Automatic Image Annotation using Neural Networks

Aanchan K Mohan
Rutgers University
Department of Electrical Engineering
aanchan@eden.rutgers.edu

Marwan A.Torki
Rutgers University
Department of Computer Science
mtorki@cs.rutgers.edu

Abstract

Automatic Image Annotation is one of the most challenging problems in Computer Vision today. The manual annotation of images is not only expensive, but also time consuming and sometimes inaccurate as well. Therefore, there is a huge urge in the Computer Vision community today to find ways to automatically annotate images and it is with this motivation that we propose a novel approach to the problem of automatic image annotation in this project. To learn the relationship between the segmented regions in the image and annotation keywords, our approach uses a series of feed-forward neural networks in parallel trained independently using the backpropogation algorithm. The output from these neural networks are then mapped to annotation keywords and are then ranked according to their corresponding magnitude at the output of the neural network.

1. Introduction

Automatic Image Annotation is one of the most challenging problems in Computer Vision today. The manual annotation of images is not only expensive, but also time consuming and sometimes inaccurate as well. One of the biggest applications of Automatic Image annotation is in the field of image search and retrieval. Some of the biggest search engines today like Yahoo! And Google Image Search rely heavily on the textual information that accompanies an image while performing an image. These search engines therefore do not really identify the 'content' of the image while performing a search. Therefore looking at manual annotations while performing an image search can sometimes yield spurious results.

Therefore a computerized system that can accurately suggest annotation keywords for images after analyzing its content can actually prove to be very useful. Though this is the ultimate goal of any image annotation system, practical and computational issues put a constraint on the number of associations that can be learned between keywords and the segments in an image. Therefore even the best automatic image annotation systems have just a limited vocabulary of words that can be learned.

1.1. Related Work

While image retrieval has been active over the years, an emerging new and possibly more challenging field is automatic concept recognition from visual features of images. The challenge is primarily due to the *semantic gap* [1] that exists between low-level visual features and high-level concepts.

The problem of automatic image annotation is closely related to that of content-based image retrieval (CBIR). Since the early 1990s, numerous approaches, both from academia and the industry, have been proposed to index images using numerical features automatically extracted from the images [15].

CBIR for general-purpose image databases is a highly challenging problem because of the large size of the database, the difficulty of understanding images, both by people and computers, the difficulty of formulating a query, and the issue of evaluating results properly. A number of general-purpose image search engines have been developed.

We cannot survey all related work in the allocated space. Instead, we try to emphasize some of the work that is most related to our work. The references below are to be taken as examples of related work, not as the complete list of work in the cited area.

The common ground for CBIR systems is to extract a signature for every image based on its pixel values and to define a rule for comparing images. The signature serves as an image representation in the view of a CBIR system. The components of the signature are called features. One advantage of a signature over the original pixel values is the significant compression of image representation. However, a more important reason for using the signature is to gain on improved correlation between image representation and semantics. Actually, the main task of designing a signature is to bridge the gap between image semantics and the pixel representation, that is, to create a better correlation with image semantics [16].

A note on the topic of concept and annotation: The primary purpose of a practical content-based image retrieval system is to discover images pertaining to a given concept in the absence of reliable meta-data. All attempts at automated concept discovery, annotation, or linguistic indexing essentially adhere to that objective more closely than do systems that return an ordered set of similar images. Of course, ranked results have their own role to play, e.g. visualization of search results, retrieval of specific instances within a semantic class of images etc. Annotation, on the other hand, allows for image search through the use of text. For this purpose, automated annotation tends to be more practical for large data sets than a manual process. If the resultant automated mapping between images and words can be trusted, then text-based image searching can be semantically more meaningful than CBIR. Image understanding has been attempted through automated concept detection. The annotation process can be thought of as a subset of concept detection, i.e., images pertaining to the same concept can be described linguistically in different ways based on the specific instance of the concept. The question is whether visual features of images convey anything about their concept or not [17].

Typical advances in the field of Automatic annotation has been achieved starting from concept detection through supervised classification, involving simple concepts such as city, landscape, sunset, and forest, have been achieved with high accuracy in [2]. An extension of multiple-instance learning has been shown effective for categorization of images into semantic classes [3]. Learning concepts from user's feedback and within a dynamically changing image database using Gaussian mixture models is discussed in [4]. An approach to *soft* annotation, using Bayes Point machines, to give images a confidence level for each trained semantic label has been explored in [5]. This vector of confidence labels can then be exploited to rank relevant images in case of a keyword search. Automated annotation of pictures with a few hundreds of words using two-dimensional multi-resolution hidden Markov models has been explored in [6]. While the classification process chooses a set of categories an image may belong to, the annotation set is chosen in a way that favors statistically salient words for a given image. A confidence based dynamic ensemble of SVM classifiers has been used for the purpose of annotation in [7].

Many of the approaches to image annotation have been inspired by research in the text domain. In [8], the problem of annotation is treated as a *translation* from a set of image segments to a set of words, in a way analogous to linguistic translation. Hierarchical statistical methods for modeling the association between image segments and words, for the purpose of automated annotation, have been proposed in [9,10]. Generative language models have been used for the task of image annotation in [11,12]. Closely related is an approach, involving coherent language models, which exploits word-to-word correlations to strengthen annotation decisions [13]. All the annotation strategies discussed so far model visual and textual features separately prior to association. A departure from this trend is seen in [14], where latent semantic analysis (LSA) is used on uniform vectored data consisting of both visual features and textual annotations.

1.2. Report Outline

This report is organized as follows: Section 2 describes the training process in great detail and the steps followed therein, Section 3 summarizes the results obtained following our proposed approach, and Section 4 concludes this report with a discussion about future work that this project could be extended to.

2. The Training Process

2.1. The Training Database

The IAPR-TC 12 Benchmark image database was used for this project. This database consists of 20000 natural still images(plus 20000 thumbnails). Each image is associated with a free-flowing text caption describing the image.

For this project, about 17558 images were provided as a part of the training set along with image keywords(nouns) extracted from the free-flowing text, and the free-flowing text itself.

2.2. The Proposed Approach

The theme of our approach follows from Li and Wang[15] in their work on the ALIPR(Automatic Linguistic Indexing of Pictures Real-time). But our approach differs from theirs fundamentally in two important aspects.

The first aspect is that the ALIPR scheme, constructs profiling models for 'concepts' or a group of keywords that are visually and semantically related. A 'concept' thus for example could be something like 'alarm,bed,clock'. Therefore,in the ALIPR scheme once the training has been completed, a trained dictionary of semantic concepts is obtained. Our approach uses the keywords themselves and not semantically related concepts, though we did make a sincere approach to work with 'concepts'. This shortcoming will be addressed on a little later.

The second aspect in which our approach differs from the ALIPR scheme is that ALIPR uses a complex statistical modeling method using D2-clustering to construct the profiling model for a given concept. We implemented a neural network instead to learn the association between the keywords and the image segments during the training phase.

Therefore the procedure that we followed consists of doing some preprocessing on the image database to extract a dictionary of keywords we wished to use. The next step involved extracting the relevant features from the image and then use these extracted features to segment the images into regions. Once this has been done the last step is to train a set of neural networks to learn the association between the segmented regions of an image and the annotation keywords for that image which are present in the dictionary that was obtained during the preprocessing step. Figure 1 summarizes the entire training approach.

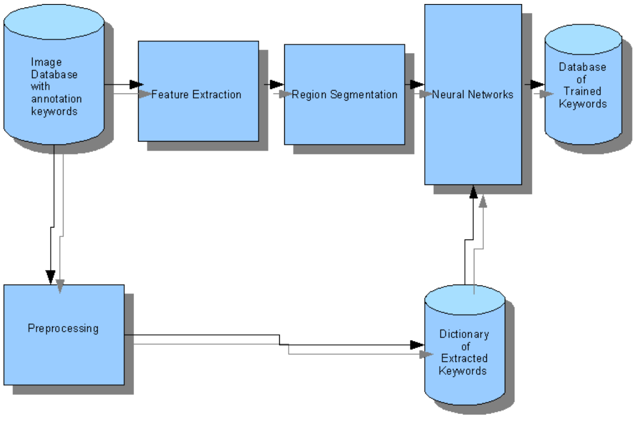


Image Database with annotation keywords
Feature Extraction
Region Segmentation
Neural Networks
Database of Trained Keywords
Preprocessing
Dictionary of Extracted Keywords

Figure 1: The Training Process

2.3. The Preprocessing

After initial preprocessing on the IAPR-TC 12 image database, it was discovered that all the images in the database could be annotated with 1927 unique keywords. Our initial approach was to determine appropriate semantic concepts that could be constructed from this dictionary of keywords. The idea was to then create profiling models for each of these concepts. After considering various approaches, finally a word co-occurrence matrix approach to determine concepts was implemented. A few of the examples of the concepts that were extracted for this database are, 'aisle,tower,cathedral', 'catamaran,boat,coast','dome,building,sky','kitchen,wall,stove' to name a few. But once the concepts were successfully extracted it was difficult to map images to concepts because it is difficult to heuristically determine as to which image can be mapped to which concept, based on the annotation keywords of the images. To overcome this difficulty we finally created a dictionary of keywords based on their frequency of occurrence in the image database. We got rid of words which appeared to be outliers, and kept only those keywords whose frequency of occurrence was greater than 10. The final dictionary of keywords we used consists of 704 keywords.

2.4. Feature Extraction

This is the first step in the training procedure once the preprocessing is complete. The process of feature extraction follows from the procedure used in Wang,Li et al.[16], in their work on the development of the SIMPLIcity system. This is also similar to the feature extraction procedure implemented by Li and Wang in [15]. The process of feature extraction involves the extraction of a 6 element vector for each block of the image after having divided the entire image into regions of blocks of size 4X4. Three elements of this vector are used to represent the color information in a 4X4 block and the other three elements are used to represent texture information over the 4X4 block.

The feature extraction for the color information is done in the L,u,v space. This involves in converting the image from the default RGB color space to the L,u,v color space. The first element of the vector consists of the average value of the L component over a 4X4 pixel region of the image, the second element consists of the average value of the u component over a 4X4 pixel region of the image and the third element similarly consists of the average value of the v component over a 4X4 pixel region of the image.

For the extraction of texture a single level Daubechies' 4 wavelet transform is performed,on each 4X4 block to obtain four frequency bands each of size 2X2 corresponding to LL,LH,HL and the HH bands.

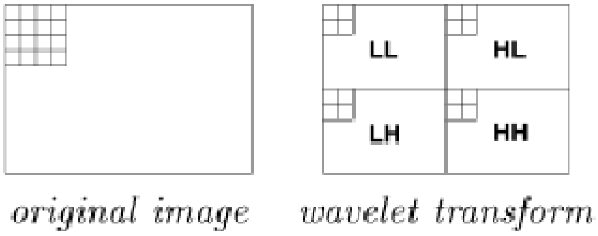


Figure 2:The Wavelet Transform

The three texture feature elements were obtained by computing the squared average of the wavelet coefficients of the HL,LH and HH bands. Thus each 4X4 block can be completely represented by a six element vector.

2.5. Region Segmentation

The second step in our training approach is to cluster the image into a set of 16 clusters that describe the image. We used the K-means clustering algorithm to perform the task of clustering to limit our output clusters to 16 clusters as we mentioned.

The idea of using 16 clusters for each image is to unify the representation of all images to have the same number of attributes to prepare a uniquely sized input vector for the next step in the training approach, which is the neural network part.

We added the spatial features (x, y coordinates) to make the clustering more robust and thus we have ended with 8 features of which three color features, three texture features and two spatial features. We then result in a vector of 128 attributes that defines an image.

We performed image reconstruction to see how much the clustering affects the image and how well the features describe the image, some results of the reconstruction are shown in Figure 3.

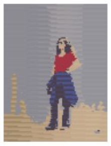
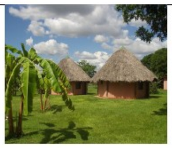
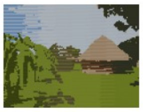
Original image (22007)	Reconstruction	Original image (27021)	Reconstruction
			

Figure 3:K-means Clustering

2.6. Artificial Neural Networks

The last step in our training approach is to build and train a neural network to learn relationship between image segments and the annotation keywords.

We chose to use neural networks to do the learning because, firstly, it typically takes a very little amount of time for simulating of the network, and using a neural network also allows us to learn more than one concept at the same time, thereby not requiring us to learn each concept alone. In our case since we have a dictionary of 704 words, it would be extremely time consuming to learn all such concepts individually. The usage of the Neural Network toolbox in MATLAB makes it extremely convenient, to make a neural network network learn such associations between keywords and image segments and also eliminates the time spent in debugging and testing unreliable code. The last and the most important reason made us use a neural network is the fact that it can be dealt with it as a 'black-box' which actually does the learning.

We processed more than 15000 images of the training set successfully and for each one of these imageswe performed the feature extraction and region segmentation steps that were mentioned earlier. Each image is thus described as an input-output vector pair. The input to the neural network is a vector obtained as the result of the performing region segmentation using K-means, so we have 128 attributes vector to describe the input of the image to the neural network, and for the target output we used a 704 element vector with binary values, so that a 1 in a particular position would indicate the presence of an annotation keyword for that image, corresponding to an entry of that particular keyword in the dictionary that was obtained during the preprocessing step. Therefore the input to the neural net is a 128x15000 matrix representing the entire training set and a corresponding 704x15000 matrix representing the output. We deleted the spatial features that were added to enhance the clustering task and we ended up with a 96x15000 input matrix instead of a matrix of size 128x15000.

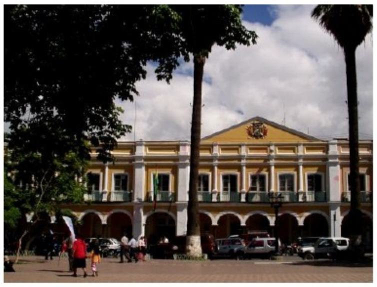
Due to the huge amount of processing required we were not able to learn the network as a whole so we decided to learn 10 different networks with each network taking an input matrix of size 96x1500 and a target output 704x1500. Therefore the input to each network would essentially consist of an input vector corresponding to every tenth vector from the original matrix which is of size 96X150000.This approach allowed us to exploit the benefits of parallel processing, thereby allowing each network to learn the association between image segments and keywords from all across the image training dataset.

Thus in the final approach we used 10 neural networks each of which are learned separately. We used feed forward networks with one hidden layer consisting of 100 neurons and we used the backpropagation algorithm to train the network. We used the option of Resilient back propagation as it is memory efficient. The transfer function used in the hidden layer neurons was tan-sigmoidal and for the output layer it was log-sigmoidal to bound the output between zero and one.

The last thing worth mentioning about the neural networks we used is the training time for learning all the networks. Using 100 neurons in the hidden layer and 1000 epochs to learn each of the networks it took about 30 minutes for each network to train and if we were able to use the parallel processing scheme we may reduce the required to about 30 minutes as a whole but it actually took about 5-6 hours to train the all neural networks.

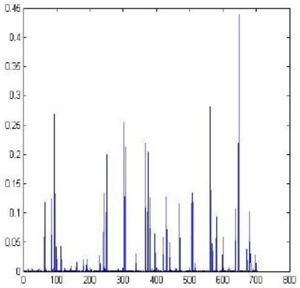
Some results obtained on the training set shows how well the neural network can do. For image 25.jpg from the training set, the image along with its annotation keywords appear in Figure 4.

Now if we take a look at the output vector learned in Figure 5 using the network we find exactly four matches from the top eight scoring words (House, building, tree and front), and two that are present but they are not of the keywords (man and sky) and two more unrelated words (meadow and horse), We thus think the results are very comprehensive to the training data.



Key Words	
Column	150
square	581
house	306
front	251
building	94
car	114
woman	698
palm	421
tree	648
child	129

Figure 4



Highest outputs of the neural net

648→Tree
562→ Sky
94 → Building
304→ Horse
366→ Man
306→ House
375→ Meadow
251→ Front

Figure 5

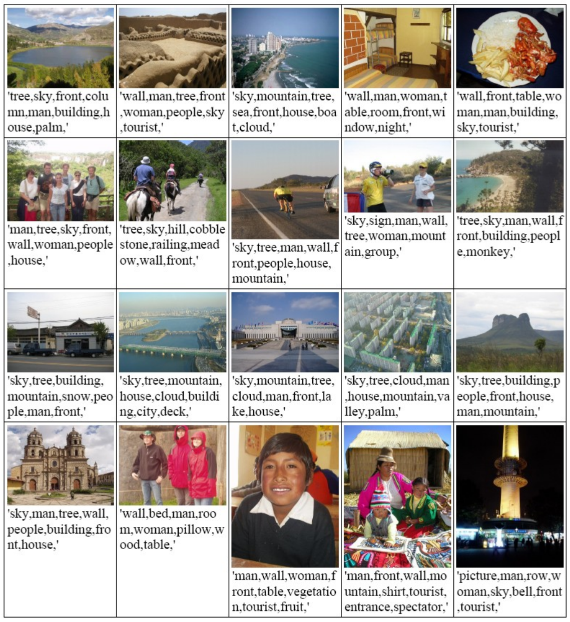
3. Results

The testing set given included 3328 images from the IAPR TC-12 database that were not included in the training set. After the training process was completed, it is worth mentioning that it took less than 4 minutes to process the entire testing set on a computer with 1GB RAM and an Intel Core Duo Processor working at a clock frequency of 1.60 GHz.

The following table consists of a list of annotation keywords obtained on some of the images which formed a part of the testing set that was provided as a part of this project. The following Figure 6 shows the image thumbnails and the corresponding annotation keywords for that image.

4. Conclusion and Future Work

The image annotation scheme developed during the course of this project certainly looks very promising. The accuracy rate and the recall rate are both greater than 19%, which is quite high. We would like to continue working on this scheme in order to refine it and improve its performance in the task of image annotation. To start with we could work with concepts themselves instead of keywords. We believe that pursuing this approach would yield far better results. Also, for the task of region segmentation be believe that instead of using a K-Means procedure, we could use the Mean-Shift Algorithm for segmentation, and choose the top 16 clusters thus obtained for training the neural network.



'tree,sky,front,column,man,building,house,palm,'	'wall,man,tree,front,woman,people,sky,tourist,'	'sky,mountain,tree,sea,front,house,boat,cloud,'	'wall,man,woman,table,room,front,window,night,'	'wall,front,table,woman,man,building,sky,tourist,'
'man,tree,sky,front,wall,woman,people,house,'	'tree,sky,hill,cobblestone,railing,meadow,wall,front,'	'sky,tree,man,wall,front,people,house,mountain,'	'sky,sign,man,wall,tree,woman,mountain,group,'	'tree,sky,man,wall,front,building,people,monkey,'
'sky,tree,building,mountain,snow,people,man,front,'	'sky,tree,mountain,house,cloud,building,city,deck,'	'sky,mountain,tree,cloud,man,front,lake,house,'	'sky,tree,cloud,man,house,mountain,valley,palm,'	'sky,tree,building,people,front,house,man,mountain,'
'sky,man,tree,wall,people,building,front,house,'	'wall,bed,man,room,woman,pillow,wood,table,'	'man,wall,woman,front,table,vegetation,tourist,fruit,'	'man,front,wall,mountain,shirt,tourist,entrance,spectator,'	'picture,man,row,woman,sky,bell,front,tourist,'

Figure 6

5. References

[1] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[2] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H.-J. Zhang, "Image Classification for Content-Based Indexing," *IEEE Trans. Image Processing*, 10(1):117–130, 2001.

[3] Y. Chen and J. Z. Wang, "Image Categorization by Learning and Reasoning with Regions," *Journal of Machine Learning Research*, 5:913-939, 2004.

[4] A. Dong and B. Bhanu, "Active Concept Learning for Image Retrieval in Dynamic Databases," *Proc. IEEE International Conference on Computer Vision*, 2003.

[5] E. Y. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines," *IEEE Trans.*

Circuits and Systems for Video Technology, 13(1):26–38, 2003.

[6] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

[7] B. Li, K.-S. Goh, and E. Y. Chang, "Confidence-based Dynamic Ensemble for Image Annotation and Semantics Discovery," *ACM Multimedia*, 2003.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In Seventh European Conference on Computer Vision, pages 97–112, 2002.

[9] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, 3:1107-1135, 2003.

[10] D. M. Blei and M. I. Jordan, "Modeling Annotated Data," *Proc. ACM Conference on Research and Development in Information Retrieval*, 2003

[11] J. Jeon, V. Lavrenko, and R. Manmatha, "AutomaticImage Annotation and Retrieval using Cross-media Relevance Models", *Proc. ACM Conference on Research and Development in Information Retrieval*, 2003

[12] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," *Proc. Advances in Neutral Information Processing Systems*, 2003.

[13] R. Jin, J. Y. Chai, and L. Si, "Effective Automatic Image Annotation Via A Coherent Language Model and Active Learning," *Proc. ACM Multimedia*, 2004.

[14] F. Monay and D. Gatica-Perez, "On Image Auto-Annotation with Latent Space Models," *Proc. ACM Multimedia*, 2003.

[15]Jia Li and James Z. Wang, ``Real-time Computerized Annotation of Pictures," *Proceedings of the ACM Multimedia Conference*, pp. 911-920, ACM, Santa Barbara, CA, October 2006. .

[16]James Z. Wang, Jia Li, Gio Wiederhold, ``SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol 23, no.9, pp. 947-963, 2001

[17]R. Datta, J. Li, and J. Z. Wang, "Content-based image retrieval - Approaches and trends of the new age," In *Proc. Int. Workshop on Multimedia Information Retrieval*, pp. 253–262, 2005.