Natural Language Description of Human Activites from Video Images Based on Concept Hierarchy of Actions

Atsuhiro Kojima and Takeshi Tamura
Library and Science Information Center, Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531, JAPAN

Kunio Fukunaga
Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531, JAPAN

Abstract. We propose a method of describing human activities from video images based on concept hierarchies of actions. Major difficulty in transforming video images into textual descriptions is how to bridge a semantic gap between them, which is also known as inverse Hollywood problem. In general, the concepts of events or actions of human can be classified by semantic primitives. By making correspondence between these concepts and the semantic features extracted from video images, appropriate syntactic components such as verbs, objects, etc. are determined and then translated into natural language sentences. We also demonstrate the performance of the proposed method by several experiments.

Keywords: natural language generation, concept hierarchy, semantic primitive, position/posture estimation of human, case frame

1. Introduction

It is becoming popular to introduce natural language concepts into a vision system. Traffic surveillance system, for instance, represents moving vehicles by series of verbs or short sentences in place of numerical expressions of the objects' location(Kollnig *et al.*, 1994; Nagel, 1994). In their method, the motion of vehicles are estimated from displacement vector field and associated with motion verbs by evaluating attributes of each trajectory segment, such as vehicle speed. Comparing with vehicle movements, description of human behavior in image sequences is more complicated. In a visual surveillance system(Thonnat and Rota, 1999), human behavior is represented by scenarios, i.e. predefined sequences of events. The scenario is then translated into text by filling a template of natural language sentence. Similarly, in an automatic annotation system for a sport scene, each formation of players is represented by belief networks based on visual evidence and temporal constraints(Intille and Bobick, 1998). These works focus on the way rather to express the contents of the image simply than to generate rich expressions using a variety of notion of verbs

On the other hand, natural language has various concepts of actions, events and states inherently; an appropriate verb for an observed event can convey

the meanings of the event effectively. To this end, contribution from artificial intelligence and natural language processing have been introduced into several works. Herzog and Rohr (1985) present three levels of representation from low-level geometrical description estimated from images to high-level textual description. Okada (1980) demonstrate textual explanation of activities of entities in a series of line drawings through simulated mind model in their pioneering works(Okada, 1996). Kitahashi *et al.* (1997) and Babaguchi *et al.* (1996) present schemes for integrating pattern information and natural language notions. They also demonstrate applications to generate an instruction for machine assembly or route guidance for a map from mechanical or geographical information.

For human activities on real video images, we have proposed a method of generating textual descriptions from position and orientation of human head in behalf of whole body posture(Kojima *et al.*, 2000). This may be sufficient for actions without any movement of hands :walking, standing and seeing. The human body, however, is highly articulated and the motion of each body parts can be expressed independently. These sub-activities of human body, such as direction of a line of sight, positions of hands, posture of body and relation to other objects, must be considered simultaneously. In particular, interaction with other objects has relevance to most of human activities in natural language notion.

In this paper, we propose a method of generating textual description which explains human behavior appeared on real video images by extracting semantic features of human motions and making correspondence with concept hierarchy of actions. We finally demonstrate textual surveillance system as an example of the application of our method.

2. Outline of the process

First of all, we will show an outline of the proposed method as shown in Figure 1. At present, it is not so easy to estimate accurate posture and motion of highly articulated object like human in realtime. For this reason, we assume the following three clues which can be obtained by relatively light-weight processes are enough for ordinary cases, e.g. office-work scenes, to detect a posture of a human.

- *Position of head* implies not only a position where the human is but also a posture whether he/she is standing or sitting.
- *Direction of head* implies what he/she is looking at.
- *Positions of hands* imply a sort of gesture and interaction with objects.

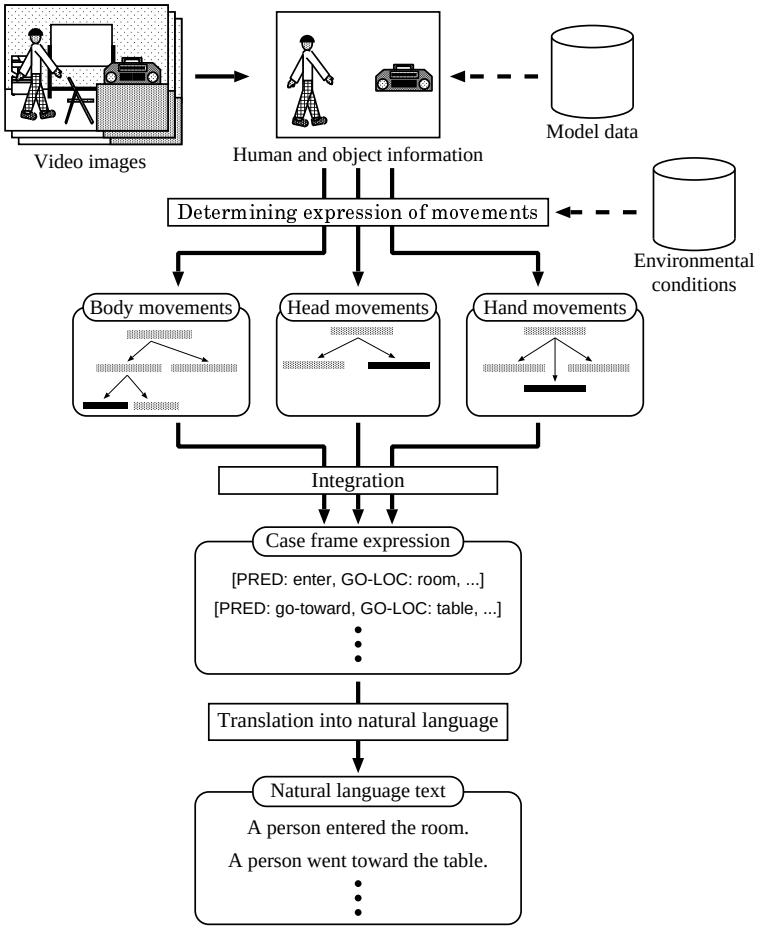


Figure 1. Outline of the proposed method.

For more accurate understanding of interaction with an object such as "pick up a cup" or "put a book on the table," it is necessary to estimate the relative position and motion of the human and the object as well as to identify the object.

For each frame of input video images, the body and skin regions of a human are extracted by calculating difference of colors between input and background images pixel by pixel. Positions of the head and the hands are found by perspective transformation. Orientation of the head is also estimated by calculating correlation between input head region and the head models with multiple aspects which are prepared beforehand.

On the other hand, action of transferring an object in particular is detected in a separate way. Examining shapes of the regions of a human and an object appeared on difference images, it can be verified whether the human bring

and put the object, or pick up and take it out. In addition, the object can be identified by comparing edges and color histograms of extracted object region with those of object models.

Next, conceptual descriptions of actions are generated for each body part by applying domain knowledge, such as allocation of equipment in a room, to the position/posture of the human obtained from the video images. In natural language, concepts of motion verbs include a number of semantic primitives. The meaning of a verb tends to be more concrete and specific as the number of semantic primitives increases. So we construct concept hierarchies of actions for each body parts classified by combination of semantic primitives. By making correspondence between a semantic primitive of action and a feature extracted from the video images, the most appropriate predicate, object, etc. are selected.

We fill these syntactic components into a case frame which is often used as a semantic representation of a sentence in the area of natural language processing. Finally, case frames of body parts are integrated into a frame expressing total body action. Applying syntactic rules and natural word dictionary, the case frame is translated into a natural language sentence.

3. Recognition of human behavior from video images

In this section, we presents a method of recognizing position and orientation of human head, position of hands and interaction with objects from video images.

3.1. Extraction of human head regions

In the area of computer vision, a number of methods have been used to detect human face and hands from video images for gesture recognition. It is, however, hard to detect them from compound background as in office room.

In this paper, we use a method we proposed previously (Asanuma *et al.*, 1999) in which human skin regions can be robustly tracked under compound backgrounds. This method is based on probability of occurrence of colors calculated from both background and skin region pixel by pixel. The distribution of a background pixel is found from pixels on the same position of sequential background images during a certain period where nobody appears, whereas the distribution of skin pixel is found from facial skin regions extracted from many human pictures. The probability that a pixel belongs to background or skin region can then be calculated from the chromaticity values of the pixel and the distributions. Consequently, changes and noises of the background are expected to be absorbed. In addition, inertia of motion of each skin region is also took into account by calculating motion of the region in recent frames.

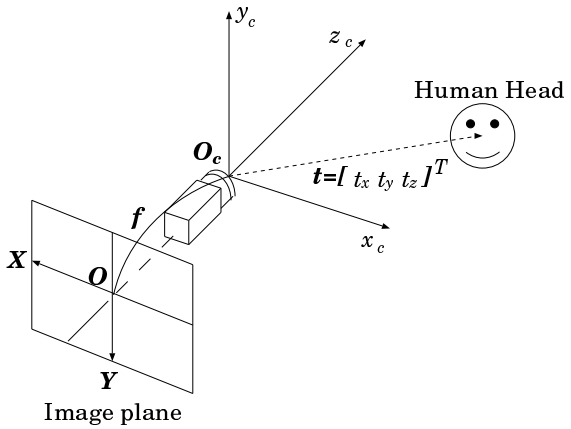


Figure 2. Coordinate system of camera and human.

Using Dempster-Shafer theory(Shafer, 1976), these three kinds of probabilities are integrated into one probability that the pixel belongs to a skin region of the human. The human skin regions are then extracted by thresholding each pixel.

At this time, we assume that there is only one human in a scene. Therefore, the human skin regions must be three at most: the face and both hands. So we regard the highest and/or largest skin region as a facial region. Now, foreground regions are extracted at the same time as the inverse of the background probability. Including foreground pixels around the facial region, i.e. the hair region, a head region are extracted.

Then the position of the head and hands are estimated in the following manner. Figure 2 shows the coordinate systems of a camera and a human. Here $O-XY$ indicates the image plane such that the center of the image captured from the camera is located on the origin. $O_c - x_c y_c z_c$ indicates coordinate system for the camera in which the view point (of the camera) is on the origin. And let $\boldsymbol{t} = [t_x\ t_y\ t_z]^T$ be the position of the human head.

A point $[x_c\ y_c\ z_c]^T$ on $O_c - x_c y_c z_c$ can be computed from projected position X, Y on the image plane, provided that the distance z_c in the depth-direction is given, using perspective transformation as follows:

$$\begin{bmatrix} x_c \\ y_c \end{bmatrix} = \frac{d}{f} \cdot z_c \begin{bmatrix} X \\ Y \end{bmatrix} \tag{1}$$

where d and f are the size of a pixel and the focal length of the camera respectively; both are calibrated beforehand.

Let us assume that the size of human head is known. Then the distance t_z from the camera on z_c axis can be computed from apparent size on the image plane. Now, we can find the head position $\boldsymbol{t}$ by assigning $z_c = t_c$. In the similar

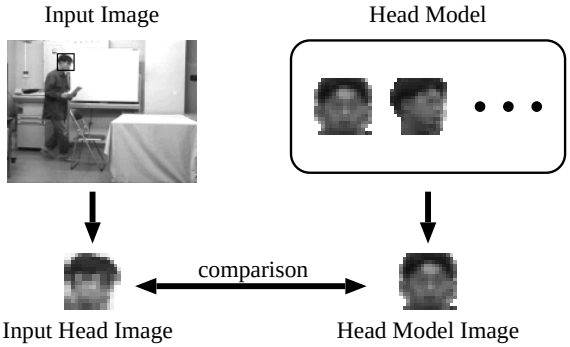


Figure 3. Orientation estimation of head.

way, each hand position can be found provided that a distance of hand equals to t_z.

On the other hand, orientation of the head is estimated by evaluating similarity between human-head image extracted from the input image with pre-recorded images of human-head model with multiple aspects as shown in Figure 3. Let I be an input head image normalized to the same size of model images, and $M_{\phi,\theta}$ be a model image viewed from pitch ϕ, yaw θ. Considering an image as a vector which consists of elements valued by chromaticity of each pixel, we define similarity as a summation of squared difference of each element on the two vectors: I and $M_{\phi,\theta}$. The similarity becomes small when correlation of two images is high; the orientation (ϕ, θ) of a model which shows the smallest similarity is supposed to be the estimated orientation.

3.2. INTERACTION WITH OBJECTS

Most of human actions have relevance to interaction with objects. In this paper, we focus on actions concerning transferring an object, such as picking up or putting an object.

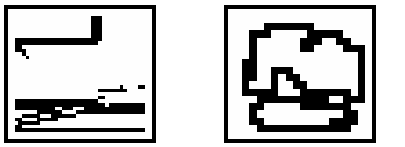
As mentioned previously, the foreground regions, i.e. the difference regions, are extracted out of the background. Let us consider the shapes of regions of a human and an object appeared on the difference image at the instant of occurring of an action. In case that an isolated region other than human region is detected, it can be regarded that either the human placed an object in the scene or took an object out.

Figure 4(a) shows difference images at the instant of appearing of isolated regions; but it is not enough to detect whether the human is picking up an object or putting one. On the contrary, on the original images at the same instant, shown in Figure 4(b), we can see apparent differences; the object is *not* in the circumscribed region when picking, while the object is in when putting. Thus we can distinguish these two types of actions by evaluating a ratio of overlapping edge points between original and difference images

(a) Difference images in which a human is picking up (left) or putting (right) an object.

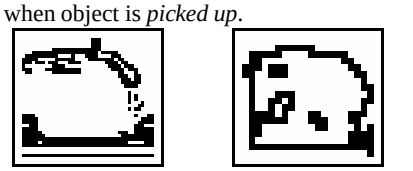
(b) Input image for each image above. Human and object regions are circumscribed in rectangles.

Figure 4. Difference region when object region appears.



input edge image difference edge image

(a) Edge of object region when object is *picked up*.



input edge image difference edge image

(b) Edge of object region when object is *put*.

Figure 5. Edge of object region when object is picked up or put.

shown in Figure 5. In case that the number of overlapping edge points is greater than a threshold, the same object must appear on the both region; we regard the human is putting the object.

Besides, we continually update the background images around object regions; an object region once placed in a scene will be treated as background after a certain seconds.

3.3. IDENTIFYING OBJECTS

Since the object region was extracted in the previous subsection, we apply two-way matching on this region with prepared object models: shape based matching and the color matching. In this paper, we presume several object models commonly used in human activities of usual office work: cup, book, radio, notebook PC, etc.

Shape information of the models consist of edge images from multiple aspects, because the shape of an object may vary by a direction of view. We prepared 4 to 8 aspects for each object. Edge extraction and scaling, preserving aspect ratio, are applied to the object region from the input image. In addition, dimming by Gaussian filter is also applied to avoid subtle gap between edge points. A similarity between edge images are then defined as a number of overlapping edge points divided by the total number of edge points.

Next, as for color matching, we use chromaticity of a^* and b^* from CIE1976$L^*a^*b^*$ UCS. For each pixel of the object region, RGB-color is converted to $L^*a^*b^*$ and accumulated into a histogram of two-dimensional $a^* - b^*$ plane. The histogram is normalized by the number of pixels and compared with that of object models. Thus, the input object is supposed to be identified as the object model which shows the highest similarity.

4. Perceiving human activities

In general, there is a semantic gap between geometric information directly obtained from images and conceptual information contained in natural language. In this section, we first clarify the correspondence between numerical/geometrical information of position and posture of human and concepts of actions of the human. Each concept of an action is expressed in the form of a case frame which consists of syntactic components of a sentence.

We also construct state transition models based on the concept hierarchies in order to deal with dynamically changing position and posture of subjected human.

4.1. Expressing human activities in case frames

Case frame is a kind of frame expression specifically representing the relationship between cases in a natural sentence. According to Fillmore (1968), cases are classified into eight categories: *agent*, *object*, *locus*, *source* and so on. A case frame is then defined as a frame consisting of slots indicating the sort of the cases and their values as follows:

[PRED:walk, AG:person, GO-LOC:by(door), SO-LOC:front(table)]

where PRED, AG, GO and SO indicate a predicate, agentive, goal and source cases respectively. The goal and source are sometimes qualified by LOC, like GO-LOC, for the purpose of qualification of semantic categories: a locus in this example.

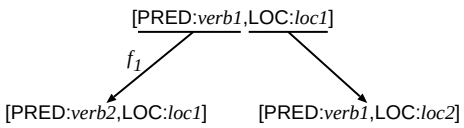


Figure 6. General form of concept hierarchy.

4.2. CONCEPT HIERARCHY OF ACTIONS

A general form of a concept hierarchy of actions is presented in Figure 6, in which each node is expressed by a case frame. Here, a parent node of the form [PRED:$verb_1$, LOC:loc_1] derives two child nodes. One is the reification about the meaning of verb: A verb $verb_2$ is derived from $verb_1$ by adding a new geometric/semantic feature f_1. For example, a verb 'move' derives more concrete verb 'walk' if the feature 'move_slow' is added, which indicates the speed of movement is relatively slow. In addition, accompanying cases may also be added according to verbs.

The other is reification about the coverage of locus: the locus loc_1 in a parent frame is narrowed to loc_2 in a child by adding information about location. In other words, the relation of loc_1 and loc_2 is inclusive; i.e. $loc_1 \supset loc_2$. For example, a locus indicating merely 'in(room)' (in a room) may be replaced with more specific denotation 'by(table)' (by a table).

The features we used here can be computed from geometric/numerical information from images and the given knowledge about the environment of a scene. Similarly, in natural language processing, the componential features about meaning of verbs are called *semantic primitives*. We consider semantic primitive as semantic feature extractable from images. Table I shows semantic primitives about the concepts of actions of human.

Basically, the meaning of an expression goes more specific as the further semantic primitives are extracted. Some of these semantic primitives, however, can not be verified simply. For example, a semantic primitive 'move_near(*ag*, *obj*)' which means an agent's approach to an object will be computed from the degree of decrement in relative distance between the agent and the object in a unit period. This can be compared numerically with conflicting siblings in a common criterion. Thus we apply Logistic function to evaluate feature value x shown in Table I. The Logistic function, defined as equation (2), has a sigmoid shape; a range of each feature value is $[0, 1]$, and the more apparent the feature is indicated, the closer to 1 the value is.

$$f(x) = \frac{1}{1 + Ae^{-Bx}} \tag{2}$$

Here A and B are constants relevant to an offset and inclination; these must be determined empirically for effective feature selection.

Table I. Semantic primitives.

semantic primitive	**feature value** (x)	**meaning**
semantic primitives about movements		
move(loc)	$\|\Delta horiz(loc)\|$	movement
move_slow(loc)	$\|\Delta horiz(loc)\|$	move slowly
move_fast(loc)	$\|\Delta horiz(loc)\|$	move fast
move_near(ag,obj)	$\Delta\|loc_{ag} - loc_{obj}\|$	move toward obj
move_apart(ag,obj)	$\Delta\|loc_{ag} - loc_{obj}\|$	move away from obj
...		
semantic primitives about states		
high(loc)	$vert(loc)$	loc is high
low(loc)	$vert(loc)$	loc is low
face(ag, obj)	$dir_{ag} \cdot dir_{ag-obj}$	ag faces obj
near(ag,obj)	$\|loc_{ag} - loc_{obj}\|$	ag is near to obj
...		
semantic primitives about attributes		
prop_operable	–	operable
prop_readable	–	readable
prop_loadable	–	loadable/supportable
...		

From this, a case frame with the highest feature value is selected out of conflicting siblings. In case that a node has unity child frame, it will be adopted if the values are higher than a threshold. As far as multiple objects are concerned such as 'move_near', all the candidates are evaluated and the highest object is selected.

4.3. GENERATING CASE FRAMES FOR EACH BODY PART

In the following, we present concept hierarchies for several body parts: body, head and hand. Figure 7 shows a concept hierarchy of body actions. The root node with verb of 'be' as a predicate expresses a human merely exists in a scene. If a semantic primitive 'move' is added, the predicate will be replaced with the verb 'move'. Similarly, some predicates are considered: walk, stand, sit, etc. As for a locus 'in(room)' in the root frame, it will be replaced with more specific denotation such as 'by(door)', (by the door), 'front(table)' (in front of the table), etc.

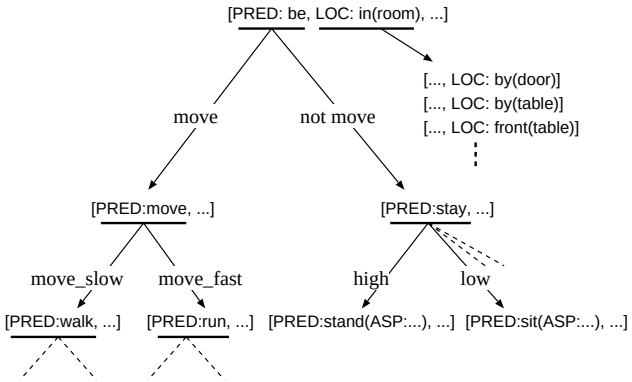


Figure 7. Concept hierarchy of body actions.

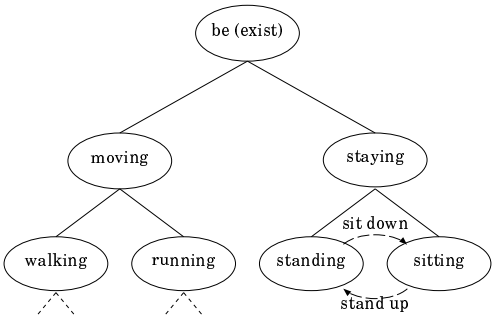


Figure 8. State transition diagram of body actions.

In practice, however, to follow the rapidly changing states of human, a dynamic model is need. As shown in Figure 7, there are two types of verbs: one is a *durative action verb* such as 'walk', and the other is an *instantaneous action verb* such as 'stand up'. By making each type of verb correspondent to a state and a state transition respectively, we constructed a hierarchical state-transition diagram (STD) as shown in Figure 8. Here, a solid line represents reification associated with a semantic primitive, and a dashed arrow represents state transition.

The following is an algorithm to generate case frames using the STD:

1. Let s be a current state. For newly estimated position/posture of human, each semantic primitive are evaluated along solid lines downward from the top of the STD, then the new state s' is determined.

2. In case the state s and s' are identical, no case frames are generated, let s' be s and go back to step 1.

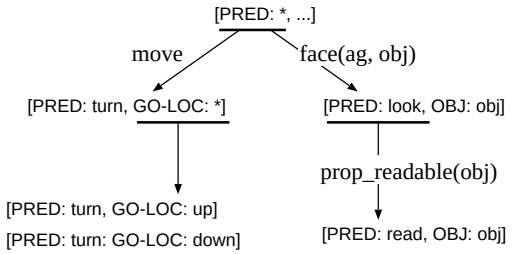


Figure 9. Concept hierarchy of head actions.

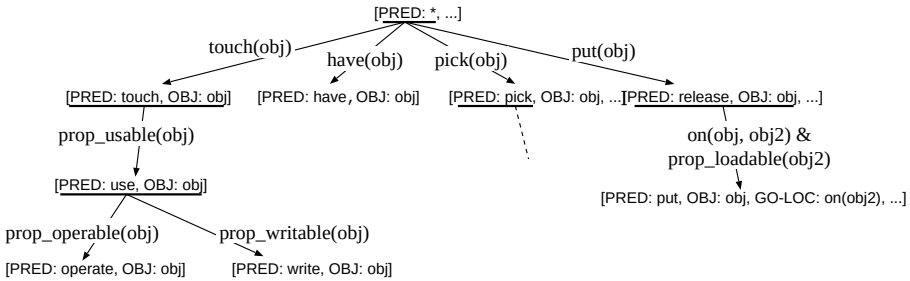


Figure 10. Concept hierarchy of hand actions.

3. Otherwise, a state transition will occur; if there is a state transition specified by a dashed arrow in the STD on the path $s \rightarrow s'$, a case frame with a verb associated with the transition is generated.

4. a case frame with a verb associated with the new state s' is generated, let s' be s and go back to step 1.

As for instantaneous actions, a resultant state must be specified explicitly in some cases: ‘be sitting’ or ‘be standing’ after ‘sit down’ or ‘stand up’. It is important for text generation to distinguish these aspects. They are specified in a case frame like ‘stand(ASP:resultant)’.

In the same manner as this, a case frame for a head action is generated. Figure 9 shows a concept hierarchy of head actions. Orientation of head is thought to be important because it indicates an object or something in which a human has interest. The root node is abstract and represents nothing. The predicate will be replaced with ‘turn’ if the human turns his/her head. Simultaneously, a goal case may be added to specify some obscure directions, such as ‘up’ or ‘down’. If an object is found on his/her line of view, the predicate will be replaced with ‘look’ and the object is specified as a goal case. In addition, in case that the object has special property ‘prop_readable(*obj*)’, which indicates object is something readable like books, the predicate will then be reified as ‘read’.

Finally, a concept hierarchy of hand actions is shown in Figure 10. Many of human actions are related with hand usage, especially handling objects.

There are various expressions in natural language depending on the attribute of objects and the direction of transfer.

In this method, the verbs of 'pick' and 'release' as stated in 3.2 are derived from the root node. As for 'release', it will be reified to 'put ... on ...' if the transfered object is placed on another object with *loadable* property like a table. The verb 'touch' may be generated if the positions of a hand and an object are overlapped. In this case, the predicate may also be replaced with 'operate' if the object have special property of 'prop_operable' like notebook PC.

4.4. BUILDING WHOLE BODY EXPRESSION

So far, we obtained three case frames from each body part. Now we will integrate them into unity case frame of whole body action. However, since these frames may be uneven in concreteness, it is not always appropriate to merging them directly. So we introduce three patterns of integration rules as follows:

1. Taking the most essential expression in behalf of other frames.

   Ex. sit + read a book $\Rightarrow$ read a book

2. Merging two expressions in parallel.

   Ex. look at a clock + walk

   $\Rightarrow$ walk ..., looking at a clock

3. Emerging a new concept from two concepts.

   Ex. have a notebook PC + move

   $\Rightarrow$ carry a notebook PC

For most of cases we suppose, the pattern 1 in which the most essential expression is selected in behalf of other frames may be suitable to natural expression. Therefore, we apply the pattern 1 by default. For specific combination of concepts of actions, other rules are applied. Table II shows the typical rules for these patterns.

5. Generating textual description

The case frame is translated into natural language sentence using case structures and verb patterns (Nishida and Takamatsu, 1982; Nishida *et al.*, 1988). A case frame represents relationship between a verb and other cases such as object, source and goal. In English, the order of phrases is dominated by the

Table II. Rules for integrating case frames of body parts.

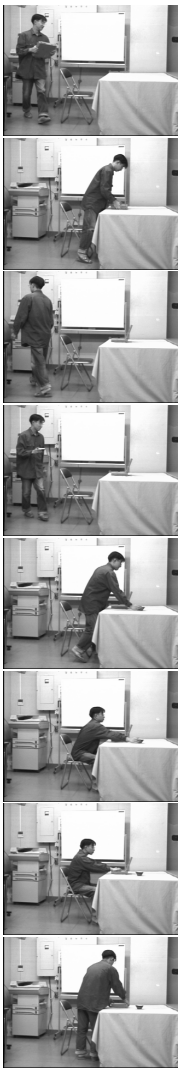
integration by pattern 1
body:[PRED:stand(ASP:resultant), AG:*ag*, LOC:*loc*, . . .] +
hand:[PRED:pick, OBJ:*obj*]
⇒ [PRED:pick, AG:*ag*, OBJ:*obj*, LOC:*loc*, . . .]
. . . . .
integration by pattern 2
body:[PRED:move, AG:*ag*, . . .] +
head:[PRED:look, OBJ:*obj*]
⇒ [PRED:move, AG:*ag*,. . . SUB:[PRED:look, OBJ:*obj*]]
. . . . .
integration by pattern 3
body:[PRED:move, AG:*ag*, . . .] +
hand:[PRED:have, OBJ:*obj*]
⇒ [PRED:carry, AG:*ag*, OBJ:*obj*, . . .]
. . . . .

verb patterns classified by Hornby (1975). For example, the intransitive verb 'walk' (VP2) is necessarily accompanied with only an agentive case, while goal and/or source case is optional: *Ex.* "A person walked". On the other hand, the verb 'put' (VP15) is necessarily accompanied with an agentive, an object and a goal cases: *Ex.* "A person put a book on a table". In this manner, according to the verb pattern, the order of verbs, adverbs and other syntactic components are determined.

Here, we considered the past tense is suitable for the situation that the text is generated one after another in realtime. For the resultant aspect of instantaneous actions, the progressive form is used exceptionally.

6. Results and discussions

In order to test the ability of our method of text generation, we have implemented a prototype system to surveillance of human activities in a machine room of our laboratory and to report in textual form. We suppose 30^+ verbs for in-door human activities and 9 object models. The allocations of a door, a table and other equipment are also given in advance. Our system is built by C++ and Prolog on a PC with dual PentiumII 266MHz. Its performance is approximately 2 frame/sec for processing images of 160×120 pixels. In practice, we first recorded a scene and then replayed it in 1/3 of normal speed so as to process at the effective rate of 6 frame/sec.



There is no one.
A/The person entered a/the room.
A/The person is standing in front of a/the door.
A/The person went near a/the table.
⋮
A/The person is standing in front of a/the table.
A/The person put a/the notebook PC on a/the table.
⋮
A/The person went toward a/the door.
A/The person exited a/the room.
There is no one.
A/The person entered a/the room.
⋮
A/The person released a/the cup.
A/The person is standing by a/the table.
A/The person is standing in front of a/the table.
A/The person sat down in front of a/the table.
⋮
A/The person picked up a/the cup.
A/The person is sitting in a/the room, having a/the cup.
A/The person operated a/the notebook PC.
A/The person released a/the cup.
A/The person operated a/the notbook PC in a/the room.
⋮
A/The person stood up in a/the room.
⋮
A/The person picked up a/the notebook PC.
⋮
A/The person carried a/the notebook PC to a/the door.
A/The person exited a/the room.
There is no one.
⋮

Figure 11. Input images and the generated text.

We performed experiments on 10 different scenes: in each scene, typically, a human brings an object, use it for a while and took it out. Figure 11 shows the results of generated text. From this, we can see that most of the generated text are almost suitable to explain actions of the human. Table III shows the results of selected verbs. The selection of verbs are performed well except for the verb ‘operate’; instead of this, the verb ‘use’ is selected. One reason

Table III. Results of generated verbs.

verb	appropriate verb (%)
stand up	93.8
sit down	100.0
pick up	87.5
release	79.4
put ... on	69.9
operate	20.0

for this is that the position of the actor's hand was not estimated correctly in practice during the action, so that a semantic primitive essential to the verb 'operate' could not be extracted.

We have introduced concept hierarchy of actions; each node is represented by a case frame and hierarchically allocated by notion of verbs and coverage of locus. As shown by experimental results, we confirmed that appropriate verb is selected according to semantic features extracted from images in most cases. For example, in our experiments, all the objects are put on the table. In 88% of the trials, a semantic primitive 'on(*obj*, table)' is correctly extracted, which represents the relationship between the object and the table. The resulted sentences are like 'put ... on the table'. The rest of the trials, the resulted sentences are merely 'release ...' because of lack of the relationship between the object and the table. We should note that it is not always the best selection for the event but still seems to be good. In other words, appropriate verbs can be selected effectively even if some of the semantic primitives are failed to extract. It is owing to concept hierarchy of actions that appropriate verbs can be selected in spite of lack of semantic primitives.

Other cases of failure are due to instability on position/posture estimation of a human or recognition of objects. This is rather technical issue on the process of perceiving image features before understanding the meanings of the image. In fact, an error of position of human head or hand may be relatively large in the depth direction because we used monocular vision in this experiment. In this paper, we focused on the transformation of geometric information into symbolic/conceptual information; this may be improved if another technique such as stereo vision is applied.

7. Conclusions

We clarified the way of transforming video images represented as geometrical/numerical information into textual descriptions as conceptual information. We applied concept hierarchy of actions to extraction of meanings of images by verifying correspondence between semantic features of human actions and the natural language concepts. Consequently, appropriate verbs and objects can be selected in a sophisticated way.

Acknowledgements

We are grateful to Dr. Tadahiro Kitahashi of Osaka University for useful discussion and advice. We are also grateful to Katsunori Asanuma for helping experiments.

References

Asanuma, K., Onishi, M., Kojima, A., and Fukunaga, K.: 'Extracting Regions of Human Face and Hands Considering Information of Color and Region Tracking', *Trans. IEEJ(C)*, 119-C(11):1351–1358, 1999 (in Japanese).

Babaguchi, N., Dan, S., and Kitahashi, T.: 'Generation of Sketch Map Image and Its Instructions to Support the Understanding of Geographical Information', In *Proc. of ICPR '96*, pp.274–278, 1996.

Fillmore, C. J.: 'The case for case', In E. Bach and R. Harms editors, *Universals in Linguistic Theory*, pp.1–88, Rinehart and Wiston, New York, 1968.

Herzog, G. and Rohr, K.: 'Integrating Vision and Language: Towards Automatic Description of Human Movements', In *Proc. 19th Annual German Conf. on Artificial Intelligence*, pp.257–268, 1995.

Hornby, A. S.: *Guide to Patterns and Usage in English*, Oxford Univ. Press, London, 1975.

Intille, S. and Bobick, A.: *Representation and Visual Recognition of Complex, Multi-agent Actions using Belief Networks*, Technical Report 454, M.I.T Media Lab. Perceptual Computing Section, 1998.

Kitahashi, T., Ohya, M., Kakusho, K., and Babaguchi, N.: 'Media Information Processing in Documents – Generation of Manuals of Mechanical Parts Assembling –', In *Proc. of 4th Int. Conf. on Document Analysis and Recognition*, Ulm, Germany, pp.792–796, 1997.

Kojima, A., Izumi, M., Tamura, T., and Fukunaga, K.: 'Generating Natural Language Description of Human Behavior from Video Images', In *Proc. of ICPR 2000 Vol.4*, pp.728–731, 2000.

Kollnig, H., Nagel, H.-H., and Otte, M.: 'Association of motion verbs with vehicle movements extracted from dense optical flow fields', In *Proc. of 3rd European Conf. on Computer Vision '94*, pp.338-347, 1994.

Nagel, H.-H.: A Vision of 'Vision and Language' Comprises Action: An Example from Road Traffic, *Artifitial Intelligence Review*, 8:189–214, 1994.

Nishida, F. and Takamatsu, S.: 'Japanese-English Translation through Internal Expressions', In *Proc. of COLING-82*, pp.271–276, 1982.

Nishida, F., Takamatsu, S., Tani, T. and Doi, T.: 'Feedback of Correcting Information in Postediting to a Machine Translation System', In *Proc. of COLING-88*, pp.476–481, 1988.
Okada, N.: 'Conceptual taxonomy of Japanese verbs for understanding natural language and picture patterns', In *Proc. of COLING-80*, pp.123–135, 1980.
Okada, N.: 'Integrating vision, motion and language through mind', *Artifitial Intelligence Review*, 8:209–234, 1996.
Shafer, D.: *A Mathematical Theory of Evidence*, Princeton Univ. Press, 1976.
Thonnat, M. and Rota, N.: 'Image Understanding for Visual Surveillance Applications', In *Proc. of 3rd Int. Workshop on Cooperative Distributed Vision*, pp.51–82, 1999.

Address for Offprints: Library and Science Information Center, Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531, JAPAN
E-mail: ark@center.osakafu-u.ac.jp
Tel: +81-722-54-9155
FAX: +81-722-54-9940