Semantic Relationships in Multi-modal Graphs for Automatic Image Annotation*

Vassilios Stathopoulos, Jana Urban, and Joemon Jose

Department of Computer Science, University of Glasgow,
17 Lilybank Gardens, Glasgow G12 8QQ, UK
{stathv,jj}@dcs.gla.ac.uk

Abstract. It is important to integrate contextual information in order to improve the inaccurate results of current approaches for automatic image annotation. Graph based representations allow incorporation of such information. However, their behaviour has not been studied in this context. We conduct extensive experiments to show the properties of such representations using semantic relationships as a type of contextual information. We also experimented with different similarity measures for semantic features and results are presented.

1 Introduction

Multimedia content, and especially image and video, is produced at highly increasing rates. This indicates the need for effective methodologies for storing and organising multimedia content in order to render it accessible and reusable. Early Content Based Image Retrieval (CBIR) systems were solely based on indexing low-level visual features. The success of such systems, however, was limited mainly due to the semantic gap [1]. A solution towards bridging the semantic gap is to index images using also semantic features, such as keywords, describing the content of the image. The majority of Automatic Image Annotation (AIA) systems incorporate statistical approaches for finding correlations between image visual features and words used to annotate images in a training set. The learnt correlations can then be used to annotate new images.

Often not all the keywords are distinguishable from the visual features alone. For example the concepts of 'meeting' and 'corporate leader' are two of the concepts used in the TrecVid 2006 evaluation campaign [2]. Contextual image information can be used to identify concepts non-distinguishable from visual features and improve object detection. Recently, it was shown that relationships between semantic features can be utilised to improve the annotation performance of existing algorithms [3]. Removing irrelevant terms and identifying others more relevant to be included in the annotation can significantly improve performance.

Graphs and graph learning algorithms provide an interesting alternative for the problem of inference using multi-modal representations of documents. Graph

* The research leading to this paper was supported by European Commission under contracts FP6-027026(K-Space) and FP6-027122(Salero).

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 490–497, 2008.

representations of image collections have been previously used for Automatic Image Annotation in [4] and Image Retrieval in [5]. In [4] only similarities between visual features are incorporated in the graph while, in [5] relationships incorporating image usage information are also considered. In this paper a graph representation is extended to incorporate semantic relationships and the effects in annotation performance as well as the properties of the correlation measure between graph nodes are investigated. Doing so we wish to study the potential of graph models and graph correlation measures for integrating contextual information in an ad-hoc manner.

2 Images and Their Captions as a Graph

An image can be represented by a number of low-level visual features which can be global (extracted from the whole image) or local (extracted from image regions after a segmentation algorithm and concatenated to a single vector describing the image region). In either case an image can be decomposed into a number of feature vectors. Images, their corresponding feature vectors and words can be represented as nodes in a graph $G =< V, E >$, where V is the set of all nodes and E is the set of all edges. A similar strategy with those in [4] and [5] is followed to construct the graph. Let W be the set of nodes representing unique words w used as captions for all the images in the collection. Also let F be the set of nodes representing all the feature vectors f extracted from all the images. Finally let I be the set of all nodes representing images i in the collection. Then the vertices of the Image Graph (IG) can be defined as $V = I \cup F \cup W$.

The relationship between images and their feature vectors can be encoded in the IG by a pair of edges (i_n, f_j) and (f_j, i_n) connecting image nodes i_n and their feature vectors f_j. In a similar way relationships between images and their caption words are encoded by a pair of edges (i_n, w_j) and (w_j, i_n). Now assume that a function $dist(f_i, f_j)$ returns a positive real value measuring the distance, or dissimilarity, between two feature vectors. This function can be the Euclidean distance or any other valid distance metric on feature vectors. Using this function, the k nearest neighbours of each feature vector f_i are selected and a pair of edges $\{(f_i, f_k), (f_k, f_i)\}$ for all the k nearest neighbors, is used to denote their similarity in IG. Similarly, assume a function $dist(w_i, w_j)$ or $sim(w_i, w_j)$ returning a positive real number quantifying the distance or similarity of two words. We will discuss these two functions in the next section. Again the k nearest neighbours of each w_i can be selected and a pair of edges $\{(w_i, w_k), (w_k, w_i)\}$, for all k, is included to the graph to represent semantic relationships between words.

2.1 Finding Correlations between Graph Nodes

One measure of correlation between nodes in a graph can be derived as follows. By performing a random walk on a graph the long term visit rate, or the stationary probability, of each node can be calculated. Random Walks with Restarts

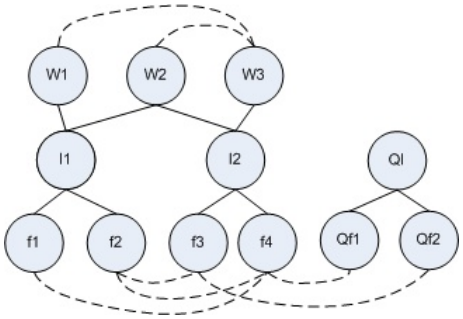


Fig. 1. Image graph with nodes corresponding to a new query image and its feature vectors, QI, Qf1 and Qf2. k = 1 for this graph

(RWR) [6] are based on the same principle but the stationary probabilities are biased towards a specific node, referred to as the restart node. Starting from a node s, a RWR is performed by randomly following a link to another node at each step with a probability a to restart at s. Let $\mathbf{x}^{(t)}$ be a row vector where $\mathbf{x}_i^{(t)}$ denotes the probability that the random walk at step t is at node i. $\mathbf{s}$ is a row vector of zeros with the element that corresponds to the starting node set to 1. Also let $\mathbf{A}$ be the row normalized adjacency matrix of the graph IG. In other words $\mathbf{A}$ is the transition probability table where the element $a_{i,j}$ gives the probability of j being the next state given that the current state is i. The next state is then distributed as

$$\mathbf{x}^{(t+1)} = (1-a)\mathbf{x}^{(t)}\mathbf{A} + a\mathbf{s} \tag{1}$$

To annotate a new image it's feature vectors are calculated and the corresponding nodes are inserted to the graph, see Fig. 1. The starting node is set to the new node corresponding to the query image (QI in Fig. 1). The stationary probability of all words are calculated by recursively applying (1) until convergence. Words are sorted in decreasing order of their stationary probability and the top, say 5 words are selected to annotate the new image.

3 Semantic Relationships

In this study we adopt two approaches for calculating the semantic similarity of words used as captions of images. The first method exploits the co-occurrence of words in the WWW assuming a global meaning of words. The second exploits the co-occurrence of words in the training set in order to identify particular uses of words in the image collection.

3.1 Normalized Google Distance

Although the WWW is not the most reliable source of information, it does reflect the average interpretation of words' meaning globally. Cilibrasi and Vitanyi

[7] propose a method for estimating a distance between words by utilizing page counts returned from web search engines. The probability of a word w_i can be taken to be the relative frequency of pages containing w_i thus $p(w_i) = f(w_i)/N$, where f is a function which returns the number of pages containing w_i in the search engine's index and N is the number of web pages indexed by the search engine. Similarly the probability of a word w_j, $p(w_j)$ as well as the joint probability of the two words $p(w_i, w_j)$ can also be obtained using a web search engine. Therefore the conditional probability $p(w_i|w_j)$ can be estimated as $p(w_i|w_j) = p(w_i, w_j)/p(w_j)$. Since $p(w_i|w_j) \neq p(w_j|w_i)$, the minimum is taken in order to calculate a distance between w_j and w_i giving $dist(w_i, w_j) = min\{p(w_i|w_j), p(w_j|w_i)\}$.

Based on this simple measure the authors in [7] develop the Normalized Google Distance (NGD) that utilizes the Google search engine to estimate the meaning and similarities of words. NGD is expressed as

$$NGD(w_i, w_j) = \frac{\max(\log(1/f(w_i)), \log(1/f(w_j))) - \log f(w_i, w_j)}{\log N - \min(\log f(w_i), \log f(w_j))} \tag{2}$$

3.2 Automatic Local Analysis

Automatic Local Analysis (ALA) is mainly used for query expansion in traditional IR [8]. It utilises documents returned as a response to a user query in order to calculate co-occurrences of words. Then the query can be expanded with highly correlated keywords. The aim of the approach followed in this study is to calculate a similarity between words regardless of the query, in order to enhance the structure of the image graph with semantic relationships. Images can be considered as documents while the frequency of a word in an image is either 1 or 0.

Let H be an $N \times M$ matrix where N is the number of unique words used to annotate the image collection and M is the number of images in the collection. An element $H_{i,j}$ is equal to 1 if and only if word w_i is in the caption of image w_j and 0 otherwise. The co-occurrence correlation between w_i and w_j is then defined as $corr(w_i, w_j) = \sum_{t=1}^{M} H_{it} \times H_{jt}$

This measure gives the number of images where the two words appear together. Words that appear very often in the collection will tend to co-occur frequently with most of the words in the vocabulary and thus the score can be normalized to take into account the frequency of the words in the collection.

$$NormCorr(w_i, w_j) = \frac{corr(w_i, w_j)}{corr(w_i, w_i) + corr(w_j, w_j) - corr(w_i, w_j)} \tag{3}$$

Using (3) the neighborhood of a word w_i can be defined as a vector $\mathbf{s}_{w_i} = \{NormCorr(w_i, w_1), \ldots, NormCorr(w_i, w_N)\}$. Words having similar neighborhood frequently co-occur with a similar set of words and thus they have some synonymic relation. The semantic relationship between two words can then be calculated using the cosine of $\mathbf{s}_{w_i}$ and $\mathbf{s}_{w_j}$.

4 Experiments and Results

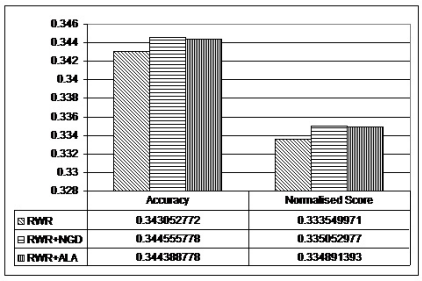
In this study a subset of the Corel image collection consisting of 5000 manually annotated images was used. The dataset is divided into a training set (4500 images) and a test set (500 images). For this dataset image regions are extracted using the NormalisedCuts algorithm and visual features extracted from each region are concatenated in a single vector. The visual features used are average and standard deviation of RGB and LUV values, mean oriented energy and 30 degrees increments, region and location of the region, region convexity, region angular mass and the region boundary length divided by the region's area. For more information about the features extraction and segmentation process refer to [9]. The dataset is available for download[1] and is extensively used in the literature [9,10,11].

For the first run (RWR) of the algorithm described in Section 2, edges between word nodes denoting semantic relationships are discarded while individual features in the region feature vectors are normalized to 0 mean and 1 variance. The values for the restart probability and the number of nearest neighbors for each region feature vector, as have been shown in [4], can be set empirically to $a = 0.65$ and $k = 3$ respectively. The second run (RWR+ALA) incorporates edges between word nodes indicated by the similarity of words as calculated by Automatic Local Analysis described in Section 3.2. Finally in the third run (RWR+NGD) the Normalized Google Distance is used to create edges between word nodes based on their semantic distances.

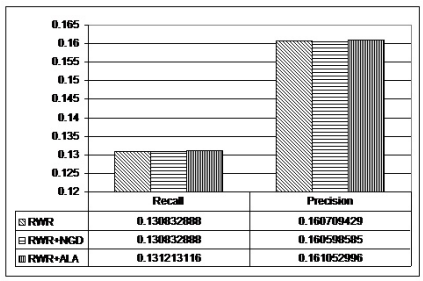
Results in Fig. 2(a) and Fig. 2(b) are reported using average Accuracy [4], Normalized Score[12] and average Precision-Recall [9,11,10] measures. Accuracy for an image is defined as the number of correctly annotated words divided by the number of the expected words for the particular image. The expected number of words is the number of words in the true annotation of the image taken from the test set. Normalized Score is defined as $NS = Accuracy - inc_i/(N_w - e_i)$ where inc_i is the number of incorrectly predicted words for the i^{th} image, N_w is the number of words in the vocabulary and e_i is the number of expected words for the i^{th} image. Averages are taken over all images in the test set. Precision and Recall are measured for each word in the vocabulary and are defined as follows. Precision is the number of correctly annotated images with a particular word divided by the number of images annotated by that word. Recall is the number of correctly annotated images divided by the number of relevant images in the test set. The relevant images are simply the images having the particular word in their true annotations. In this study we report average Precision Recall values over all the words in the vocabulary.

Despite the small increase in performance, the differences in the Accuracy and Normalized Score averages between the RWR and RWR+ALA runs are statistically significant using a paired t-test with a 0.05 threshold. On the other hand, the average differences between the RWR and RWR+NGD runs are not statistical significant. This indicates that semantic relationships calculated using

[1] `http://kobus.ca/research/data/eccv_2002/`



(a) Avg. Accuracy and Normalised Score



(b) Avg. Precision and Recall

Fig. 2. Results obtained from the three runs of the algorithm. See text for description.

Automatic Local Analysis in the image graph can improve the annotation performance although the increase is not dramatic.

The not statistically significant improvement of results obtained by RWR+NGD can have two possible explanations. Firstly, in contrast to the assumption of the authors in [7], NGD is not symmetric $NGD(w_i, w_j) \neq NGD(w_j, w_i)$. In [7] is assumed that the Google search engine returns the same number of pages regardless of the order of the words in the query. Thus the second term in the numerator of (2) is assumed to be symmetric. However, during this study it was found that $f(w_i, w_j) \neq f(w_j, w_i)$ which was probably due to changes in the implementation of the Google search engine. Secondly, NGD reflects the co-occurrence of words in the WWW. While for some words these can be beneficial for image annotation, for others might lead to the opposite results.

The improvement in performance in the RWR+ALA run is due to the improved detection accuracy of particular words. Studying the raw results we found that only 5 words are affected by the semantic relationships and the corresponding Precision Recall values are given in Fig. 4. Studying the Precision Recall measures for each individual word we also found interesting properties of the stationary probability obtained by RWR. Firstly, we found that although for some words there are significantly more training images in the training set than for other words, most of the time the more frequent words are erroneously predicted. For example, for the word 'water' there are 1004 images in the training set. The Precision and Recall for this word is 0.269 and 0.931 indicating that this word is erroneously predicted mostly due to its frequency in the training set. On the other hand, for the word 'jet' the corresponding Precision and Recall values are 0.705 and 0.63 while there are only 147 images in the training set.

Secondly we found a relationship of the restart probability with the number of words having at least one image correctly annotated. As the restart probability increases the number of words with at least one image correctly annotated increases. For a small restart probability only the most frequently occurring words in the training set are predicted, while for a larger restart probability the stationary probability favours word nodes closer to the query image node. In this context the distance between nodes is the geodesic distance in the graph. In Fig. 4 we show how

the number of words with positive recall behaves for different values of the restart probability. For this experiment we did not use semantic relationships; however the behavior is similar to when edges between word nodes are incorporated regardless of the method used (ALA or NGD).

	RWR		RWR+ALA	
Word	Precision	Recall	Precision	Recall
grass	0.22929	0.70588	0.23076	0.70588
rocks	0.16279	0.31818	0.16666	0.31818
ocean	0.35714	0.55555	0.38461	0.55555
tiger	0.62532	0.5	0.66666	0.6
window	0.33333	0.125	0.52356	0.125

Fig. 3. Precision Recall values for the five words which are affected from semantic relationships

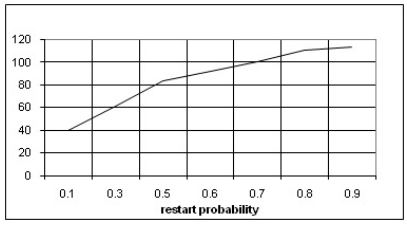


Fig. 4. Number of words with positive recall for different values of restart probability

These findings suggest that the stationary probability obtained by RWR is mostly affected by the frequency of occurrence of the words in the training collection. In other words, it is affected by the connectivity of the nodes in the graph. The notion of distance between nodes in the graph is reflected by the restart probability, although there is not an explicit relation. We conclude that both connectivity of nodes and the geodesic distances in the graph are important properties that must be explicitly considered in the correlation measure.

5 Discussion and Conclusion

Although we achieved a small statistically significant improvement in annotation performance, we have identified two drawbacks to the stationary probability as a correlation measure. First, in contrast to traditional machine learning techniques, the number of training samples for a particular word had negative effect on annotation performance. Second, although the geodesic distances of nodes in the graph are of significant importance in order to facilitate inference, the stationary probability does not define a distance in the graph. The notion of distance is encoded by the restart probability, but the relation is not clear.

One of the most successful applications of the stationary probability obtained by Random Walks with Restarts is the so called PageRank[13] measure of web page relevance. PageRank is an indicator of relevance based on the quality of web page citation measuring in-links of each web-page. In such application the edges in the graph denote attribute value relationships of the form "page A suggests/links to page B". For AIA, the majority of edges in the image graph denote similarities or distances between nodes. We conclude that for such type of edges a correlation measure based mostly on the connectivity of nodes in the graph is not appropriate, leading to the problems above mentioned.

The graph representation, however, provides an interesting approach for integrating contextual information which is vital for improving performance. There are number of different graph correlation measures that can be defined which might be more appropriate for the Automatic Image Annotation problem. We are currently experimenting with other measures such as the Average First Passage Time [14,15].

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE TPAMI 22(12), 1349–1380 (2000)
2. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR 2006, Santa Barbara, CA, USA, pp. 321–330. ACM Press, New York (2006)
3. Wang, Y., Gong, S.: Refining image annotation using contextual relations between words. In: CIVR, Amsterdam, The Netherlands, ACM, pp. 425–432 (2007)
4. Pan, J.Y., Yang, H.J., Faloutsos, C., Duygulu, P.: Gcap: Graph-based automatic image captioning. In: CVPRW 2004, vol. 9, p. 146. IEEE Computer Society Press, Los Alamitos (2004)
5. Urban, J., Jose, J.M.: Adaptive image retrieval using a graph model for semantic feature integration. In: MIR 2006, Santa Barbara, CA, USA, pp. 117–126. ACM Press, New York (2006)
6. Lova'sz, L.: Random walks on graphs: A survey. In: Combinatronics Bolyai Society for Mathematical Studies, Budapest, vol. 2, pp. 353–397 (1996)
7. Cilibrasi, R., Vitanyi, P.M.: Automatic meaning discovery using google. In: Hutter, M., Merkle, W., Vitanyi, P.M. (eds.) Kolmogorov Complexity and Applications. Number 06051 in Dagstuhl Seminar Proceedings, Schloss Dagstuhl, Germany, IBFI (2006)
8. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Reading (1999)
9. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 97–112. Springer, Heidelberg (2002)
10. Yavlinsky, A., Schofield, E.J., Rüger, S.: Automated image annotation using global features and robust nonparametric density estimation. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 507–517. Springer, Heidelberg (2005)
11. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: SIGIR 2003, Toronto, Canada, pp. 119–126. ACM Press, New York (2003)
12. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. JMLR 3, 1107–1135 (2003)
13. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998), `http://dbpubs.stanford.edu:8090/pub/1999-66`
14. Kemeny, J.G., Snell, J.L.: Finite Markov Chains. Springer, New York (1960)
15. Gallager, R.G.: Discrete Stochastic Processes. Kluwer Academic, Boston (1996)