Volume 2, Issue 8, August 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Analytical Study on Image Parsing

Anurag Kumar Mishra
Software Engineering,CSVTU
India

Dr. Sipi Dubey
CSE,CSVTU
India

Abstract— Image parsing is the problem of assigning an object label to each pixel. It unifies the image segmentation and object recognition problems. For instance, for a database of horse images, image parsing can be thought of as the task of classifying each pixel as part of a horse or non horse. In more complicated problems, image parsing might require multiple labels, e.g. roads, cars, houses etc. in outdoors scenes. Clearly, pixels cannot be classified in this manner based only on their intensities or even local feature descriptors. Contextual information plays a critical role in Resolving ambiguities. Image parsing can be posed as a supervised learning problem where a classifier is learnt from training data consisting of images and corresponding label maps. Auto context and convolution networks are two promising approaches that apply context to image parsing in the supervised learning setting. Convolution networks are a type of artificial neural network (ANN) in which each processing element carries out a convolution followed by nonlinearity.

Keywords— Query By Image Content, Chabot, ImageParser,

1. Introduction

Fast growth of public photo and video sharing websites, such as "Flickr" and "YouTube", provides a huge corpus of unstructured image and video data over the Internet. Searching and retrieving visual information from the Web, however, has been mostly limited to the use of meta-data, user-annotated tags, captions and surrounding text (e.g. the image search engine used by Google [1]). In this paper, we present an image parsing to text description framework that generates text descriptions in natural language based on understanding of image and video content. Fig. 1 illustrates two major tasks of this framework, namely image parsing and text description. By analogy to natural language understanding, image parsing computes a parse graph of the most probable interpretation of an input image. This parse graph includes a tree structured decomposition the contents of the scene, from scene labels, to objects, to parts and primitives, so that all pixels are explained. It also has a number of spatial and functional relations between nodes for context at all levels of the hierarchy.

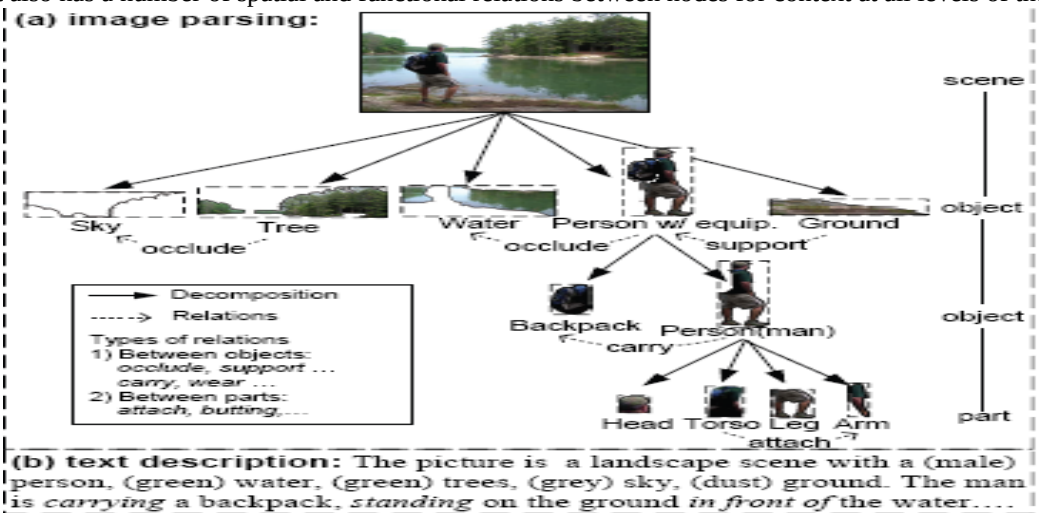

Figure 1: Image to Text Description

There are several application areas of using words and images in parallel. The following example applications extensively utilize the joint distribution of images and words:

- Browsing support: Museums release their collections partially in the web to attract visitors. Typically users who don't know the collection well, prefer to browse [12]. Therefore, it is attractive to organize the collection to

support browsing. Collecting together images that look similar and similarly annotated is a good start. Using image and text together, clustering performance is improved, hence browsing becomes more practical.

- Auto-illustrate: A tool that automatically suggests images to illustrate blocks of text (auto-illustration) would be interesting for many users. Auto-illustration is possible if the joint probability of text and image can be learned. Then, to illustrate a text, one can obtain images with high probability given a text.

- Automated image annotation: Archivists receive pictures and annotate them with words that are likely to be useful keys for retrieving the pictures; journalists then search the collection using these keywords. Annotation have a procedure that annotate images automatically. Annotation process can be done automatically, by learning the joint probabilities of words and images. It is possible to auto-annotate the images by predicting words with high posterior probability given an image. For auto-annotation, words are predicted for a given image. Although, in that sense, auto-annotation can be considered as a suggestive strategy for recognition, it doesn't identify which image region corresponds to which word. The models proposed for annotation can be adapted for solving the correspondence problem [11]. However, since these models are trained to learn the relationships between the whole image and words, the specific relationships between the image regions and words are not explicitly learned.

II. Related Work

- QBIC : QBIC (Query By Image Content) [13] is one of the first content based image retrieval systems developed by IBM. The queries in QBIC are based on sample images, user-constructed sketches and drawings, and selected color and texture patterns. The on-line QBIC demo is at http://wwwqbic.almaden.ibm.com.

- RetrievalWare : RetrievalWare [14] is a content-based image retrieval engine developed by Excalibur Technologies Corp. It searches the images according to their color, shape and texture content, brightness and color structure and aspect ratio. It supports the combination of these features and the weights associated with each feature can be adjusted by the users. More information is available at http://vrw.excalib.com.

- Chabot : Chabot [61], developed at UC Berkeley, retrieves images based on both their content information and associated meta-data. Chabot support queries by color and by text, in addition to some limited domain concept queries like \sunset" or \snow". Queries based on color are of the form \find me the image that has color mostly blue" and are performed on color histograms. More information can be found at http://elib.cs.berkeley.edu/ ginger/chabot.html.

- Photobook and FourEyes : Photobook is a tool developed at MIT Media Lab [65] for searching and browsing images. Features are compared using the matching algorithms that Photobook provides. Photobook includes FourEyes [56], which is an interactive, power-assisted tool for segmenting and annotating images based on the examples from the user. More information can be found at http://vismod.www.media.mit.edu/vismod/demos/photobook/

- .ImageRover : ImageRover [11], which is developed at Boston University, combines textual and visual statistics for searching the images from web. The user initializes a search by specifying a few keywords describing the desired images. The user can then refine this initial query through relevance feedback using both visual and textual cues. The on-line demo can be found at http://cs-www.bu.edu/groups/ivc/ImageRover.

- Netra : Netra is a prototype image retrieval system, developed in the UC Santa Barbara [15]. Color, texture, shape and spatial location information of segmented image regions are used to search and retrieve similar regions from the database. It allows the user to compose queries like "retrieve all images that contain regions that have the color of object A, texture of object B, shape of object C, and lie in the upper one-third of the image" where the individual objects could be regions belonging to different images. The on-line demo is at http://maya.ece.ucsb.edu/Netra.

- MARS : MARS (Multimedia Analysis and Retrieval System) is a system developed at University of Illinois at Urbana-Champaign [6]. MARS organizes various visual features into a meaningful retrieval architecture which can dynamically adapt to different applications and different users. The on-line demo is at http://www-db.ics.uci.edu/pages/demos/.

- Color-WISE and Web-WISE : Color-WISE [7] is a color based image retrieval system, developed in Wayne State University. Dominant hue and saturation values are determined for different parts of an image through a

process of block-based histogram building and peak detection. Web-WISE [12] is designed for content based seeking and retrieval of images on the web. More information can be found at http://www.cs.wayne.edu/ ilc/vision/wise.html.

- Surfimage : Surfimage [8] is a prototype software for the IMEDIA (Image and multimedia indexing, browsing and retrieval) project, developed at INRIA. It uses the query-by-example approach for retrieving images and integrates advanced features such as image signature combination, classification, multiple queries and query refinement with relevance feedback. The on-line demo can be found at http://www-rocq.inria.fr/cgi-bin/imedia/surfimage.cgi.

- PicToSeek : PicToSeek [13] is an image retrieval system for the web, developed at ISIS Research Group at University of Amsterdam. The basic idea is to extract invariant features (independent of the imaging conditions) from each of the images in the database, which are subsequently matched with the invariant feature set derived from the query image. The on-line demo can be found at http://www.science.uva.nl/research/isis/zomax/.

- Blobworld : Blobworld [1], which is developed at UC Berkeley, is a system for image retrieval, based on coherent image regions which roughly correspond to objects. Each image is automatically segmented into regions (blobs) with associated color and texture descriptors. Query is based on the attributes of one or two regions of interest, rather than a description of the entire image. The on-line demo can be found at http://elib.cs.berkeley.edu/vision.html. There are many other image retrieval systems in the literature including Virage [6], VisualSEEK [7] and WebSEEK [7] , and WebSeer [9].

III. Image Database and Interactive Image Parsing

Building an image dataset with manually annotated parse graphs provides training examples needed for learning the categorical image representation in the AoG model. Properly annotated datasets also provide training examples needed for learning semantic relationships. This dataset must be large-scale in order to provide enough instances to cover possible variations of objects. Functional relationships such as "carry" and "eat" are also specified manually. To cope with the need of labeling tens of thousands of images, an interactive image parsing software, named Interactive ImageParser (IIP), was developed to facilitate the manual annotation task. As stated in a report [2], this dataset provides ground truth annotation for a range of vision tasks from high level scene classification and object segmentation to low level edge detection and edge attributes annotation. Comparing with other public datasets collected in various groups, such as MIT LabelMe [6], ImageNet [3], the MSRC dataset [4], Caltech 101 and 256 [5], [6] and Berkeley segmentation [7], [8], the LHI dataset not only provides finer segmentation but also provides extra information such as compositional hierarchies and functional relationships.

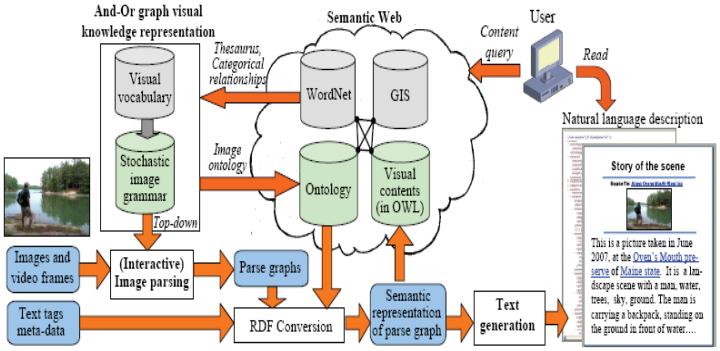

Figure 2: Image Parsing with database

IV. Image Parsing Methods

- **Segmentation by Edge Detection:** The edge-based methods make use of various edge operators to produce an "edginess" value at each pixel. The values are then threshold to obtain the edges. The regions within connected edges can be considered as different segments because they lack continuity with adjacent regions. The Sobel operator was studied and implemented to find edges in images. The edges thus found could also be used as aids by other image segmentation algorithms for refinement of segmentation results.
- In simple terms, the operator calculates the gradient of the image intensity at each point, giving the direction of the largest possible increase from light to dark and the rate of change in that direction. The result therefore shows how "abruptly" or "smoothly" the image changes at that point, and therefore how likely it is that that part of the image represents an edge, as well as how that edge is likely to be oriented.
- In theory at least, the operator consists of a pair of 3 × 3 convolution masks. This is very similar to the Roberts Cross operator

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Figure 3 : Sobel convolution masks.

- These masks are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid, one mask for each of the two perpendicular orientations. The masks can be applied separately to the input image, to produce separate measurements of the gradient component in each orientation (call these G_x and G_y). These can then be combined together to find the absolute magnitude of the gradient at each point and the orientation of that gradient. The gradient magnitude is given by:

Figure 4: Image Segmented by Sobel Operator

Segmentation by Grouping: Image segmentation can be related to perceptual grouping and organization in vision and several key factors, such as similarity, proximity, and good continuation, lead to visual grouping [1]. However, many of the computational issues of perceptual grouping have remained unresolved. In this report, a graph theoretic approach to this problem is adopted, focusing specifically on the case of image segmentation. Since there are many possible partitions of an image into subsets, how do we pick the "right" one? This is illustrated in Figure 5 where there are multiple groupings possible. There are two aspects to be considered here. The first is that there may not be a single correct answer. A Bayesian view is appropriate - there are several possible interpretations in the context of prior world knowledge. The difficulty, of course, is in specifying the prior world knowledge. Some of it is in the form of local properties, such as coherence of brightness, color, texture, or motion, but equally important are global properties about symmetries of objects or object models. The second aspect is that the partitioning is inherently hierarchical. Therefore, it is more appropriate to think of returning a tree structure corresponding to a hierarchical partition instead of a single "flat" partition.

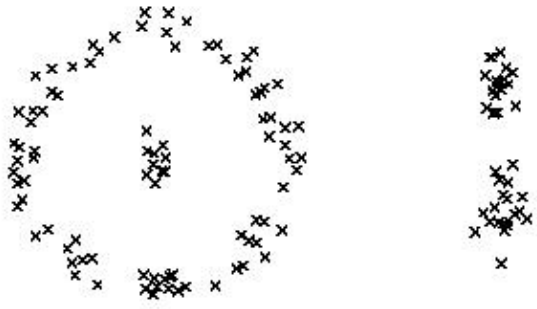

Figure 5 Points in a plane - what is the "correct" grouping?

V. Conclusion & Future Work

On a large test set, the method predicts numerous words with high accuracy. Simple methods are proposed to identify words that are not predicted well and the system is retrained on a reduced vocabulary consisting of words with better prediction rates. Individual words are grouped into word clusters to improve the performance for the words that cannot be distinguished using the current set of features.

Therefore, it leaves a number of issues open-ended for future research;

- The segmented regions of the images are represented by a set of simple basic features. We make no claim that the image features adopted are canonical. They are chosen to be computable for any image region, and be independent of any recognition hypothesis. Construction of a feature set that can over a better performance for the proposed set remains an open question.

- The feature vectors of the regions are clustered using the k-means algorithm, where number of classes is set to a predefined value. It is likely that better clustering can improve the performance of the system, and hence needs to be investigated.

- The proposed system can be a useful tool in evaluating segmentation and feature extraction algorithms on the large-scale. In [10], a number of features and segmentation algorithms are compared based on their word prediction performances.

- The annotations in the Corel data set are relatively simple in the sense that they consist of individual keywords and the vocabulary is relatively small. Many data sets, mentioned above, contain free text annotations. In such data sets, natural language processing is required to identify candidate annotations that appear to refer to the picture.

References:

[1] Calora. http://www.calora.org.

[2] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In Int. Conf. on Computer Vision, pages 408{15, 2001.

[3] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report ICSI-TR-97-021, University of Berkeley, 1997.

[4] Hulton Getty Archive. http://search.hultongetty.com/.

[5] Informedia Project. http://informedia.cs.cmu.edu.

[6] TV archive. http://televisionarchive.org.

[7] Web archive. http://www.archive.org.

[8] Yahoo News. http://news.yahoo.com.

[9] L.-M. Albiges. Remote public access to picture databanks. Audiovisual Librarian, 18(1):22{27, 1992}

[10] L.H. Armitage and P.G.B. Enser. Analysis of user need in image archives. Journal of Information Science, 23(4):287{299, 1997}.

[11] K. Barnard, P. Duygulu, N. de Freitas, D. A. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. Journal of Machine Learning Research, 3:1107{ 1135, 2003}

[12] K. Barnard, P. Duygulu, and D. A. Forsyth. Clustering art. In IEEE Conf. on Computer Vision and Pattern Recognition, volume 2, pages 434{441, 2001}

[13] K. Barnard, P. Duygulu, and D. A. Forsyth. Modeling the statistics of image features and associated text. In Document Recognition and Retrieval IX, Electronic Imaging, 2002.

[14] K. Barnard, P. Duygulu, and D. A. Forsyth. Recognition as translating images into text. In Internet Imaging IX, Electronic Imaging, 2003.

[15] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. A. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. In IEEE Conf. on Computer Vision and Pattern Recognition, 2003.

