Automatic Object Identification in an Annotated Image using Latent Semantic Analysis

Asst. Lecturer Haithem Kareem Abass*

Abstract

Object identification is the process of mapping visual areas within an image to human been concepts, this is a vital approach to unify image understanding between machines and human. Anyway, images in the Internet are retrieved using natural language query where search engines implement text match to the textual description attached to images. Current image retrieval engines produce huge disturbance in fulfilling user query due to the nature of text matching technique where many un-wanted results are obtained.

In this paper a new model is presented primarily to automatically identify objects composing an image and map identified objects to attached annotations; this is implemented in this paper by using LSA (Latent Semantic Analysis). This approach has lead to secondary revenue where annotations have been de-noised from an-wanted words. A real case study has been taken to verify the hypothesis of this paper and resultant calculations approved the approach.

Keywords: Object Identification, Image mining, LSA, SVD, annotation

* Software Engineering Department, Al-Mansour University College

1- Introduction

From the inspection of popular image search engines such as Google, Bing and Baidu, the retrieval paradigm employed by these search engines is still based on the keywords composing the query; this query is formulated by users to initiate image search process. Users use natural language words to describe requested image, or other multimedia contents, and the responsibility of a search engine is to scan databases for a proper match. The most crucial element is the search scenario is the indexing of images, or other multimedia contents, where natural language is demanded to achieve the labeling of available images with textual description; this process is called image annotation [1,2].

Content-based image retrieval, the problem of searching large image repositories according to their content, has been the subject of a significant amount of computer vision research in the recent past. While early retrieval architectures were based on the query-by-example paradigm, which formulates image retrieval as the search for the best database match to a user-provided query image, it was quickly realized that the design of fully functional retrieval systems would require support for semantic queries. These are systems where the database of images are annotated with semantic keywords, enabling the user to specify the query through a natural language description of the visual concepts of interest. This realization, combined with the cost of manual image labeling, generated significant interest in the problem of automatically extracting semantic descriptors from images [1,2,3].

Images are annotated using different methodologies, some are manually; this when clients comment on certain images and automatically such as mining the textual text in internet pages that hold that image. Crucial challenge in image annotation is the redundant words that increase false results such as the irrelevant images returned by Google search engine [3].

The earliest efforts in the area were directed to the reliable extraction of specific semantics, e.g. differentiating indoor from outdoor scenes, cities from landscapes, and detecting trees, horses, or buildings, among others. These efforts posed the problem of semantics extraction as one of supervised learning: a set of training images with and without the concept of interest was collected and a binary classifier trained to detect the concept of interest.

The classifier was then applied to all database of images which were, in this way, annotated with respect to the presence or absence of the concept [2,3].

More recently, there has been an effort to solve the problem in its full generality, by resorting to unsupervised learning. The basic idea is to introduce a set of latent variables that encode hidden states of the world, where each state defines a joint distribution on the space of semantic keywords and image appearance descriptors (in the form of local features computed over image neighborhoods). After the annotation model is learned, an image is annotated by finding the most likely keywords given the features of the image [1, 2, 3].

2- Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the meaning of words. Meaning is estimated using statistical computations applied to a large corpus of text [4].

The corpus embodies a set of mutual constraints that largely determine the semantic similarity of words and sets of words. These constraints can be solved using linear algebra methods, in particular, singular value decomposition [4].

LSA has been shown to reflect human knowledge in a variety of ways. For example, LSA measures correlate highly with humans' scores on standard vocabulary and subject matter tests; it mimics human word sorting and category judgments; it simulates word-word and passage-word lexical priming data; and it accurately estimates passage coherence [4, 5].

3- Singular Value Decomposition (SVD)

The core processing in LSA is to decompose A using SVD (Singular Value Decomposition); SVD has designed to reduce a dataset containing a large number of values to a dataset containing significantly fewer values, but which still contains a large fraction of the variability present in the original data [3, 4, 5].

$$A - U\Sigma V^{T} \quad \text{---------------}(١)$$

Where

1- $EigenVector(AA^{T}) \rightarrow Columns(U)$
2- $EigenVector(A^{T}A) \rightarrow Columns(V)$
3- $EigenValue(A^{T}A)\ OR\ EigenValue(AA^{T}) \rightarrow \Sigma$

The first structure is the single pattern that represent the most variance in the data, after all, SVD is an orthogonal analysis for dataset, U is composed of eigenvectors of the variance-covariance matrix of the data, where the first eigenvector points to the direction which holds the most variability produced by all other vectors jointly. U is an orthogonal matrix where all its structures are mutually uncorrelated. Eignevalues are representing scalar variance of corresponding eigenvectors; this way total variation exhibited by the data is the sum of all eigenvalues and singular values are the square root of the eigenvalues [4, 6].

4- Objectivity

Using Latent Semantic Analysis (LSA) to reduce the redundant annotation of an image by truncating less variant key words of the annotation, and deploying the fact that Visual Blobs are correlated to Annotation concepts (i.e., natural language words), to investigate the theory that variation in variance-covariance natural language semantic space is analogues to visual semantic space.

5- The Proposed Object Identification and Indexing Scheme

In this proposal images are represented by concepts it hold. Image concept is the projection of human interpretation to the visual structures within an image, hence:

$$I = \sum_{i=1}^{N} C_i \cdot \vec{r}_i \qquad \text{---------------} (2)$$

Where I is any image and C_i is the i^{th} concept recognized with that image

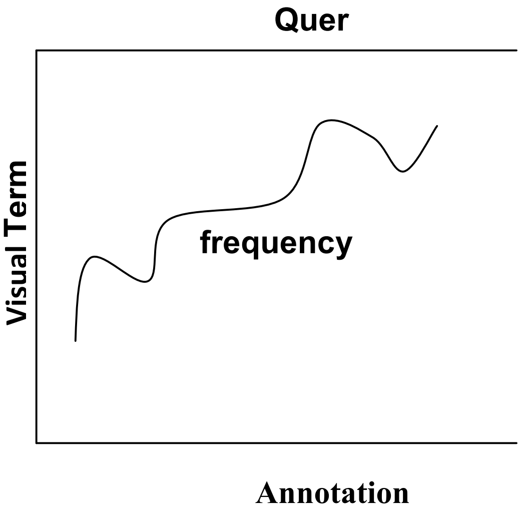
$$q = \sum_{i=1}^{K} w_i \cdot \vec{u}_i \qquad \text{---------------} (3)$$

Where q is the query entered by the user, w_i is the i^{th} word within the query and $\vec{u}_i$ is the semantic unit vector. Semantic meaning for image's concept should correlate human's interpretation for that concept; hence, eq.3 is a prerequisite

$$\vec{v}_i \cdot \vec{w}_i = 1 \qquad \text{---------------} (4)$$

The knowledge space of the proposed system is constructed through two matrices: first one is composed of natural language concepts (i.e., query and returned image ontologies), while the second matrix is composed of visual concepts (i.e., significant objects in a certain image and other significant objects within other images).

Both matrices are to be decomposed using eq.1 , the correlation between these two matrices extend LSA analysis from 2-Dimension to 3-Dimension, where Three dimensions LSA is presented in this proposal where the frequency of visual words are calculated in two axis: first one is query and the second is the annotation, as it is presented in the following figure:

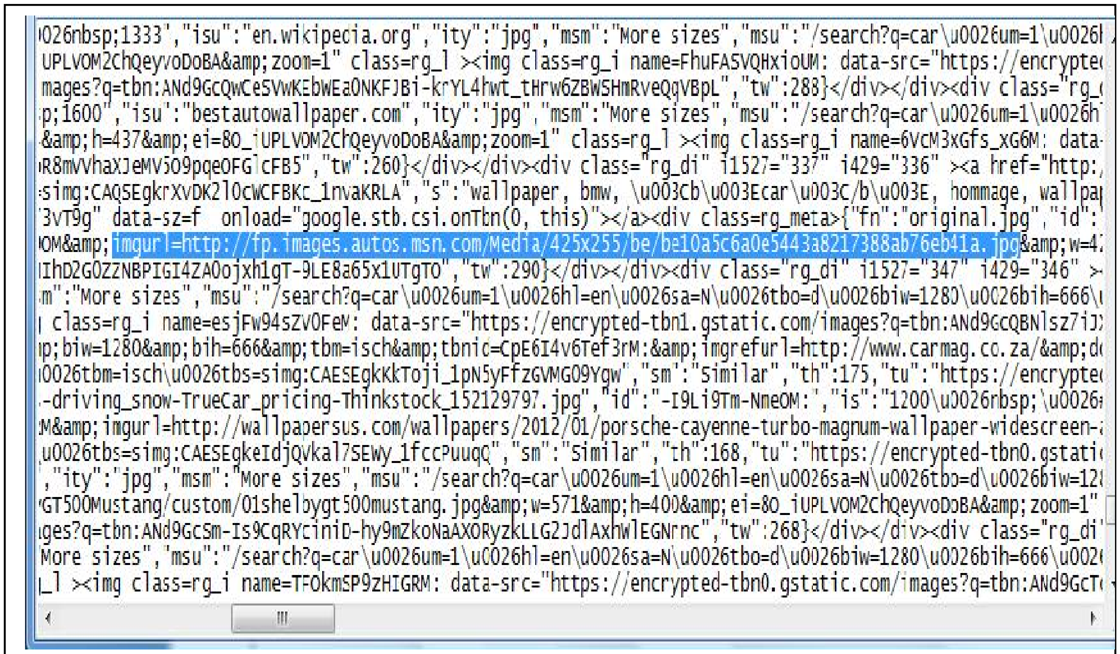


In this paper Java program has been designed to work as web client (i.e., analogous to internet explorer) and user queries for image repositories are sent to the internet as the algorithm (1) presents

```
Algorithm 1:   ReadImageAnnotation
Input : string query;
Output: List<string> imgurl_list;

 Step 1:  Initialize user Query = query;
 Step 2: Initialize Go Search Connection as URL Connection  to Google URL + user Query;
 Step 3: Set  GoSearchConnection  Properties as
         Method  = 'GET';
         Char-set  = 'utf-8';
     User-Agent = 'Mozila-4.0';
     GoSearchConnection. Open;
 Step 4: Get input Stream from GoSearchConnection to stream Reader;
    while stream Reader has imgurl do
           add current imgurl to imgurl_list;
step 5: return imgurl_list;
```

The resultant outcome due to executing algorithm (1) is presented in figure (1).



```
)026nbsp;1333","isu":"en.wikipecia.org","ity":"jpg","msm":"More sizes","msu":"/search?q=car\u0026um=1\u0026h
UPLVOM2ChQeyvoDoBA&amp;zoom=1" class=rg_l ><img class=rg_i name=FhuFASVQHxioUM: data-src="https://encryptec
mages?q=tbn:ANd9GcQwCeSVwKEbWEaONKFJBi-krYL4hwt_tHrw6ZBWSHmRveQqVBpL","tw":288}</div></div><div class="rg_
p;1600","isu":"bestautowallpaper.com","ity":"jpg","msm":"More sizes","msu":"/search?q=car\u0026um=1\u0026h
&amp;h=437&amp;ei=8O_iUPLVOM2ChQeyvoDoBA&amp;zoom=1" class=rg_l ><img class=rg_i name=6VcM3xGfs_xG6M: data-
R8mVVhaXJeMV5O9pqeOFGlcFB5","tw":260}</div></div><div class="rg_di" i1527="337" i429="336" ><a href="http:
simg:CAQSEgkrXVDK2l0cWCFBKc_1nvaKRLA","s":"wallpaper, bmw, \u003Cb\u003Ecar\u003C/b\u003E, hommage, wallpa
3vT9g" data-sz=f  onload="google.stb.csi.onTbn(0, this)"></a><div class=rg_meta>{"fn":"original.jpg","id":
OM&amp;imgurl=http://fp.images.autos.msn.com/Media/425x255/be/be10a5c6a0e5443a8217388ab76eb41a.jpg&amp;w=4
IhD2G0ZZNBPIGI4ZA0ojxh1gT-9LE8a65x1UTgTO","tw":290}</div></div><div class="rg_di" i1527="347" i429="346" >
m":"More sizes","msu":"/search?q=car\u0026um=1\u0026hl=en\u0026sa=N\u0026tbo=d\u0026biw=1280\u0026bih=666\
 class=rg_i name=esjFw94sZVOFeM: data-src="https://encrypted-tbn1.gstatic.com/images?q=tbn:ANd9GcQBNlsz7iJ
p;biw=1280&amp;bih=666&amp;tbm=isch&amp;tbnid=CpE6I4v6Tef3rM:&amp;imgrefurl=http://www.carmag.co.za/&amp;d
0026tbm=isch\u0026tbs=simg:CAESEgkKkToji_1pN5yFfzGVMG09Ygw","sm":"Similar","th":175,"tu":"https://encrypte
-driving_snow-TrueCar_pricing-Thinkstock_152129797.jpg","id":"-I9Li9Tm-NmeOM:","is":"1200\u0026nbsp;\u0026
M&amp;imgurl=http://wallpapersus.com/wallpapers/2012/01/porsche-cayenne-turbo-magnum-wallpaper-widescreen-
u0026tbs=simg:CAESEgkeIdjQVkal7SEWy_1fccPuuqQ","sm":"Similar","th":168,"tu":"https://encrypted-tbn0.gstati
,"ity":"jpg","msm":"More sizes","msu":"/search?q=car\u0026um=1\u0026hl=en\u0026sa=N\u0026tbo=d\u0026biw=12
GT500Mustang/custom/01shelbygt500mustang.jpg&amp;w=571&amp;h=400&amp;ei=8O_iUPLVOM2ChQeyvoDoBA&amp;zoom=1"
ges?q=tbn:ANd9GcSm-Is9CqRYciniD-hy9mZkoNaAXORyzkLLG2JdlAxhWlEGNrnc","tw":268}</div></div><div class="rg_di"
More sizes","msu":"/search?q=car\u0026um=1\u0026hl=en\u0026sa=N\u0026tbo=d\u0026biw=1280\u0026bih=666\u002
_l ><img class=rg_i name=TFOkmSP9zHIGRM: data-src="https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcT
```

Figure 1: Image Server Resultant HTML due to Custom Query

As figure (1) presents, text accompanied the image is retrieved and analyzed for annotation. After retrieving the annotation, the system then starts the identification process by decomposing target image into its kernel objects as presented in algorithm (2). This is done by using BlobCounter class shipped with 'AForge.net' package, the extracting procedure starts by creating alpha mask for the image, then using created alpha mask to extract objects as it is presented in figure (2).

```
Algorithm 2: DecomposeImage2Blob
Input : int blobCount, Bitmap orgImage
Output: List<Bitmap> blobImage

Step 1: define blob counter
         BlobCounter blobCounter  = new BlobCounter();
Step 2: Initialize blobCounter
         blobCounter.FilterBlobs     = true;
         blobCounter.MinHeight      = 15;
         blobCounter.MinWidth       = 15;
step 3: Process the input Image
         blobCounter.ProcessImage(orgImage);
step 4:  Get blob information  using blob counter
         Blob[] blobs = blobCounter.GetObjectsInformation();
Step 5:  Extract Images of each blob
        For each tmpBlob in  blobs
         Begin
                 blobCounter.ExtractBlobsImage(orgImage,tmpBlob,true);
                 Bitmap image = tmpBlob.Image.ToManagedImage();
                blobImage.add(image);
         End
```

Figure (2) presents decomposing an image to a list of Blobs; these Blobs represent the significant visual objects (i.e., colored areas with high discrimination factor than the background)within that image, anyway, Blobs are the first detail that human eye capture at the first glance to the image. Further details of the image need more sophisticated techniques to be revealed such pattern recognition, image semantics and other image mining techniques.

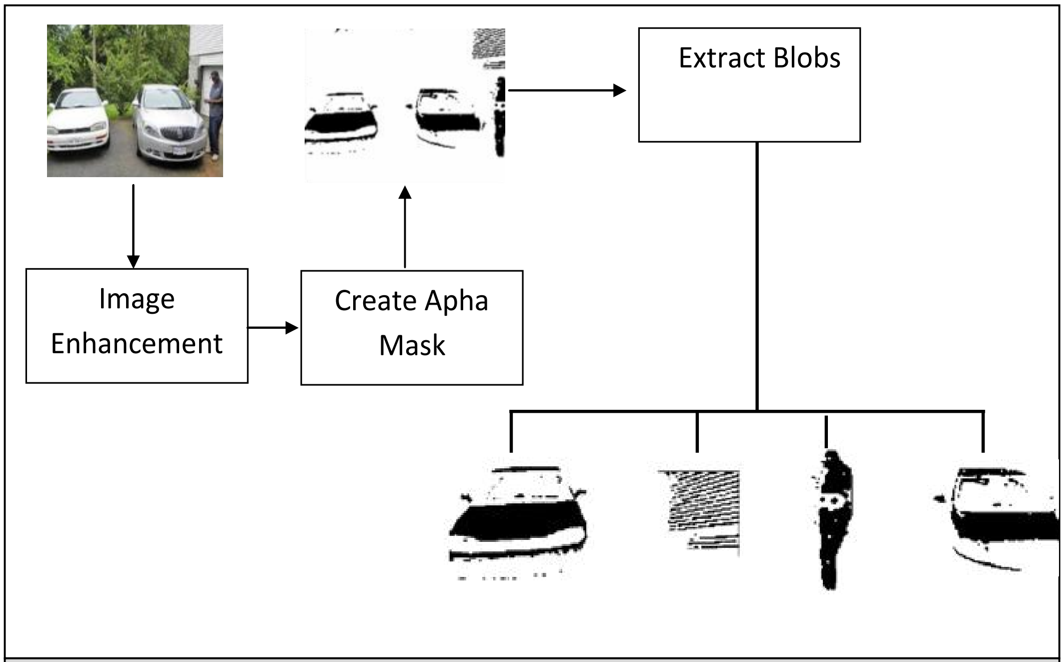


Figure 2: Process of Extracting Image Blobs using Alpha Mask

The following is a query and its resultant images, which are presented in figure(3); these resultant images have the following annotations set:

S1: instead-of-mowing-grass-the-plains-man-wins-car

S2: Oregon_state_police_investigating_fatal_car_crash_west_of_valley

S3:pb_man_lying_on_grass

S4: free_ems_mini_plant_cut_hair_man_grass_doll

S5: vin_diesel_actor_man_car_wheel_serious_bald

S6: two_people_car_race_arrested_grass

Q1: car_man_grass

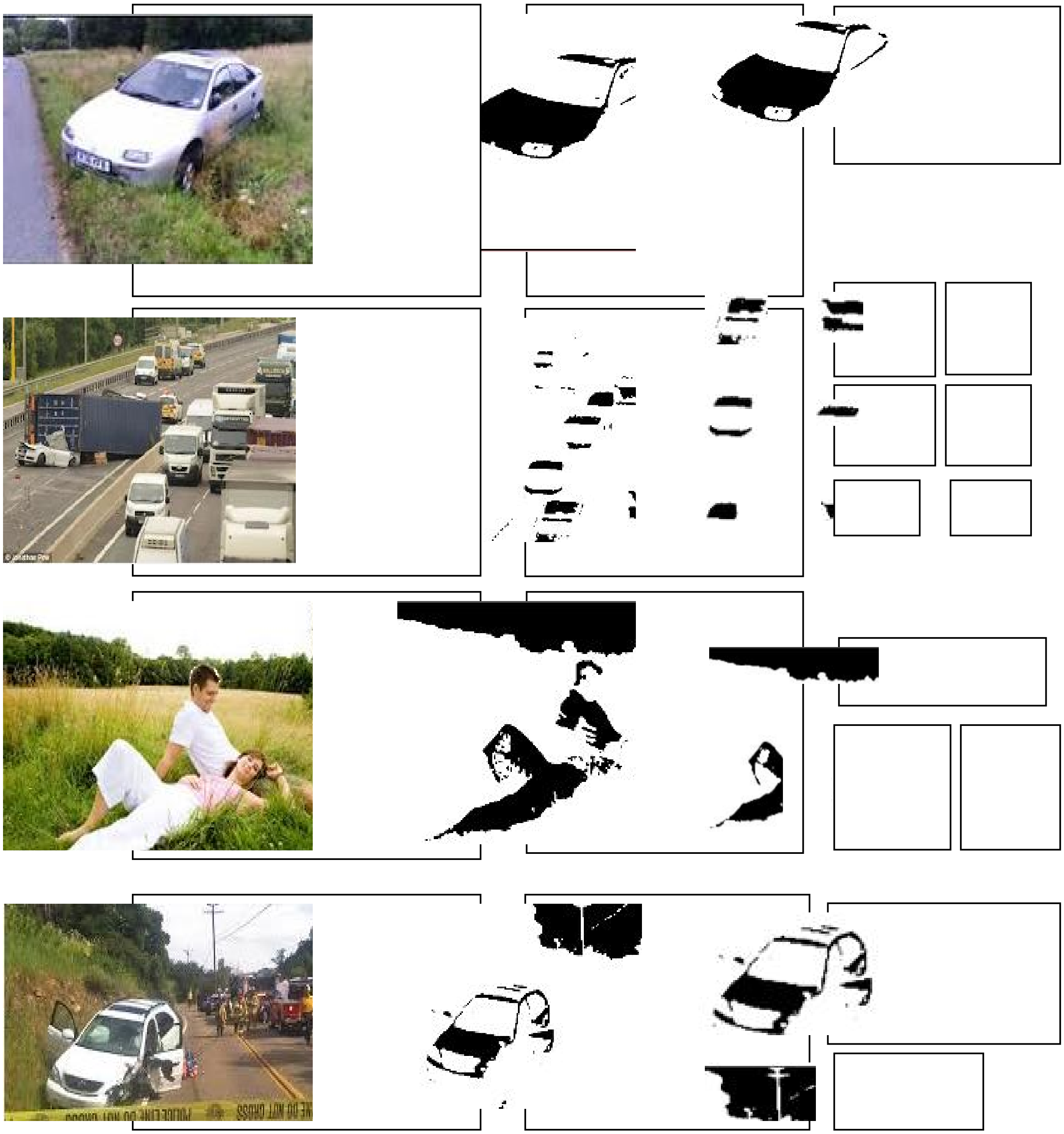


Figure 3: List of Images and their visual analysis (Blob decomposition)

LSA is applied to the annotations and the query to construct the semantic space matrix as it is presented in table (1):

Table 1: Semantic Space of LSA based on word repetition in Annotation

I	query	S1	S2	S3	S4	S5	S6	Q
1	Car	1	0	1	1	1	0	1
2	Man	1	1	0	0	1	1	1
3	Grass	1	0	1	1	0	0	1
4	Crash	0	1	0	0	0	0	0
5	Race	0	0	0	0	0	1	0

The analysis steps are shown below:

U =

-0.6462	-0.3039	-0.0000	-0.4397	0.5447	0.0000	0.0000
-0.5368	0.7337	-0.0000	-0.1821	-0.3747	0.0000	0.0000
-0.5368	-0.4298	0.0000	0.6219	-0.3747	0.0000	0.0000
-0.0547	0.3039	-0.7071	0.4397	0.4597	0.0000	0.0000
-0.0547	0.3039	0.7071	0.4397	0.4597	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.000

Σ=

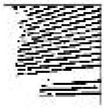
3.2886	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	1.8478	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.7654	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.4300	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

V=

0.5230	-0.0000	0.0000	-0.0000	-0.4759	0.7071	0.0000
-0.1799	0.5615	-0.7071	0.3366	0.1977	0.0000	-0.0000
-0.3597	-0.3971	0.0000	0.2380	0.3953	0.0000	-0.7071
-0.3597	-0.3971	0.0000	0.2380	0.3953	0.0000	0.7071
-0.3597	0.2326	-0.0000	-0.8125	0.3953	-0.0000	0.0000
-0.1799	0.5615	0.7071	0.3366	0.1977	0.0000	0.0000
-0.5230	0.0000	0.0000	0.0000	-0.4759	-0.7071	0.0000

Form Figure (3) and Blob decomposition of the target image we construct the semantic space for the visual objects as table (2) presents:

Table 2: Semantic Space of Visual Object level

I	Blob	S1_Blobs	S2_Blobs	S3_Blobs	S4_Blobs	S5_Blobs	S6_Blobs
1		1	1	0	1	0	0
2		0	0	0	0	0	0
3		0	0	1	0	0	0
4		1	1	0	1	0	0

U=

-0.7071	0.0000	-0.7071	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
0.0000	-1.0000	0.0000	0.0000	0.0000	0.0000
-0.7071	0.0000	0.7071	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

Σ=

2.4495	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

V=

-0.5774	0.0000	0.8165	0.0000	0.0000	0.0000
-0.5774	0.0000	-0.4082	0.7071	0.0000	0.0000
0.0000	-1.0000	0.0000	0.0000	0.0000	0.0000
-0.5774	0.0000	-0.4082	-0.7071	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	-1.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-1.0000

SVD is a geometrical analysis tool that is used to study system dynamic according to the variance to dependant variables; this is done by sampling system states in term of vector of attributes, and then constructs the problem space as a two dimension matrix [A].

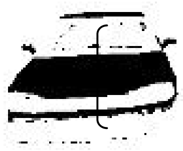
The result of the SVD is a decomposing of the original matrix in a way that demonstrates the relation between variable variance and the system dynamic in the direction of this variable. This concept has been deployed in this research to prove the hypothesis that states, **“Variant in annotation concept is corresponding to variant of visual concept in well described images”.**

From the results, corresponding function is validated in statistical manner, where confidence is promoted when system matrix is increased horizontally.

The output Eigen value matrixes that result of using SVD for both above two matrices shows correlation between words (i.e., concepts) and visual Blobs as the following:

From the first matrix the following results:

{car, man, grass, crash, race} due to **{3.2886, 1.8478, 1.0, 0.7654, 0.43}** and from visual matrix the following results:



,

due to {2.4495, 1.0}

Thus, from mapping the analysis in natural language domain (i.e., image annotations) to the visual domains (i.e., extracted Blobs), the following can be concluded:

Is a car

Is a man

6- Conclusions

From the results obtained by appling LSA as semantic analysis to the images, the following conclusions can be presented as follow:

1- For an image repository, certain query results in a collection of images accompanied by its annotations, analysis and mining process of these annotations is isomorphic to the analysis and mining process in visual domain.

2- Images can be re-indexed according to optimized annotations obtained by using LSA, where annotations are de-noised from linearly correlated words into linearly un-correlated.

3- Mining of image annotations is a potential approach to reduce the gap between low level features and high level semantics and machine understanding.

4- Articulate materials published over the internet can be used to semantically annotate images contained within those materials using LSA. The annotation will be rich due to using intensive concepts to describe the image or the image itself is used as a illustration to the concepts introduced by the text.

5- Identifying objects within an image introduces new potential in constructing semantic network that binds images semantically to build larger subjective connections.

6- SVD possess a great potential in finding isomorphic relationships between different domains with versatile representations. In this research two domains (natural language and visual domains)are investigated against isomorphism and the results shows covenant isomorphism relationship.

7- References

1- NursuriatiJamil and SitiAisyahSa'adan, Proceeding, “Visual Informatics: Bridging Research and Practice”, visual informatics conference, IVIC 2009, Kuala Lumpur, Malaysia, 2009.

2- Masashi Inoue, “On the Need for Annotation-Based Image Retrieval”, National Institute of Informatics, Tokyo, Japan, 2006

3- Reinhard Koch and Fay Huang (Eds), “ Computer Vision-ACCV 2010 Workshops: ‘Two-Probabilistic Latent Semantic Model for Image Annotation and Retrieval’ ”, Springer, USA, 2011.

4- Thomas K. Landauer, Danielle S. McNamara, Simon Dennis and Walter Kintsch,”Handbook of Latent Semantic Analysis”, Lawernce Erlbaum Associates, Inc, USA, 2011,ISBN:978-1.4106-1534-3

5- NursuriatiJamil and SitiAisyahSa'adan, Proceeding, “Visual Informatics: Bridging Research and Practice”, visual informatics conference, IVIC 2009, Kuala Lumpur, Malaysia, 2009.

6- Panagiotis Symeonidis, Ivaylo Kehayov, and Yannis Manolopoulos, “ Text Classification be Aggregation of SVD Eigenvectors”, Springer, USA, 2012

تعريف الكيانات أوتوماتكيا في الصورة الموصوفة بأستخدام خوارزمية LSA

م.م.هيثم كريم عباس*

الخلاصة

تعريف الكيانات في صورة هي عملية ربط المساحات البصرية بما يقابلها من مفاهيم معروفة للانسان، هذه العملية حيوية جدا لتوحيد عملية فهم الصورة بين الماكنة والانسان، حيث ان الصور في الانترنيت يتم استرجاعها عن طريق استعلام معتمد على اللغات الطبيعية حيث ان محركات البحث تنفذ عملية المطابقة النصية للنصوص المرافقة للصور في عملية الاسترجاع.

ان محركات البحث الحالية تنتج الكثير من الفوضى في الاستجابة لطلب المستخدم وذلك بسبب طبيعة عملية مطابقة النصوص مما ينتج عنه الكثير من النتائج الغير مطلوبة.

في هذا البحث تم عرض نموذج جديد لعملية تعريف المكونات التي تكون الصورة وربط هذه المكونات مع مفردات توصيف الصورة، هذا الربط تم من خلال استخدام خوارزمية LSA والتي عند استخدامها تم الحصول على فوائد اخرى ثانوية تتمثل في ازالة الكلمات الفائضة من التوصيف.

تم اخذ حالة حقيقية للتحقق من الفرضيات المطروحة في هذا البحث حيث اثبتت الحسابات والنتائج صحة هذه الفرضيات

* قسم هندسة البرمجيات/كلية المنصور الجامعة