A graph theory-based online keywords model for image semantic extraction

WANG, Jing <http://orcid.org/0000-0002-5418-0217> and XU, Zhijie

A Graph Theory-based Online Keywords Model for Image Semantic Extraction

Jing Wang; Zhijie Xu
Visualisation, Interaction and Vision Research Group
School of Computing and Engineering, University of Huddersfield
Queensgate, Huddersfield, West Yorkshire, UK, HD1 3DH
+44 (0)148447 2156
jing.wang@hud.ac.uk; z.xu@hud.ac.uk

ABSTRACT
Image captions and keywords are the semantic descriptions of the dominant visual content features in a targeted visual scene. Traditional image keywords extraction processes involves intensive data- and knowledge-level operations by using computer vision and machine learning techniques. However, recent studies have shown that the gap between pixel-level processing and the semantic definition of an image is difficult to bridge by counting only the visual features. In this paper, augmented image semantic information has been introduced through harnessing functions of online image search engines. A graphical model named as the "Head-words Relationship Network" (HWRN) has been devised for tackling the aforementioned problems. The proposed algorithm starts from retrieving online images of similarly visual features from the input image, the text content of their hosting webpages are then extracted, classified and analysed for semantic clues. The relationships of those "head-words" from relevant webpages can then be modelled and quantified using linguistic tools. Experiments on the prototype system have proven the effectiveness of this novel approach. Performance evaluation over benchmarking state-of-the-art approaches has also shown satisfactory results and promising future applications.

CCS Concepts
•Information systems ➞ Data management systems ➞ Information integration

Keywords
graphical model; linguistic hyponym trees; image semantic; online search engine

1. INTRODUCTION
Image scene understanding has been recognised as one of the most difficult computer vision tasks. Although many machine learning algorithms have been developed for bridging gaps between image pixels and their semantic meanings, recent studies have shown that visual features alone are usually not sufficient for knowledge acquisitions from images [6]. To tackle the problem, many researchers have extended their effort through the route of information integration such as using WWW.

For example, Zhang [14] has established a model for solving annotation problem for scalable images. The model uses webpage image samples when defining a group of image patches based on calculating their co-occurrence relationships. The model can be used for large-scale image database annotation and content-based image retrieval (CBIR) applications.

In many other applications, online search engines have played important roles for information classification. For example, Bollegala [2] has introduced a semantic relationship measuring algorithm based on page counts and text snippets retrieved from web searching results. Liu [10] has improved CBIR accuracy by enhancing image contextual information through exploring social networks.

It has been widely accepted that online sources have become the most comprehensive information repository for extracting, abstracting, defining and representing complex forms of human knowledge. Web information is generally composed of multimedia data formats such as text, images, audios, and streamed videos, whose semantic information can be highly related. These potential correlations inside and between webpages have stimulated this research to explore semantic information stored in images through an automated and synchronised mechanism. Comparing with the "pure" computer vision-based "internal" approaches, this research aims at investigating "external" and information-synthesis-based methods for image scene understanding, with anticipated applications in auto-image annotation, image tagging and semantic computing.

Given an image in question, the proposed method starts from retrieving a group of online images samples sharing similarly visual features with the targeted image. Based on human intuition, it can be assumed that images containing visually-similar objects are likely to share related semantic meanings. Those retrieved images are part of some webpages, where extra semantic information can be acquired from their titles, tags, and other content and context descriptions. One of the main goals of this research is to investigate the relationships in-between the extracted keywords and the targeted image.

Many pilot works, such as [9, 12, 13], have used large-scale and online CBIR techniques to annotate uncaptioned images, *i.e.* clustering algorithms. The works often shown promising results if most candidates share identical keywords. But the clustering

SAC 2016, April 04-08, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2851613.2851633

approaches treated the image keywords as independent members in computation, the semantic similarity information was lost during the process. These algorithms were quite sensitive to the image retrieval outputs. The accuracy can drop significantly if the retrieved online sources contain too many unrelated words.

While it is almost impossible to retrieve a webpage of a hundred percent matching accuracy with a query image in terms of their semantic meanings. The keywords shared by a set of "similar" webpages with the targeted image often possess valuable clue about the true meaning of the image, although a large amount of unrelated semantic interpretation also existed – the so-called semantic noise.

In this paper, a novel algorithm named as the "Head-words Relationship Network" (HWRN) has been introduced for establishing the semantic relationship between the keyword candidates, and to remove semantic noises. In this model, the similarity between two keywords is firstly quantified as a measureable index for representing and distinguishing the semantic distance between keywords and their accompanying noises. A graphical model then integrates the words' statistical features and their semantic relationships through graph theory-based analytical methods such as path-finding and degree comparisons.

The developed algorithm can be treated as a self-learning method using graphical models and linguistic networks. There are two major contributions from this study: firstly, an innovative algorithm for recognising image contents through harnessing rich online resources has been devised. Secondly, the semantic meaning of an image can be modelled through quantifiable linguistic analysis on keywords retrieved from a group of web sources containing severe semantic noise. The research has enabled a system prototype being developed to test the effectiveness of the model for harmonising an image's visual and semantic information.

This paper is structured as follows: Section 2 elaborates the methodology of the core studies carried out in this research, including HWRN construction and image keywords extraction. In Section 3, the system has been tested against benchmarks set to the status quo in the field. Section 4 summarises the contributions of the work, which leads to a list of anticipated future improvements.

2. THE HEAD-WORDS RELATIONSHIP NETWORK

2.1 Problem Definition and Abstraction

Image keywords contain semantic meanings of major visual contents in the captured image scenes. Through generic computer vision algorithms (F), knowledge-level information can be extracted from an image (I) and described as semantic keywords (S), which can be modelled as $S = F(I)$. Modern web search engines can readily retrieve a group of visually similar images $\tilde{I}_k (k = 1 \cdots N)$ using colour, texture and shape features where $I \approx \tilde{I}_k$. Those images are located in online data repositories such as news webpages, image albums, and even social medial networks, which are usually "noisy", "asynchronous" in terms of semantic definitions, and are often dispersed across webpage titles, XML tags, introductions, and paragraphs. It can be denoted as $\tilde{T}_k$ for each set of image searching results.

Based on human intuition when observing images, if $I \approx \tilde{I}$, the semantic keywords between the sample image and the retrieved set should also be identical, that is $S \approx \tilde{S}$. In this research, an accumulation and approximation algorithm has been developed to formulate the transformation function expressed as

$$S \approx \phi_{\tilde{I}}(\tilde{T}) \quad (1)$$

The approximation function ϕ is formed by semantic relationship associated with online image searching results. $\tilde{I}$ and $\tilde{T}$ are defined as random variables for denoting the observed images, whose keywords domain are set up by $(\tilde{I}_k, \tilde{T}_k)$ instance pairs.

In order to accurately assign semantic information to a targeted image, it is assumed that, when $I \approx \tilde{I}_k$, the image searching results will contain relevant profiling information in the keywords domain, where each image instance will carry visual-similarity to an extent. Based on the preliminary study [8], it is safe to assume that online data sources contain sufficient information and semantic knowledge for every image instance $\tilde{I}_k$ with matched content pair $\tilde{T}_k$. The main task of this research is to develop the algorithm for the approximation function ϕ that can summarise common keywords from a large set of $\tilde{T}$ instances.

2.2 Online Image Search Mechanisms

The instance pairs $(\tilde{I}_k, \tilde{T}_k)$ are built up based on visually similar images searched and retrieved from online resources using reverse image searching mechanisms equipped by web search engines such as Google's "Search-by-Image". The so-called "reverse" operation is based on various content-based image retrieval (CBIR) algorithms measuring image colours, textures, and shapes-based features such as the scale-invariant feature transform (SIFT) and the speeded up robust features (SURF). The algorithm starts from bag-of-word (BoW) representation and spatial pyramid matching (SPM), followed by a support vector machine (SVM) classifier. Figure 1 illustrates an example of searching an input image online with its top 10 rated searching results, where each matched image is an instance of $\tilde{I}$. In reality, those retrieved images are used for different online purpose, such as retailing, news reports, art exhibitions and personal blogs, so the titles are semantically independent. However, it is noticed that certain keywords are "similar" and related to the input image.

Theoretically, $\tilde{T}_k$ is defined by the entire searched webpages' contents. However, most webpages contain large amount of "noisy" semantic information such as sponsor information, advertisements, and hybrid links. The identification and extraction of the "key messages" through searching the main webpage text bodies present inherited practical problems [1]. In this research, a simplified definition of $\tilde{T}_k$ is given by the "surrounding-and-intimate" text of images, such as captions, page titles and text bodies, since they are most likely to contain the significant semantic information of the entire webpage.

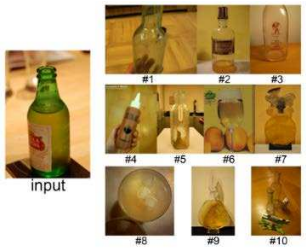

Figure 1: A "Search-by-Image" result

The $\tilde{T}_k$ can be further refined by extracting only the "head-words" (HWs, denoted as t_p) from the theme, which are the core semantic elements in the titles. The structure of each $\tilde{T}_k$ can be defined as

$\tilde{T}_k = \{t_p\}\ p = 1,2,....$ The image keywords usually come from the dominant objects in image scenes, where HWs are suitably defined by the noun phrases (NPs) of each title by using the NP tagging techniques [7]. The NP is a fundamental grammatical structure acting as a noun in a sentence. NPs can usually be located as verb's subjects and objects. For example, in a webpage title "*BBC News - Ecuador tackles bootleg alcohol after wave of deaths*" contains NPs "*BBC News*", "*Ecuador*", "*bootleg alcohol*" and "*wave of deaths*". The head-words can be identified from each NP as t_1= "*news*", t_2="*Ecuador*", t_3="*alcohol*" and t_4="*death*".

2.3 HWRN Formulation and Operations

After identifying all members of $\tilde{T}$ in a search, the input image I will be mapped to a huge group of HWs. Although the collection of HWs contains likely candidates of the image keywords, since the HWs are independently searched and located from different webpages, those keywords' statistical features are often undistinguishable from the noises (irrelevant words) based on the intrinsic characteristics of the reverse searching mechanism. It is therefore necessary to introduce the pair-wised relationships of words and extract the most representative keywords under the global threshold settings. In this research, the work has been benefitted by recent developments in linguistic studies and the graphical theories.

A Head-words Relationship Network (HWRN) is defined as a graph $G = (V, E)$, where V denotes the collection of vertices $\{v_i\}$ and E denotes the collection of edges $\{e_{ij}\}$ between two vertices v_i and v_j. In this application, each HW is defined as a vertex that $v_i = t_p$. The task is to calculate the value of pair-wised edge weights that represents the similarity between two words.

The semantic relationship calculation algorithm developed in the research is based on the linguistic hyponym relations [5], where a word A is a hyponym of word B if A is a subtype or instance of B. For example, "*ink*" is a hyponym of "*liquid*", "*news*" is a hyponym of "*information*". This hierarchical "A *is a kind of* B" logic relationship can be formulated as a tree structure where the ancestor word is a more generic semantic description of each of its descendant. Many semantic taxonomies and thesauri project such as WordNet [4] are concentrated on providing structured information to declare semantic relationships between words. Figure 2 illustrates an example to organise hyponyms into a tree structure, where lines represent the hyponym links.

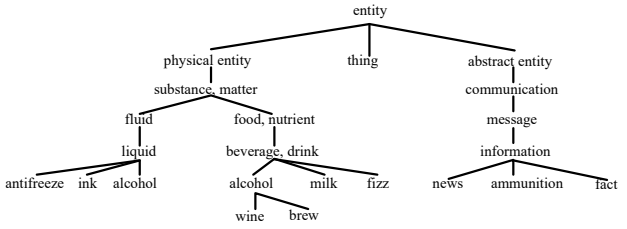

Figure 2: Sample "entity" structure of the hyponym relationships

In general, most semantic distance calculation methods can be used in the HWRN process. In this paper, an updated Resnik algorithm [11] has been applied, which can be summarised as following three steps:

Step 1: HW's probability estimation

Defining a probability function $p: W \rightarrow [0,1]$, where W is the total collection of English nouns observed, such that $t_i \in W$, $p(t_i)$ is referred as a HW's probability. Practically, W is estimated by using a large collection of text samples (more than 1×10^6 words) from broad range of English newspapers and books. Those text samples are then located in the tree structure to calculate each $p(t_i)$ by counting the occurrence of t_i based on their hyponym relationship. An occurrence is counted if the word or any descendants of the word in the tree is observed. For example, in the Figure 2, an occurrence of "*news*" is also counted towards the occurrence of "*information*", "*message*", "*communication*", "*abstract entity*" and "*entity*"; formally,

$$p(t_i) = \frac{count\ (t_i)}{N} \tag{2}$$

where N is the size of the English words space. It is worth noting that $p(t_i)$ is monotonically non-decreasing value along the tree's leaves to the root. If A is a hyponym of word B (A is-a-kind-of B), then $p(A) \leq p(B)$.

Step 2: HW's information capacity level

HW's information capacity level is intrinsically decided by the fact that the semantic information decreases if the word is more abstractive. For example, in the Figure 2, the semantic meaning of the word "*milk*" is less ambiguous than "*nutrient*" in human language, which means more semantic information has been carried in the former. The HW's information level is commonly modelled as

$$info(t_i) = -\ln p(t_i) \tag{3}$$

Step 3: Pair-wise HWs' relationship

Similarity of two HWs is measured by the overlapping information they shared. In a hyponym tree, the set of common ancestors contains the shared information between two words. A pair-wised HWs' relation is defined by the largest value of the information capacity level of their common ancestors. Due to the non-increasing nature of $info(t_i)$, the similarity of two HWs can be rated by $info(t_{ij})$, where t_{ij} denotes the first intersected node of t_i.and t_j found in the tree.

When dealing with the so-called polysemy situation, such as the word "*alcohol*", which could mean "a kind of beverage" or even "non-drinkable hydroxyl compounds", more than two nodes for the same HW can be located in the tree. It is possible to eliminate a "less-likely" node based on the context of the HW, for computation robustness, each pair of the HW can still be calculated independently, with the pair having larger value to be kept as the HW's "true" relationship.

2.4 HWRN-based Keywords Extraction

After establishing pair-wised HWs relationships through hyponym tree, the quantified HWRN frames a semantic similarity feature space. Because each $\tilde{T}_k$ is extracted from different webpages independently, only limited numbers of HWs are likely candidates of the keywords for a targeted image that will be used to form the feature space. The constructed HWRN and its feature space will then transform the image keywords extraction problem into a graph nodes recognition task.

The pair-wised HWs' relationship forms the foundation for the construction of HWRN. In this research, the graph of the HWRN is formulated by a $m \times m$ adjacency matrix R where m is the total size of HWs summed up by the number of $\tilde{T}_k$ instances. The weights, $w(e_{ij})$, of each edge is defined as

$$R_{ij} = w(e_{ij}) = -\ln\left[\frac{info(t_{ij})}{c}\right] \tag{4}$$

where C is the normalisation factor with linearly scales $info(t_{ij})$. The matrix of HWRN is symmetrical since $info(t_{ij}) = info(t_{ji})$. R_{ij} can be treated as the distance between two HWs, which measures their semantic distance. The main diagonal elements are not calculated since no self-comparisons for HWs are performed.

However, different HWs containing repeated words are included, which is expressed as $R_{ij} = -\ln\left[\frac{info(t_i)}{C}\right]$, if and only if $t_i = t_j, (i \neq j)$. For simplification, words that cannot be found from hyponym tree will not be counted as the nodes of a HWRN.

Based on the Equation 2 and 3 in Section 2.3, the HWs' information level can be calculated using a look-up-table based on the WordNet approach. The information can then be used for calculating the HWRN's weight adjacency matrix. The process can be illustrated in the following pseudocode. Since this adjacency matrix is symmetrical, only lower-triangular part is used for calculation.

START HWRN Construction Version 1	
1	*m* = Get size of the entire collection of HW's
2	Initialise a *m*×*m* matrix *R*
3	Loop-1: *I* = 1 to *m*
	Loop-2: *j* = 1 to *i* -1
	R_{ij} = *info*(t_{ij})
	End Loop
4	[Normalisation]
	C = maximum member of *R*
5	*R* = -ln(R/C) [See Equation 4]
END HWRN Construction Version 1	

In the experiment, if the HW's information level shared by two HWs is less than a threshold, the two words are considered dissimilar and will be ignored. In practice, this threshold always produces a sparse adjacency matrix, which reduces the time consumption of the corresponding graph operations. Aforementioned operation has been upgraded to the following optimised process:

START HWRN Construction Version 2	
1-3	Same as Version 1
4	[Normalisation]
	T = 0.3[threshold]
	C = maximum member of *R*
	R = *R/C*
5	*if* $R_{ij} \leq$ T
	$R_{ij} = \infty$
	else
	R_{ij} = (R_{ij}-T)/(1-T) [Re-equalise R in (0,1] range]
	R_{ij} = -ln(R_{ij})
	End if-else
END HWRN Construction Version 2	

The sub-graph keeps the stronger links between HWs based on the global topology of the HWRN, which mimics the skeleton of the HWRN. A node having more links than others in the sub-graph is then selected as the keyword. In another term, a node having maximum degrees of HWRN will be chosen as the keyword for targeted image.

In addition, the fully connected HWRN contain a large amount of noisy information that can be filtered out by the weak semantic links and nodes. For a more flexible representation, weighted degree has been used for each node. The degree of each node are then be counted and sorted in descending order. The top three candidates can then be selected as the primary keywords for the image. It is also observed that repeated words from different HWs can use to gain graph degrees, which rise its possibility to be chosen as the image keywords.

It is also worth noting that the threshold operation may cause fragmentation and splitting the HWRN graph into many unconnected sub-graphs. Each sub-graph will need to be operated independently. Given a unconnected graph G containing n connected sub-graphs $\{G_1, G_2, \dots, G_n\}$, the algorithm can be formulated as:

$$\deg_i = \text{weighted degree}(G_i) \quad (5)$$

The $\deg_i$ is a list of the weighted degrees of sub-graph G_i. All $\deg_i$ will be conjoined into a complete list of $deg = \{\deg_1, \dots, \deg_n\}$, which is then being sorted to a descending order for selecting the "top-rated" keywords from HWs.

3. TEST AND EVALUATION

3.1 Experimental Design

In this research, a HWRN-based image keywords auto-extraction system has been developed and tested on a PC platform of Intel CORE i7 CPU and 4G RAM.

The image database used for the experimental design in this research is from the "Visual Object Classes (VOC) Challenge 2012" introduced by the PASCAL project [3]. The database has over 10,000 images and 20 categorised object classes. Every image contains at least one tagged object. During the test, a random image was sent to the Google image search engine. The top 100 searching results were used for the visual similarity test. Their websites' text body, including title, image captions and text body, have been processed for extracting the head-words. Overall, the average head-words number for each test was between 500 and 800 words. The 3 top ranked and primary keywords were then extracted based on the algorithm and the system.

An evaluation over the effectiveness of the proposed algorithm and its operational model was carried out following two experimentation methods:

- **Human intuition test**

The human intuition test (I-test) invites people to rank the matching results based on their own observations. Only the test image (input) and the detected keywords (output) are provided. For quantifying the response on matching accuracy, five correlation levels (CL) from "0" (strong relevance) to "4" (no relevance) were defined in this test for evaluating the correlation between an input image and the detected keywords.

- **Hyponym tree-based test**

In the hyponym tree-based test (HT-test), the image annotations provided by the PASCAL project were treated as the ground truth. A HT-test was carried out by quantifying the CL, for example, marked as "strong relevance" if it matches to the ground truth of one of its descendants in the tree. Alternatively, a path will be located between the result and the ground truth where the CL is equal to the steps between the two points. For example, the path between "*furniture*" (annotation) and "*chair*" (ground truth) involves 2 steps, which is equivalent to the correlation level 2. While the steps larger than 4 will be recognised as "no relevance". If there is no path found between the detected keywords and the ground truth, such as the words "*furniture*" and "*bottle*", the result will also be labelled as "no relevance". To be reckoned as "correct" image keywords, the CL should be no more than 2 in both the I- and HT- tests.

3.2 Feasibility Analysis through HT-test

The result of the feasibility analysis is summarised in Table 1. The average precision (AP) rate was measured during the HT-test. The AP scores were calculated by using the system outputs and the images list in Table 1. Table 1 also shows a like-to-like comparison with state-of-the-art clustering-based approaches [9, 13]. All experiments have been carried out under the same HT-test settings described in Section 3.1.

Table 1: Accuracy test using the PASCAL VOC dataset

Annotation	Number of Images	HWRN AP (%)	SRC[9] AP(%)	CMRM[13] AP(%)
aeroplane	670	**94.2**	70.2	81.5
bicycle	552	**85.2**	45.2	75.6
bird	765	**80.2**	67.6	69.3
boat	508	**65.2**	55.3	60.2
bottle	706	**70.4**	32.6	62.5
bus	421	**84.1**	80.8	83.9
car	1161	**85.5**	63.3	80.5
cat	1080	**90.1**	85.2	88.6
chair	1119	**79.8**	65.3	68.1
cow	303	**78.1**	69.6	75.6
table	538	**70.1**	32.9	62.3
dog	1286	**89.3**	62.1	74.2
horse	482	**92.7**	70.5	81.2
motorbike	526	**95.1**	66.9	76.6
person	4087	**43.3**	21.5	40.6
pot plant	527	**69.4**	43.5	59.3
sheep	325	**93.8**	67.6	72.3
sofa	507	**82.3**	68.6	79.2
train	544	**94.5**	56.2	82.5
monitor	575	**48.7**	40.5	46.6

Since the proposed approach can be seen as a post-processing step for image knowledge-level operations, it has shown better performance over clustering-based approaches which ignored the semantic relationship between clusters. In the developed system, the superior performance is more evident in the vehicles group such as "*aeroplane*", "*train*" and "*motorbike*". Those objects usually dominate an entire image, which made it easier to be detected by online search engines. The webpages "owning" those images are usually focused on explicitly defined semantic topics subject as "*travel*", "*safety*", "*vehicle*" and "*traffic*". The strong semantic connections between those keywords are more distinguishable from the background noises.

The human detection test has resulted the lowest AP score. It is due to the inherent and more diversified nature of those images. Even with a clear shot on a person, the theme of the images is very difficult to summarise, such as in advertisements for beauty products, snapshots from a fashion show, athletes playing outdoor sports, and people congregating in social events.

In general, the HT-test has shown strong correlation of the proposed technique to the nature of images. The operation is more accurate if the objects in the tested images having more explicit semantic meanings.

3.3 Robust analysis through I-test

The hyponym tree model focuses on the "A is a kind of B" relationship, which over-simplifies the real human logic when dealing with semantic relationship in real-life. Different from the HT-test, people usually concentrate on the "dominant" image areas when looking at a photo, and making more complex and comprehensive judgment on the image nature and contents.

For example, in Figure 4(a), the image is tagged by the keyword "*boat*" during the HT-test, but the theme of the image is actually the "*city skyscape*" of Melbourne. Because the image contains landmark buildings, the searched websites had focused on the tourist sites and city introductions. During the test, the developed system even identified the name of the city, Melbourne. However, it missed out the keyword "*boat*". Similar errors occurred under circumstances as shown in Figure 3.

output= "*Melbourne*" Annotation = "*boat*" | output = "*night*" Annotation = "*bicycle*" | output = "*city*" Annotation = "*train*"

Figure 3: Accurate results based on I -test

5 volunteer students were invited for the I-test. The people involved in the experiment did not know the original object annotations and made judgement purely based on their intuitions. The CL scores from each person have been averaged. For comparing with the HT-test results, the original annotations were used as labels for each image groups. Since some images from the PASCAL VOC 2012 database contain multiple objects and have more than one labels, only the main object label has counted in the test. The 20 AP scores from the I-test are listed in Figure 4.

As shown in Figure 4, the I-test shows higher scores across board than the HT-test. The scores grew significantly higher when dealing with images labelled as TV/monitor, person and boat. For example, the "TV/monitor" images often contain certain interactions with other objects, such as "people operating a computer", the extracted keywords such as "learning" and "office" are intuitively close to those images scenarios.

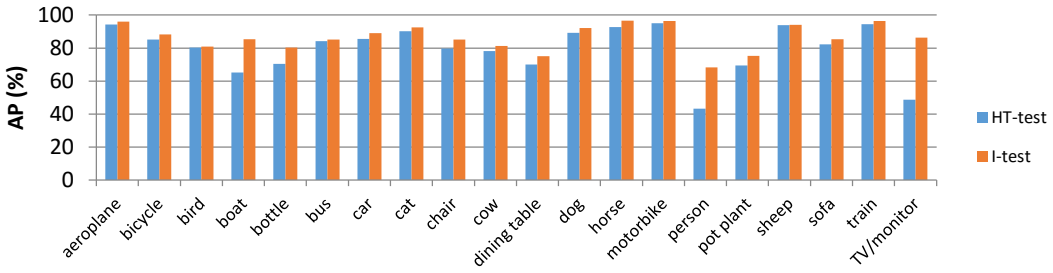

Figure 4: AP score comparison between I-test and HT-test

It is also worth noting that some images containing interactions between "tagged" objects can be used for detecting activities. During the test, many extracted keywords contain the names of certain activities creating a rich source for further semantic analysis (for instance, the verb-based process). For example, in Figure 5, a woman holding a book with a child had retrieved a group of semantically similar images with identical actions. The captions of each image has been quoted in the Figure. By using the proposed method for the top 10 results, keywords such as "development" and "education" are spotted, which indicates theme-rich connections to other HWs such as "child", "educator" and "center".

Input image | "*JBAB Child Development Center children learn through play*" | "*International Newsroom: An initiative: Against child abuse*"

Figure 5: Images containing theme-specific activities

4. CONCLUSIONS AND FUTURE WORK

In this research, a novel graphical model-based image keywords selection algorithm - the "Head-words Relationship Network" (HWRN) - has been devised. A prototype system has been developed to extract image keywords from online resources using online image search engines. The system operates on visually identical images identified by various CBIR techniques using colours, shapes, and texture features. Keywords coming from the host webpages are then used as inputs for the HWRN processes.

The devised HWRN graphical model is composed of a group of noun phrases as candidate words (head-words). The quantified semantic relationships among candidates can be calculated through using the linguistic hyponym trees. Those pair-wised semantic relationships are then used for composing a weighted undirected graph that annotates the global relationships of those candidates. Through applying classic operations such as weighted degree counting, the most representative nodes in the graph will be selected as the keywords for describing the targeted images.

This approach has shown very promising feasibility and robustness during the system test using the quantified hyponym tree (HT-test) and human intuition (I-test) assessments. The analyses on the test results have proven that the developed algorithm is a valid and effective approach for automating the gathering and analysing of online image resources using the rich web "knowledge repositories", hence facilitating the effort in image understanding, machine learning and knowledge acquisition.

Currently, the prototype system can only process semantic information in English. During the test, it is noticed that many highly ranked hosting webpages were various foreign languages - a potential fertile ground for further investigation with many important applications.

REFERENCES

[1] Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M. and Etzioni, O., Open information extraction for the web. in *IJCAI*, (2007), 2670-2676.

[2] Bollegala, D., Matsuo, Y. and Ishizuka, M. A web search engine-based approach to measure semantic similarity between words. *Knowledge and Data Engineering, IEEE Transactions on*, *23* (7), (2011), 977-990.

[3] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111* (1), (2014), 98-136.

[4] Hearst, M.A. Automated discovery of WordNet relations. *WordNet: an electronic lexical database*, (1998), 131-153.

[5] Hearst, M.A., Automatic acquisition of hyponyms from large text corpora. in *Proceedings of the 14th conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, (1992), 539-545.

[6] Henry, P., Krainin, M., Herbst, E., Ren, X. and Fox, D. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The International Journal of Robotics Research*, *31* (5), (2012), 647-663.

[7] Hepple, M., Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, (2000), 278-277.

[8] King, G., Ensuring the data-rich future of the social sciences, Science, 331(6018), (2011), 719-721.

[9] Li, X., Chen, L., Zhang, L., Lin, F. and Ma, W.-Y., Image annotation by large-scale content-based image retrieval. in *Proceedings of the 14th annual ACM international conference on Multimedia*, ACM, (2006), 607-610.

[10] Liu, S., Cui, P., Luan, H., Zhu, W., Yang, S. and Tian, Q. Social-oriented visual image search. *Computer Vision and Image Understanding*, *118*, (2014), 30-39.

[11] Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, *11*, (1999), 95-130.

[12] Wang, C., Jing, F., Zhang, L. and Zhang, H.-J., Content-based image annotation refinement. in *Computer Vision and Pattern Recognition, CVPR'07. IEEE Conference on*, IEEE, (2007), 1-8.

[13] Wang, C., Jing, F., Zhang, L. and Zhang, H.-J., Image annotation refinement using random walk with restarts. in *Proceedings of the 14th annual ACM international conference on Multimedia*, ACM, (2006), 647-650.

[14] Zhang, S., Tian, Q., Hua, G., Huang, Q. and Gao, W. ObjectPatchNet: Towards scalable and semantic image annotation and retrieval. *Computer Vision and Image Understanding*, *118*, (2014), 16-29.