Integrating Vision and Language: Towards Automatic Description of Human Movements

Gerd Herzog[1], Karl Rohr[2]

[1] SFB 314 — Projekt VITRA, FB Informatik, Universität des Saarlandes,
Im Stadtwald 15, D-66041 Saarbrücken, herzog@cs.uni-sb.de
[2] Arbeitsbereich Kognitive Systeme, FB Informatik, Universität Hamburg,
Vogt-Kölln-Str. 30, D-22527 Hamburg, rohr@informatik.uni-hamburg.de

Abstract. The integration of vision and natural language processing increasingly attracts attention in different areas of AI research. Up to now, however, there have only been a few attempts at connecting vision systems with natural language access systems. Within the SFB 314, special collaborative program on AI and knowledge-based systems, the automatic natural language description of real world image sequences constitutes a major research goal, which has been pursued during the last ten years. The aim of our approach is to obtain an incremental evaluation and simultaneous description of the perceived time-varying scenes. In this contribution we will report on new results of our joint efforts at combining the natural language access system VITRA with a vision system. We have investigated the problem of describing the movements of articulated bodies in image sequences within an integrated natural language and computer vision system. The paper will focus on our model-based approach for the recognition of pedestrians and on the further evaluation and language production in VITRA.

1 Motivation

The ability to talk about what we see constitutes an important aspect of our communicative capabilities. Consequently, the interplay between perception and language in human-machine interaction constitutes a prominent issue in many potential application areas of language-oriented AI research. Only recently, the integration of vision and natural language processing has attracted growing attention in the field (see [2], [7], [16], [23], [34]).

Leaving out the problem of machine perception, the relationship between natural language and visual accessible information has mainly been studied in the context of natural language interfaces to graphical systems (e.g. ANIMNL [3], SWIM [5], VIENA [33], WINTOP [26]) and to geographical information systems (e.g. GEOSYS [10] and LEI [8]). Only a few approaches have been concerned with a natural language access to camera-recorded images. LANDSCAN [4] answers questions about aerial images, which were taken from a miniaturized model of a city. Image analysis in LANDSCAN is guided by the natural language queries and tries to focus on the recognition of the relevant objects, mentioned in the questions. Dynamic traffic scenes have been investigated in the dialog-system HAM-ANS [35] and in the system NAOS [25], which generates retrospective natural language descriptions. Despite first promising results for rigid objects [9], the connection to the vision component could not be achieved at that time and the geometric descriptions of the analyzed time-varying scenes had to be prepared manually from the underlying image sequences.

In the context of the SFB 314, special collaborative program on AI and knowledge-based systems, the integration of image sequence analysis and natural language production has been further investigated (see [15]). As a first step, trajectories of the centroids of object candidates in the image plane could be provided and utilized to answer simple questions about observations in a short traffic scene [31]. The same image analysis method, which considers object candidates to be essentially rigid, has been applied to short sections of an image sequence recorded from a soccer game. The calibration of the camera allowed for the transformation of trajectories from the image plane into world coordinates and the transformed trajectories served as input for the VITRA system, which generates a running report for the scene under consideration [14]. The more recent model-based approach described in [20] accomplishes the automatic 3D-reconstruction of vehicles in traffic scenes and provides more reliable trajectories of rigid objects for the natural language system.

In this contribution we will report on new results of our joint efforts, which extend the current framework to cope with the automatic perception and verbal description of the movements of articulated bodies, namely pedestrians, in image sequences. In the long term, our ongoing investigations in the following application areas will benefit from a proper treatment of the recognition and natural language description of human movements:

- Intelligent multimedia systems (see [1])
- Driver support systems in road vehicles (see [21])
- Autonomous mobile robot systems (see [32])

In addition, the systems envisaged here could serve practical purposes in other areas as well, e.g., in traffic control and in medical technology.

2 Model-Based Recognition of Human Movements

Within computer vision the automatic interpretation of human movements is one of the most challenging tasks. A central problem in recognizing such movements is due to the fact that the human body consists of body parts linked to each other at joints to enable different movements of the parts and therefore, in general, has to be treated as a non-rigid (or more precisely as an articulated) body. In addition, for general camera positions always some of the body parts are occluded. Although occlusion can give important cues in a recognition task, the automatic interpretation gets more difficult. Another problem that has to be dealt with is the clothing which can have a large influence on the appearence of a person (wide or tight trousers, different jackets, etc.). Clothing can also cause complex illumination phenomena that, in addition, change during movement.

Because of these difficulties most existing approaches assume the joints of the human body to be marked or they investigate synthetic images. Approaches for real-world images often analyze gymnastic movements but not locomotion which in general simplifies the interpretation because the effect of self-occlusion is diminished. Other approaches restrict their analysis of locomotions to certain parts of the body (for references see, for example, [6] and [30]).

In this section we describe our model-based approach for recognizing human movements (for details see [28], [29], [30]). With this approach the human body as well as its movement is represented explicitly. Given a monocular real-world image sequence recorded with a stationary camera our algorithm determines the 3D positions as well as the postures of moving persons. The algorithm is designed for analyzing the movement of human walking (which is the most frequent type of locomotion of persons) but could also be generalized to other movements. In comparison to [17], [18] where also images of walking persons were analyzed, in our approach we employ data from medical movement studies. This data represents an average over a relative large number of test persons. We also use a Kalman filter for incrementally estimating the model parameters yielding smooth and robust results. An additional advantage is that our system runs completely automatic.

2.1 Overview of our Approach

Our algorithm can be subdivided into two phases: initialization and incremental estimation. Whereas in the first phase the images are evaluated in a batch type manner, in the second phase processing is done on an incremental basis. The main parts can be summarized as follows.

1. *Initialization*

   Independent evaluation of about 10-15 images:
   - Detection of image regions corresponding to moving persons
     (using a change detection algorithm and binary image operations)
   - Estimation of the movement states, i.e. 3D positions and postures
     (using a calibration matrix for central projection and matching
     contours of the 3D model with grey-value edges)
   - Determination of starting values for the Kalman filter
     (using linear regression and all the estimates from above)

2. *Incremental estimation*

   After initialization the following Kalman filter scheme is applied to each image:
   - Prediction of the movement state
     (using estimation results from previous images)
   - Determination of measurements
     (using matching results of the 3D model to the current image)
   - Estimation of the current movement state
     (using the predicted movement state and the measurements)

2.2 Human Body and Movement Model

We represent the human body by a cylindrical volume model as suggested in [22] (see Fig. 1). For modelling the movement of walking we use a kinematic approach exploiting

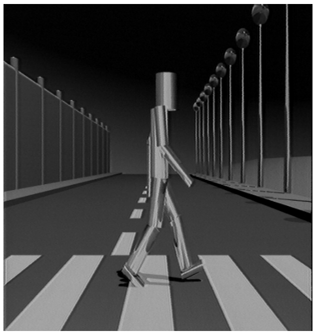


Fig. 1. 3D model of the human body

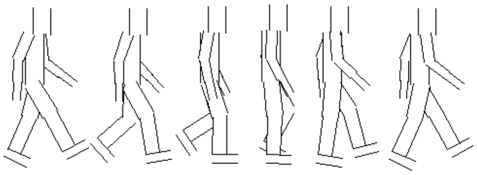


Fig. 2. Movement states of walking represented by the visible contours of the 3D model ($pose = $ 0.0, 0.1, 0.2, 0.3, 0.4, 0.5)

data from the medical movement study described in [24]. The study demonstrates that the movement curves of the body parts of different persons are very similar. This fact opens us the possibility to use this data as a knowledge source. Note, however, that the similarity for different persons is very astonishing if one imagines that it is often possible to identify persons by their gait.

In our approach we have used the angle values from the movement curves for each of the joints at the shoulder, ellbow, hip and knee and have interpolated them by periodic cubic splines. Analogously, we have modelled the vertical displacement of the whole body. Since walking is a symmetric movement, the movement curves of the joints are only needed for one side of the human body. A nice property is that we need only one parameter, named $pose$, to specify the relative positions of all body parts (see Fig. 2).

2.3 Incremental Estimation Using a Kalman Filter

We apply a Kalman filter to incrementally estimate the 3D position $\mathbf{X} = (X, Y, Z)$ as the center of the torso as well as the posture $pose$ of a person in succesive images:

$$\begin{aligned} \mathbf{p}_k &= \underline{\Phi}_{k,k-1}\,\mathbf{p}_{k-1} + \mathbf{w}_{k-1} \\ \mathbf{z}_k &= \underline{\mathbf{H}}_k\,\mathbf{p}_k + \mathbf{v}_k \end{aligned}$$

where the searched model parameters are repesented by the state vector $\mathbf{p}_k$ at the point of time k. $\underline{\Phi}_{k,k-1}$ is the transition matrix and $\mathbf{w}_k$ represents modelling errors. $\underline{\mathbf{H}}_k$ denotes the measurement matrix and $\mathbf{v}_k$ represents measurement errors. Predicting the parameters and the covariance matrix is done by :

$$\begin{aligned} \mathbf{p}_k^* &= \underline{\Phi}_{k,k-1}\,\hat{\mathbf{p}}_{k-1} \\ \underline{\mathbf{P}}_k^* &= \underline{\Phi}_{k,k-1}\,\underline{\hat{\mathbf{P}}}_{k-1}\,\underline{\Phi}_{k,k-1}^T + \underline{\mathbf{Q}}_{k-1} \end{aligned}$$

With these predictions and the current measurement $\mathbf{z}_k$ the estimates $\hat{\mathbf{p}}_k$ and $\underline{\hat{\mathbf{P}}}_k$ in the current image can be computed as:

$$\begin{aligned} \hat{\mathbf{p}}_k &= \mathbf{p}_k^* + \underline{\hat{\mathbf{K}}}_k(\mathbf{z}_k - \underline{\mathbf{H}}_k\,\mathbf{p}_k^*) \\ \underline{\hat{\mathbf{P}}}_k &= (\underline{\mathbf{I}} - \underline{\hat{\mathbf{K}}}_k\underline{\mathbf{H}}_k)\underline{\mathbf{P}}_k^* \\ \underline{\hat{\mathbf{K}}}_k &= \underline{\mathbf{P}}_k^*\underline{\mathbf{H}}_k^T(\underline{\mathbf{H}}_k\underline{\mathbf{P}}_k^*\underline{\mathbf{H}}_k^T + \underline{\mathbf{R}}_k)^{-1} \end{aligned}$$

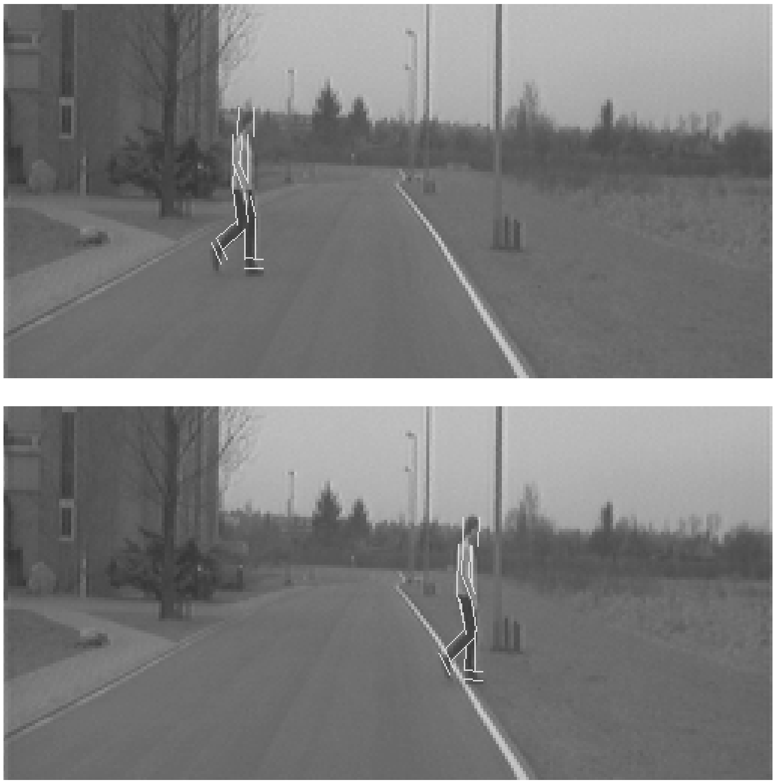


Fig. 3. Estimated movement states superimposed onto the original images (images number 20 and 80; for visualization purposes the upper and lower part of the original images have been cut)

In the following we assume the velocity of the movement to be constant and that the pedestrian moves parallel to the image plane. In our approach the problem of self-occlusion is treated by fitting the model as a whole (using a hidden-line algorithm to remove occluded model contours). Note, that based on the predictions the search space for matching the model contours with grey-value edges can considerably be reduced.

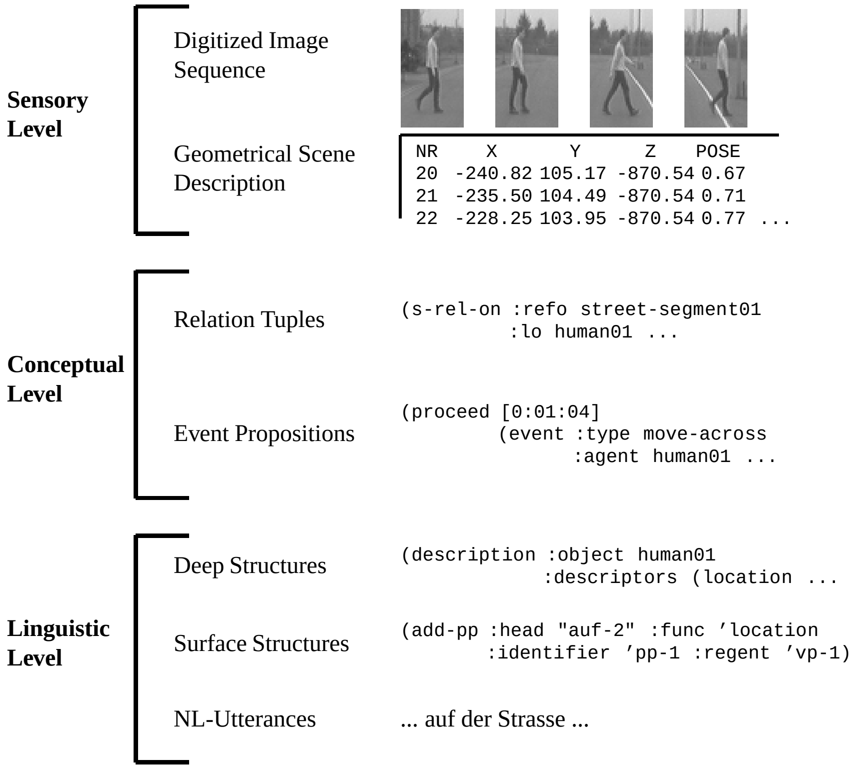
2.4 Experimental Results

Our approach has been applied to a real-world image sequence which shows a walking person crossing the street. The whole sequence consists of 80 images corresponding to about 3 seconds observation time. Estimated movement states superimposed onto the original images are shown in Fig. 3. Note, that for visualization purposes we have cut out the upper and lower part of the original images. The algorithm, however, has been applied to the whole images. For each image our algorithm provides estimates for the 3D position and the posture. These values are transferred to the natural language access sytem.

3 High-Level Scene Analysis

Information concerning visible objects and their locations over time, together with additional world knowledge about the objects, constitutes the *geometrical scene description* (GSD). This intermediate representation has been proposed in [25] as an idealized interface between a vision system and a natural language access system.

Further interpretation of the GSD is required in order to translate the results of low-level vision processes into a natural language description. High-level scene analysis aims at recognizing conceptual units at a higher level of abstraction, including spatial relations for the explicit characterization of spatial arrangements of objects as well as motion events for the qualitative representation of object movements. These conceptual structures bridge the gap between visual data and natural language concepts, such as spatial prepositions, motion verbs, and temporal adverbs (see Fig. 4).



```
NR     X       Y       Z    POSE
20  -240.82 105.17 -870.54 0.67
21  -235.50 104.49 -870.54 0.71
22  -228.25 103.95 -870.54 0.77 ...
```

```
(s-rel-on :refo street-segment01
          :lo human01 ...
```

```
(proceed [0:01:04]
         (event :type move-across
                :agent human01 ...
```

```
(description :object human01
             :descriptors (location ...
```

```
(add-pp :head "auf-2" :func ’location
        :identifier ’pp-1 :regent ’vp-1)
```

... auf der Strasse ...

Fig. 4. Levels of representation in the transformation from visual data to verbal descriptions

In the GSD spatial information is encoded quantitatively. In analogy to prepositions, their linguistic counterparts, spatial relations provide a qualitative description of spatial arrangements of objects. Each spatial relation characterizes a class of object configurations by specifying conditions, such as the relative position of objects or the distance between them. The detailed geometric knowledge, encoded in the GSD, can be exploited for the definition of a reference semantics, that does not assign simple truth values to spatial predications, but instead introduces a measure of degrees of applicability that expresses the extent to which a spatial relation is applicable (see [13]). Since different degrees of applicability can be expressed by linguistic hedges, such as *'directly'* or *'more or less'*, more exact scene descriptions are possible. Furthermore, the degree of applicability can be used to select the most appropriate reference objects and relations if an object configuration can be described by several spatial predications.

Our system (for details see [11], [13], [32]) is capable of computing topological (e.g. *in*, *near*, etc.) as well as orientation-dependent relations (e.g. *left-of*, *over*, etc.). Since the frame of reference is explicitly taken into account, the system can cope with the *intrinsic*, *extrinsic*, and *deictic* use of directional prepositions. A continuous gradation of the applicability of a relation is achieved by mapping the relevant factors, i.e., the scaled local distance and if necessary a local deviation angle, onto appropriate cubic spline functions.

Header:	(MOVE-ACROSS ?o*mobile-object ?s*surface)
Subconcepts:	(MOVE ?o) [I1]
	(LOC-INTERIOR ?o ?s) [I2]
	(LOC-EXTERIOR ?o ?s) [I3]
Temporal-Relations:	[I2] :during [I1]
	[I2] :meets [I3]
	[I2] :equals [MOVE-ACROSS]

Fig. 5. Event model for the concept *'move-across'*

The interpretation of object movements in terms of motion events serves for the symbolic abstraction of the temporal aspects of a time-varying scene. In VITRA, the recognition of motion events is based on generic event models, i.e., declarative descriptions of classes of interesting object movements. These event concepts are organized into an abstraction hierarchy, grounded on specialization (e.g., *running* is a *moving*) and temporal decomposition (see Fig. 5). This conceptual hierarchy can be utilized in the language production process in order to guide the selection of the relevant propositions.

If a real-world image sequence is to be described simultaneously as it is perceived, one has to talk about object movements even while they are currently happening and not yet completed. Thus, event models are translated automatically into labeled directed graphs, which allow an incremental event recognition by traversing one edge after the other. These *course diagrams* model the prototypical progression of an event and they rely on a discrete model of time, which is induced by the underlying image sequence. Since a distinction between events that have and those that have not occurred is insufficient, the additional predicates *start*, *proceed*, and *stop* have been introduced, which can be used to characterize the progression of an event. The edges in the course diagrams are typed in order to define these basic event predicates [12], [13].

In general, even a high-level analysis as it is depicted here, may not be sufficient for the selection of an adequate description. Besides spatio-temporal aspects, non-visual concepts like the presumed intentions, plans, and plan interactions of the observed agents can play an important role (e.g., in *"The old lady is waiting at the pedestrian crossing for the traffic lights to turn green."*). In the soccer domain of VITRA, a module for such an intentional interpretation has already been realized [27].

4 Generating Natural Language Descriptions

The incremental high-level scene analysis continuously provides information as the image sequence progresses. Simultaneously, natural language utterances have to be generated in order to provide a running report of the time-varying scene. In VITRA, this comprises (1) the selection of currently relevant propositions, (2) their ordering into a linear text structure, and (3) the successive encoding of the selected propositions [15].

Because of the strong temporal restrictions, the description must focus on the most relevant facts, which are determined by salience, topicality, and recognition state of the corresponding events. The taxonomic structure formed by the generic event concepts can be exploited for the valuation of the importance of an event while selecting among different propositions. Generally, more complex motion events are preferred since they provide a higher degree of condensation of visual information. The salience of an event is also determined by the frequency of occurrence. Topicality decreases continuously for terminated movements and actions as the scene progresses. To avoid redundancy, an occurrence will not be mentioned if it is implied by some other proposition already verbalized. The linearization of propositions depends primarily on the temporal ordering of the corresponding events and actions; secondarily, focusing criteria are employed to maintain discourse coherence.

After relevant propositions are selected and ordered, they are passed over to the encoding component. In the process of transforming symbolic event descriptions into natural language utterances, first a verb is selected and the case-roles associated with it are instantiated. In our approach, lexical choice relies on a rule-based conceptual lexicon, which constitutes the connection between non-linguistic and linguistic concepts. Considering the contents of the text memory and the partner model additional selection processes decide which information concerning the case-role fillers should be conveyed. The choosen information is then translated into natural-language expressions referring to objects, locations, and time.

Internal object identifiers are transformed into noun phrases by the selection of attributes that enable the listener to uniquely identify the intended referent. Anaphoric expressions are generated if the referent is in focus and no ambiguity is possible. Spatial prepositions and appropriate objects of reference are selected to refer to spatial relations; time is indicated by the verb tense and by temporal adverbs.

The verbalization process, which includes grammatical encoding, linearization, and inflection, receives preverbal messages in a piecemeal fashion (see Fig. 6). It is based on the formalism of Lexicalized LD/LP Tree Adjoining Grammar (LTAG), which associates lexical items with syntactic rules, permits flexible expansion operations and allows the description of local dominance to be separated from linear precedence rules [19].

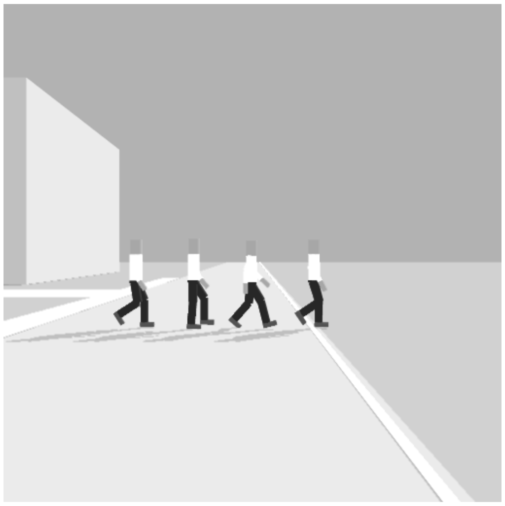
```
(add-utt-par :identifier ’utt-par-1 :intention ’declarativ)
(add-vp :head "geh" :identifier ’vp-1 :mood ’indicative)
(add-np :head "fussgaenger" :identifier ’np-1
        :specifier ’definite)
(add-np :head "strasse" :regent ’pp-1 :func ’prepobject)
(add-pp :head "ueber" :func ’location :identifier ’pp-1
        :regent ’vp-1)
```

Der Fussgaenger geht ueber die Strasse. *(The pedestrian walks across the street.)*

Fig. 6. Preverbal messages and generated utterance

These characteristics make LTAG’s a good candidate for incremental natural language generation [1].

Fig. 7 shows the GSD as it has been derived from the real-world image sequence, presented in Fig. 3, and the corresponding natural language description. In this short example, the location of the only visible mobile object had to be described first, since the listener must be enabled to construct a visual conceptualization of the scene. For the synthesis of the localization expression several spatial predications are combined in this case. Despite the high degree of applicability for the topological relation *‘on’* this spatial reference is not specific enough, because of the extent of the reference object *‘street’*. Two projective relations have to be composed in order to relate the pedestrian to the partly visible building, which constitutes a salient landmark in the scene. Taking



Auf der Strasse in etwa rechts vor dem Osttrakt befindet sich ein Fussgaenger.
Er geht ueber die Strasse.
(There is a pedestrian on the street nearly in front and to the right of the eastern part of the building.
He walks across the street.)

Fig. 7. Visualization of the GSD (for the images number 20, 40, 60, and 80 of the image sequence shown in Fig. 3) and the corresponding verbal description

into account the prominent front of the building, an intrinsic orientation of the reference object has been selected instead of the deictic frame of reference, as it is established by the perspective view. In our example, the crossing of the street constitutes the salient motion event, which is verbalized as the movement continues.

5 Conclusion

One of the goals of language-oriented AI research is to attain a completely operational form of referential semantics that reaches down to the sensoric level. In this contribution we have reported on current progress in our attempts at integrating vision and natural language processing. These results constitute a first step towards automatic simultaneous description of human movements in real-world image sequences. An advantage of our approach is the emphasis on concurrent image sequence evaluation and natural language processing, carried out on an incremental basis. These features constitute an important prerequisite for real-time performance.

So far, from the point of view of natural language generation, image analysis is still restricted to rather short and relatively simple image sequences. For more complex scenes the vision system would require an extension for the classification of different kinds of mobile objects and different movement types. In addition, image analysis may not be restricted to the recognition of moving objects, but could provide the 3D reconstruction of the stationary background as well. Because of the simplicity of the investigated time-varying scene, which contains a single mobile object and only a few landmarks, there is no need to fully exploit the 3D reconstruction of the articulated body in the course of high-level analysis and language production. In order to investigate the generation of more complex textual descriptions, appropriate synthetic scenes have to be studied as well. For methodical reasons, however, it seems to be decisive to cope with available data from real-world image sequences if current limitations in the integration of vision and language processing are to be overcome.

6 Technical Notes

The computer vision system has been implemented in ADA on a DEC Workstation. The VITRA system is written in Common Lisp and CLOS, with the graphical user interface implemented in CLIM, and has been developed on Symbolics UX1200S Lisp Coprocessors, and on HP 9720 as well as on SUN Workstations.

Acknowledgements

We would like to thank the anonymous referees for valuable comments and suggestions. The support of the DFG, SFB 314 “Künstliche Intelligenz und Wissensbasierte Systeme”, is gratefully acknowledged.

References

1. E. André, G. Herzog, and T. Rist. Von der Bildfolge zur multimedialen Präsentation. In *Integration von Bild, Modell und Text '95*, pages 129–142, Madgeburg, 1995. ASIM, Techn. Univ. Wien.
2. Artificial Intelligence Review Journal, 8, Special Volume on the Integration of Natural Language and Vision Processing, 1994.
3. N. I. Badler, B. L. Webber, J. Kalita, and J. Esakov. Animation from Instructions. In N. I. Badler, B. A. Barsky, and D. Zeltzer, editors, *Making Them Move: Mechanics, Control, and Animation of Articulated Figures*, pages 51–93. Morgan Kaufmann, San Mateo, CA, 1991.
4. R. Bajcsy, A. Joshi, E. Krotkov, and A. Zwarico. LandScan: A Natural Language and Computer Vision System for Analyzing Aerial Images. In *Proc. of the 9th IJCAI*, pages 919–921, Los Angeles, CA, 1985.
5. X. Briffault and M. Zock. What do we Mean when we Say "to the Left" or "to the Right"? How to Learn about Space by Building and Exploring a Microworld. In P. Jorrand and V. Sgurev, editors, *Artificial Intelligence: Methodology, Systems, Applications (AIMSA'94)*, pages 363–371. World Scientific, Singapore, 1994.
6. C. Cédras and M. Shah. Motion-based Recognition: A Survey. *Image and Vision Computing*, 13(2):129–155, 1995.
7. Centre National de la Recherche Scientifique. *Images et Langages: Multimodalité et Modélisation Cognitive, Colloque Interdisciplinaire du Comité National de la Recherche Scientifique*, Paris, 1993.
8. D. N. Chin, M. McGranaghan, and T.-T. Chen. Understanding Location Descriptions in the LEI System. In *Proc. of the 4th Conf. on Applied Natural Language Processing*, pages 138–143, Stuttgart, Germany, 1994.
9. L. Dreschler and H.-H. Nagel. Volumetric Model and 3D-Trajectory of a Moving Car Derived from Monocular TV-Frame Sequences of a Street Scene. *Computer Graphics and Image Processing*, 20:199–228, 1982.
10. M. Fürnsinn, M. Khenkhar, and B. Ruschkowski. GEOSYS — Ein Frage-Antwort-System mit räumlichem Vorstellungsvermögen. In C.-R. Rollinger, editor, *Probleme des (Text-) Verstehens, Ansätze der künstlichen Intelligenz*, pages 172–184. Niemeyer, Tübingen, 1984.
11. K.-P. Gapp. Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space. In *Proc. of AAAI-94*, pages 1393–1398, Seattle, WA, 1994.
12. G. Herzog. Utilizing Interval-Based Event Representations for Incremental High-Level Scene Analysis. In M. Aurnague, A. Borillo, M. Borillo, and M. Bras, editors, *Proc. of the 4th International Workshop on Semantics of Time, Space, and Movement and Spatio-Temporal Reasoning*, pages 425–435, Château de Bonas, France, 1992.
13. G. Herzog, T. Rist, and E. André. Sprache und Raum: Natürlichsprachlicher Zugang zu visuellen Daten. In C. Freksa and C. Habel, editors, *Repräsentation und Verarbeitung räumlichen Wissens*, pages 207–220. Springer, Berlin, Heidelberg, 1990.
14. G. Herzog, C.-K. Sung, E. André, W. Enkelmann, H.-H. Nagel, T. Rist, W. Wahlster, and G. Zimmermann. Incremental Natural Language Description of Dynamic Imagery. In C. Freksa and W. Brauer, editors, *Wissensbasierte Systeme. 3. Int. GI-Kongreß*, pages 153–162. Springer, Berlin, Heidelberg, 1989.
15. G. Herzog and P. Wazinski. VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions. *Artificial Intelligence Review*, 8(2/3):175–187, 1994.
16. B. Hildebrandt, R. Moratz, G. Rickheit, and G. Sagerer. Integration von Bild- und Sprachverstehen in einer kognitiven Architektur. *Kognitionswissenschaft*, 4(3):118–128, 1995.

17. D. Hogg. Model-based Vision: A Program to See a Walking Person. *Image and Vision Computing*, 1(1):5–20, 1983.
18. D. Hogg. *Interpreting Images of a Known Moving Object*. PhD thesis, University of Sussex, Brighton, UK, 1984.
19. A. Kilger. Using UTAGs for Incremental and Parallel Generation. *Computational Intelligence*, 10(4):591–603, 1994.
20. D. Koller. *Detektion, Verfolgung und Klassifikation bewegter Objekte in monokularen Bildfolgen am Beispiel von Straßenverkehrsszenen*. Infix, St. Augustin, 1992.
21. W. Maaß, P. Wazinski, and G. Herzog. VITRA GUIDE: Multimodal Route Descriptions for Computer Assisted Vehicle Navigation. In *Proc. of the Sixth Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE-93*, pages 144–147, Edinburgh, Scotland, 1993.
22. D. Marr and H. K. Nishihara. Representation and Recognition of the Spatial Organization of three-dimensional Shapes. In *Proc. Royal Society B200*, pages 269–294, London, 1978.
23. P. McKevitt, editor. *Proc. of AAAI-94 Workshop on Integration of Natural Language and Vision Processing*, Seattle, WA, 1994.
24. M. P. Murray, A. B. Drought, and R. C. Kory. Walking Patterns of Normal Men. *Journal of Bone and Joint Surgery*, 46-A(2):335–360, 1964.
25. B. Neumann and H.-J. Novak. NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen. *Informatik Forschung und Entwicklung*, 1:83–92, 1986.
26. P. Olivier, T. Maeda, and J. Tsujii. Automatic Depiction of Spatial Descriptions. In *Proc. of AAAI-94*, pages 1405–1410, Seattle, WA, 1994.
27. G. Retz-Schmidt. *Die Interpretation des Verhaltens mehrerer Akteure in Szenenfolgen*. Springer, Berlin, Heidelberg, 1992.
28. K. Rohr. Auf dem Wege zu modellgestütztem Erkennen von bewegten nicht-starren Körpern in Realweltbildfolgen. In H. Burkhardt, K. H. Höhne, and B. Neumann, editors, *Mustererkennung 1989, 11. DAGM Symposium*, pages 324–328. Springer, Berlin, Heidelberg, 1989.
29. K. Rohr. Incremental Recognition of Pedestrians from Image Sequences. In *Proc. of IEEE Conf. on Computer Vision & Pattern Recognition*, pages 8–13, New York, NY, 1993.
30. K. Rohr. Towards Model-based Recognition of Human Movements in Image Sequences. *Computer Vision, Graphics, and Image Processing (CVGIP): Image Understanding*, 59(1):94–115, 1994.
31. J. R. J. Schirra, G. Bosch, C.-K. Sung, and G. Zimmermann. From Image Sequences to Natural Language: A First Step Towards Automatic Perception and Description of Motions. *Applied Artificial Intelligence*, 1:287–305, 1987.
32. E. Stopp, K.-P. Gapp, G. Herzog, T. Längle, and T. C. Lüth. Utilizing Spatial Relations for Natural Language Access to an Autonomous Mobile Robot. In B. Nebel and L. Dreschler-Fischer, editors, *KI-94: Advances in Artificial Intelligence*, pages 39–50. Springer, Berlin, Heidelberg, 1994.
33. I. Wachsmuth and Y. Cao. Interactive Graphics Design with Situated Agents. In W. Strasser and F. Wahl, editors, *Graphics and Robotics*. Springer, Berlin, Heidelberg, 1994.
34. W. Wahlster. Text and Images. In R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, editors, *Survey on Speech and Natural Language Technology*. Kluwer, Dordrecht, 1994.
35. W. Wahlster, H. Marburger, A. Jameson, and S. Busemann. Over-answering Yes-No Questions: Extended Responses in a NL Interface to a Vision System. In *Proc. of the 8th IJCAI*, pages 643–646, Karlsruhe, FRG, 1983.

This article was processed using the LATEX macro package with LLNCS style