MMID: Multimodal Multi-view Integrated Database for Human Behavior Understanding

Yuichi NAKAMURA   Yoshifumi KIMURA   Ye YU   Yuichi OHTA
Institute of Information Sciences and Electronics
University of Tsukuba
Tsukuba 305, JAPAN

Abstract

This paper introduces the Multimodal Multi-view Integrated Database (MMID), which holds human activities in presentation situations. MMID contains audio, video, human body motions, and transcripts, which are related to each other by their occurrence time. MMID accepts basic queries for the stored data. We can examine, by referring the retrieved data, how the different modalities are cooperatively and complementarily used in real situations. This examination over different situations is essential for understanding human behaviors, since they are heavily dependent on their contexts and personal characteristics. In this sense, MMID can serve as a basis for systematic or statistical analysis of those modalities, and it can be a good tool when we design an intelligent user interface system or a multimedia contents handling system. In this paper, we will present the database design and its possible applications.

1 Introduction

Face and gesture recognition is becoming more and more important for HCI (human computer interaction), human activity measurements, personal identification and so on. Many researches are devoted to (1) measuring positions of hands or feet as pointing devices, (2) recognizing gesture patterns which have predefined meaning such as sign languages, and (3) surveying human activities by counting or measuring humans. So far, gesture recognition problems have been considered as the measurements of body position or the recognition of body motion.

Humans, however, use gestures in much more flexible way, often cooperatively with speech or other modalities. If we want to understand human gestures as an effective means of communications, we have to know how gestures are being used in real situations. Moreover, since human behaviors are heavily dependent on environments around him/her, we need to investigate a variety of situations.

For this purpose, we gathered the records of human behaviors in multimodal way, and built a prototype database named Multimodal Multi-view Integrated Database (MMID). Because of our resource limitations, we are now concentrating on presentation situations such as lectures or demonstrations. In those situations, a human has a clear intention to inform something to other people, and the human tries to effectively organize speech, gesture, and other modalities for explanation. MMID contains these activities in terms of video, audio, motion captured data, and speech transcripts, all of which are related by their occurrence time. The contents can be retrieved by specifying an example or a template in one of the modalities. From the retrieved data, we can easily overlook how multiple modalities are cooperatively used in human communications. Thus if enough variety of data are stored in MMID, it can be a good tool for designing user interface or multimedia contents handling system.

In the following sections, we will present the basic idea, the design of MMID, and the experiments with MMID.

2 MMID Objective

2.1 Necessity of Multimodal Approach

Humans often use visual expressions, gestures or diagrams, language expressions, speeches or texts, and other modalities. In these expressions, their contents refer to each other and explain each other explicitly or implicitly. This multimodal method greatly helps the audience to understand the presentation or explanation.

Any single modality is not enough to communicate with humans. Let us consider the gesture shown in Figure 1, which is too ambiguous to grasp his intension. He may express, by the gesture, the length or

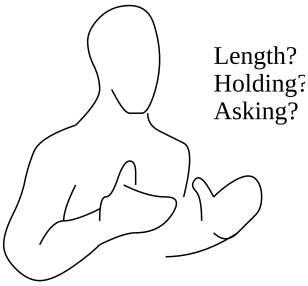


Figure 1: Ambiguity in gesture interpretation

something, the pose to hold something, or an intention of asking something from you. Similarly, when a human raise his hand, it may be interpreted as the sigh of picking up a taxi, a motion of taking a book from a bookshelf, or a gesture of pointing upward. Only the way to clarify this ambiguity is to consider the context or the environment of the gesture. Speech is a modality which gives this context information. A small number of words, such as "about this size" or "when I pick up the stone...", clearly disambiguate the situation.

In this sense, it is clear that the integration of multiple modalities, especially visual and language modalities, is an essential key to advanced HCI. However, it has been very difficult to investigate a variety of multimodal cooperation in detail, since multimodal data are huge and multimedia handing systems were not common. Recently, these limitations are gradually dissolving. We can accumulate large amount of data, for example, several hours of video, in one hard disk drive. It is, therefore, the time to gather a variety of human activities in a digital database.

2.2 Functions of MMID

Here we give a brief explanation of MMID with an example of pointing gesture analysis, *i.e.* deictic movement analysis, which will be further described in Section 4.1. As a typical deictic motion, we often image a person pointing something with his hand stretched, and saying "this" or "it" at the same time. However, there are a wide variety of pointing motions in our actual life. For example, a pointing with both hand bended, a pointing without any demonstrative pronoun, etc. Furthermore, the gesture with one stretched arm may not represent a pointing motion as the "picking up a taxi" example described above. To deal with this problem, we might ask the following questions:

- How many kinds of gestures are used for pointing?
- In what rate does the stretched hand gesture represent pointing?
- Which words frequently appear in pointing?
- What is the sufficient combination of a gesture and spoken words for detecting pointing?

We have to examine large amount of multimodal data to get good answers for these questions. For this purpose, MMID can potentially give the following information.

1. Frequency of a specific gesture or a speech
2. Variations of gestures or speeches used for a specific purpose
3. Cooccurrence of gestures and speeches in a specific situation
4. Individuality or differences of gestures and speeches among persons

The basic idea for obtaining this information is to retrieve similar portions in one modality, and to compare with the other corresponding modalities. The actual experiments will be given in Section 4.

2.3 Human Behaviors in MMID

Before describing the actual inside of MMID, we briefly mention the kinds of human behaviors stored in MMID. Ekman's classification is well known and often used in many researches [EF69] [Kur94]. Nonverbal behaviors are firstly classified into 6 classes, one of which is further classified into 6 subclasses:

emblem: Motions or postures which can be translated into language. Sign languages are classified into this class.

illustrator: Movements which complementarily support speech. This class is divided into 6 subclasses: *deictic movements* for specifying a position or object; *spatial movements* for spatial display of analog value; *pictograph* for shape description drawing by hands, etc.; *kinetograph* for movement display, *ideograph* for expressing a thinking process; *baton* for emphasizing or making a rhythm of speech.

affect display: Emotional expressions such as facial expressions.

regulator: Movements which regulate conversations such as switching a speaker.

adaptor: Movements for adapting a person to the environment. Mostly, personal habits without relations to speech contents.

Emblem and *illustrator* are the human movements which we usually consider as gestures. They are the main target of HCI researches, since they convey important information in conversation or presentation. *Affect display* is becoming an important target of recent HCI researches. The users' feeling and emotion are good information to control machine process. *Regulator* is also important to control machine display or audio.

In our research, *emblem* and *illustrator* are mostly dominant, since our current target is a presentation scene in which one person is giving a presentation to the audience. However, we often need to deal with all the above kinds of gestures, since *emblem* and *illustrator* should be discriminated from the others in actual recognition stages. For this purpose, all movements in presentations are recorded in MMID.

3 MMID Construction

3.1 Contents Selection

Currently, MMID has two kinds of data. One is a collection of original presentations, and the other is a collection of cooking shows from TV broadcast programs.

For the original presentations, presenter's body motion, views from multiple cameras with audio, and transcripts of his speech are recorded. The scenarios currently stored are shown in Table 1. In these scenarios, the presenter demonstrates many kinds of gestures which we usually see in actual presentations; for example, deictic movement (pointing gestures), spatial movements, and pictographs. Each scenario has from 30sec to 2min length, and was played by 6 different people. The total length is about 50min (8.5min/person × 6 person).

The cooking shows are recorded from TV broadcast, and transcripts are manually added. Eight scenarios are stored, each of which is 25min in length (total 200min). Motions contained in those data are mainly operations by hands such as cutting some materials. The cooking show data lack the motion data and simultaneous multiple views compared with the original presentation data. However, they are good sources because they are easy to record, and speeches and acts inside them seem to be natural. Moreover, a demonstrator usually describe his movements by his speech. For example, a demonstrator puts an egg into a bowl saying "then, put this into water" at the same time. In this sense, a cooking show is one of the most useful data in which speeches and motions are mostly synchronous.

3.2 Data Acquisition

Videos for our original presentations are taken from multiple (currently 6) cameras. Each camera aims at a different portion of the presenter scene: the whole scene; the upper half of the body; the right hand; the left hand; the objects on a table; a stage view from the left side. The angle of the cameras for the upper body, hands, and objects are controlled by a host computer by using the position of those parts measured by the magnetic sensors. In this way, the multiple cameras shoot at all important portions which potentially attract viewer's attention. An example of the views is shown in Figure 2.

Table 1: Contents of original presentations

Scenario	Contents	Length
scenario1	cooking show (cutting)	1min
scenario2	presentation	2min
scenario3	a visitor at the front door	30sec
scenario4	assembling a video equipment	1min
scenario5	assembling a camera set	2min
scenario6	talk of two persons	2min

Videos for cooking shows are taken from broadcast programs. We can consider that the views of the show were taken from multiple points, the view on the broadcast is already edited and switched according to the editor's or producer's intention.

Both kinds of videos are digitized into MPEG or Motion JPEG[1]. The 6 videos for each scenario of the original presentations are completely synchronized.

Transcripts of the original presentations were prepared beforehand. Speeches in cooking shows are manually transcribed, since Japanese TV broadcast system has not established a closed-caption system yet. They contain an average of 2845 words. An example of the transcript is shown in Table 2. Each line is separated such that it forms a phrase or a sentence. Each line contains a frame number, which is the time code in the corresponding video(s), and several words which the presenter spoke at that moment.

Motions are measured by 6 magnetic sensors (Ascension Tech's A Flock of Bird). Each sensor measures 6 degrees of freedom: the position (x, y, z) and the orientation $(rotation, role, azimth)$. The sensors are attached to the presenter's head, both hands, both shoulders, and the back. Accuracy is from 0.3–1.0in for position, 1 degree for orientation if we remove all metals from the measuring space. Since this restriction is too severe and it loses generality, we allow small metals in the measuring space. To compensate this drawback, we calibrate sensor outputs and actual positions linearly by using 27 points distributed in a 1m^3 cube before actual measurements. The measurement range is about 5m in radius, and the sampling rate is set to 30Hz. Since we currently attach 6 sensors on a speaker's body, 6 sequences of position and orientation values are obtained, each of which includes 6 values at every 1/30 second. Each measurement is associated with the frame number of the corresponding video(s).

[1] We are currently using SGI Movie format.

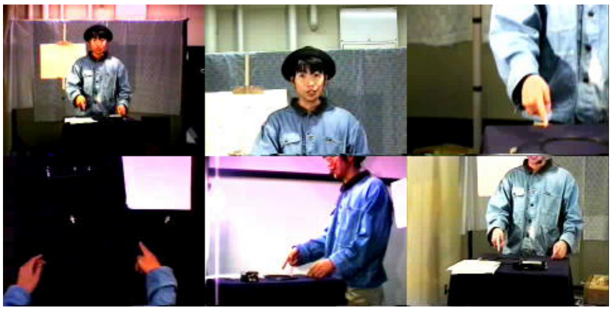


Figure 2: Example of multi-view video data

Table 2: Example transcript

7020	これは（あの）温サラダにする分です。 (These are for hot salad.)
7101	すでに洗ってありますので、 (There are already washed)
7215	はい、（えーっと）まず葉先と軸とを分けます。 (Then, (well), first split the leaves and the stem)

Each important movement is manually detected, and the category of the movement is added as a motion label to the data. The duration of each motion is recorded in terms of its start and end time. Table 3 shows the motion categories we are currently dealing.

3.3 Content Retrieval and Display

Each of the stored datum can be denoted by a tuple:

$$< scenario\_id, time\_code, mode\_id, data\_value >$$

where *mode_id* represents the modality of the datum, which is video, motion (raw data or motion label), or transcript (sentence, phrase, or word). Given a query for similar datum, the system searches all data in the same modality by checking the *data_value*, then searches the corresponding data in other modalities. Currently, *time_code* is used for finding the potential correspondence across the modalities. For this retrieval, we can think query and retrieval schema as follows.

Motion: Gesture or posture detection and retrieval by searching similar motion sequence.

Transcript: Retrieval for word, morphological form, and case. Semantic retrieval for specific situations, such as cutting or assembling.

Video: Similar scene retrieval, face detection, object detection, etc.

Currently, motion, posture, and transcript retrieval have been implemented. Others and query by a combination of multiple modalities are under development.

Table 3: Motion label

put, put-in, take-out, pick-up, cut, stab, push, hit, twist, fold, pull, rub, shake, knead, stir, scoop-up, turn-back, fix-to, turn-up

For motion, the system does not segment the stored sequence into pieces, since gesture segmentation is not an established technique. Instead of segmentation, the system provide an exact template matching of postures and the Continuous Dynamic Programming (CDP) [Oka96] matching for gestures. In other words, since the system can try to find good match continuously from every starting point by CDP, we do not need to segment the motions. For transcripts, the system first performs morphological analysis of the input transcripts, then makes an index by words. By looking up a word from the index, the system can retrieve every situation in which the word is used. The system also accepts a query by morphological form, for example, "set" used as a verb.

The data inside MMID or the retrieved results can be shown with GUI.

Comprehensive Viewer: As shown in Figure 3, data in the video, the transcript, motions can be seen in one window: video view potion at the upper left; transcript view potion at the left center; motion view portion in the right; the query portion in the bottom. Videos with audio, transcripts, motions can be replayed synchronously.

Cataloged Viewer: Multiple pieces of the retrieved data can be seen in one window. As shown in Figure 4, a combination of a still image taken from video, the display of a posture or one frame of the motion, and a sentence from the transcripts is shown for each piece. Then, we can easily see multiple data by placing two or more such combinations in one window.

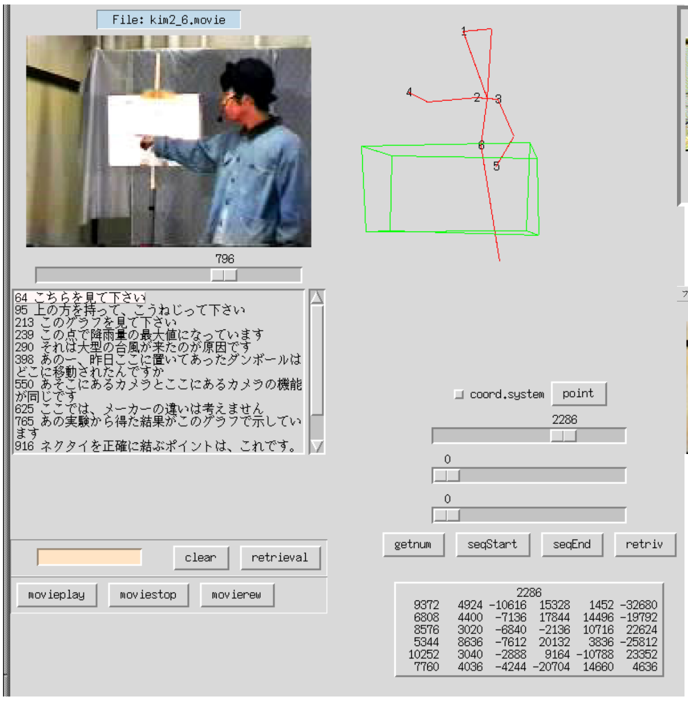
4 Experiments: Actual Usage of MMID

We will introduce two ongoing researches in which MMID plays an important role[2]. One is a research of pointing gesture detection, the other is video summarization by detecting focused points.

4.1 Pointing Gesture Detection

The target is to recognize the gestures of pointing an object, a direction, or a location. The most typical case is the situation in which a person is saying "this" or "that" with his hand stretched. The problem is,

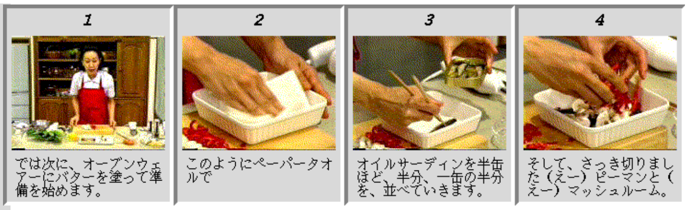
[2] We have not published any papers about these.



```
64 こちらを見て下さい
95 上の方を持って、こうねじって下さい
213 このグラフを見て下さい
239 この点で降雨量の最大値になっています
290 それは大型の台風が来たのが原因です
398 あのー、昨日ここに置いてあったダンボールは
どこに移動されたんですか
550 あそこにあるカメラとここにあるカメラの機能
が同じです
625 ここでは、メーカーの違いは考えません
765 あの実験から得た結果がこのグラフで示してい
ます
916 ネクタイを正確に結ぶポイントは、これです。
```

```
                2286
 9372   4924 -10616  15328   1452 -32680
 6808   4400  -7136  17844  14496 -19792
 8576   3020  -6840  -2136  10716  22624
 5344   8636  -7612  20132   3836 -25812
10252   3040  -2888   9164 -10788  23352
 7760   4036  -4244 -20704  14660   4636
```

Figure 3: Example view of comprehensive display



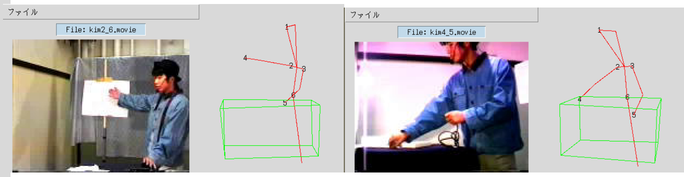
1 では次に、オーブンウェアーにバターを塗って準備を始めます。

2 このようにペーパータオルで

3 オイルサーディンを半缶ほど、半分、一缶の半分を、並べていきます。

4 そして、さっき切りました（えー）ピーマンと（えー）マッシュルーム。

Figure 4: Example view of cataloged display



(a) deictic case (b) simple operation

Figure 5: Example of typical gestures

however, not so simple. As we already discussed in Section 2.2, we have to check a variety of situations.

We have 207 deictic motions labeled "deictic movements" in the database. Checking the body motion, language data, we can obtain useful statistics as follows.

From the language point of view, we can check how demonstrative pronouns are used. An example of the retrieval results for "これ (KORE)" is shown in Table 4. "KORE" in Japanese usually has the same meaning as "this", and it is often used in deictic sense. Our statistics show 61% of the cases with "KORE" are deictic situations. In contrast to the above case, " それ (SORE)" has different characteristics, though "SORE" is often considered as "that" or "it". Only 3 % of the cases are deictic, and the predominant usage is anaphoric use[3]. Typical usage is shown in Table 5. This means that we need deeper analysis for deictic situation detection with "SORE" if we do not use gesture information. More detailed statistics can be obtained by checking the adjective use (modifier) of the pronouns, the modified nouns, or the verbs.

As for motion, situations with stretched hand are shown in Figure 5. The statistics of the distance between one of the hands and the body, that is the degree of the stretching, the duration of the motion pose can be gathered, for both of deictic cases and non-deictic cases. In our original presentations, only 33% of the cases with a stretched hand are for pointing. We have to think other conditions, for example, a condition that the hand moving speed should be local minimum at the moment of pointing.

With additional motion constraints and language information, the accuracy and recall rate of deictic movement detection was drastically improved in our experiment.

As shown above, MMID provides useful statistics to look over the usage of speeches, postures, or gestures, and to determine the system parameters such as thresholds. Moreover, MMID will contribute to clarify which combinations are effective for this purpose.

4.2 Summarizing Presentation

Video summarization is an important topic for multimedia computation. The target of the research is to pick up important segments which efficiently represent the whole story.

For this purpose, there have been researches based on scene change detection (for example, [Z+95] [YYL96]). A representative frame is chosen from each segment, *i.e.* cut/shot. Then representative frames are arranged in a window which gives an overview of the whole video. This approach, however, is often ineffective. For example, a cooking show contains many similar scenes that are difficult to distinguish without language explanations. We need more intelligent approach to pick up appropriate portions accompanied with language explanation.

In cooking shows, there are typical important operations such as cutting, putting, boiling, and so on. Basically, these situations can be associated typical words, such as "入れる (IRERU)" corresponding to "put in",

[3] referring a word or a sentence previously given.

Table 4: Example of “KORE” usage

1. これは置いておきましょう (leave this here)
2. これを1cm幅ぐらいに切っていきます (cut this with 1cm interval)
3. これぐらいで (with this amount)
4. これで終りにしましょう (finish at this moment)

Table 5: Example of “SORE” usage

1. それをざるに上げます (put it on the basket)
2. それと, お酒ですね。(alcohol in addition to that)
3. それで, . . . (then, ...)

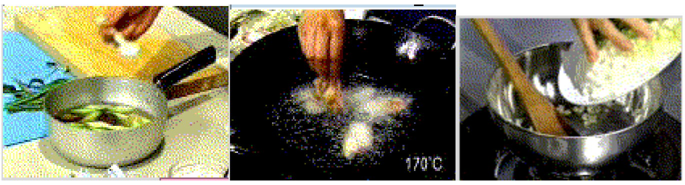
“加える (KUWAERU)” corresponding to “add”. 167 cases with these two words are found. The classification of the situations situations with those words are Table 6. As we can see in this table, if we gather all scenes with “IRERU”, 46% of them will represent putting situations.

On the other hand, if we query by motion labels, we get similar gestures. For example, we have 375 put-in situations in the database. We can easily understand there are several kinds of put-in motions, which are different in terms of human body motion. Three of them are shown in Figure 6.

By analyzing the characteristics of the retrieved words and motions, we can check how many words or how many motion patterns are required to detect put-in situations. Finally, by combining motion detection and natural language analysis, we will obtain a small set of relevant segments from a large sequential medium.

Table 6: Classification of put-in situations

Word	frequency	put-in motion
入れる (put-in)	60	28
入る (get-in)	34	9
置く (place)	29	14
盛り付ける (arrange)	13	7
others	33	19



(a) Put/pour something with a spoon

(b) Pour something by hand

(c) Put something by taking out

Figure 6: Example of put-in situations

5 Conclusion

In this paper, we introduced MMID, Multimodal Multiview Integrated Database. MMID stores human activities in presentations in terms of audio, video, human body motion, and transcripts. We can examine, by referring the data, how different modalities are cooperatively and complementarily used in real situations. MMID can serve as a basis for systematic or statistical analysis of those modalities, and can be a good tool when we design an intelligent user interface system or a multimedia contents handling system. We introduced two ongoing researches to show how MMID helps actual applications.

MMID is still at the prototyping stage, and it requires more powerful additional functions: more accurate motion sequence matching; retrieval by querying flexible modality combinations; a variety of image analysis including motion detection from broadcast programs. Finally but not least, it requires much more data, since human behaviors depend heavily on individuals, situations, and cultures.

References

[EF69] P. Ekman and W. Friesen. “The Repertoire of Nonverbal Behavior : Categories, Origins,Usage,and Coding”. *Semiotica*, Vol. 1, pp. 49–98, 1969.

[Kur92] T. Kurokawa. “Gesture Coding and a Gesture Dictionary for a Nonverbal Interface”. *IEICE Trans. Fundamentals*, Vol. E75-A, pp. 112–121, 1992.

[Kur94] Takao Kurokawa. *Nonverbal Interface (in Japanese)*. Ohm, 1994.

[McN87] D. McNeil. *“Psycholinguistics”*. Harper & Row, 1987.

[Oka96] Ryuichi Oka. Spotting method approach towards information integration. *RWC Technical Report (TR-96001)*, 1996.

[YYL96] M. Yeung, B. Yeo, and B. Liu. Extracting story units from long programs for viso browsing and navigation. *IEEE Int. Conference on Multimedia Computing and Systems*, pp. 296–305, 1996.

[Z$^+$95] H. Zhang, et al. Automatic Parsing and Indexing of News Video. *Multimedia Systems*, Vol. 2, No. 6, pp. 256–266, 1995.