Extracting Visual Information From Text: Using Captions to Label Human Faces in Newspaper Photographs

Rohini K. Srihari and William J. Rapaport

Department of Computer Science
State University of New York at Buffalo
Buffalo, New York 14260 USA

ABSTRACT

There are many situations where linguistic and pictorial data are jointly presented to communicate information. A computer model for synthesising information from the two sources requires an initial interpretation of both the text and the picture followed by consolidation of information. The problem of performing general-purpose vision (without apriori knowledge) would make this a nearly impossible task. However, in some situations, the text describes salient aspects of the picture. In such situations, it is possible to extract visual information from the text, resulting in a relational graph describing the structure of the accompanying picture. This graph can then be used by a computer vision system to guide the interpretation of the picture. This paper discusses an application whereby information obtained from parsing a caption of a newspaper photograph is used to identify human faces in the photograph. Heuristics are described for extracting information from the caption which contributes to the hypothesised structure of the picture. The top-down processing of the image using this information is discussed.

INTRODUCTION

There are many situations where words and pictures are combined to form a communicative unit; examples in the print media include pictures with captions, annotated diagrams, and weather charts. In order for a computer system to synthesise the information from these two diverse sources of information, it is necessary to perform the preliminary operations of natural-language processing of the text and image interpretation of the associated picture. This would result in an initial interpretation of the text and image, following which an attempt at consolidation of the information could be made. Although vision and natural-language processing are challenging tasks, since they are severely under-constrained, natural-language processing can more easily exploit constraints posed by the syntax of the language than vision systems can exploit constraints about the physical world. This fact, combined with the observation that the text often describes salient features of the accompanying picture in joint communicative units, leads to the idea of using the information contained in the text as a guide to interpreting the picture. This paper focuses on a method of extracting visual information from text, which results in a relational graph describing the hypothesised structure of the accompanying picture (in terms of the objects present and their spatial relationships). The relational graph is subsequently used by a vision system to guide the interpretation of the picture. We describe the implementation of a system which labels human faces in a newspaper photograph, based on information obtained from parsing the caption. A common representation, namely a semantic network, is used for the knowledge contained in both the picture

and the caption. The theory is general enough to permit construction of a picture when given arbitrary descriptive text (without an accompanying picture).

Newspaper photographs have all the elements required for a true integration of linguistic and visual information. Accompanying captions usually identify objects and provide background information which the photograph alone cannot. Photographs, on the other hand, provide visual detail which the captions do not. Newspaper captions often identify people in a picture through visual detail such as "Tom Jones, wearing sunglasses ...". In order for a computer system to be able to identify Tom Jones, it is necessary to understand the visual implication of the phrase "wearing sunglasses". The face satisfying all the implied visual constraints could then be labeled accordingly.

The idea of integrating natural language and vision has been relatively unexplored. [Abe *et al.*, 1981; Yokota *et al.*, 1984] are two systems which consider the bidirectional flow of control from text to a related picture and vice versa. Several systems have implemented a portion of the task. Generating natural-language descriptions of results obtained from a vision system is considered in [Maddox and Pustejovsky, 1987; Neumann and Novak, 1983]. The reverse process of generating pictures based on natural-language input is considered in [Adorni *et al.*, 1984; Waltz and Boggess, 1979]. [Herskovits, 1986] discusses a theory of encoding and decoding English expressions of location, which focuses on the meaning of prepositional phrases. In the research being presented here, the emphasis is on generating a description of a picture (rather than a picture itself), such that the description can be used by a vision system to actually find the required objects and relationships in an associated picture. [Zernik and Vivier, 1988] attempts a similar task when given locative expressions pertaining to airport scenes.

This paper describes a three-stage process used to identify human faces in newspaper photographs. We consider only those photographs whose captions are factual but sometimes incomplete in their description of the photograph. In the first stage, information pertaining to the story is extracted from the caption, and a structure of the picture in terms of the objects present and spatial relationships between them is predicted. The information contained in this structure would be sufficient for generating a picture representing the meaning of the caption. Using this information to label faces in an existing picture however, entails further processing. The second stage which constitutes the vision component, calls on a procedure to locate human faces in photographs when the number of faces and their approximate sizes are known. Although the second stage locates faces, it does not know whose they are. The last stage establishes a unique correlation between names mentioned in the caption and their corresponding areas in the image. These associations are recorded in a semantic network and enable us to selectively view human faces as well as obtain information about them. Input to the system is a digitized image of a newspaper photograph with a caption, as in Figure 1a. The system returns a labeling of parts of the image corresponding to the faces of the people mentioned in the caption, as in Figure 2a and Figure 2b.

PROCESSING THE CAPTION

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 relational graph representing the expected structure of the picture. The relational graph includes information such as the objects hypothesised to be in the picture, their physical appearance, and spatial relationships between them. This is similar to dynamic schema construction

[Weymouth, 1986]. We use the SNePS knowledge-representation and reasoning system to represent both factual information and the relational graph derived from the caption [Shapiro and Rapaport, 1987]. A common representation facilitates the integration of information from both sources. 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 [Srihari, 1981], and it also provides a natural-language parsing and generating facility [Shapiro, 1982].

Figure 3 illustrates a small portion of the output of the parser on processing the caption of Figure 1a. It postulates that two humans, namely Diandra (**m12**) and Michael (**m11**), are present in the picture and that Diandra is to the left of Michael (**m41**). We separate factual information obtained from the caption (**m6**) from derived visual information (**m7**). The hypothesised presence of an object in the picture is represented by a node such as (**m30**). This node represents the proposition that an object (whose visual model is represented by **m31**) is present in the picture (whose visual model is **m7**) and that the object is explicitly mentioned in the caption. Node (**m61**) associates the visual model of an object (**m31**) with the node representing the individual mentioned in the caption (in this case, Diandra). For the visual model represented by node **m31**, the model description is contained in the sub-network represented by nodes **m32** and **m34**. The model description of objects reflects configuration details of the object which can be obtained from the caption. Currently, the model description for a human consists of one default configuration containing the single component "face". If the situation arises where some people are standing and others seated, the parser inserts "top-of" relationships to represent the height discrepancies. This information is vital to the face labeling process. We indicate that sofas may be in the picture by using the value "inferred" for the relation "type", as in node (**m42**). At present, we restrict our search for objects to faces only.

Predicting Objects

There are three classes of heuristics used to extract information from the caption: rules that predict the presence of objects, rules that predict spatial relations between objects and rules that predict configurations of objects. Depending on the type of sentence, several rules are used to predict the presence of objects in a picture. We have observed that many captions are of the form "<subject list> <prepositional-phrase list>". A <prepositional-phrase list> is a series of preposition + noun-phrase pairs, as in the caption of Figure 1a. In such sentences, we propose that each of the subjects in <subject list> is present in the picture. A more interesting question is which of the noun phrases in <prepositional-phrase list> are present in the picture. We can judge whether the entire object is present in the picture based on its scale. We carefully avoid predicting objects which are mentioned in the caption but are not present in the picture. In Figure 1a, although we expect to see components of an apartment such as a sofa (Figure 3), we do not expect to see anything pertaining to New York. In many phrases of the form "<subject> <verb-phrase> <direct-object>", the verb-phrase (e.g., wearing, holding, greeting) indicates the presence of the direct object in the picture. The notion of time is very important here. Captions are traditionally in the present tense even though they refer to events in the past. We have observed that any object referred to at a time previous to the current time is not in the picture.

We also stress the importance of correctly predicting the class of an object. A recent photograph in *The Buffalo News* depicted a horse and her trainer. The caption was "Winning Colors being grazed by her trainer, Wayne Lucas, yesterday morning before the running of the Kentucky Derby". A simplistic parser might conclude that the name "Winning Colors" referred to a human, based on the fact that it was a sequence of two capitalized words serving the subject role in a sentence. This

would predict the presence of two human faces and thus provide incorrect data to the face-location module. A more sophisticated parser would realise that the object of "grazing" is usually an animal such as a cow or a horse. If the parser had access to information about the Kentucky Derby, it could conclude definitely that "Winning Colors" was a horse. If this information were not known, the vision component of the system would be called on to disambiguate between the possiblities of a horse and cow. The last case illustrates the bi-directional flow of information from the caption to the picture and vice versa.

Predicting Spatial Relations Between Objects

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. However, it is not as simple in the case of captions which combine implicit and explicit means of identification. Consider the caption "The All-WNY boys volleyball team surrounds the coach of the year, Frontier's Kevin Starr. Top row, from left, are ..." accompanying a group photograph. The spatial location of the coach must be inferred first through a detailed understanding of the word *surrounds*. The row and column relationships can then be correctly interpeted. Such examples provide are providing a real challenge to both the language-parsing as well as the face-locating stages of our system.

An implicit method of identification frequently used is the ordering of the subjects in the caption to reflect their order in the picture. Our grammar has been designed to assert the spatial relation *left-of* when parsing a list of subjects. This heuristic was used in generating the network of Figure 3. There is frequent departure from the above convention in pictures depicting well-known subjects or in male-female pairs since it is assumed that the reader can disambiguate. Our grammar is designed to raise a flag whenever it encounters such cases, indicating further evidence is required before identification can be made. The labeling procedure uses world knowledge (such as relative heights) to establish a unique correlation between names mentioned in the caption and faces located in the photograph.

Some caption types use detailed information from the picture pertaining to an object in order to uniquely identify that object. For instance, consider the caption "Joseph Crowley, holding the pennant, Thomas Jones ...". The only way to identify Joseph Crowley is first to identify a pennant and then to determine which person in the picture is holding it. Here, identification is not achieved through the constraints posed by spatial relations, but through identification of another object followed by a test for proximity.

Predicting Configurations of Objects

The model description for a class of objects contains a description of a prototype for that class. In the absence of any further information, only the default portions of the model will be instantiated. However, there is often detailed information in the caption pertaining to the specific configuration of an object, which can be represented by instantiating optional components (or configurations) in the model description of the particular object. A simple example of this is the use of words such as "sitting" or "standing" which express different configurations of human body parts. The phrase "shaking hands with" implies a configuration of the arms perpendicular to the body, and the hands of the individuals touching. Consider a caption which refers to a baseball player "diving" for a ball. The use of the word "dive" along with the context of baseball suggest a more horizontal configuration of the body. This information can be valuable if it becomes necessary to detect the entire body (torso, legs etc.), rather than just the face. Jackendoff [Jackendoff, 1987] summarises this idea by

saying that many verbs of station and locomotion are used more to express 3D configurations of objects than to express action. Our grammar has been designed to add to or change the default configurations of humans if the text suggests it.

PROCESSING THE PICTURE

Picture processing in this project is the process of using the information in the hypothesised structure to find relevant objects and spatial relationships in the picture. Currently, we only deal with human faces. Since the caption often gives us spatial constraints on the location of objects, it is frequently sufficient to use crude object-detection modules. In this application, we use a face-locator module which generates candidates for faces. It is often the case that spatial constraints alone are sufficient for eliminating false candidates.

Using caption information and heuristics from photojournalism [Arnold, 1969], the possible range of face sizes appearing in a newspaper photograph can be narrowed. 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, which works in three stages: feature selection, feature detection, and grouping. We have selected as features the two arcs corresponding to the hair-line and the chin-line, and the two lines corresponding to the sides of the face. These features seem to be robust, since they are not greatly affected by factors such as scale, viewing position, or resolution. Furthermore, they are relatively easy to detect. A first-level Hough transform detects the arcs and collinear edge-elements in the image. A line-finder then uses back-projection from the accumulator array to the original image to extract line segments. The curves and line segments are grouped together by a modified Hough transform to generate candidate regions for locations of faces [Govindaraju *et al.*, 1989].

For each image area hypothesised to be a face, this module returns the coordinates of a bounding rectangle for that area. This facilitates the representation of image data in the semantic network. Figure 1b illustrates the performance of the face-locator on the image shown in Figure 1a.

REFINING CANDIDATES AND LABELING FACES

This section describes how faces can be labeled by using the spatial information contained in the caption and heuristics obtained from photojournalism. In general, the location procedure generates more candidates than required (Figure 1b). We have already shown how linguistic heuristics can be used to derive spatial constraints from the caption when they are not explicitly given. These constraints are applied to the candidates generated by the face-locator in an attempt to first reduce the number of candidates and eventually produce a unique labeling of faces (Figures 2a and 2b).

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. We employ additional refinement rules to reduce the number of possibilities. Some of these update the confidence of a candidate pair satisfying a spatial relation, while others update the confidence of the candidate itself. An example of the former is a rule which decreases the confidence of a pair of candidates satisfying a *left-of* or *right-of* relationship where there is a significant vertical difference between the two candidates (in captions where no height discrepancy is indicated). Examples of the second type of rule include one which uses intrinsic image features to update the confidence of a candidate and another which favours centrally located (in the image) candidates.

Identification rules currently operate on pairs of candidates. They contain world knowledge such as "Reagan is taller than Carter" and allow us to further reduce the candidate set. After all the rules have been applied, we employ a procedure which selects the globally best binding based on the confidences of pairs as well as confidence of the candidates comprising the pair. Labeling information is represented in the semantic network by asserting nodes which associate concepts of people with the corresponding areas in the image. In cases where the system cannot uniquely identify faces, all possible candidates for each person appearing in the caption are recorded.

SUMMARY

Our system for understanding newspaper pictures with captions consists of a three-stage process whereby the caption is first parsed with the goal of predicting the structure of the picture. The second stage uses information from the first stage in a top-down processing of the image. The final stage, labeling, is the process of matching pictures of objects with the words representing them in the caption.

The next step in this research is generating visual models for expressions containing certain verb phrases which have a similar visual implication to everyone (e.g. wearing hat, shaking hands). Such phrases are frequently used in captions to identify people in the photograph.

This work was supported by National Science Foundation grants IRI-8613361 and IRI-8610517.

References

[Abe *et al.*, 1981] N. Abe, I. Soga, and S. Tsuji. A Plot Understanding System on Reference to Both Image and Language. In *Proceedings of IJCAI-81*, pages 77–84, 1981.

[Adorni *et al.*, 1984] Giovanni Adorni, Mauro Di Manzo, and Fausto Giunchiglia. Natural Language Driven Image Generation. In *Proceedings of COLING-84*, pages 495–500, 1984.

[Arnold, 1969] Edmund C. Arnold. *Modern Newspaper Design.* Harper and Row, New York, 1969.

[Govindaraju *et al.*, 1989] Venu Govindaraju, David B. Sher, Rohini K. Srihari, and Sargur N. Srihari. Locating Human Faces in Newspaper Photographs. In *Proceedings of CVPR*, 1989.

[Herskovits, 1986] Annette Herskovits. *Language and spatial cognition.* Cambridge University Press, 1986.

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

[Maddox and Pustejovsky, 1987] Anthony B. Maddox and James Pustejovsky. Linguistic Descriptions of Visual Event Perceptions. In *Proceedings of the 9th Annual Cognitive Science Society Conference*, pages 442–454, Seattle, 1987.

[Neumann and Novak, 1983] B. Neumann and H. Novak. Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences. In *Proceedings of IJCAI-83*, pages 724–726, 1983.

[Shapiro, 1982] Stuart C. Shapiro. Generalized Augmented Transition Network Grammars For Generation From Semantic Networks. *American Journal of Computational Linguistics*, 8(2):12–25, 1982.

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

[Srihari, 1981] Rohini K. Srihari. *Combining Path-based and Node-based Reasoning in SNePS.* Technical Report 183, SUNY at Buffalo, 1981.

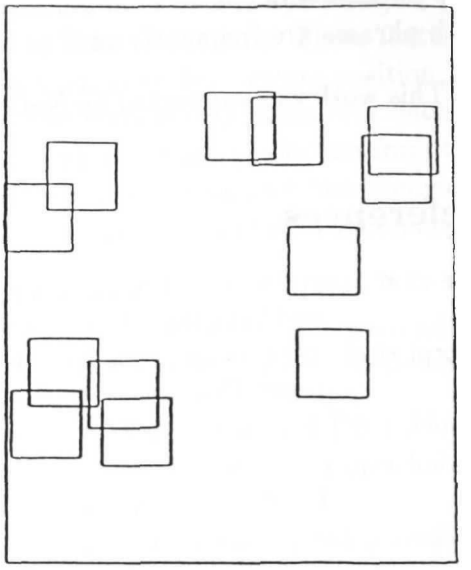
[Waltz and Boggess, 1979] David L. Waltz and L. Boggess. Visual Analog Representation for Natural Language Understanding. In *Proceedings of IJCAI-79*, pages 926–934, 1979.

[Weymouth, 1986] T.E. Weymouth. *Using Object Descriptions in a Schema Network for Machine Vision.* PhD thesis, University of Masschussetts at Amherst, 1986.

[Yokota *et al.*, 1984] Masao Yokota, Rin-ichiro Taniguchi, and Eiji Kawaguchi. Language-Picture Question-Answering Through Common Semantic Representation and its Application to the World of Weather Report. In Leonard Bolc, editor, *Natural Language Communication with Pictorial Information Systems*, Springer-Verlag, 1984.

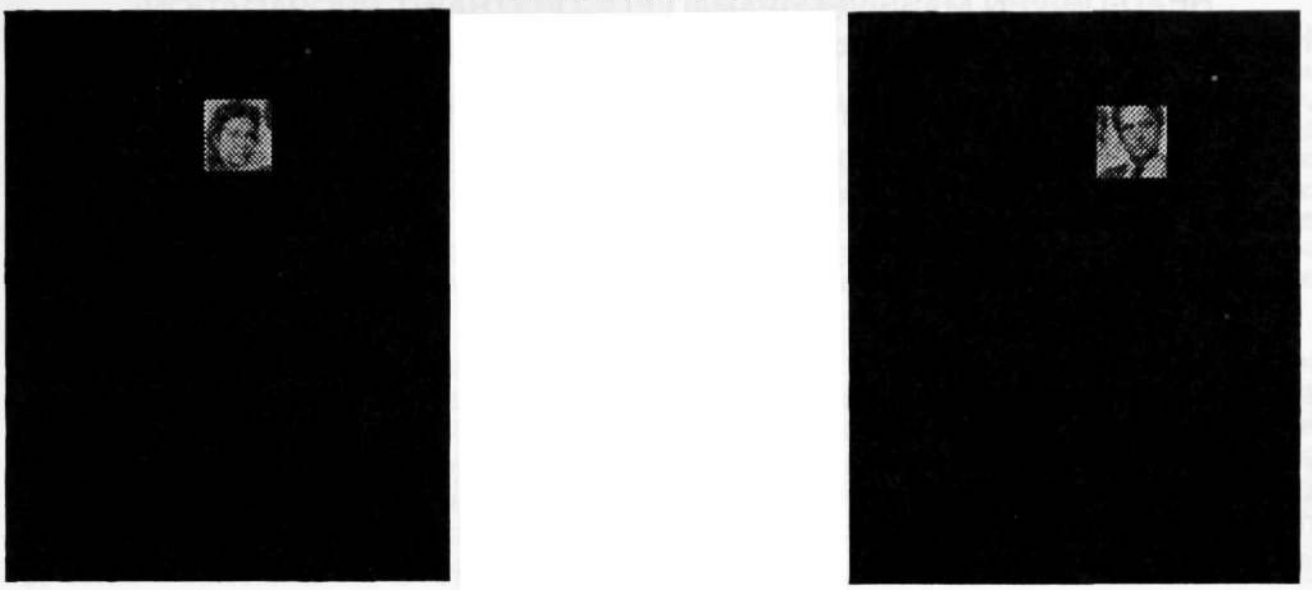
[Zernik and Vivier, 1988] Uri Zernik and Barbara J. Vivier. How Near Is Too Far? Talking about Visual Images. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pages 202–208, Lawrence Erlbaum Associates, 1988.

(a)



(b)

Figure 1: *(a) a newspaper photograph with caption "Diandra and Michael Douglas at their New York apartment (b) candidates generated by the face-locator module*



(a) (b)

Figure 2: *(a) output of system when asked to display Diandra (b) output of system when asked to display Michael*

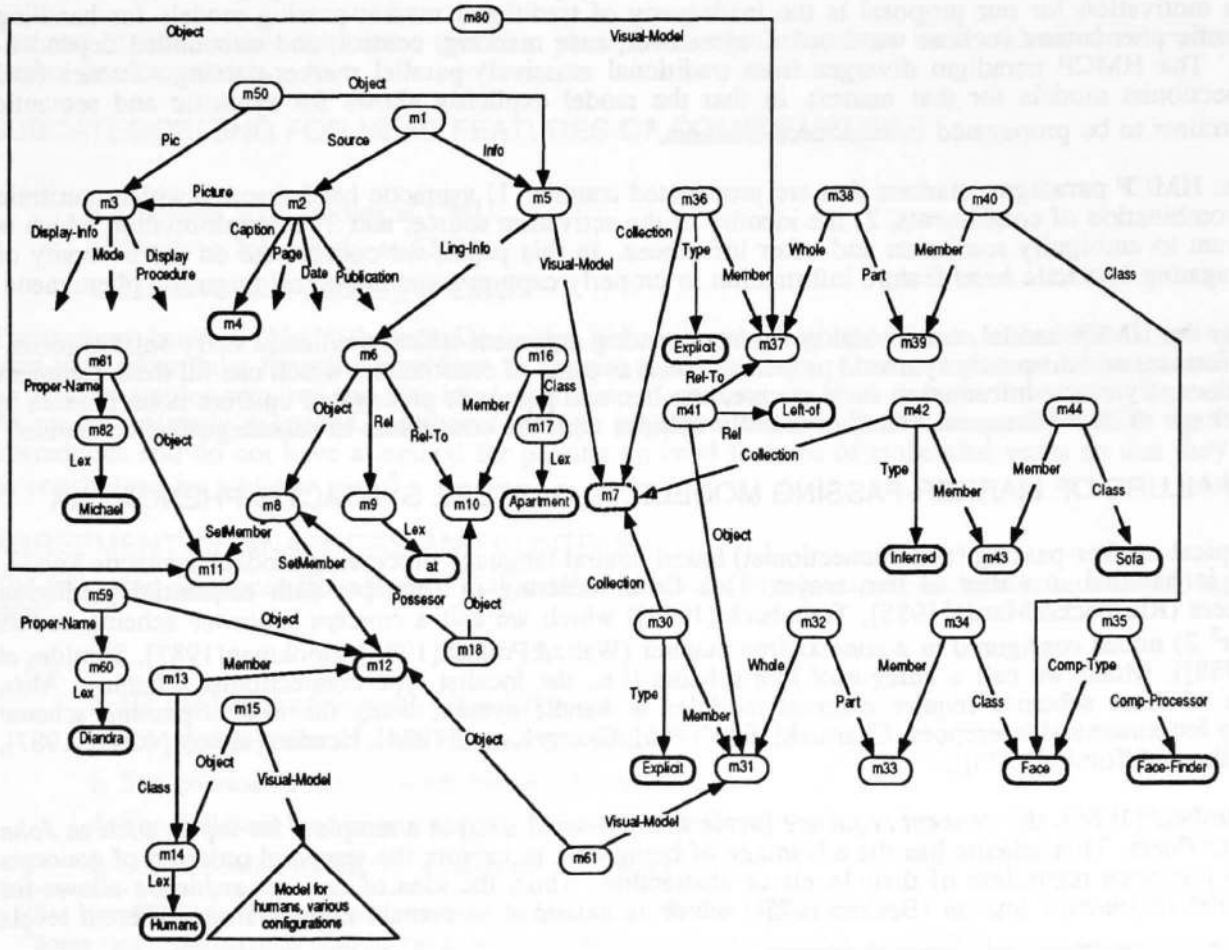


Figure 3: *partial output of the parser on caption of Figure 1a*