbia, Canada
22. Park K-W, Lee D-H (2006) Full-automatic high-level concept extraction from images using ontologies and semantic inference rules. In: ASWC, pp 307–321
23. Recommendation WC (10 February 2004) OWL: Web Ontology Language Overview http://www.w3.org/TR/owl-features/
24. Sanjiv K, Martial H (2005) A hierarchical field framework for unified context-based classification. Paper presented at the Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2
25. Singhal A, Luo J, Zhu W (2003) Probabilistic spatial context models for scene content understanding. Paper presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA
26. Thies C, Herzog H, Schmitz-Rode T, Deserno TM (2007) Bridging the semantic gap for object extraction from biomedical images by classification. Biomed Tech 52
27. Wu Y, Tseng BL, Smith JR (2004) Ontology-based multi-classification learning for video concept detection. In: IEEE International Conference on Multimedia and Expo, ICME '04, pp 1003–1006
28. Ying L, Dengsheng Z, Guojun L, Wei-Ying M (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recogn 40(1):262–282. doi:10.1016/j.patcog.2006.04.045
29. Zlatoff N, Tellez B, Baskurt A (2004) Image understanding and scene models: a generic framework integrating domain knowledge and Gestalt theory. In: International Conference on Image Processing, ICIP '04, Vol. 2354, pp 2355–2358



Ahmad Adel Abu Shareha received a B.Sc. degree in Computer Science from Al al-Bayt University in 2004. He obtained his M.Sc in Computer Science from Universiti Sains Malaysia (USM) in 2006. Since 2007, he has been research assistance in Computer Vision Research Group (CVRG) at the School of Computer Sciences, USM. Currently he is a PhD student at the School of Computer Sciences, USM. His research interests in multimedia semantic, data mining and machine learning.

Mandava Rajeswari is a lecturer at the School of Computer Sciences, Universiti Sains Malaysia. She received the B.Sc. degree from the University of Madras, India. She obtained her M.Sc. from M.Tech, IIT Kanpur, India. PhD from University of Wales. Her research interests analyze and to extract contents and information from the images; derive knowledge from the extracted information; to represent the knowledge and use the knowledge in various applications in addition to using it to guide the information extraction from the images. In the early stages of this research the focus was to extract information from the images and put into several applications that include automated visual inspection, and real time process control in industry; robot vision for intelligent assembly; image database retrieval and image segmentation. While these areas remain as research interests, focus is now shifted to image Knowledge extraction and representation and Knowledge guided image segmentation analysis and visualization. The major domain of research is in medical images and natural images.

Dhanesh Ramachandram is a lecturer at the School of Computer Sciences, Universiti Sains Malaysia. He received the B.Sc. and his PhD degress from Universiti Sains Malaysia, Malaysia. Dr. Dhanesh's research focuses on the delineation and segmentation of various structures of interest from "slices" of CT and MRI images of patients. Various segmentation techniques have been investigated to allow accurate delineation and quantification of volume. 3D visualization of the 2D modalities provides new insight into treatment planning of patients. Another aspect of current active research is the semantic understanding of image content. Attaching generic labels to image content by relating low-level image features to high level semantic concepts is of interest. In this respect, various approaches to pattern recognition and machine learning, ex: clustering are being actively pursued.

Dr. Latifur R. Khan is currently an Assistant Professor in the Computer Science Department at the UTD, where he has taught and conducted research since September 2000. His research work is currently supported by grants from the Air Force Office of Scientific Research (AFOSR), National Science Foundation (NSF), the Nokia Research Center, Alcatel, Raytheon, and the SUN Academic Equipment Grant program. Dr. Khan is one of the principal investigators at the CyberSecurity and Emergency Preparedness Institute at UTD, where he is involved with finding solutions to deal with the rapidly growing Homeland Security problems in cybercrime, information assurance, and emergency preparedness. In addition, Dr. Khan is the director of the state-of-the-art DBL@UTD, UTD Data Mining/ Database Laboratory, which is the primary center of research related to data mining and image/video annotation at University of Texas–Dallas. Dr. Khan's research areas cover data mining, multimedia information management, semantic web and database systems with the primary focus on first three research disciplines. He has served as a committee member in numerous prestigious conferences, symposiums and workshops including the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Dr. Khan has published over 80 papers in prestigious journals and conferences.