International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 1, January - 2014

Automatic Caption Generation for Multilevel Annotating Images

Mrs. C. Kavitha M.E.,(Ph.D.) S. Revathi M. Nivvetha

SAP, CSE@CCET,Pondy CSE@CCET,Pondy CSE@CCET,Pondy

ABSTRACT:

The ratio of digital images increases rapidly, hence to categorize these images and to retrieve the relevant web images we use annotation mechanism. Thus to encourage the accuracy of image categorization and retrieval of relevant images, multilevel annotation has received a higher scope in research attention. Meanwhile ,a single image has to be labeled and annotated for categorizing it with other similar images. The labeled images has to be prioritized ,to predict the similar images of its kind. Hence we use two mechanism: text based search, the image is to be annotated and then the text is used to describe the concept of the image. content based search consists of two stages: mapping of images and hash code based image retrieval which is done to categorize the labeled images. Also, we are going to generate caption for the web images that has been annotated in multilevel here we are using SRC algorithm to generate captions for the images.

Index Terms- multilevel annotation, SRC algorithm, hash codes

1 INTRODUCTION

A web service is a method of communication between two electronic devices over the World Wide. The web service architecture includes Web Service Roles that is Service provider, which is the provider of the web service. The service provider also implements the service and makes it available on the Internet. The Service requestor, which is provided as the consumer of the web service. The requestor utilizes an existing web service by opening a network connection and sending an XML request. Service registry, which logically centralized the directory of services. The registry provides a central place where developers can publish new services or finding the existing ones. It therefore serves as a centralized clearing house for companies and their services. The characteristics of web services are: xml-based, Loosely coupled, Coarse-grained, Ability to be synchronous or asynchronous, Supports RPC. Recently we have witnessed an enormous growth in the amount of digital information which is available on the Internet. Flickr, one of the best known photo sharing websites available, hosts more than 3 billion images, with approximately 2.5 million images being uploaded every day on the internet. Many online news sites like CNN, Yahoo!, and BBC publish images with their stories and even provide photo feeds related to current events happening in the day today lifestyle. Browsing and finding pictures in large-scale and varieties of collections are an important problem that has attracted much interest within information retrieval.

Many of the search engines deployed on the web related images without verifying their content, simply by matching user questions against collocated textual information. example includes metadata , user-annotated tags, captions, and, generally, text surrounding the image. As this stage the applicability of search engines , a great deal of work has been focused on the development of methods that generate description of the words for a picture automatically. The typical approach to image description generation adopts a two-stage framework consisting of content selection and surface realization. The first stage analyzes the content of the image and identifies “what to say” , whereas the latter stage determines “how to say it” . Both stages are usually developed manually. Content selection makes use of dictionaries that specify a mapping between words and images available on the internet, and surface realization uses human written templates or grammars for generating the output. This approach can create sentences of

high quality that are both meaningful and understandable . however, the present manually created resources limits the deployment of the existing methods to real-world applications. Developing dictionaries that specify comprehensively image-to-text correspondences is a hard and much of time-consuming task that must be frequent for new domains and languages. the related problem of generating captions for news images given is discussed . Our approach leverages the vast resource of pictures available on the web and the fact that many of them naturally co-occur with topically related documents and are generating captions for the images[1]. We focus on captioned images embedded in web articles, and learn the models of text based search and the content based search . At the training phase, our models learn from images, their captions, and associated documents, while at test time they are given an image and the document is embedded in and generate a caption for that particular image[2]. Compared to most of the work which is done on the image description generation, our approach is to generate captions for the web images.

Hereby we are going to discuss about the related work, experimental results, future enhancement and conclusion

2 RELATED WORK

Although image understanding is a popular subject within the computer vision, relatively little work has focused on Generating captions for the images. As mentioned earlier, a handful of approaches has been done to create image descriptions automatically following the stages in the generation. The picture is first annotated using image processing techniques into an abstract illustration, which is then rendered into a natural language depiction with a text generation engine. The main of this paper is to give articles or description about the query image by using search engine. The task of generating captions for news images is narrative to our knowledge. Based on manual annotation or ontology information we exploit different modal database of news articles, images, and their captions[3].The caption generation task includes some resemblance to headline generation , where the aim is to create a very short abstract for a document. However, we wish to create a caption that not only summarizes the document but is also useful for the image retrieval[4].

In our paper we are going to annotate the given query image. Thus there are many ways of annotations available they are indicative, informative, evaluative and combination. Here we are going to use informative annotation[5],[6],[7]. Probably informative annotation will annotate the given image partially we are redefining the informative annotation to multilevel annotation.

Multilevel annotation annotates the whole images detail and list out all the features that are available on the query image .thus after annotating we are going to prioritized to the features of the image using hash code based on the image retrieval process. Thus according to the priorities ,the images get searched in the database and retrieve the similar images that are available on the database to the user.

3 EXPERIMENTAL RESULTS

Our experiments used web articles accomplished by generating captioned images. Many of the image based datasets are used in the computer vision and the image retrieval are not suitable for caption generation since they were developed using different formats. the given query image is first annotated by identifying the objects present in the image[8]. Then it will give priority based on the image stored in the database. Search engines are programs that search documents for specified keywords and returns a list of the documents where the keywords were found. A search engine is really a general class of programs, however, the term is often used to specifically describe systems like Google, Bing and Yahoo! Search that enable users to search for documents on the World Wide Web[9],[10]. The images are stored in the database according to the category which are already defined by the user .the given query image is first annotated using multilevel annotation. Then the annotated image is labeled and prioritizing the main object in the image[11].

There are three modules discussed below:

They are

1. Text based search
2.Content based search
3.Learning annotations by clustering

3.1 Text Based Search

In our approach, we collected flickr data set contains over 2,000 downloaded images from 52 different groups. These descriptions capture the corresponding images. In this type the image are searched based on the given text.

Fig-3.1 text based search

3.2 Content Based Image

3.2.1 Mapping visual features to hash codes

The visual features are mapped into bit streams, with higher bits which represent the more important of an image, which is used to speed up the searching process by comparing only the value of higher bits of the given images. This idea is to encode image visual features to so-call hash codes. Images are divided into even no of blocks and standard luminance of each block is extracted as visual features. These features are changed by a PCA mapping matrix learned , and then quantized into hash codes[12],[13]. The quantization approach is that if a feature component is larger than the mean of this vector, it is quantized to 1, otherwise to 0.

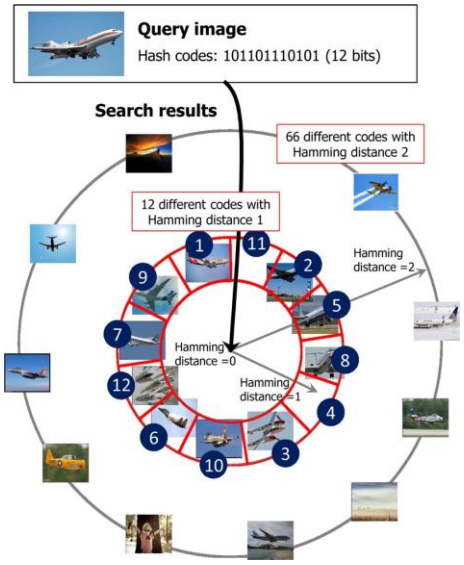

Fig 3.2.1 query adaptive image search represented by 12-bit hash code

3.2.2 Hash code-based Image retrieval.

Four distance measures are proposed and compared.

1) Hash code filtering plus Euclidean distance measure: The advanced bits of the hash codes contain the greater part of energy of an image. Hence if the advanced bits of both hash codes match, possibly they are more same than only lower bits match. This measure is proposed based on these analysis of image retrieval[14]. Images whose advanced n bits of hash codes match accurately those of the query image are kept, and then ranked according to Euclidean distances based on Correlogram features[15].

2) Hamming distance: It measures the number of bits of two different hash codes.

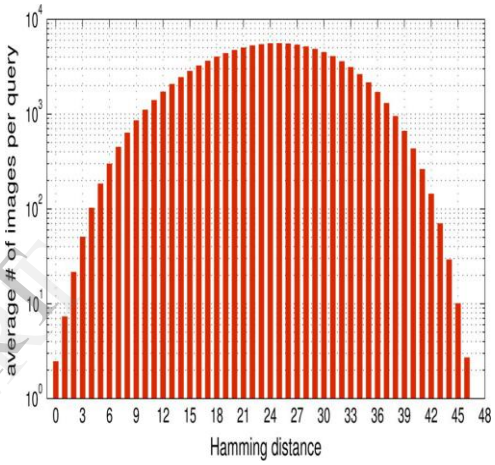

Fig.3.2.2 avg no of returned images at each hamming distances

3) Weighted Hamming distance: Intuitively, since advanced bits are more important, difference in advanced bits should be larger-weighted. This measure evenly separates the 32-bit hash codes into 8 bits, and weights correspond the major Hamming distance [16].

4) Euclidean distance on color Correlograms : This measure is used as a baseline to assess the effectiveness of the hash code based methods for searching the image.

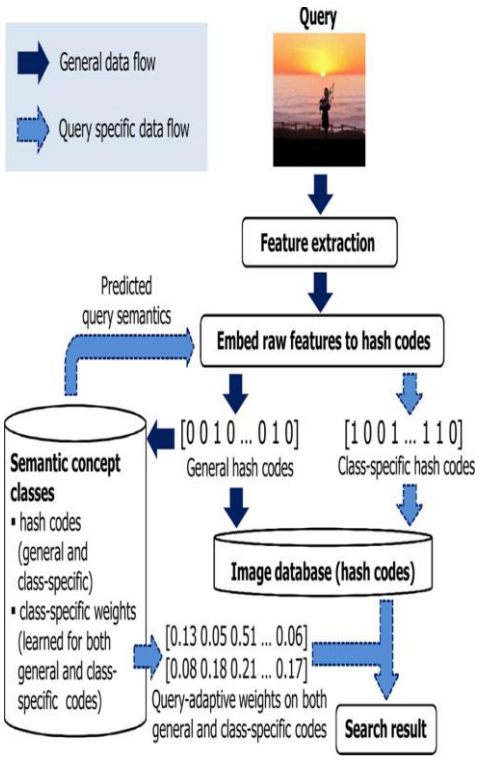

Fig 3.2.3 query adaptive hash code selection

3.2 Learning Annotations By Clustering

3.3.1 The Search Result Clustering algorithm

The Search Result Clustering (SRC) algorithm is one of the effective clustering technique that can generate clusters with highly privileged readable names[17].

It clusters documents by ranking salient phrases. Given a number of documents, it extracts all possible phrases and calculates several properties for each phrase such as phrase frequencies, document frequencies, etc. Then a pre-learned regression model is applied to combine these properties into a single salience score[18],[19],[20]. The top-ranked phrases are taken as the names of the candidate clusters, which are further merged according to their member documents. This method is more suitable for Web applications than other traditional clustering algorithms because it emphasizes the efficiency of identifying relevant clusters.

3.3.2 Annotation prediction

We use SRC algorithm to cluster the retrieved semantically and visually related images according to their text, content and the related URLs . However, it is still an major problem to determine the overall number of clusters for SRC as well as many well-known clustering algorithms, such as k-means[21],[22]. SRC algorithm is used to group all the images in one or two clusters and hence images inside one cluster will cover the major headings such that the learned cluster names are meaningless. This is a trade-off and we enhance the algorithm to output at least four clusters. Moreover, to ensure both the effectiveness and the efficiency for the given image.

We calculate a score for each cluster based on two main possibilities below respectively, and the names of the clusters whose scores are more a certain threshold are extracted. After removing the duplicate words and phrases, the output are as the learned annotations.

The two scoring strategies evaluated are:

1) Maximum cluster size criterion : A cluster's score equals to the number of its member images[23]. This is just the Maximum a Posteriori estimation (MAP). It assumes that the key concepts are the dominant ones.

2) Average member image score criterion : The average similarity of the member images to the query image is used as the score of the corresponding cluster. The reason is obvious[24]. the more relevant the member images of a cluster are to the query, the more probably the concepts learned from this cluster represents the content of the query image[25].

.

4 FUTURE ENHANCEMENT

The query image given is first annotated and then prioritized and then the image is matched with the article and finally generate caption for the given image. The future work can be done by using the bar code .Using barcode the image can be searched according to the code which is used to know the name of the image and the type of the image.

5 CONCLUSION

This paper contains the system of annotating images and generating captions to the preferred images. The motivation of creating this system is to provide an affordable and reliable articles. The captions are used to enhance the information of the image. By this method we can also retrieve the details of the barcode by giving the barcode image. Further the advancement of barcode called as QR code can also be applied in this technique and its details can be retrieved. Hence, to enhance the accuracy and to display the relevant images with captions we can use

automatic caption generation using multilevel annotation.

6 REFERENCES

[1] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image
Classification for Content-Based Indexing," IEEE Trans. Image
Processing, vol. 10, no. 1, pp. 117-130, 2001.
[2] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia
information retrieval: State of the art and challenges," *ACM*
Trans. Multimedia Comput., Commun. Appl., vol. 2, no. 1, pp. 1–19,
Feb. 2006.
[3] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object
Recognition as Machine Translation: Learning a Lexicon for a
Fixed Image Vocabulary," Proc. Seventh European Conf. Computer
Vision, pp. 97-112, 2002.
[4] D. Blei, "Probabilistic Models of Text and Images," PhD
dissertation, Univ. of Massachusetts, Amherst, Sept. 2004.
[5] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M.
Jordan, "Matching Words and Pictures," J. Machine Learning
Research, vol. 3, pp. 1107-1135, 2002.
[6] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous Image Classification
and Annotation," Proc. IEEE Conf. Computer Vision and Pattern
Recognition, pp. 1903-1910, 2009.

[7] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the
Semantics of Pictures," Proc. 16th Conf. Advances in Neural
Information Processing Systems, 2003.
[8] S. Feng, V. Lavrenko, and R. Manmatha, "Multiple Bernoulli
Relevance Models for Image and Video Annotation," Proc. IEEE
Conf. Computer Vision and Pattern Recognition, pp. 1002-1009, 2004.
[9] L. Ferres, A. Parush, S. Roberts, and G. Lindgaard, "Helping
People with Visual Impairments Gain Access to Graphical
Information through Natural Language: The igraph System," Proc.
11th Int'l Conf. Computers Helping People with Special Needs,
pp. 1122-1130, 2006.
[10] A. Abella, J.R. Kender, and J. Starren, "Description Generation of
Abnormal Densities Found in Radiographs," Proc. Symp. Computer
Applications in Medical Care, Am. Medical Informatics Assoc.,
pp. 542-546, 1995.
[11] A. Kojima, T. Tamura, and K. Fukunaga, "Natural Language
Description of Human Activities from Video Images Based on
Concept Hierarchy of Actions," Int'l J. Computer Vision, vol. 50,
no. 2, pp. 171-184, 2002.
[12] A. Kojima, M. Takaya, S. Aoki, T. Miyamoto, and K. Fukunaga,
"Recognition and Textual Description of Human Activities by
Mobile Robot," Proc. Third Int'l Conf. Innovative Computing
Information and Control, pp. 53-56, 2008.

[13] P. He´de, P.A. Moe¨llic, J. Bourgeoys, M. Joint, and C. Thomas,
"Automatic Generation of Natural Language Descriptions for
Images," Proc. Recherche d'Information Assiste´e par Ordinateur, 2004.
[14] B. Yao, X. Yang, L. Lin, M.W. Lee, and S. Chun Zhu, "I2T: Image
Parsing to Text Description," Proc. IEEE, vol. 98, no. 8, pp. 1485-
1508, 2009.
[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L.
Berg, "Baby Talk: Understanding and Generating Image Descriptions,"
Proc. IEEE Conf. Computer Vision and Pattern Recognition,
pp. 1601-1608, 2011.
[16] A. Farhadi, M. Hejrati, A. Sadeghi, P. Yong, C. Rashtchian, J.
Hockenmaier, and D. Forsyth, "Every Picture Tells a Story:
Generating Sentences from Images," Proc. 11th European Conf.
Computer Vision, pp. 15-29, 2010.
[17] V. Ordonez, G. Kulkarni, and T.L. Berg, "Im2Text: Describing
Images Using 1 Million Captioned Photographs," Advances in

Neural Information Processing Systems, vol. 24, pp. 1143-1151, 2011.
[18] A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for Image
Annotation," Int'l J. Computer Vision, vol. 90, no. 1, pp. 88-105,
2010.
[19] C.-F. Chai and C. Hung, "Automatically Annotating Images with
Keywords: A Review of Image Annotation Systems," Recent
Patents on Computer Science, vol. 1, pp. 55-68, 2008.
[20] J.-Y. Pan, H.-J. Yang, and C. Faloutsos, "MMSS: Multi-Modal
Story-Oriented Video Summarization," Proc. Fourth IEEE Conf.
Data Mining, pp. 491-494, 2004.
[21] T. Hofmann, "Unsupervised Learning by Probabilistic Latent
Semantic Analysis," Machine Learning, vol. 41, no. 2, pp. 177-196,
2001.
[22] F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for
Cross-Media Image Indexing," IEEE Trans. Pattern Analysis and
Machine Intelligence, vol. 29, no. 10, pp. 1802-1817, Oct. 2007.

[23] T.L. Berg, A.C. Berg, J. Edwards, and D. Forsyth, "Who's in the
Picture," Advances in Neural Information Processing Systems, vol. 17,
pp. 137-144, 2005.
[24] M. O ¨ zcan, L. Jie, V. Ferrari, and B. Caputo, "A Large-Scale
Database of Images and Captions for Automatic Face Naming,"
Proc. British Machine Vision Conf., pp. 1-11, 2011.
[25] J. Luo, B. Caputo, and V. Ferrari, "Who's Doing What: Joint
Modeling of Names and Verbs for Simultaneous Face and Pose.