Subjective Explanation of Pictures Considering Feelings of Objects in Pictures

Shigeru Kato*, and Takehisa Onisawa**

* Onisawa Lab., Institute of Engineering Mechanics and Systems, University of Tsukuba, Japan
(Tel : +81-298-53-5060; E-mail: shigeru@fhuman.esys.tsukuba.ac.jp)
** Institute of Engineering Mechanics and Systems, University of Tsukuba, Japan
(Tel : +81-298-53-5060; E-mail: onisawa@esys.tsukuba.ac.jp)

Abstract: This paper aims at the construction of the system that outputs subjective and consistent linguistic expressions of some pictures. Input to this system is information on some pictures and this system outputs explanations of basic contents of pictures and consistent connective relationships between pictures. The present system consists of the explaining part of basic contents and the explaining part of connective relationship. The former part explains behaviors and feelings of objects drawn in pictures. The latter one explains the matters not drawn in pictures by guessing them from pictures. This part considers consistency of connective relationships. From the viewpoint that the interpretation of pictures is dependent on individual subjectivity, the present system has individual database for individual subjectivity. In order to confirm the usefulness of the present approach, simulation experiments are performed. In the experiments the individual databases of subjects are constructed and the outputs of the system are evaluated. Experiment results show that good evaluation is obtained.

Keywords: image understanding, subjectivity, consistent explanation of pictures, case-based reasoning

1. INTRODUCTION

Since a computer was invented, many studies on natural language understanding by a computer have been performed, and various methods have been proposed in the field of cognitive science and artificial intelligence [1]. Context analysis and understanding of matters not expressed in text are necessary for text understanding. In the field of artificial intelligence, there are many researches on story understanding or story generation. These studies mainly aim at story understanding by the analysis of sentence structures or story generation by the use of the planning techniques [2] [3]. On the other hand, when human understands and/or makes stories, human has visual images of stories. Therefore, it is important to consider visual image in order to realize story understanding and generation on a computer. This paper aims at the construction of the system that outputs subjective and consistent linguistic expressions of pictures when some pictures are given in any order.

From the viewpoint that the interpretation of pictures is dependent on individual subjectivity, the present system has individual database for individual subjectivity. Although our previous study [4] mainly describes only the object's behavior as the explanation of pictures, the present system explains the feelings of objects as well as object's behavior. Such a system is applicable to story creation [5][6] or to the development of automatic explanation system of pictures for eye-handicapped persons. In order to confirm the usefulness of the present approach, simulation experiments are performed. In the experiments the individual databases of subjects are constructed and the outputs of the systems are evaluated.

2. STRUCTURE OF SYSTEM

The input to the system is information on some pictures called *picture information* [7]. System outputs are linguistic expressions explaining pictures contents called *state description* and linguistic expressions explaining the connection between pictures called *event description*. As shown in Fig.1, this system consists of a *basic contents explanatory part* explaining contents of respective pictures, and a *connective relationship explanatory part* explaining the consistent connection between pictures.

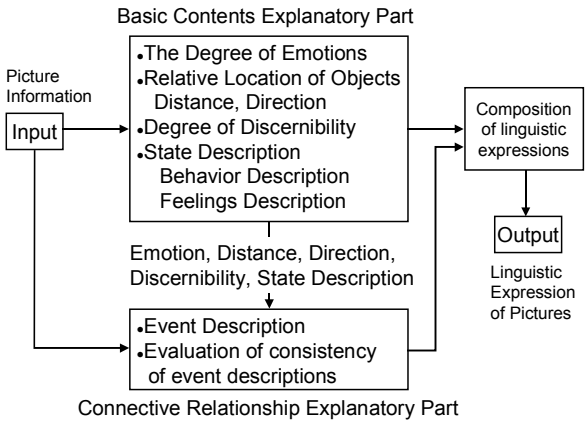

Fig. 1 Structure of system

3. PICTURE INFORMATION

It is assumed that numerical information on positions and sizes of objects in a picture is obtained beforehand by ideal image processing because this study focuses on not image recognition but image understanding. Picture information is objective information on objects in a picture as shown in Table 1 [7], where if an object is not human, the facial expression EyeSize, EyeShape, EyebrowSlant, EyebrowShape, and MouthSize are not considered.

Table 1 Picture information

Item	Meaning
Name	Name of an object
Location	Position of an object
Size	Size of an object
Direction	Direction of object's body
FaceDirection	Direction of object's face
EyeSize	Size of eyes
EyeShape	Shape of eyes
EyebrowSlant	Slant of eyebrows
EyebrowShape	Shape of eyebrow
MouthSize	Size of mouth
MouthShape	Shape of a mouth

4. BASIC CONTENTS EXPLANATORY PART

As shown in Fig.2, the basic contents explanatory part processes information on each object in a picture, which includes the degree of emotions recognized from a face, the distance between two objects, the direction of an object toward other object or the degree that an object discerns another. These pieces of information are dealt with as fuzzy numbers numerically [7].

This part infers object's behavior and feelings toward another object by the case-based reasoning method searching cases similar to information obtained from the picture information. For example, "A boy rides the boat" and "A boy likes the girl" are the explanation of the boy's behavior and that of the boy's feelings toward the girl, respectively. The cases about object's behavior and feelings descriptions are preserved in the behavior description case database and the feelings description case database, respectively, and are increased by a user's adding a new case to the database. The behavior description and the feelings description are called the state description.

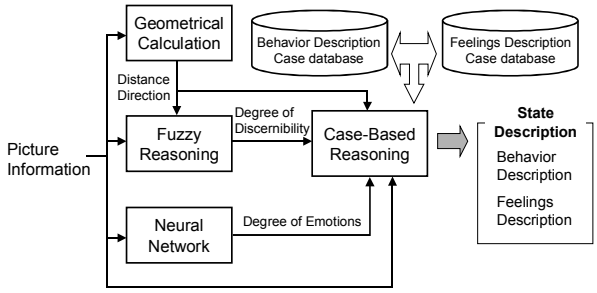

Fig. 2 Basic contents explanatory part

5. CONNECTIVE RELATIONSHIP EXPLANATORY PART

In the connective relationship explanatory part, the explanation of the connection between pictures is generated. As shown in Fig.3 when three pictures are inputted to the system, event descriptions between the first and the second pictures, and those between the second and the third pictures are obtained independently by case-based reasoning. The event description is the explanation of matters not drawn in pictures by guessing to happen between pictures. The event descriptions are obtained by searching cases similar to the outputs of the basic contents explanatory part for two pictures. Event description cases are preserved in the event description case database and are increased by a user's adding a new case to the database [7].

When some event descriptions between pictures are obtained, consistencies are evaluated for all combinations of event descriptions. Only the consistent combinations of event descriptions are outputted as the event descriptions of pictures.

6. EVALUATION OF CONSISNTENCY

As shown in Fig.4, in the connective relationship explanatory part the evaluation of the consistency between event descriptions (A,B) is performed by considering correspondences between the final state of an object in event description A and the initial state of the one in event description B, where the initial state is the object's state before the event in event description B and the final state is the one after the event in event description A [4].

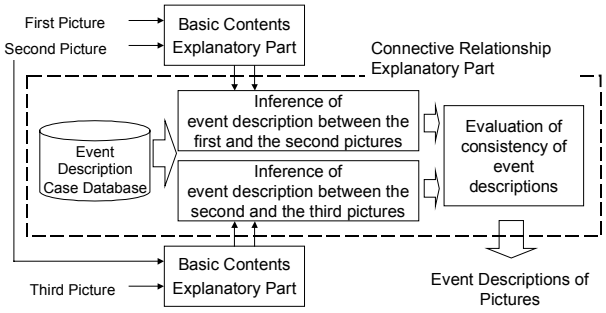

Fig. 3 Connective relationship explanatory part

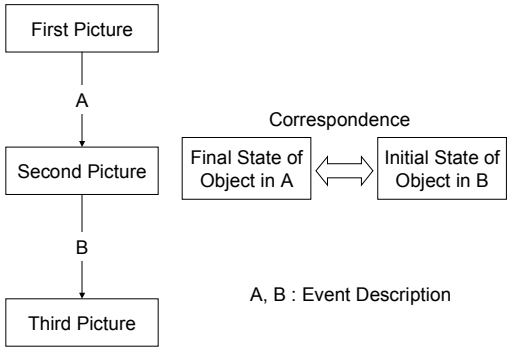

Fig. 4 Evaluation of consistency

6.1 Object's state

The object's final and initial states are expressed by the body state and position. Users give this piece of information when they add a new event description case to the event description case database. This piece of information is expressed as the slot of the event description case.

(i) Body state

Let the set of objects in event E be O_E. And let the final body state and the initial body state of object $o_i \in O_E$ be $H_E^I(o_i)$ and $H_E^F(o_i)$, respectively. $H_E^I(o_i)$ and $H_E^F(o_i)$ are expressed by one of the following expressions.

- An object can act now.
- An object can't act now but can do soon.
- An object can't act.

For example, when an object is sleeping or unconsciousness, the body state of the object is expressed by "An object can't act now but can do soon", and when an object is dead, the body state is expressed by "An object can't act".

(ii) Position

The initial and final positions of object o_i $(\in O_E)$, $P_E^I(o_i)$ and $P_E^F(o_i)$ are represented by formula (1).

$$P_E^I(o_i) = \bigcup_{o_j \neq o_i, o_j \in O_E} \{(L_E^I(o_i, o_j), o_j)\}$$
$$P_E^F(o_i) = \bigcup_{o_j \neq o_i, o_j \in O_E} \{(L_E^F(o_i, o_j), o_j)\} \quad (1)$$

where $(L_E^I(o_i,o_j),o_j)$ and $(L_E^F(o_i,o_j),o_j)$ are binomial terms of object o_j and the initial relative position between o_i and o_j, $L_E^I(o_i,o_j)$ and object o_j and the final relative position between o_i and o_j, $L_E^F(o_i,o_j)$, respectively, where $o_j \in O_E$. $L_E^F(o_i,o_j)$ and $L_E^I(o_i,o_j)$ are expressed by linguistic labels prepared beforehand such as *near, far from, inside of, outside of, on, under, in front of* and *behind*. Since the position of o_i is expressed in the form of the relative position to o_j, the number of positions of o_i taken into consideration is the same as the number of objects o_j. Therefore $P_E^I(o_i)$ and $P_E^F(o_i)$ are expressed by the union of $(L_E^I(o_i,o_j),o_j)$ and the union of $(L_E^F(o_i,o_j),o_j)$, respectively.

For example, let us consider event description E = *"A boy puts an apple on a box"*. In this example, O_E is $\{boy, apple, box\}$ and the initial position and the final position of each object in Event E are expressed as follows.

$$P_E^I(boy) = \{(near, apple), (near, box)\}$$
$$P_E^I(apple) = \{(near, boy), (near, box)\}$$
$$P_E^I(box) = \{(near, boy), (near, apple)\}$$
$$P_E^F(boy) = \{(near, apple), (near, box)\}$$
$$P_E^F(apple) = \{(near, boy), (on, box)\}$$
$$P_E^F(box) = \{(near, boy), (under, apple)\}$$

6.2 Evaluation of consistency

The consistency of the connective relation of event descriptions (A,B) is evaluated by the degree of the correspondence between the final state in event description A and the initial state in event description B. The degree is calculated by formula (2). The value of $f((A,B))$ expresses the consistency degree. If $f((A,B)) \geq 0.6$, then the connective relation of (A,B) is evaluated to be consistent and event descriptions (A,B) are outputted as the explanation of pictures. If $O_A \cap O_B = \phi$, $f((A,B))$ is not calculated and event descriptions (A,B) are outputted.

$$f((A,B)) = \min_{\substack{o_i \in O_A \cap O_B \\ O_A \cap O_B \neq \phi}} h(H_A^F(o_i), H_B^I(o_i)) \times g(P_A^F(o_i), P_B^I(o_i)) \quad (2)$$

where

$f((A,B)) \in [0,1]$

O_A : the set of objects in event A

O_B : the set of objects in event B

$$h(H_A^F(o_i), H_B^I(o_i)) = \begin{cases} 1 & (H_A^F(o_i) = H_B^I(o_i)) \\ 0 & (H_A^F(o_i) \neq H_B^I(o_i)) \end{cases} \quad (3)$$

$$g(P_A^F(o_i), P_B^I(o_i)) = \begin{cases} \min\limits_{(x,y) \in P_A^F(o_i) \times P_B^I(o_i)} k((x,y)) \\ \qquad (P_A^F(o_i) \neq \phi, P_B^I(o_i) \neq \phi) \\ 1 \qquad (P_A^F(o_i) = \phi \text{ or } P_B^I(o_i) = \phi) \end{cases} \quad (4)$$

$h(H_A^F(o_i), H_B^I(o_i))$ in formula (3) expresses the degree of correspondence between the final body state of o_i in event description A, $H_A^F(o_i)$ and the initial body state of o_i in event description B, $H_B^I(o_i)$. And $g(P_A^F(o_i), P_B^I(o_i))$ in formula (4) expresses the degree of correspondence between $P_A^F(o_i)$ and $P_B^I(o_i)$. $k((x,y)) \in [0,1]$ expresses the degree of correspondence between x and y which are the elements of $P_A^F(o_i)$ and $P_B^I(o_i)$, respectively. $k((x,y))$ is calculated by formula (5).

$$k((x,y)) = \begin{cases} 1 & (x = y) \\ S((x,y)) & (x \neq y) \\ \overline{S} & (x \neq y \text{ and } \textit{no corresponding cases}) \end{cases} \quad (5)$$

where $S((x,y))$ is the degree of correspondence between x and y preserved in the case database of degree of correspondence between positions. Examples of the case database are shown in Fig.5. When there is no case corresponding with (x,y) in the database, $k((x,y))$ takes $\overline{S}$, the average of all degrees of correspondence of similar cases, where the similar case of (x,y) is the following case. Let (x,y) be expressed in the form of $((L_1,o_1),(L_2,o_2))$. (1) L_1 and L_2 are the same linguistic labels of positions as those of x and y (2) o_1 is the same object as that of x or they are in the same group (3) o_2 is the same object as that of y or they are in the same group. The group of objects is given in the thesaurus as shown in Fig.6. When there is no similar case, $k((x,y))$ takes 1.0 and a new case of $S((x,y)) = 1.0$ is added to the database.

Position1	Position2	Degree of Correspondence
(on, box)	(under, box)	0.0
(near, tree)	(near, car)	0.7
…	…	…
(inside of, boat)	(outside of, boat)	0.0

Fig. 5 Examples of case database of degree of correspondence between positions

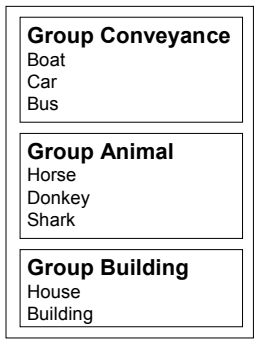

Fig. 6 Thesaurus

If there is only one object o_i in event description A or B, it is assumed that $g(P_A^F(o_i), P_B^I(o_i))$ takes 1.0.

6.3 Learning of position cases

This system interacts with users in order to obtain the proper degree of correspondence between positions in the database.

When three pictures are inputted to the system, the system outputs all combinations of event descriptions (A,B) and then users evaluate them whether the interpretation of the position in the second picture is consistent or not. The evaluation is performed with a 3-point scale, *extremely inconsistent*, *a little inconsistent*, and *not inconsistent*. The degree of correspondence is modified according to the users evaluation for case (x, y) satisfying $(x, y) \in P_A^F(o_i) \times P_B^I(o_i)$ and $o_i \in O_A \cap O_B$. The modification of the degree of correspondence is defined by formula (6).

$$\textit{Modification value} = \begin{cases} -0.4 & (\textit{extreamly inconsistent}) \\ -0.3 & (\textit{a little inconsistent}, S((x,y)) < 0.6) \\ -\dfrac{0.1}{S((x,y))^2} & (\textit{a little inconsistent}, S((x,y)) \geq 0.6) \\ +\dfrac{S((x,y))^2}{3} & (\textit{not inconsistent}) \end{cases} \quad (6)$$

7. COMPOSITION OF LINGUISTIC EXPRESSIONS

The explanations of pictures are the state descriptions obtained from the basic contents explanatory part and the event descriptions obtained from the connective relationship explanatory part. Fig.7 shows the order of linguistic expressions in the explanations of three pictures.

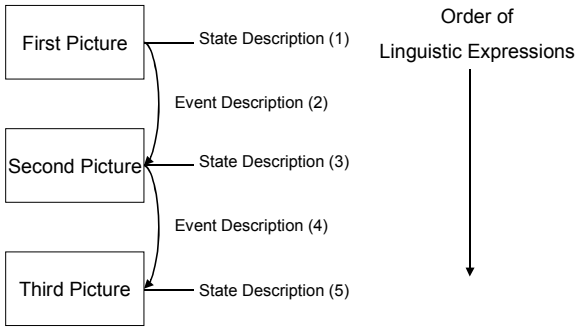

Fig. 7 Order of linguistic expressions

When many state descriptions are obtained in a picture, these descriptions are sometimes redundant. For example, let us consider the state descriptions in Fig.8. "A boy is talking with a girl" and "A girl is talking with a boy" can be expressed in a simple expression "A boy and a girl are talking". "A cat is looking at a boy" and "A cat is looking at a girl" also can be expressed in a simple expression "A cat is looking at a boy and a girl". It is necessary to paraphrase such a redundant linguistic expressions for generating natural linguistic expression [8]. This system changes the redundant state descriptions into a simple expression.

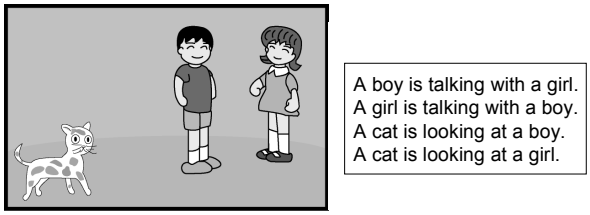

Fig. 8 Example of the state descriptions

Let the state description be expressed in the form of (S,V,O), where S is the subjective object, V is the verbal expression of the behavior or feelings of S and O is the target object of V. The paraphrase is performed according to the following rules.

(i) If the same S and V or the same O and V are used in two descriptions, the two descriptions are paraphrased to one description according to the following rules.

(A, V, **B**) + (A, V, **C**) → (A, V, **B and C**)
(**B**, V, A) + (**C**, V, A) → (**B and C**, V, A)

For example, the state descriptions "A cat is looking at a boy" and "A cat is looking at a girl" are paraphrased to "A cat is looking at a boy and a girl".

(ii) If S and O in one description correspond to O and S in another description, two descriptions are paraphrased to one description according to the following rule.

(A, V, B) + (B, V, A) → (A *and* B, V)

For example, the state descriptions "A boy is talking with a girl" and "A girl is talking with a boy" are paraphrased to "A boy and a girl are talking".

8. EVALUATION OF SYSTEM

Since the interpretation of pictures is dependent on individual subjectivity, it is necessary that the system outputs the explanations of pictures reflecting individual subjectivity. This paper considers the individual database for individual subjectivity. In the experiments individual databases are constructed and outputs of the system obtained by the simulation experiments are evaluated whether it is comprehensive or not.

8.1 Individual Database

Six subjects, undergraduate or graduate students, perform the experiments. The individual database means each individual subject's behavior description case database, feelings description case database, event description case database and case database of degree of correspondence of positions, which are obtained from some pictures by interacting between subjects and the system.

8.2 Simulation Results

The simulation experiments are performed using four pictures shown in Fig.9, which are not used for obtaining the individual database. Twenty-four kinds of all permutations of three pictures chosen from four pictures are inputted into the systems, and the systems with individual databases A-F output 42, 40, 29, 18, 36 and 59 linguistic expressions.

When pictures 2, 3 and 4 in Fig.9 are inputted to the system in order of pictures 4, 2 and 3, the system has the linguistic expressions dependent on the individual database as shown in Fig.10. The descriptions with underlines are event descriptions. The other descriptions are state ones. For example, the system having database A outputs behavior description "A boy is frightening a dog" as explanation of the first picture. In addition, the system outputs the feelings description "A boy hates a dog" and "A dog is interested in a boy". It is found that the explanation is subjective because of feelings description.

As for the event description, the systems with individual databases A-F output various event descriptions depending on individual subjective such as "A boy bumps against a bear", "A boy and a dog are found out by a bear", "A dog is eaten by

a bear", "A boy meets a bear", "A dog calls a bear" and "A boy is found out by a bear", between the first and the second pictures.

When pictures 2, 3 and 4 are inputted in order of pictures 2, 4 and 3, linguistic expressions shown in Fig.11 are obtained by the systems with individual databases A-F. It is found that event descriptions in Fig.11 are not necessarily the same as the ones in Fig.10.

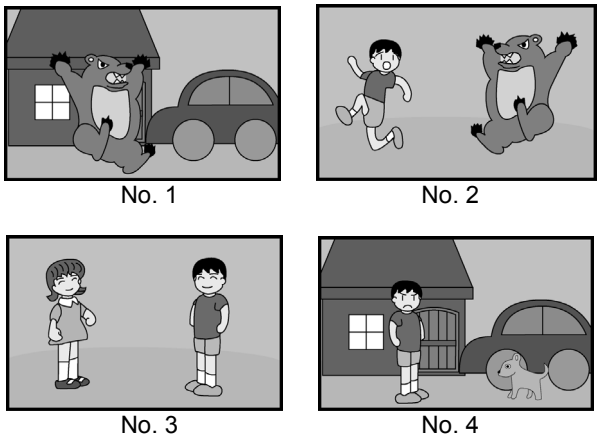

Fig. 9 Pictures used for simulation experiments

8.3 Evaluation of Systems

Linguistic expressions obtained in the simulation experiments are evaluated from the two points of view. The one is that subjects A-F evaluate the outputs of system having their own individual databases, and the other is that five other subjects evaluate outputs of the systems with individual databases A-F, where five other subjects are undergraduate or graduate students.

Subjects are shown system outputs one by one and then asked to evaluate them with a 5-point scale evaluation whether explanation of pictures is understandable or not;

1: incomprehensible
2: rather incomprehensible
3: neutral
4: rather comprehensible
5: comprehensible

The evaluation results are shown in Table 2. The evaluation averages by subjects A-F are 4.7, 5.0, 4.7, 4.1, 5.0 and 4.7, respectively. Evaluation averages among five subjects are higher than 3.9. It is found that system outputs are comprehensible for themselves, and that system outputs are rather comprehensible objectively.

A
A boy is frightening a dog.
A boy hates a dog.
A dog is looking at a boy.
A dog is interested in a boy.
A boy bumps against a bear.
A bear is attacking a boy.
A bear is angry with a boy.
A boy is afraid of a bear.
A boy runs away from a bear.
A girl is smiling with a boy.
A girl and a boy like each other.
A boy is smiling to a girl.

B
A boy is glaring at a dog.
A boy is angry with a dog
A dog is walking up to a boy.
A boy and a dog are found out by a bear.
A bear is attacking a boy.
A boy is running away from a bear.
A boy is afraind of a bear.
A boy runs away from a bear.
A girl is smiling to a boy.
A boy is similing at a girl.

C
A boy is glaring at a dog.
A boy hates a dog.
A dog is running up to a boy.
A dog is eaten by a bear.
A bear is attacking a boy.
A boy is running away from a bear.
A boy is afraid of a bear.
A boy meets a girl.
A girl and a boy are smiling at each other.
A girl and a boy like each other.

D
A boy is glaring at a dog.
A dog is approaching a boy.
A boy meets a bear.
A bear is attacking a boy.
A bear hates a boy.
A boy is running away from a bear.
A boy is afraid of a bear.
A boy runs away desperately.
A girl and a boy are smiling at each other.

E
A boy is glaring at a dog.
A boy is angry with a dog.
A dog is looking at a boy.
A dog calls a bear.
A bear is chasing a boy.
A bear is angry with a boy.
A boy is running away from a bear.
A boy is afraid of a bear.
A boy runs away from a bear.
A girl and a boy are looking at each other.
A girl and a boy like each other.

F
A boy is glaring at a dog.
A boy hates a dog.
A dog is approaching a boy.
A boy is found out by a bear.
A bear is frightening a boy.
A boy is running away from a bear.
A boy is afraind of a bear.
A boy runs away from a bear.
A girl and boy are smiling at each other.
A girl and a boy like each other.

Fig. 10 Simulation results 1

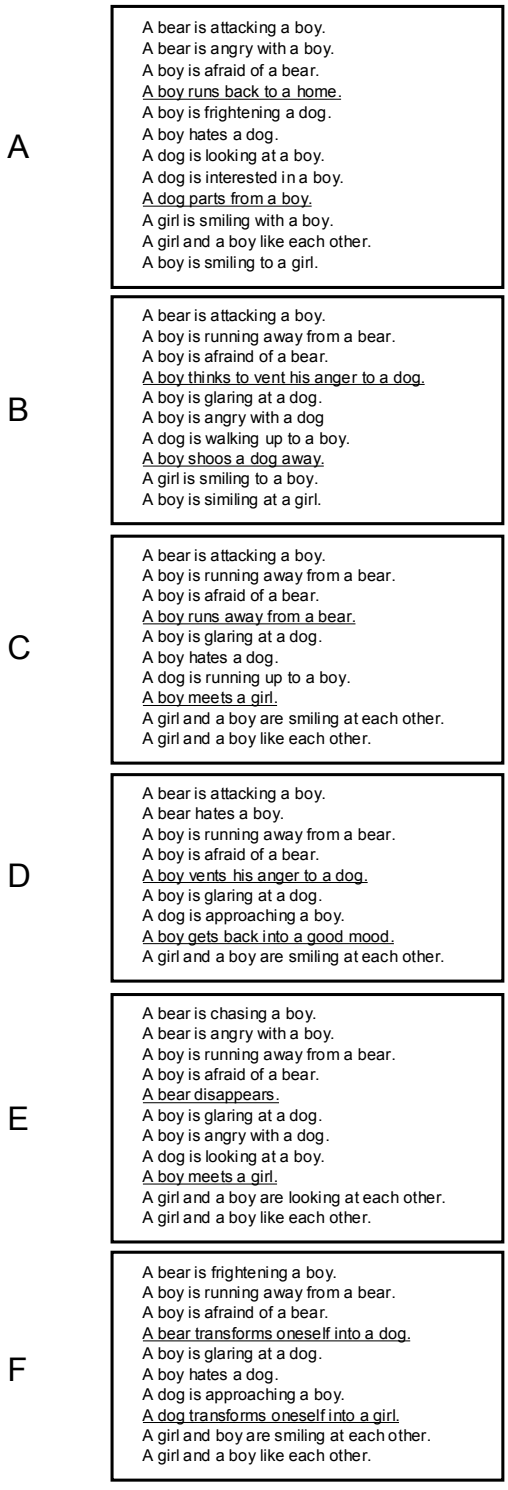

Fig. 11 Simulation results 2

Table 2 Evaluation results of system outputs

Individual database	Evaluation average by subjects A-F (Evaluation average among 5 subjects)
A	4.7 (3.9)
B	5.0 (4.3)
C	4.7 (4.0)
D	4.1 (4.3)
E	5.0 (4.2)
F	4.7 (4.1)

9. CONCLUSIONS

This paper describes the system that outputs subjective and consistent linguistic expressions of some pictures. The system explains behaviors and feelings of objects in pictures and events between pictures, which are expressed subjectively. And in order to obtain understandable explanation of pictures, this paper considers consistency of explanation. In addition, this system paraphrases the redundant state descriptions for natural linguistic expression. From the viewpoint that the interpretation of pictures is dependent on individual subjectivity, the present paper considers individual databases obtained by the interpretation of pictures.

In order to confirm the usefulness of the present approach, simulation experiments are performed. In the experiments the individual databases are constructed for individual subjects and the outputs of the system are evaluated. From the simulation results, it is confirmed that system outputs subjective explanation of pictures because of feelings description and outputs comprehensible explanation of pictures for the user and also outputs rather comprehensible explanation objectively.

In a future, it is necessary to generate more interesting linguistic expressions. For that purpose, it is important to consider the relationship between descriptions for example, the causality between the state description and the event description and to compare the system output with story written by human to find what kind of contents must be included in the story.

REFERENCES

[1] M. Mateas and P. Sengers: “Narrative Intelligence,”Proceedings of AAAI Fall Symposium, AAAI Press, Technical Report FS-99-01, pp. 1-10, 1999.

[2] Marc Cavazza, Fred Charles, and Steven Mead: “Characters in Search of an Author: AI-Based Virtual Storytelling,” Proceedings of the International conference on Virtual Storytelling 2001, pp145-154, France, 2001.

[3] Yunju Shim and Mink Kim: “Automatic Short Story Generator Based on Autonomous Agents,” Proceedings of PRIMA 2002, pp.151-162, Japan, 2002.

[4] Shigeru Kato and T.Onisawa : "Generation of Consistent Explanation of Pictures Considering Their Connection", Proc. of 2002 IEEE International Conference on SMC, Hammamet, Tunisia, Oct., 2002.

[5] K.Kuriyama, T. Terano and M. Numao: “Story Composition Support by IGA and CBR,” Proc. of the Third Asian Fuzzy System Symposium, pp. 485-488, Korea, 1998.

[6] Spierling, U, Grasbon, D, Braun, N, & Iurgel, I, “Setting the scene: playing digital director in interactive storytelling and creation,” *Computers&Graphics*, 26, pp.31-44, 2002.

[7] M. Iwata and T. Onisawa: ”Linguistic Expression Generation Model of Subjective Content in a Picture”, Journal of Advanced Computational Intelligence, Vol.3, No.1 , pp. 56-65, 1999.

[8] Charles Callaway and James Lester: ”Narrative Prose Generator”: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, WA, pp. 1241-1248, 2001.