PICTION: A System that Uses Captions to Label Human Faces in Newspaper Photographs*

Rohini K. Srihari
Department of Computer Science
State University of New York at Buffalo
Buffalo, New York 14260 USA
e-mail: rohini@cs.buffalo.edu

Abstract

It is often the case that linguistic and pictorial information are jointly provided to communicate information. In situations where the text describes salient aspects of the picture, it is possible to use the text to direct the interpretation (i.e., labelling objects) in the accompanying picture. This paper focuses on the implementation of a multi-stage system *PICTION* that uses captions to identify humans in an accompanying photograph. This provides a computationally less expensive alternative to traditional methods of face recognition. It does not require a pre-stored database of face models for all people to be identified. A key component of the system is the utilisation of spatial constraints (derived from the caption)in order to reduce the number of possible labels that could be associated with face candidates (generated by a face locator). A rule-based system is used to further reduce this number and arrive at a unique labelling. The rules employ spatial heuristics as well as distinguishing characteristics of faces (e.g., male versus female). The system is noteworthy since a broad range of AI techniques are brought to bear (ranging from natural-language parsing to constraint satisfaction and computer vision).

Introduction

The idea of integrating natural language and vision has been relatively unexplored. Yet there are frequent situations where text and pictures are jointly presented to communicate information; we shall refer to these as *communicative units*. In such situations, it is possible to use the information contained in the text (e.g., spatial constraints) to guide the interpretation (i.e., labelling objects) of the accompanying picture. This helps to overcome many of the problems associated with general-purpose vision. If the ultimate goal is to develop a natural-language system that can visualise the world that it deals with (in either a discourse or narrative domain), real examples are required on which the system's performance can be tested. Newspaper photographs with captions provide such an example.

This paper discusses the design and implementation of a system called *PICTION* [Srihari and Rapaport, 1989], that identifies human faces in newspaper photographs based on information contained in the associated caption. Most front-page newspaper photographs tend to have captions that are factual and descriptive, qualities required for this task. The system is based on a newly developed theory of extracting visual information from text [Srihari and Rapaport, 1989], that is, information useful in identifying objects in an accompanying picture. The focus of this paper however, is on the implementation of the multi-stage system *PICTION*. It is noteworthy since it provides a computationally less expensive alternative to traditional methods of face recognition in situations where pictures are accompanied by descriptive text. Traditional methods such as that presented in [Weiss *et al.*, 1986] employ model-matching techniques and thus require that face models be present for all people to be identified by the system; our system does not require this. It will be shown that spatial constraints obtained from the caption, along with a few discriminating visual characteristics (e.g., male/female) are sufficient for eliminating false candidates and correctly identifying faces.

To date there has been little work in the area of using linguistic information to guide the interpretation of an accompanying picture. Most work in this area relates to diagram understanding, a task which differs from the present one since (i) line-drawings, rather than images are involved and (ii) the text usually consists of single words or short phrases rather than complete sentences. [Jackendoff, 1987] addresses the general problem of establishing correspondence between language and pictures. The research presented in [Zernik and Vivier, 1988] comes closest to that being described here; the authors describe a system that interprets locative expressions thus enabling a vision system to detect objects in airport scenes.

Figure 1 is an example of a digitised newspaper pho-

*This work was supported in part by grants from the National Science Foundation (NSF IRI-86-13361) and Eastman Kodak Company.

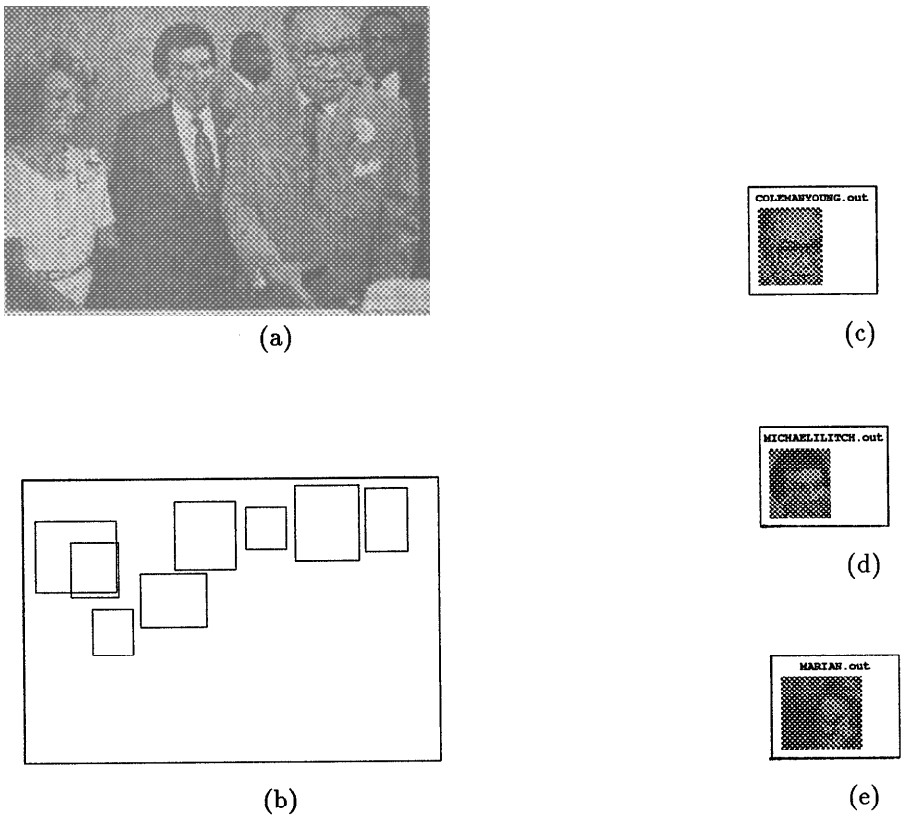


Figure 1: (a) Digitised photograph whose caption is "Mayor Coleman Young of Detroit, right, at an N.A.A.C.P. dinner last week. With the Mayor were Michael Illitch, owner of the Detroit Red Wings, and his wife, Marian" (*The New York Times*, May 1, 1989). (b) face candidates (c)-(e) output of *PICTION*.

tograph and accompanying caption[1] that the system can handle. The system attempts to isolate and label those parts of the image that correspond to the faces of Mayor Coleman, Michael Illitch and his wife Marian. It should be noted that in this example, the caption assumes that the reader can distinguish between Michael and his wife, and hence they are not identified explicitly. Furthermore, there are two other people in the picture who are not identified in the caption and it is a challenge for the system to select the correct three people.

System Overview

A multi-stage system (illustrated in 2) is employed in order to carry out the task of identification. The system consists of three main processing modules, namely the Vision module, the NLP (Natural Language Processing) module and the Interpretation module. The first step consists of parsing the caption and producing the *conceptualised graph* which reflects factual information contained in the caption as well as predictions regarding the composition of the picture. The interpretation module (i) calls on the vision module to first generate candidates for faces in the image (Figure 1b), (ii) uses the spatial constraints in the conceptualised graph to form initial hypotheses about name-face correspondences and (iii) invoke a rule-based system to narrow down the set of hypotheses based on visual characteristics of the people being identified. Eventually, the interpretation module generates a list of coordinate-name pairs, representing name-face correspondences. This information is consolidated into the original graph producing the consolidated graph. From this, the system is able to produce the desired pictorial output. Each of these stages are discussed in greater detail.

[1]To simplify the parsing process, the two sentences of the caption were combined into one (not affecting the meaning): "Mayor Coleman Young of Detroit, right, with Michael Illitch, owner of the Detroit Red Wings, and his wife Marian at a dinner last week".

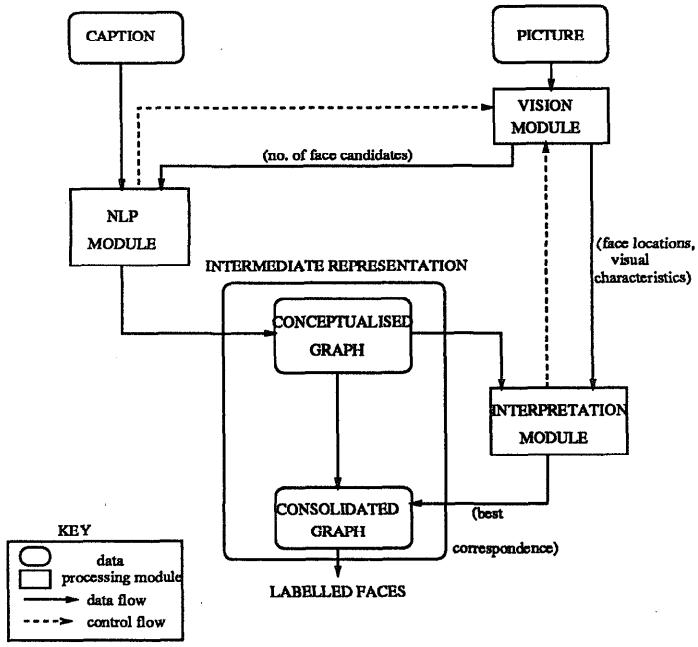


Figure 2: *PICTION*: System Overview

Processing the Caption (NLP Module)

The process of interpreting the caption has two main goals. The first is the representation of the factual information contained in the caption. This is explicit information provided by the caption, namely the identification of the people in the photograph and the context under which the photograph was taken. More important for our application, however, is the second goal, the construction of a conceptualised graph representing the expected structure of the picture. The conceptualised graph includes information such as the objects hypothesised to be in the picture, their physical appearance, and spatial relationships between them. We use the SNePS (Semantic Network Processing System) knowledge-representation and reasoning system to represent both factual information and the conceptualised graph derived from the caption [Shapiro and Rapaport, 1987]. SNePS is a fully intensional, propositional, semantic-network processing system in which every node represents a unique concept. It can perform node-based and path-based inference and it also provides a natural-language parsing and generating facility.

We are interested in two aspects of visual information provided by the caption. These are (i) information allowing an NLP system to predict which people mentioned in the caption are present in the picture, and (ii) information necessary for identifying these individuals. Captions frequently refer to people who are not present in the picture as in "Gov. Cuomo and his wife, Matilda, walk up the church steps to attend the marriage of their son Andrew to Kerry Kennedy, daughter of Ethel Kennedy and the late Robert F. Kennedy" (*The Buffalo News*, June 10, 1990). Sentence type plays an important role in determining which people are present in the picture. We have also observed that any person referred to using a time other than the event time (the time picture was taken) is not in the picture. This phenomena is marked linguistically through the introduction of relative clauses beginning with the words 'before', 'after', 'as', etc. We also stress the importance of correctly predicting the class of an object. This is illustrated by the caption "Marge Schott, ...gets a farewell hug from Buster T. Bison, the Buffalo Bisons' mascot ..." (*The Buffalo News*, May 15, 1990) accompanying a picture of a woman and a toy bison. A simplistic parser could mistake Buster T. Bison to be a human causing the face identification system to be led astray.

Specifying spatial relations between objects is the principal method of identification. The caption often explicitly specifies the spatial relation, as in "Thomas Walker, left, John Roberts, center ..." thus making the task relatively simple. A more complex form of identification involves the use of visual detail (e.g., "Tom Smith, wearing a top hat ..."). A subtle version of this type of identification can be observed in captions that require the system to distinguish between one or more people based on gender or class information (e.g. "Amy Jones, 9, poses with a clown at the circus"). It is assumed that the reader can distinguish between the clown and Amy Jones.

Many captions combine implicit and explicit means

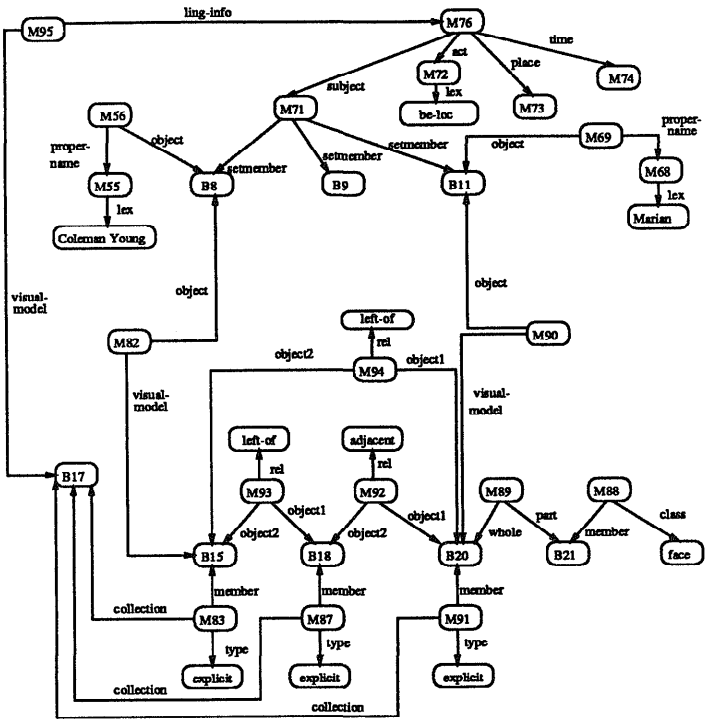


Figure 3: Partial output of the parser on caption of Figure 2

of identification such as that in Figure 1. The system is able to handle such cases correctly. Figure 3 illustrates a portion of the SNePS network resulting from the processing of the caption in Figure 1 upto and including this stage. In Figure 3, node M95 serves as a link between the visual-model describing the composition of the picture (node B17), and factual information contained in the caption (node M76). Considering the linguistic information first, node M76 asserts that the people in the set denoted by node M71, (comprised of nodes B8, B9 and B11), are present at a certain place (node M73) and time (node M74). Furthermore, node B8 represents the concept of the person named Coleman Young, node B9 represents Michael Illitch (not illustrated), and node B11 represents Michael's wife Marian.

Turning to the portion of the SNePS network representing derived visual information, nodes B15, B18 and B20 are the individual visual models for Coleman Young, Marian and Michael Illitch respectively. Nodes M83, M87 and M91 assert the presence of these objects in the visual-model for the entire picture (node B17). Nodes M92, M93 and M94 assert spatial relations between the visual-models for the 3 people. Nodes M93 and M94 specify that Marian and Michael Illitch are to the left of Coleman Young. Node M92 asserts that Marian and Michael are adjacent to each other. The latter is a weak spatial constraint but is the only one that can be inferred since we cannot assume that the order of mention in the caption reflects the left-to-right ordering in pictures (where there is a combination of a male and female).

Vision Module

The vision module in this system performs two main functions. The first is the task of locating candidates for faces, and the second is the characterisation of faces. From the caption, we are able to determine the number of faces and some weak bounds on the size of faces. These constitute parameters to the face-location module [Govindaraju *et al.*, 1989]. Since the face-locator is still under development at this stage, the output is simulated based on a thinned-edge image. For each image area hypothesised to be a face, this module returns the coordinates of a bounding rectangle for that area. The simulated output exhibits realistic problems since false candidates are generated due to incidental allignment of edges and true candidates are occasionally missed due to poor edge data. Simulated output on the image of Figure 1(a) is shown in Figure 1(b). The spurious candidates (c4 and c8), are due to incidental alignment of edges, in the desired pattern, along the sleeve lengths of Marian and Michael Illitch.

In general, the vision module is required whenever it becomes necessary to examine in detail the portion of the original image (i.e., the photograph) corresponding to a box representing a face candidate. The vision module is expected to make some qualitative judgement about this area. The process used to make each of these judgements will be referred to as a "filter". Specifically, the vision module is called on to verify visual characteristics that may be useful in the identification process. The visual characteristics that are being used currently are male versus female, baldness and colour (black/white).

It is these filters that enabled the successful processing of the example presented in Figure 4. The system identifies and labels Bush and Ford- even though the caption might lead a machine to believe that Ford was on the left. The system is told in advance that Ford is bald. It applies the "baldness" filter to all face candidates and is able to select the one which best matches. This filter is based upon finding minima (of a specified depth, width and position) in vertical profiles (of the top one-third rows only) of face candidates. The male/female filter was applied successfully in the picture of Figure 1 in order to distinguish between Michael Illitch and Marian.

Interpretation Module

An interpretation strategy for the picture is a systematic method of using information in the hypothesised structure to find relevant objects and spatial relationships in the picture. At this point, the interpretation strategy is implicitly embedded in the main control structure of the system and consists of three steps: (i) face location, (ii) constraint satisfaction and (iii) rule-based identification. Each of the above three steps necessitates the use of the vision module. We have already discussed the function of the face locator. The final output of the rule-based identification system is incorporated into the intermedidate representation (the SNePS network), thus completing the face-identification task.

Constraint Satisfaction In general, the location procedure generates more candidates than required (Figure 1b). Spatial constraints (obtained from the caption) are applied to the candidates generated by the face-locator to produce all possible *bindings*. A labeling algorithm [Haralick and Shapiro, 1979] is employed which uses a look-ahead operator in order to eliminate backtracking in the treesearch. A "binding" represents the assignment of correspondence between face candidates and people predicted to be in the picture.

Rule-Based Face Identification Because a large number of candidates are generated by the face locator, spatial constraints alone cannot produce a unique binding between candidates and people mentioned in the caption. Furthermore, a spatial constraint such as "adjacent" (used frequently in male-female pairs) will produce at least two possibilities. The rule-based face identification module evaluates the bindings produced by the constraint satisfier and selects the best one(s).

We refer to each of the tuples in a binding as a "match". The refinement and identification rules fall into three categories: (i) those that update the confidence of a candidate being a face (irrespective of which person that face is associated with), (ii) those that update the confidence of a "match" , i.e. the confidence that a particular face candidate corresponds to a named person and (iii) those that update the confidence of an entire binding. An example of a candidate rule is one that examines the centrality of the face candidate in the image. Most of the match rules involve checking for distinguishing visual characteristics of a face. For example, if the match associates a face candidate with a person who is known to be bald, it invokes the "baldness" filter on this face candidate in order to verify the match. An example of a binding rule is one which examines height relationships between people in the same row. It favours those bindings where the vertical positions of faces do not differ significantly.

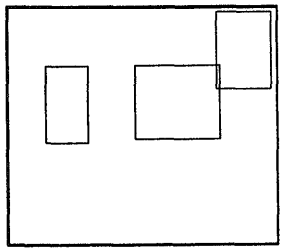
Currently, we are using a hierarchical Bayesian updating scheme to combine evidence from rules supporting the same hypothesis. Assuming that we are able to generate the three types of confidences described above, the weight of a match, $weight_{match_i}$ is computed as $confidence_{c_{x\,i}} * confidence_{match_i}$. We assign a weight W to every binding B, such that $W = [\prod_{i=1}^{k} weight_{match_i}] * confidence_B$ where $confidence_B$ is the confidence associated with the binding itself. If the weight of bindings differ by less than a preset threshold, *thresh*, then they are considered equivalent. This creates multiple correspondences for some or all of the faces. Based on experimental results, a value of *thresh* optimising both accuracy and uniqueness was selected. In cases where the system cannot uniquely identify faces, all possible candidates for each person appearing in the caption are recorded.

Testing System and Evaluating Results

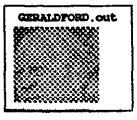
The system was trained on a dataset of 50 pictures and captions obtained from *The Buffalo News* and The New York Times. There are three success codes used to evaluate the system: (a) SU (success), indicating that everyone identified in the caption is identified by the system correctly and uniquely; (b) PS (partial success), indicating multiple possibilities for one or more people where the actual face is included and (c) E (error), indicating that one or more people was identified incorrectly (i.e., true face not included). An overall success rate (i.e. SU only) of 65 percent was obtained. The most common reason for a result of PS or E was the failure of the face locator to locate one or more of the identified faces. In only one case, the error was due to an incorrect parsing. Other reasons for a result code of PS or E included (i) the failure of spatial

Figure 4

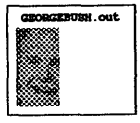
4(a)



4(b)



4(c)



4(d)

Figure 4: (a) photograph with caption "Former President Gerald Ford joins George Bush at rally in Royal Oak Michigan (*The Buffalo News*, Oct. 19, 1988). (b) output of face-locator (c,d) output of *PICTION*

heuristics (ii) inability to properly characterise faces (e.g., male/female, young/old).

Summary

We have presented a new method of face recognition in situations where pictures are accompanied by descriptive text. It is based on a new theory addressing the issue of extracting visual information from text. A multi-stage system *PICTION* has been described, which uses an intermediate representation referred to as a conceptualised graph in order to consolidate information (i) obtained from parsing the caption and (ii) information obtained from picture processing. The system is noteworthy since it does not require a prestored database of face models for all faces that are to be recognised. The results based on a limited testing of the system, illustrate the viability of this approach as an alternative to traditional methods of face-recognition. Future work includes the enhancement of filters that characterise faces, the ability to process more complex type of visual information and finally, extensive testing of the system.

References

Govindaraju, Venu; Sher, David B.; Srihari, Rohini K.; and Srihari, Sargur N. 1989. Locating human faces in newspaper photographs. In *Proceedings of CVPR.* 549–554.

Haralick, Robert M. and Shapiro, Linda G. 1979. The Consistent Labeling Problem: Part 1. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1(2):173–184.

Jackendoff, Ray 1987. On Beyond Zebra: The Relation of Linguistic and Visual Information. *Cognition* 26(2):89–114.

Shapiro, Stuart C. and Rapaport, William J. 1987. SNePS Considered as a Fully Intensional Propositional Semantic Network. In Cercone, Nick and McCalla, Gordon, editors 1987, *The Knowledge Frontier, Essays in the Representation of Knowledge.* Springer-Verlag, New York. 262–315.

Srihari, Rohini K. and Rapaport, William J. 1989. Integrating Linguistic and Pictorial Information: Using Captions to Label Human Faces in Newspaper Photographs. In *Proceedings of the 11th Annual Conference of the Cognitive Society.* Lawrence Erlbaum Associates. 364–371.

Weiss, Richard; Kitchen, Leslie; and Tuttle, Julianne 1986. Identification of Human Faces Using Data-driven Segmentation, Rule-based Hypothesis Formation and Iterative Model-based Hypothesis Verification. COINS Technical Report 86-53, University of Mass. at Amherst.

Zernik, Uri and Vivier, Barbara J. 1988. How Near Is Too Far? Talking about Visual Images. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum Associates. 202–208.