

Motion Verbs and Vision

¹Ioanna Malagardi and ²John Kontos

Abstract--In the present paper we aim at the use of the analysis of the definitions of the Motion Verbs for the application to the understanding and description of action sequences such as those recorded in a video. The main points of computer processing of verbs of motion involve that the definitions are given as input to a system that produces an output that gives a grouping of the verbs and synthesized definitions of these verbs using primitives. In the system presented here the input action sequence is analyzed using the semantics of primitive motion verbs and the way they combine for the synthesis of complex verbs that summarize the action sequence. A future application of this work could be in the automatic text generation of descriptions of motion images obtained by artificial vision systems. These texts may be helpful for people with vision disabilities.

Index Terms – Cognitive Vision, Moving Images, Motion Verbs, Semantic Ontology, Video

1 INTRODUCTION

IN the present paper we aim at the use of the analysis of the definitions of the Motion Verbs for the application to the understanding and description of action sequences such as those recorded in a video. Motion Verbs are analyzed using primitive verbs as described below using definition chains. A primitive motion verb can be classified according to pictorial criteria that may be obtained by the comparative analysis of a sequence of images. This classification can be inherited by non primitive verbs in accordance with their dependence on primitive verbs.

A variety of approaches have been proposed for the processing of action sequences recorded in a video. Most of these approaches refer to one or more of three levels of representation, namely, image level, logic level and natural language level. Motion verbs are useful for the third level representation.

In previous work [1] and [2] we presented a system of programs that concerns the processing of definitions of 486 verbs of motion as they are presented in a dictionary. This processing aimed at the exploitation of dictionaries for Natural Language Processing systems. A recent proposal for the organization of a Machine Readable Dictionary is based on the structure and development of the Brandeis Semantic Ontology (BSO), a large generative lexicon ontology and lexical database. The BSO has been designed to allow for more widespread access to Generative Lexicon-based lexical resources and help researchers in a variety of natural language computational tasks [3].

The semantic representation of images resulting from knowledge-assisted semantic image analysis e.g. [4] can be used to identify a primitive motion verb describing the sequence of a few images. Other efforts for the semantic annotation and representation of image sequences aim at the building of tools for the pre-processing of images and are briefly present below.

2 RELATED WORK

Cho et al. [5] propose to measure similarity between trajectories in a video using motion verbs. They use a hierarchical model based on *is_a* and *part_of* relation in combination with antonym relations. The ontological knowledge required by their system is extracted from WordNet.

They created five base elements to represent motion of moving objects namely “approach”, “go to”, “go into”, “depart” and “leave”. They use motion verbs to represent moving of objects as high level features from the cognitive point of view. According to this paper the problem of bridging the

¹Department of Informatics & Telecommunications
Educational & Language Technology Laboratory
NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
HELLAS E-mail: imalagar@di.uoa.gr

²Department of Philosophy & History of Science
NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
HELLAS E-mail: ikontos2003@yahoo.com

gap between high level semantics and no level video features is still open. The method proposed in the present paper is a novel contribution towards the solution of the bridging problem mentioned above.

The University of Karlsruhe group Dahlkamp and Nagel, [6], [7], is developing a system for cognitive vision applied to the understanding of inner-city vehicular road traffic scenes. They argue that an adequate natural language description of developments in a real-world scene can be taken as a proof of "understanding what is going on". In addition to vehicle manoeuvres the lane structure of inner-city roads and road intersections are extracted from images and provide reference for both the prediction of vehicle movements and the formulation of textual descriptions. Individual actions of an agent vehicle are associated to verb phrases that can be combined with a noun phrase referring to the agent vehicle to construct a single sentence in isolation. The next step is to concatenate individual manoeuvres into admissible sequences of occurrences. Such knowledge about vehicular behaviour is represented internally as a situation graph formed by situation nodes connected by prediction edges. They organize situation nodes not only according to their temporal concatenation but also according to a degree of conceptual refinement. For example an abstract situation node called "cross" (for cross an intersection) is refined into a sub graph that consists of a concatenation of three situation nodes, namely, (1) *drive_to_intersection*, (2) *drive_on_intersection*, (3) *live_intersection*. Such a refinement can take place recursively. A subordinate situation node inherits all predicates from its superordinate situation nodes. A path through a directed situation graph tree implies that the agent executes the actions specified in the most detailed situation node reached at its point in time during traversal semicolon, that is, such a path implies the behavior associated with a concatenation of actions encountered along such a path.

Using these situation graphs trees the above mentioned system generates a list of elementary sentences describing simply events. These sentences are analogues to the sentences we input to our system which instead of traffic scenes analyses office scenes and recognizes higher level event structures.

Recently learning systems have been developed for the detection and representation of events in videos. For example A. Hakeem, M. Shah [8] who propose an extension of CASE representation of natural languages that facilitates the interface between users and the computer. CASE is a representation proposed by Fillmore [9]. They propose two critical extensions to CASE that concern the involvement of multiple agents and the inclusion of temporal information.

M. Fleishman et al. (2006) [10] present a methodology to facilitate learning temporal structure in event recognition. They modeled complex events using a lexicon of hierarchical patterns of movement, which were mined from a large corpus of unannotated video data. These patterns act as features for a tree kernel based Support Vector Machine that is trained on a small set of manually annotated events. To distinct types of information are encoded by these patterns. First, by abstracting to the level of events as opposed to lower level observations of motion, the patterns allow for the encoding of more fine grained temporal relations than traditional HMM approaches. Additionally, hierarchical patterns of movement have the ability to capture global information about an event.

3 GREEK MOTION VERB PRIMITIVES

The main points of computer processing of verbs of motion involve that the definitions are given as input to a system of Prolog programmes and an output is produced that gives a grouping of the verbs and synthesized definitions of these verbs.

The set of 486 verb entries related to motion and were used as input to a system that produced groups of them on the basis of chains of their definitions. The verb at the end of a chain was used as the criterion of verb grouping. Prior to using these chains it was necessary to eliminate the cyclic parts of the definition chains which were also automatically detected by the system. The definitions of the verbs in each definition are in turn retrieved from the lexicon and in this way chains of definitions are formed. These chains end up in circularities that correspond to reaching basic verbs. The elimination of circularity that occurs in certain chains requires the choice of suitable verb as terminal the chain. The choice for each case of elimination of circularity requires the adoption of some "ontology".

The results of automatic grouping were compared with groupings in Greek, German and English language that were done manually. The construction of chains was then applied to the automatic construction of definitions using a small number of verbs that appeared in at the end of the definition chains and were named "basic" representing primitive actions. The English

translation of some of the primitive Greek motion verbs obtained by our system are: Touch, take, put, stir, raise, push, walk that are used for the system presented here in this paper in order to make it intelligible to a wider audience.

4 UNDERSTANDING AND DESCRIPTION OF MOVING IMAGES

The choice of one or more primitive verbs for the automatic description of a motion sequence is based on the abstract logical description of this sequence. The abstract logical description is supposed to contain declarations of the position and state of different entities depicted in the images from which the action sequence is extracted. The comparative logical analysis of the semantic representation of images resulting from knowledge-assisted semantic image analysis is used to identify a primitive motion verb describing the sequence of a few images. The synthesized definitions of more complex motion verbs together with other domain knowledge is used to generate text that describes a longer part of the action sequence with these verbs. A system that we implemented for the description in English of action sequences is described below.

5 SYSTEM DESCRIPTION

The system consists of an input module that accepts formal descriptions of action sequences. These sequences are analyzed by a primitive action recognizer module that provides input to a complex verb recognizer and finally an output module generates the sequence description. The system was implemented in Prolog and two examples of its operation are given below. The three main modules of our system are briefly described below giving indicative Prolog rules that are used for the accomplishment of its basic function. Finally a simple example is given of how the system could be augmented in order to be able to answer natural language questions about the evolution of the action sequence that was input. This constitutes an image grounded human computer interface for multimodal natural language question answering systems. Such an interface could be a test of whether "Cognitive Vision" is achieved. An early system for the generation of visual scenes with a multimodal interface was reported in [11].

5.1 The Input Module

The Input Module accepts a sequence of facts representing images using a predicate herewith named "frame" that constitute abstract descriptions of the images. The "frame" predicate records information concerning the time of taking the image, location of the acting agent and the state of the agent and all other entities of interest.

The following are examples of rules defining the semantics of the primitive verbs "take" and "put" and which are used for the recognition of the occurrence of primitive actions by combining successive image formal descriptions:

THE TAKE RULE:

```
take(A,X,L1,C5):-frame(T1,L,_L1,_),
frame(T2,L,_hand,_),T2=T1+1,L1<>"hand",
entities(A,_X,_),
c(" the ",A,C1),c(C1," took the ",C2),
c(C2,X,C3),c(C3," from the ",C4),c(C4,L1,C5).
```

THE PUT RULE:

```
put(A,X,D):-frame(T1,L,_hand,_),
frame(T2,L,_D,_),T2=T1+1,D<>"hand",!,
entities(A,_X,_),write(" the ",A," put the ",X," at the ",D).
Where :
```

c(X,Y,Z):-concat(X,Y,Z) that constructs the concatenation Z of the strings X and Y.

The positions of the objects in the microcosm are stated as below:

```
position(desk,1).
position(door,3).
position(bookcase,6).
```

The possible states of the door and the book are given by:

state(opened,open).

The states of the agent are classified as stationary and moving. E.g. the stationary states are defined by:

stationary(sitting).

stationary(standing).

5.2 The Complex Verb Recognizer Module

The Complex Verb Recognizer Module uses the semantics of the complex verbs used in the action sequence descriptions expressed in terms of the primitive verbs used for the low level description of the actions depicted by short sequences of images. The following is an example of a rule defining the semantics of complex verb such as "transport" that is used by the Complex Verb Recognizing Module.

THE TRANSPORT RULE

```
transport(A,X,L1,L2):-write(" because "),nl,
take(A,X,L1,C),!,write(C),
write(" and "),nl,put(A,X,L2),
L1<>L2,write(" it follows that "),nl,
write("The ",A," transported the ",X," from the ",L1, " to the ",L2),nl.
```

5.3 The Action Sequence Description Generation Module

The output of our system is a sentence describing briefly the input action sequence and an explanation giving the reasons that support this description using primitive verbs. The operation of this module is closely related to the operation of the complex verb recognizer module and generates a single sentence description together with an explanation that supports the description generation.

6 EXAMPLES USED FOR THE EVALUATION OF THE SYSTEM

6.1 The First Example of Action Sequence Description

A simple example is presented here that was used for the evaluation of the feasibility of our approach. Consider the microcosm of an office environment. A video taken of an agent acting on such an environment may depict the agent approaching a book case in another room taking a book from it and placing the book on her desk. The sequence of images may show the following sequence of actions:

1. The agent is sitting at her desk.
2. The agent is getting up and walking to the door of her room.
3. The agent opens and goes through the door of her room.
4. The agent approaches the bookcase.
5. The agent takes a book from the bookcase.
6. The agent approaches her desk and puts the book on it.
7. The agent sits at her desk and opens the book.

This action sequence may be finally described by the sentence "The agent transported a book from the bookcase on her desk".

The above action sequence is represented first as a set facts using the "frame" predicate as follows:

```
frame(1,1,sitting,closed,bookcase,closed).
frame(2,1,standing,closed,bookcase,closed).
frame(3,2,walking,closed,bookcase,closed).
frame(4,3,walking,closed,bookcase,closed).
frame(5,3,standing,open,bookcase,closed).
frame(6,3,walking,open,bookcase,closed).
frame(7,4,walking,open,bookcase,closed).
frame(8,5,walking,open,bookcase,closed).
frame(9,6,standing,open,bookcase,closed).
frame(10,6,standing,open,hand,closed).
```

```

frame(11,5,walking,open,hand,closed).
frame(12,4,walking,open,hand,closed).
frame(13,3,walking,open,hand,closed).
frame(14,2,walking,open,hand,closed).
frame(15,1,standing,open,hand,closed).
frame(16,1,standing,open,desk,closed).
frame(17,1,sitting,open,desk,open).

```

6.2 The Second Example of Action Sequence Description

The second example concerns the same microcosm as above but with a different action sequence. This action sequence of the second example may be described by the sentence "The agent transported a book from her desk to the bookcase". This action sequence is represented as a set of facts using the "frame" predicate as follows:

```

frame(1,1,sitting,closed,desk,closed).
frame(2,1,standing,closed,hand,closed).
frame(3,2,walking,closed,hand,closed).
frame(4,3,walking,closed,hand,closed).
frame(5,3,standing,open,hand,closed).
frame(6,3,walking,open,hand,closed).
frame(7,4,walking,open,hand,closed).
frame(8,5,walking,open,hand,closed).
frame(9,6,standing,open,hand,closed).
frame(10,6,standing,open,bookcase,closed).
frame(11,5,walking,open,bookcase,closed).
frame(12,4,walking,open,bookcase,closed).
frame(13,3,walking,open,bookcase,closed).
frame(14,2,walking,open,bookcase,closed).
frame(15,1,standing,open,bookcase,closed).
frame(16,1,standing,open,bookcase,closed).
frame(17,1,sitting,open,bookcase,open).
entities(agent,door,book,book).

```

7 QUESTION ANSWERING FROM ACTION SEQUENCES

A future expansion of our system is the implementation of question answering module that may answer questions about the evolution of action in an action sequence. For example when the question "when is the door opened?" the time is given as output. The processing of such a question can be accomplished by rules like:

```

q1:-q(1,Q),f(Q,when,R1),f(R1,is,R2),f(R2,the,R3), f(R3,door,R4),f(R4,QS,""),state(QS,S),
ans(S,T),write("The time is ",T),nl.
ans(S,T):-frame(T,_,_,S,_,_).

```

Where:

f(X,W,Z):-fronttoken(X,W,Z) that puts the first word of X in W and the rest in Z.

The processing of such questions involves the syntactic and semantic analysis of the questions. The semantic analysis is grounded on the input formal representations of the images depicting an action sequence.

8 DESCRIPTION OF VISUALIZATIONS OF BRAIN FUNCTIONS

Using modern technology some cognitive functions of the human brain can now be visualized and observed in real time. One example is the observation of the reading process of a human using a MEG (Magnetoencephalogram) [12]. The MEG is obtained by a system that collects magnetic signals from over 200 points on the skull of a human that are processed by computer to give values for the electrical excitation at different areas inside the brain. These deduced excitations are supposed to correspond to activations of the corresponding point of the brain during the performance of a cognitive function. A strong advantage of a MEG system is its time resolution which is about 4msecs and provides the capability of detailed observations. Some Brain MEG Data from an experiment during which reading and saying a word is performed by a human is given in Table 1. The Data were provided by Prof. Andreas Papanikolaou, University of Texas. These data

result when a human is reading aloud a word projected to him while being monitored by a MEG system. This human is supposed to perform the cognitive functions of first reading silently the word, then processing it for recognition and finally saying it aloud.

N	TIME	X	Y	Z	BRAIN AREA
					VISUAL
1	256.71	-3.23	-3.80	6.92	
2	260.64	-2.90	-3.34	6.71	
3	264.58	-2.60	-3.05	6.51	
4	268.51	-2.31	-2.84	6.33	
5	272.44	-2.03	-2.63	6.18	
6	276.37	-1.78	-2.45	6.04	
7	280.31	-1.53	-2.27	5.86	
8	284.24	-1.28	-2.08	5.61	
9	288.17	-0.85	-1.76	5.08	
10	343.22	-3.23	-2.67	3.13	
11	347.15	-3.26	-2.74	3.25	
12	351.09	-3.64	-3.04	3.18	
13	355.02	-3.80	-3.24	3.09	
14	358.95	-3.80	-3.35	2.97	
					SPEECH
15	370.75	1.51	5.78	6.02	
16	374.68	1.60	5.62	5.82	
17	378.61	1.75	5.64	5.67	
18	382.54	1.83	5.70	5.53	
19	386.48	1.91	5.75	5.40	
20	390.41	2.03	5.85	5.31	
21	394.34	2.18	5.94	5.23	
					MOTION
22	465.12	-1.35	3.74	7.89	
23	469.05	-1.46	3.35	7.35	
24	472.98	-1.50	2.83	6.84	
25	555.56	-0.59	1.57	7.65	
26	559.49	-0.69	1.93	7.60	
27	563.42	-0.65	1.89	7.19	
28	567.36	-0.68	1.85	6.94	

TABLE 1.: The MEG data from one experiment.

Every cognitive action consists of a number of point activations N. The activations of the different brain areas during the above cognitive actions are supposed to be as follows:

The Visual (V) area activated for silent reading of the word.

The Speech (S) area activated for word processing

The Motion (M) area activated for saying the word.

In Table 2 the Brain Areas activated by the Cognitive Actions are shown. The Areas are activated in the order Visual, Speech and Motion

COGNITIVE ACTIONS	N	BRAIN AREA ACTIVATED				Name
		Coordinates				
		X1	Y1	X2	Y2	
Reading	14	0	0	-4	-4	Visual
Processing	7	0	0	+3	+6	Speech
Saying	7	0	0	-2	+6	Motion

TABLE 2. : Brain Areas activated by the Cognitive Actions.

We may use motion verbs to describe the dynamics of the real time visualization of such cognitive phenomena considering the MEG point activations as elementary events. Using a logical representation of these events high level descriptions of the cognitive actions observed can be described in natural language using the system presented in the present paper. Such a description will require the use of an anatomical database that provides the correspondence of the numerical coordinates of the points of activation with the medical names of the anatomical regions of the brain that these points lie. Example descriptions are: "Activation of the speech area follows activation of the vision area" and "Activation of the motion area follows activation of the speech area".

9 CONCLUSION

In the present work we aim at the use of the analysis of the definitions of the Motion Verbs for the application to the understanding and description of action sequences such as those included in a video.

The evaluation of the feasibility of useful performance of our system was presented using two examples of the processing of action sequences and explaining how the output descriptive sentence is generated by the system and an example of work in progress for the description of brain MEG imaging sequences.

A future application of this work could be in the automatic text generation of descriptions of motion images obtained by artificial vision systems. These texts may be helpful for people with vision disabilities.

Finally it had shown how the system could be augmented in the direction of multimodal question answering.

ACKNOWLEDGMENT

We thank Prof. Andreas Papanikolaou, University of Texas for the provision of the MEG data.

REFERENCES

- [1] J. Kontos, I. Malagardi and M. Pegou, "Processing of Verb Definitions from Dictionaries" *3rd International Conference of Greek Linguistics* pp. 954-961, 1997. Athens (in Greek).
- [2] I. Malagardi, "Grouping of Modern Greek Verbs related to Motion using their Definitions" *Journal of Glossologia*, Athens Greece. 11-12 2000. pp. 282-294 (in Greek).
- [3] J. Pustejovsky, C. Havasi, R. Saur, P. Hanks, and A. Rumshisky, "Towards a generative lexical resource: The Brandeis Semantic Ontology" Submitted to LREC 2006, Genoa.
- [4] P. Panagi, S. Dasiopoulou, G.Th. Papadopoulos, I. Kompatsiaris and M.G. Strintzis, "A Genetic Algorithm Approach to Ontology-Driven Semantic Image Analysis" *3rd IEE International Conference of Visual Information Engineering (VIE), K-Space Research on Semantic Multimedia Analysis for Annotation and Retrieval special session*, 2006. Bangalore, India.
- [5] M. Cho, C. Choi and P. Kim, "Measuring Similarity between Trajectories using Motion Verbs in Semantic Level", *ICACT2007*, pp. 511- 515, 2007, Korea.
- [6] H.-H. Nagel, "Steps toward a Cognitive Vision System" *AI Magazine* 25(2), pp.31-50, 2004.
- [7] H. Dahlkamp, H.-H. Nagel, A. Ottlik, P. Reuter, "A Framework for Model- Based Tracking Experiments in Image Sequences. *International Journal of Computer Vision*. 73(2), pp. 139-157, 2007.
- [8] A. Hakeem, M. Shah, "Learning, detection and representation of multi- agent events in videos" *Artificial Intelligence*, 2007 Elsevier, (in press).
- [9] C.J. Fillmore, "The case for CASE", in : E. Bach, R. Harms (Eds), *Universals in Linguistic Theory*, Holt, Rinehart and Winston, New York, pp. 1-88, 1968.
- [10] M. Fleischman, P. Decamp and D. Roy, "Mining temporal patterns of movement for video content classification", *International Multimedia Conference. Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. Poster Session. 2006. Santa Barbara, California USA.
- [11] J. Kontos, I. Malagardi and D. Trikkalidis, "Natural Language Interface to an Agent". *EURISCON '98 Third European Robotics, Intelligent Systems & Control Conference Athens. Published in Conference Proceedings "Advances in Intelligent Systems: Concepts, Tools and Applications"* (Kluwer) pp.211-218, 1998.
- [12] R. Salmelin, "Clinical Neurophysiology of Language: The MEG Approach" (Invited Review), *Clinical Neurophysiology, ELSEVIER Ireland Ltd*, 118, pp. 237-254, 2007.