Integrating Computer Vision in Web Based Context Recognition

Aviv Segev
Technion - Israel Institute of Technology
Haifa 32000
Israel
asegev@technion.ac.il

Ilan Shimshoni
University of Haifa
Haifa 31905
Israel
ishimshoni@mis.haifa.ac.il

ABSTRACT

The analysis of documents necessitates context recognition for diverse purposes such as classification, performance analysis, and decision making. Traditional methods of context recognition have focused on the textual part of documents. Images, however, provide a rich source of information that can support the context recognition process. The paper proposes a method for integrating computer vision in context recognition using the Web as a knowledge base. The method is implemented on medical case studies to determine main symptoms or achieve possible diagnoses. Initial experiments show better results than corresponding text-only methods.

1. INTRODUCTION

The field of document analysis requires context recognition for classification, performance analysis, and decision making. Traditional methods of context recognition focus on the textual part of documents. However, images constitute a rich source of information that can complement the context recognition process and computer vision techniques can supply textual information about images for diverse uses.

Previous work explored the interaction of textual and photographic information in document understanding. The use of complementary information in scene understanding has been explored in computer vision systems that use scene context in the task of object identification. Work has been done on extracting picture-specific information from text accompanying a photograph [10].

Another application is a model of object recognition as machine translation described in [5]. In this model, recognition is a process of annotating image regions with words. First, images are segmented into regions, which are classified into region types using a variety of features. Then, a mapping between region types and keywords supplied with the images is learned.

Word sense ambiguity problems are addressed in [2]. The approach bases on a method for automatically annotating images by using a statistical model for the joint probability for image regions and words. The model is learned from a database of images with associated text. To use the model for word sense disambiguation, the predicted words are constrained to be possible senses for the word considered.

The problem of modeling annotated data, data with multiple types such as image and text captions, is addressed in [4]. The work describes three hierarchical probabilistic mixture models that aim to describe such data, culminating in correspondence latent Dirichlet allocation, a latent variable model effective at modeling the joint distribution of both types and the conditional distribution of the annotation given the primary type.

Similarly, computer vision can provide information for integration with other systems. The integration of speech and image using Bayesian networks is found in [11]. Object recognition errors are taken into account by conditional probabilities estimated on test sets. The Bayesian network is dynamically built up from verbal object description and is evaluated by an inference technique combining bucket elimination and conditioning.

In this paper, we investigate a method for integrating context recognition and computer vision using the Web. The proposed method improves context recognition, based on both non-structured textual analysis and semi-structured computer vision data. The proposed model integrating computer vision and context recognition takes a document from the Web containing text and pictures and returns a set of words representing the document context.

The paper employs a model of context recognition using the Web as a knowledge base, thus giving a context recognition model immediate access to a nearly infinite amount of data in a multiplicity of fields [8]. Context is represented as a set of descriptors and a set of weights to describe a given situation. The model does not require large training sets and since the Web is used as the database, there is no database maintenance.

The proposed model was tested in the field of medicine. The model input consists of medical case studies taken from the Clinico-Pathological Conference of the Johns Hopkins University School of Medicine. These case studies, consisting of text and images, present the clinical course, radiological findings, and relevant laboratory results for a particular patient and present the medical analysis [1]. The model output is a list of words that represent major symptoms or possible diagnoses and these words are checked against the solutions in the medical case studies.

2. MODEL AND PROCESSING

This section presents the proposed Web-based model for the integration of context recognition and computer vision. First, the two main components are presented, namely the context recognition model for the detection of possible topics and the computer vision model for the definition of images. Then the integration of the two is described: the image related information is compared with the generated contexts

WWW2006, May 22–26, 2006, Edinburgh, UK.
.

and then with the text of the document for further verification.

2.1 The Context Recognition Model

Several methods have been proposed in the literature for extracting context from text. A set of algorithms was proposed in the IR community, based on the principle of counting the number of appearances of each word in the text, assuming that words with the highest number of appearances serve as the context. Variations on this simple mechanism involve methods for identifying the relevance of words to a domain, using methods such as stop-lists and inverse document frequency.

The context recognition model we use is based on the definition of context as first class objects formulated by McCarthy [7]. McCarthy defines a relation $ist(\mathcal{C}, P)$, asserting that a proposition P is true in a context $\mathcal{C}$. We use this relation when discussing context extraction.

A *context* $\mathcal{C} = \left\{\{\langle c_{ij}, w_{ij}\rangle\}_j\right\}_i$ is a set of finite set of descriptors c_{ij} from a domain $\mathcal{D}$ with appropriate weights w_{ij} that define the importance of c_{ij}. For example, a context $\mathcal{C}$ may be a set of words (and hence, $\mathcal{D}$ is a set of all possible character combinations) defining a document *Doc*, and the weights could then represent the relevance of a descriptor to *Doc*.

Let $P_1, , P_m$ be a series of textual propositions representing a document, when $\forall P_i$ there exists a collection of sets of contexts C_{ij} so that: For each i, $ist(C_{ij}, P_i)\forall j$ meaning that the textual proposition P_i is true in each of the set of contexts C_{ij}. The context recognition algorithm [8] identifies the outer context set C defined by

$$ist(\mathcal{C}, \bigcap_{i=1}^{m} ist(\mathcal{C}_{ij}, P_i))\forall j.$$

The input to the algorithm is a stream, in text format, of information. The context recognition algorithm output is a set of contexts that attempts to describe the current scenario most accurately. The set of contexts is a list of words or phrases, each describing an aspect of the scenario. The algorithm attempts to reach results similar to those achieved by the human process of determining the set of contexts that describe the current scenario.

The context recognition algorithm consists of the following major phases: collecting data, selecting contexts for each text, ranking the contexts, identifying the current contexts, and obtaining the multiple contexts.

- Collecting Data - The information from the information sources is decomposed into words and the keywords are extracted from them.
- Selecting Contexts for Each Text (Descriptors) - For each keyword a set of preliminary contexts is extracted from the Web, which is used as a context database.
- Ranking the Contexts - Each preliminary context is ranked according to the number of references it receives in the context database and the number of appearances it has in the text.
- Identifying the Current Contexts - The preliminary contexts that have significantly higher numbers of references and higher numbers of appearances are included in the current set of contexts.
- Obtaining the Multiple Contexts - The current contexts are examined for synonyms and synonymous contexts are united.

The success of the algorithm depends, to a great extent, on the number of documents retrieved from the Web. With more relevant documents, less preprocessing (using methods such as Natural Language Processing) is needed in the data collection phase.

2.2 The Computer Vision Model

The field of computer vision explores automatic pattern or object recognition. Methods are developed to discover which features distinguish objects and to design classification algorithms. The concepts of feature extraction that respond to characteristic changes in brightness, color, and texture associated with natural boundaries, used in the computer vision method integrated in the present model, are based on the concepts of Martin et al. [6].

Computer vision models usually compute a set of real-valued features that represent visual properties such as size, position, color, and texture. For example, consider in the field of medicine a simple computer vision model that analyzes four types of images: blood, brain, chest, and liver.

- Blood is recognized by parsing the image into three color based segments and checking whether one is similar to blood color. This yields perfect recognition of blood images.
- The brain is identified in two steps. First, the brain is recognized by checking roundness of the biggest object in the image. Then, the shadowed image of the main object in the picture is covered by a brain shadow and the difference is examined. The brain recognition method yields 38% accuracy.
- The method of the identification of the chest and liver is based on locating the bones, which are the lightest objects in the picture. The biggest white object is checked to see whether it matches the location (bottom center) and is proportional in size to the image of a vertebra.
- Once the vertebra is located, the method searches for the front rib bones. If the front rib bones are located, then the image is identified as chest. Otherwise, it is identified as liver.

The result of the procedure below is the identification of one of the above body parts.

Algorithm 2.1: SIMPLE VISION MODEL($Image$)

if color image { **if** contains blood color
else { check **if** brain
else if vertebra { **else if** front ribs
then identify as chest
else identify as liver

2.3 Integrating the Vision into the Context

The context recognition model and the computer vision model are integrated according to the following algorithm.

- Each possible image X_i has prior knowledge about identification failure rate (X_{i1}, X_{i2}) according to previous domain knowledge. X_{i1} defines the false negative: the algorithm fails to identify the image. X_{i2} defines the false positive: the algorithm erroneously identifies an image.
- For each domain for all contexts prior knowledge about identification failure rate (Y_1, Y_2) exists and is similarly defined.
- For each context recognition result (CRR) compare to computer vision result (CVR). The algorithm searches for the possible results of the integration in the text. Once the possible contexts are defined, they can be checked and verified against a local domain knowledge base. This was performed in the examination of medical case studies.

Algorithm 2.2: INTEGRATION ALGORITHM(CRR, CVR)

if CVR = CRR
{ Verify against domain information.
Add context with probability
0.5(1-X_{i1}) (1-Y_1)+ 0.5(1-X_{i2}) (1-Y_2)
else
{ search text (T) for CVR or dictionary synonym (DS)
if T = CVR or DS = CVR
{ Verify against domain knowledge and
rank CVR as context with probability
0.5(1-X_{i1})Y_1+ 0.5X_{i2}(1-Y_2)
else { add CVR as context with probability
0.5X_{i1}(1-Y_1) + 0.5(1-X_{i2})Y_2

3. MEDICAL CASE STUDY

Medical case studies are used to test the algorithm performance. Domain information is obtained from the Merck Manual [3], which provides medical information lists for the classification of possible diseases and diagnostic procedures using imagery techniques such as X-ray, CT, Radionactive imaging, and microscopic examination. This information was organized to correspond to the present method of image and text integration. A sample of possible diagnoses based on the image and textual information is displayed in Table 1.

The textual part of the medical records is fed into the context recognition model and processed. A list of possible contexts, which can include possible diagnoses and main symptoms, is obtained. The images from the case are analyzed by the computer vision model. Next, according to the algorithm, the results of the context recognition and of the computer vision are compared. If the results match, then they are checked against the medical information lists constructed from the Merck Manual. Then, according to the medical information presented in the table, the results represent main symptoms or possible diagnoses. If the results of the computer vision and the context recognition do not match, then a search is performed on the case study text for words that match the results of the computer vision. If no match is found, then the results of the two processes are weighted.

Table 1: Possible Diagnosis Based on Integration of Image and Text

Image Type	Body Area Tested	Procedure	Possible Diagnosis
X-ray	Any artery in the body; commonly in brain, heart, kidneys, aorta or legs	Arteriography (angiography)	blockage or defect of artery
X-ray	Liver	Percutaneous transhepatic cholangiography	Tumor
X-ray	Veins	Venography	Blockage of Vein
Ultrasound	Liver	Ultrasonography (ultrasound scanning)	Tumor, Jaundice
Positron emission tomography (PET)	Brain	Radioactive imaging to detect abnormality of function	Epilepsy, brain tumor, stroke
Radionactive Radionuclide imaging	Many organs	Radioactive imaging to detect abnormality of blood flow, structure, or function	Heart attack, coronary artery disease (CAD), valvular/ congenital disease, and other cardiac disorders

Table 2 presents the computer vision algorithm results according to the following breakdown.

The brain training was performed on 7 images extracted from the Web. The testing was performed on 13 images and their related text extracted from medical case studies. The identification of the brain related images had 38% success rate.

The chest training was performed on 8 images extracted from the Web. The testing was performed on 3 images and their related text extracted from medical case studies. The chest related images were not identified correctly. The context recognition algorithm also did not provide useful results, since it identified in these cases blood as a possible context. These results can be attributed either to problems with the chest image identification part of the computer vision algorithm or to difficulties with the context recognition algorithm implementation, since blood was obtained as a

Table 2: Computer Vision Algorithm Results

Image Type	Training	Testing	Identification
Blood	12	8	100%
Brain	7	13	38%
Chest	8	3	-
Liver	10	-	-

context because of the patients' medical testing.

The liver training was performed on 10 images, but testing was not performed on the liver. It was used to define another category.

The blood color training was performed on 12 images selected from sample blood images taken from the Web. The result of the training was a color range of the blood that appears in 12 images. To analyze the algorithm performance, initial testing for blood images was performed on 10 images. 8 images were extracted from medical case studies and another 2 were images used for control. 5 of the images were blood related. The algorithm identified correctly all 5 of the blood related images.

In one sample test case, according to the integrated medical information displayed in Table 1, the results yielded by the algorithm following the identification of blood in one sample test case include the possible diagnoses of: `blockage or defect of artery`, `heart attack, coronary artery disease (CAD)`, `valvular or congenital disease`, and `other cardiac disorders`.

Then the algorithm again searches in the text for possible contexts from the list of diagnoses in the table. The word `heart` appears since the patient had heart failure and previous heart attack. Similarly, the patient has suffered from `coronary artery disease (CAD)`. The word `cardiac` appears three times in the text.

The medical case study states that the patient expired from Coronary Allograft Vasculopathy (Accelerated Graft Arteriosclerosis), which resulted in cardiogenic shock. The algorithm yielded heart and artery related results. Therefore, the integration of the images and the web based contexts can assist in the analysis of case studies.

In our initial series of tests, a total of 20 medical case studies were examined by the integration algorithm. The Web based context recognition achieved a success rate of 40% in identifying the correct diagnosis or in supplying information about the patient medical symptom. Overall, the algorithm for the integration of context recognition and computer vision improved the results by 15%, in comparison to the Web based context recognition algorithm alone [9].

4. CONCLUSIONS AND FURTHER RESEARCH

The paper presents a Web-based technique of integrating context recognition and computer vision and demonstrates how this method can be implemented. The paper uses real medical case studies to experiment with the proposed method.

Usually document analysis focuses on the text part of a document, but we propose an idea of text understanding by understanding image first, since image can constitute a rich source of information. This idea is based on the assumption that the accuracy of computer vision is high enough to provide a useful hint for context recognition, since an inaccurate computer vision system might also mislead the overall context recognition.

Initial findings show that the proposed integration method yields improved results in comparison to the separate use of context recognition or computer vision. Additionally, use of state-of-the-art as opposed to simple computer vision algorithms can improve the results. Full scale experiments are currently underway.

The main advantage of the proposed model for the integration of computer vision into context recognition is its use of the Web as a knowledge base for data extraction. A further advantage of the use of the Web is that it minimizes model training and maintenance. Additionally, the information provided by the computer vision model complements and augments the context recognition process.

Directions of further research include extending the present method to other fields of image and text such as newspaper articles and Web pages. The addition of other inputs, such as voice recognition, may provide further improvements.

5. REFERENCES

[1] Clinico-pathological conference. Johns Hopkins University School of Medicine, 2004. http://oac.med.jhmi.edu/cpc.

[2] K. Barnard and M. Johnson. Word sense disambiguation with pictures. *Artificial Intelligence*, 167:13–30, 2005.

[3] M. Beers, editor. *The Merck Manual of Medical Information*. Merck Research Laboratories, second edition, 2003.

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of SIGIR'03*, 2003.

[5] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. *Computer Vision - ECCV 2002: 7th European Conference on Computer Vision*, volume 2353, chapter Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, pages 97–112. Springer-Verlag GmbH, 2002.

[6] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color and texture cues. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(5):530–549, May 2004.

[7] J. McCarthy. Notes on formalizing context. *In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1993.

[8] A. Segev. Identifying the multiple contexts of a situation. In *Proceedings of IJCAI-Workshop Modeling and Retrieval of Context (MRC2005)*, 2005.

[9] A. Segev, M. Leshno, and M. Zviran. Internet as a knowledge base for medical diagnostic assitance. *Expert Systems With Applications*, 33(1), 2007. to appear.

[10] R. Srihari and D. Burhans. Visual semantics: Extracting visual information from text accompanying pictures. In *National Conference on Artificial Intelligence (AAAI-94)*, pages 793–798, 1994.

[11] S. Wachsmuth and G. Sagerer. Bayesian networks for speech and image integration. In *National Conference on Artificial Intelligence (AAAI-02)*, pages 300–306, 2002.