Virtual cognitive model for Miyazawa Kenji based on speech and facial images recognition

HAMIDO FUJITA, JUN HAKURA, AND MASAKI KUREMATSU
Faculty of Software and Information Science,
Iwate Prefectural University
020-0193, Iwate
JAPAN

Abstract: - In this paper we a representing a virtual interactive model based on cognitive model of Miyazawa Kenji. We created a computer model based on cognitive thinking of Kenji literature on story telling. The user can interact in real time with Virtual Kenji. The facial gestures been collected and analyzed through Motion capture system of six camera. These six camera has been used to collect all emotional facial gestures of several people who read one assigned Kenji experiment. These people have 50 markers of 3 mm size attached to all parts of face (lips, mouth, eyebrow, cheek, moustache, eyelash, forehead), the emotional linkage between these facials parts and cognitive emotion been analyzed and recorded. We created a database we called Facial recognition database based on Miyzawa Kenji cognitive model, Also we have speech synthesis part who also analyzed the emotional part of speech. These two parts is been presented on hologram that represents the actual character of Kenji who has face with gestures harmonize with speech and facial images generated by the system and interact with the human user based on collected response from the human user and inference by the system in real time.

Key-Words: - Cognitive modeling, speech and image recognition, motion capture, Maya software, Motion Builder

1 Introduction

Virtual modeling has been active in the past years, thanks to computer graphics and virtual reality systems. This paper contribute to present an experimental work on building a virtual system based on Miyzawa Kenji art work represented by his writing. Miyazawa Kenji born in Iwate, Japan on 1896 http://www.kenji-world.net/english/who/who.html . He has famous children narrative stories with unique characters. There is rhythm in kenji's stories. Kenji's stories are set against the whole of the universe, a world replete with people, animals, plants, the wind, clouds, light, the stars and the sun. All hold discourse together. All are in empathy with one another. This free association between the elements and living things that make up our world is one of the distinguishing features that predominates Kenji's works. The interaction he portrays is never nonsensical, but always animated with an authenticity that rings true to the reader. Such rhythm represents or reflects certain cognitive behavior inside the stories through which the user may interact and feel the emotional or living part that he/she may interact. http://www.kenji-world.net/english/who/rhythm.html Most of the work of Kenji had been translated into English. http://www.kenji-world.net/english/translat/translat.html In this paper we are building a virtual work based on Kenji artwork. It is an experimental system based on hologram that interact with the human user based on the cognitive based built-up system.

Based on Kenji scripts we have extracted patterns that reflect the emotional feature behind them. Those features can be classified into and represented by facial images based on Ekman FACS system (section 2). Those patterns reflect the facial(Sec.2) and sound extracted patterns(Sec.3). Those patterns can be specialized and instantiated the data stored in the database. Those generated graphics animated files reflect the facial cognition and sound data reflect the voice emotional model. The extracted patterns come from examining and analysis of the masterpiece of Kenji work. Such analysis is based on ontological reasoning of the text. Such reasoning reflects the recognition of Kenji work by expertise and people who are aware of the Kenji experience and background. Through such analysis, Kenji words and extracted narrative story patterns reflect the emotional representation of the context in the story. Each world reflects its domain and its connection with other words reflect the semantical representation of that context.

There are two parts in the system the namely, computer based graphics representing the animated facial image generated by the system and interacted with the user based on the latter cognitive facial recognition. There is also speech synthesis part which make and generate the voice of Kenji in harmony with the generated facial

images. Also this will interact with the human user based on collected sound patterns that been reflected with reasoning process capturing the emotional recognition of human user voice that interacting with Kenji based on the latter speech and face generated scenarios.. The general view of our system design can be seen at the figure of system outline.

2 The Facial synthesis part of the system Recognize and Animate Human Facial Expressions.

This section is assigned for a construction of a database based on an automated facial expression analysis method, together with the its uses in facial expression recognition and making facial animation of with character's emotional expressions. The automated facial expression analysis method allows users to define the emotional expressions from video sequence. Namely, the defined emotional expressions are automatically identified from the observations of a subject. The actions of the feature points in the defined facial expressions are assumed as caused by a set of system(s) as in [Nishiyama et. al. 2005]. The system identification method (i.e., LSM: Least-Squares Method) is adopted to identify the system. The identified systems are stored in the database, and used both in the recognition and exhibition of the emotional expression by the virtual Kenji. For the sake of the project, we assume that the user is accustomed to the FACS (Facial Action Coding System) [Ekman and Friesen 1975], and defines the facial expressions as from the FACS point of view. The main reason for choosing the FACS as the definition base is that we are conducted to use the database to exhibit emotional expression of the virtual Kenji. Because the Virtual Kenji ought to face with general public, the expressions should also be comprehensive to general public. The FACS helps to judge which of the six basic emotions is exhibited by the subject. It is considered as universal judgments that can be available across persons, sexes, races, and so on. The FACS itself is originally provides a set of still images that depicts the typical examples of facial expressions to be used as the templates for psychologists. More recently, however, it is extended to the automatic and spontaneous facial action recognition(e.g., [Movellan et. al. 2003]). Our approach differs from that of Movellan et. al. in that we can utilize the actions of the feature points for making animated expressions of the virtual Kenji. When the virtual Kenji should exhibit certain emotional expressions, the set of systems that is labeled as the emotional expression is invoked form the database. The set of systems can estimate the temporal movements of the feature points so that the movements of the points that constitute the facial emotional expression come to be available. The movements of the points are now applied to generate the facial movements of the virtual Kenji. The emotional expressions and the facial motion to generate certain utterances are blended in the exhibited face. This will enable the virtual Kenji to be natural. The detailed mechanisms of the facial expression recognition and animation are described in the next sub-section.

Fig 1 shows the overview of our approach. As shown in the figure, the first step to detect emotional facial expressions defined by a user based on FACS is assigning the sequences of frames that express desired emotions from a sample sequence of facial expressions observed in the subject's daily life. This step is followed by the automated clustering by means of the algorithm proposed in [Nishiyama et. al. 2005]. The algorithm identifies the systems that produce the every movement of the facial points. The duration of the motion that can be identified by the same identifier is called mode. The algorithm, then, tries to merge a couple of durations that the same identifier can estimate the movements with the minimum error. This bottom-up clustering continues until the uniform system identifier is determined. Namely, the algorithm as it is finally generates the identifier that identifies not only the user defined duration, but the outside the duration as the same mode.

To extract templates for the particular facial expressions, this paper uses the user definition of the desired expressions. Namely, the abstraction process is continued until identifier(s) that uniquely identify the duration that contains the user defined facial expressions. Thus, we can get the set of unique identifiers to represent the expressions. We store the (set of) identifiers together with user defined labels of the emotions as the templates. With these templates, we can detect the user defined expressions from the user's facial expression under consideration.

To extract the user defined facial expressions, template for the particular expression is automatically generated in the proposed approach. The template here is a set of the transition matrices described in the previous section. For this aim, an appropriate level of the cluster is required. Namely, a mode, or a set of modes, that is necessary and sufficient to identify the expression should be defined automatically. This section is assigned to describe the generation method of templates, and the scene extraction based on the generated templates.

2.1 A Template Generation Method

The basic idea underlying for the template generation is defining a termination condition on the clustering. Because, the proposed method is assigned the user defined sequence of frames to be extracted as the typical facial expression, the mode(s) that is necessary and sufficient to identify the expression can be determined.

Fig.2 depicts the basic idea together with the difference between the proposed method and in the Facial Score. As

shown in the figure, the clustering continues until the appearance of the mode that estimates the movements of the facial points across the user defined duration. The mode(s), belongs to the user defined duration, just before the termination condition has invoked, is/are defined as the template. In the case of depicted in the figure, modes $<M_2, M_3, M_2>$ are the template. As shown in this example, the sequence of the modes can also be treated as a template. Each template is assigned the user definition of the expression as a label. A set of labeled templates is then treated as the Facial Expression Database. Namely, the data structure of the expression elements in the database is as follows:

$E_i = <\mathbf{M_i}, l_i>$.(1)

where, $\mathbf{M_i} = <M_j, \mid j \in D_i>$, is a mode in the user defined duration, l_i and D_i is a label for the expression.

2.2. Recognizing Facial Expression with Facial Expression Database

The template here is the transition matrix or a set of the transition matrices that corresponds to the mode(s). The extraction process is executed by estimating the movements of the facial points from the previous coordinates of the facial points with the template by using Equation (2).

$$\mathbf{FP}^{esti}\mathbf{(t)= M_jFP(t-1)+f_j+ \omega_j(t)} \quad (2)$$

Where $\mathbf{FP}^{esti}\mathbf{(t)}$ is a set estimated x-y coordinates of facial points at time t by means of identifier $\mathbf{M_j}$, $\mathbf{f_j}$ is a bias term,$\omega_j\mathbf{(t)}$ is a process noise of the system. To recognize facial expressions, every facial element in the database is treated as a candidate. The movements of the facial points in the expression under consideration are estimated by means of every E_i in Equation (1). The estimation errors between the actual movements of each facial point and the estimated ones are compared. The label(s) coupled with the mode(s) that can estimate the expression with the error value lower than a certain threshold is given to the exhibited expression. Namely, the expression is recognized as the labeled emotion.

2.3 Animating Facial Expression with Facial Expression Database

When the Kenji system interacts with the system user, the system is to infer the reaction to the user. The reaction contains not only the contents of dialog, but the facial expression suitable for the situation. The Facial Expression Database is diverted in making animation with the facial model of the virtual Kenji. Namely, the mode(s) labeled as the facial expression is reused. With Equation (2), the movements of the facial points can be estimated. This means that the mode can provide the movements of the facial points of the Kenji to represent the labeled emotion. The movements of the facial points will then be applied to the computer graphically modeled surface of the Kenji's face.

2.4 An Experiment on Recognition Ability

To confirm the possibilities of the proposed methods for recognizing facial expressions, a brief experiment is conducted. The database is constructed with facial expressions by a single subject and a single emotion, i.e., happy, is aimed to be recognized. The subject is asked to watch a Japanese comedy on DVD with makers for facial feature points on his face. Although the motion capture system can observe movements of the points in 3-dimensional space, the depth information is neglected for simplicity. The experiment has two phases: template acquisition phase, and recognition phase using the acquired templates. In template acquisition phase, the observation is carried out for about four minutes, and we manually choose 200 frames out of the four minutes observation where we can consider the subject is expressing some emotion. Within the 200 frames, we define the frames where the subject is expressing the emotion, ghappy h, as shown in Fig. 3. With the proposed method, we have obtained two expression elements: one consists of two modes, and the other consists of single mode. In our experiment, some expressions that ought to be classified as the same emotion, i.e., ghappy h, according to FACS, are distinguished. The reason for the distinction is that we would like to confirm the identification ability for slight change in the expression of the proposed method. This leads the two expression element for the same facial expression. As shown in Fig.3, the modes of the defined duration are nicely distinguished from the other durations.

Then, we have applied the expression elements in the recognition phase. In the recognition phase, 30 minutes of observation is conducted, and we have obtained 32,000 frames of the facial point movements by means of the motion capture system. To recognize the objective expression from the observation, an estimation of the movements of the facial feature points is calculated by adopting Equation (2) to the acquired templates. The estimation error for each frame and observed value are compared, and the label of the expression element with lower estimation error than certain threshold is the result of recognition. A result of the experiment is depicted in Fig. 4. In the figure, the dots on the line at estimation error 1 are the frames that a human observer judged as ghappy h. The estimation errors around the dots are abruptly decreased (circled area in the figure). This means that the method can recognize the facial expression ghappy h almost as same as human observer. Although there is another duration that exhibits row estimation errors (area surrounded by a rectangle), they can not be stand as ghappy h because Template 1 consists of two modes, namely M1 and M2. Moreover, we can distinguish the ghappy h expressions by using the templates with appropriate thresholds, because the

peaks of the two templates are distinguishable. We are planning to apply this method to recognize the other emotions in the next step of facial expression recognition part of the project.

3 The speech synthesis part:

There are some speech synthesis systems in the world. Speech translated by a speech synthesis system have include accent or utterance. But it's other attributes of speech, for example, volume, speed , tone ,does not change. If a computer can change attributes, a computer will be able to speak more natural and users will find emotion in speech. Some systems try to change volume and so on. But user should set attributes before a computer speak. While a computer translates text to speech, we try to change attributes automatically. We change attributes of speech based on properties of text. We focus on linguistic information. There are some keywords in texts. They are important for a writer and readers. If a computer speaks these keywords, a computer turns up the volume. If we know keywords, we give them to a computer, If we don't know them, we extract keywords from a text using information retrieve technique. Using this technique, for example Term Frequency, we can add a weight to words. If a computer tries to speak words, which have a weight more than threshold, it turns up the volume and add pause before speaking them. And we change attributes of words, which appear in same sentence with keywords. We change attributes a sentence includes a keyword, too. But changing attribute of words except keywords are more week than keywords We extract some words, which express emotion from text. We change attributes based on emotion expressed by a word. We have rules about relationships between emotion and voice based on experimental analysis. So we change voice attributes of words based on these rules. And we change attributes of do for a sentence includes emotional words like processing a sentence includes keywords. We extracts to onomatopoeic from a text. We estimate something that onomatopoeic express and change voice like it. And we focus on the result of parsing, feature in grammar. We change attributes first and last sentence in a paragraph. Because these sentence often have most important information. We turn up the volume and slow down. There are some sentences or some words match some rules for changing attributes, we synthesis them. It synthesizes animated, life-like, facial expressions of an individual in synchrony with that individual's speech. The system is speech driven, that is, as an individual speaks the appropriate facial expressions are generated simultaneously. In the database here we collected sound synthesis system that represent the emotional state of sound based on the pitch and amplitude analysis. All sound will be analyzed and categorized according to the emotional state.

A system can read out scripts using speech synthesis technique. But we don't set parameters for reading out, a system reads out in monotone. So, we have to put accent, intonation, stress, speed, tone and some attribute on scripts before reading out. It is a hard work. So we need a system that supports setting parameters to scripts or sets parameters to scripts automatically. In addition to do, the system doesn't change parameters without stopping. It is good for the system to be able to change parameters during reading out scripts. Because how to speak depends on the environment of this system, user's emotion and so on. So we need a method that a system changes parameters for reading out.

Reading out module in Pseudo Miyazawa Kenji System reads out scripts written by Kenji Miyazawa. In order to polish up communication with users, it is necessary to change parameters for reading out automatically. In order to change parameters, we focus on viewpoints of speech. There are following seven viewpoints for speech, Accent, Articulation, Intonation, Phrasing, Prominence, Pause and Rhythm. Speech synthesis technique can support accent and articulation. So we focus on Prominence and Phrasing. Phrasing means that we divide a sentence to some phrases based on syntax and phonology. It looks like chunking. When the system reads out a phrase, it put pause before and after the phrase. Prominence means that the system puts stress on some phrases at reading out scripts. Reading out with prominence helps us to understand what the system says. Fig 5 shows overview of reading out module. This module has phrasing process and prominence process. We describe them as follow.

First, we describe Phrasing process. If the number of moras in a sentence is more than threshold, this module tries to divide it to some phrases. We don't divide conversation sentences and short sentences. We regard conversation sentences as good for hearing. And reading out with a lot of pause is not good to hear and understand what a system says. So it is not worth to divide short sentences to phrase. This module divides a sentence to some phrases using rules based on syntax. The reason why we define the threshold is as follow. It is not natural that the system reads out a long time without breath. So the module divides a long sentence to some phrase and put pause before and after them at reading out. The threshold is defined based on the number of moras in Tanka, Tanka is a short poem in Japanese and usually has 31 moras. We decided the threshold is 30. We made 5 simple rules for dividing a sentence based on Japanese grammar. For example, we regard conjunction as end of phrase. A rule has a condition part and a conclusion part.

A condition part consists of morphological strings. A conclusion part of rules says that picks up the phrase (put pause). The system tries to match the condition part of rules to result of morphological analysis. If the condition part is true, it does the action defined in a conclusion part. This module reads out scripts written by Japanese. So we make rules based on Japanese grammar. If the system reads out scripts written by other language, we should make rules based on that language's grammar.

Next, we describe Prominence process. In order to put stress on a phrase, the module selects a phrase using rules based on syntax. It picks up onomatopoeia, rheme phrase and theme phrase to put stress. Longman Dictionary says that Onomatopoeia is the use of words that sound like the thing that they are describing. We think that onomatopoeia is an important point for hearing. So the module picks up onomatopoeia using a dictionary and puts stress on them. Rheme and Theme are important phrases to understand what someone says. We put stress on these phrases at speaking. So we put stress on rheme and theme phrase. The module tries to find rheme and theme using rules based on syntax. The structure of these rules is same as one of rules for phrasing. But a conclusion part says that puts stress on a phrase defined in a condition part. The module uses outputs of morphological analyzer at phrasing and prominence process. So the accuracy of phrasing and prominence is influenced the accuracy of morphological analysis

In order to show the effectiveness of this idea, we think two experiments as follow.

In one experiment, participants hear two kinds of utterances by the system. One is that the module reads out without phrasing and prominence. The other is that the module reads out with phrasing and prominence. Before hearing, we don't say which utterances by a module with phrasing and prominence. After hearing, we want participants to say which is easy to hear, which is more emotional and which is good. If most participants select utterances by the module with phrasing and prominence, we consider our idea is effectiveness. In the other experiment, participants divide a script to phrases and put stress on phrases. We compare the output of the module and scripts changed by them. If there are a few difference points between them, we consider our idea is good.

We describe the module does phrasing and prominence. Phrasing is how to divide a sentence to some phrases and prominence is how to put stress on some phrase in a sentence. This idea is based on the viewpoints for speech. We think it is good for reading out. But we have not evaluated it. So we have to show the effectiveness of this idea with experimental results. We have to feed back the analysis of experimental results and refine this system. In addition to, we have to think how to change parameters based on the situation around this system automatically.

4 Conclusion

This work presents the preliminary results of our cognitive virtual interactive model based on Kenji cognitive model. This is the 1st stage outcome.

Using motion Capture system installed in our laboratory, we have extracted emotions characteristics of Kenji scripts based on emotional analysis of many users who are reading Kenji scripts with emotional harmony with it as if Kenji is reading it. This basically, has been used to extract emotional feature based on common sense knowledge and human science and other analysis parameters. Also, [INTERACT] software has been used to confirm the collected or captured emotional features have the same labeling, (label reflect the emotional feature of user or Kenji extracted behavior), as user him/her self does on the video observed/recorder on them when reading the Kenji scripts as shown on Fig.6, which shows a simple labeling for the emotional feature by the users.

Scripts text analysis also has been done using emotional analysis. Also the speech of the users, so that to extract his emotional feature as well from it.

We think this system will be used to show

The motional feature between any system and human can be feasible, that the user interface enhancement to approach the human level of emotion make computer more friendly. Such softness and flexibility, being reflecting to be more human, can contribute to enhance the interaction between computer and human and make more ubiquitous.

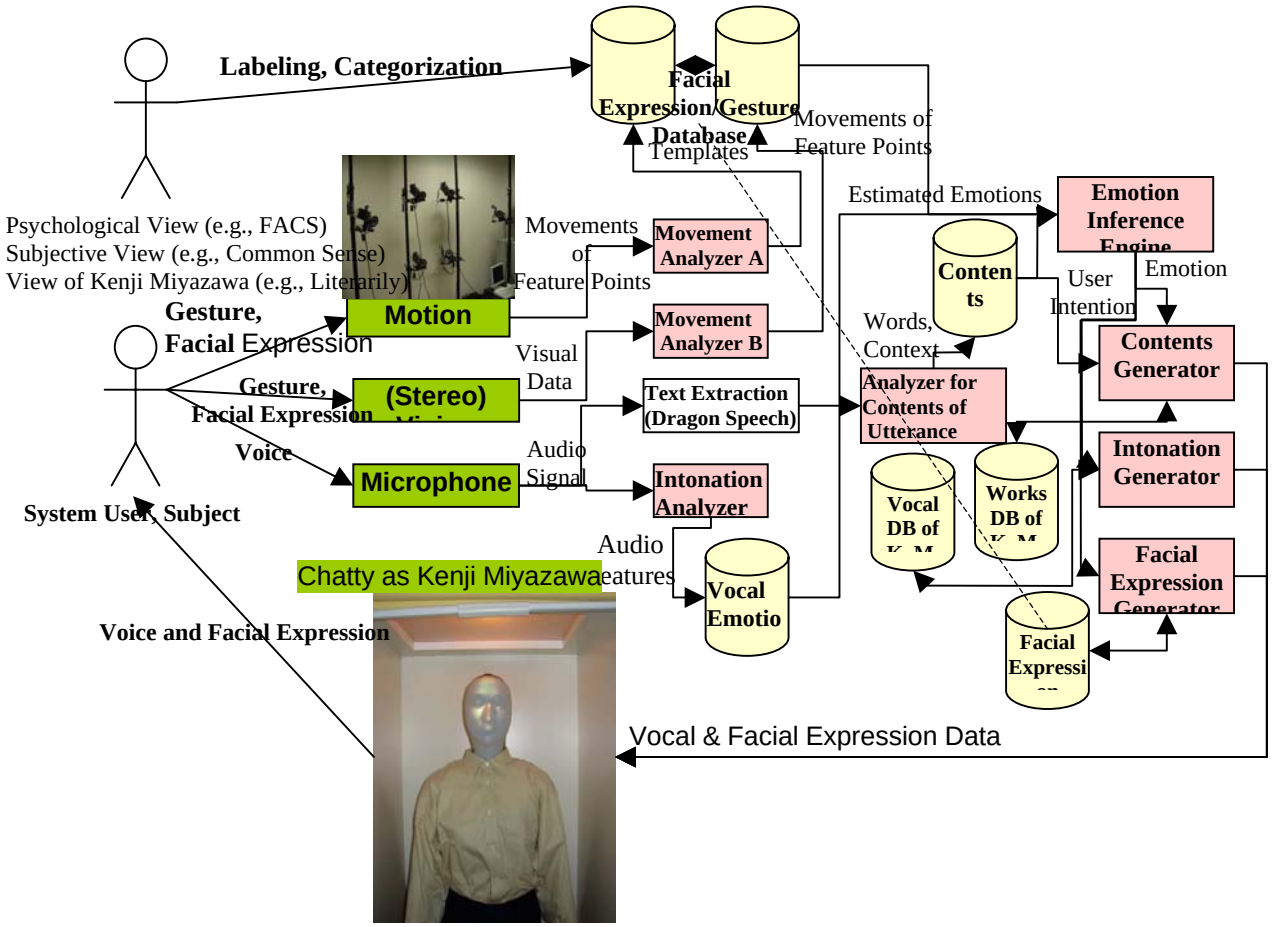
References:

[Movellan et. al., 2003] J. R. Movellan and M. S. Bartlett. The next generation of automatic facial expression measurement. In Paul Ekman, editor, What the Face Reveals. Oxford University Press, 2003.

[Ekman & Friesen 1975] P. Ekman and W.V. Friesen: Unmasking the Face. Prentice Hall, NY.,1975

[Nishiyama et. al. 2005] M Nishiyama, H Kawashima, T Hirayama and T Matsuyama: Facial Expression Representation based on Timing Structures in Faces. IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2005):140-154,2005

[INTERACT] Trademark software by Manglad co. http://www.mangold-international.com/



The outline of Virtual Kenji System

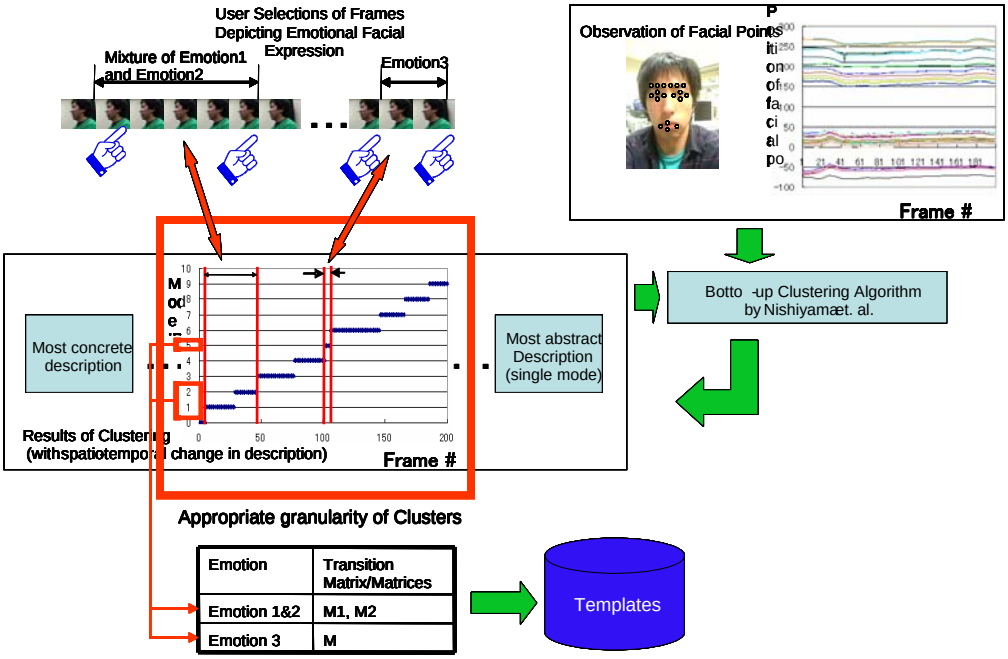
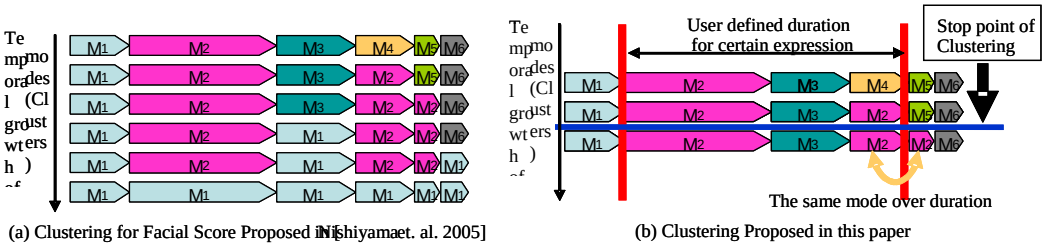


Figure 1 Overview of Approach to Automated Template



(a) Clustering for Facial Score Proposed in [Nishiyamaet. al. 2005] (b) Clustering Proposed in this paper

Figure2 Difference in Clustering Method

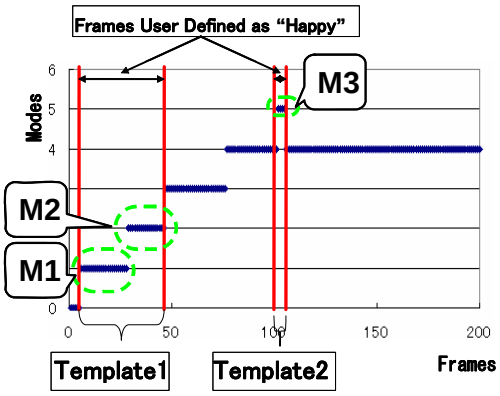


Figure 3 Acquired Templates from User Definition

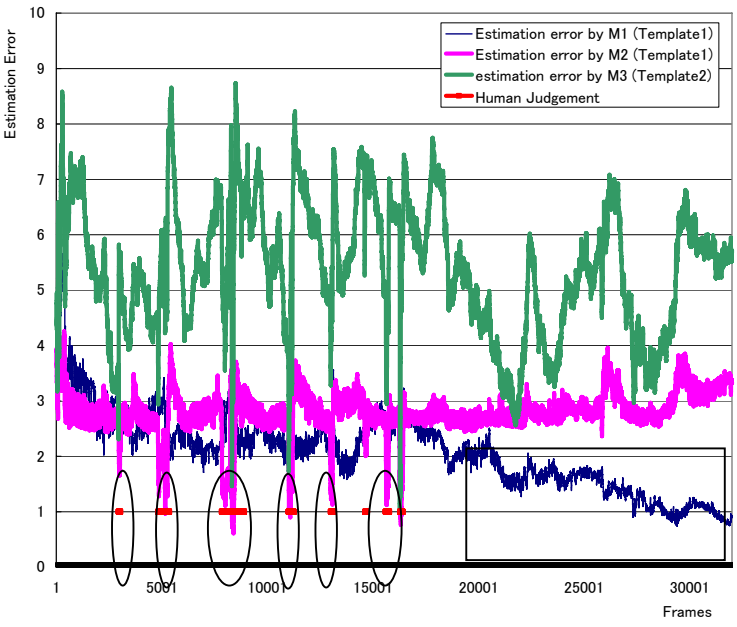


Figure 4, Experimental Results

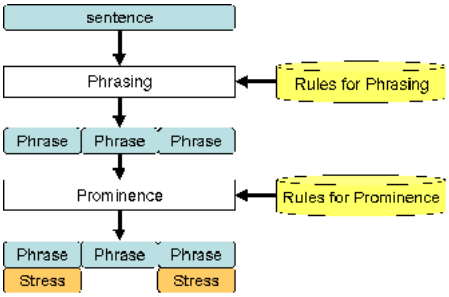


Figure.5: Overview of reading out module

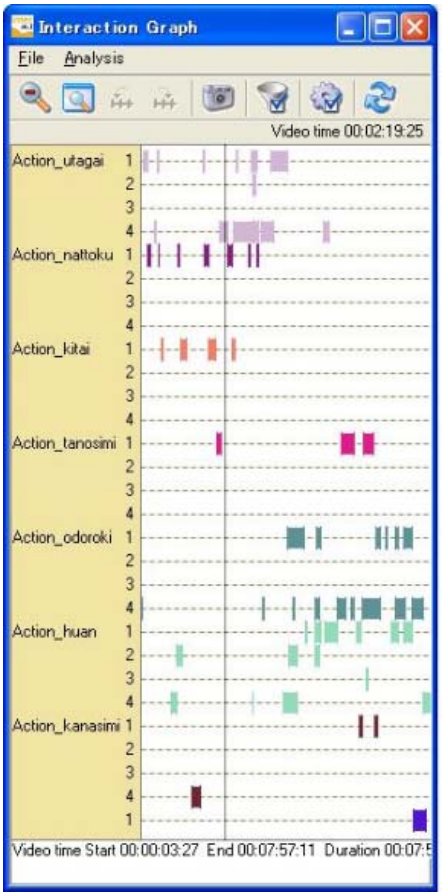


Fig.6. Emotional labeling using Interact