Context- based Conversational Hand Gesture Classification in Narrative Interaction

Shogo Okada
Tokyo Institute of Technology
Yokohama, Japan
okada@dis.titech.ac.jp

Mayumi Bono
National Institute of Informatics
Tokyo, Japan
bono@nii.ac.jp

Katsuya Takanashi
Kyoto University
Kyoto, Japan
takanasi@ar.media.kyoto-u.ac.jp

Yasuyuki Sumi
Future University Hakodate
Hakodate, Japan
sumi@acm.org

Katsumi Nitta
Tokyo Institute of Technology
Yokohama, Japan
nitta@dis.titech.ac.jp

ABSTRACT

Communicative hand gestures play important roles in face-to-face conversations. These gestures are arbitrarily used depending on an individual; even when two speakers narrate the same story, they do not always use the same hand gesture (movement, position, and motion trajectory) to describe the same scene. In this paper, we propose a framework for the classification of communicative gestures in small group interactions. We focus on how many times the hands are held in a gesture and how long a speaker continues a hand stroke, instead of observing hand positions and hand motion trajectories. In addition, to model communicative gesture patterns, we use nonverbal features of participants addressed from participant gestures. In this research, we extract features of gesture phases defined by Kendon (2004) and co-occurring nonverbal patterns with gestures, i.e., utterance, head gesture, and head direction of each participant, by using pattern recognition techniques. In the experiments, we collect eight group narrative interaction datasets to evaluate the classification performance. The experimental results show that gesture phase features and nonverbal features of other participants improves the performance to discriminate communicative gestures that are used in narrative speeches and other gestures from 4% to 16%.

Categories and Subject Descriptors

I.5.2 [Pattern Recognition]: Design Methodology—Feature evaluation and selection; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Motion

General Terms

Algorithms, Experimentation

ICMI'13, December 9–13, 2013, Sydney, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522898.

Keywords

Hand gesture recognition; Contextual information; Small group; Conversation analysis

1. INTRODUCTION

Face-to-face conversations represent a fundamental social interaction. The automatic understanding of face-to-face conversational scenes by using the multimodal signals of participants assists conversation analyses [1], the implementation of conversational artifacts [20] including conversational agents [4]. Toward this goal, many studies focus on spoken language recognition from speech signals and nonverbal behavior recognition from visual signals. In nonverbal communication, nonverbal behaviors have established a key role in the formation, maintenance, and evolution of many fundamental social constructs. From this background, the analysis of automatic nonverbal behaviors in face-to-face interactions has been extensively researched. The voice tone and prosody are typical aural features that are useful in understanding the emotional state and agreement as well as turn taking, scene detection, and floor detection, among others [3]. Eye gaze, head direction, and head gestures are used to identify addressees and backchannel information [10], [21], [18].

On the other hand, extensive researches have been conducted regarding hand gestures in Social-linguistics and Psycholinguistics. However, in comparison to other nonverbal behavior (prosody, gaze, head gestures), there are fewer researches on the autonomous analysis and modeling of hand gestures used in small groups or multiparty conversations. Gestures play roles in communications at various times during a conversation. [5] shows that observing gestures is important for the tracking floor control structure.

Our goal is to automatically analyze the roles of hand gestures in small group conversations and explore effective features that discriminate these roles. For this purpose, we set the following hypothesis. If a speaker' s gestures have visual significance to a listener or other participants, nonverbal behaviors of other participants (e.g., gazing at the speaker, nodding to the speaker) are also important gestures. For example, a speaker is likely to use narrative gestures (cospeech gestures) to make other participants understand them well if the participants gaze at the speaker. On the other hand,

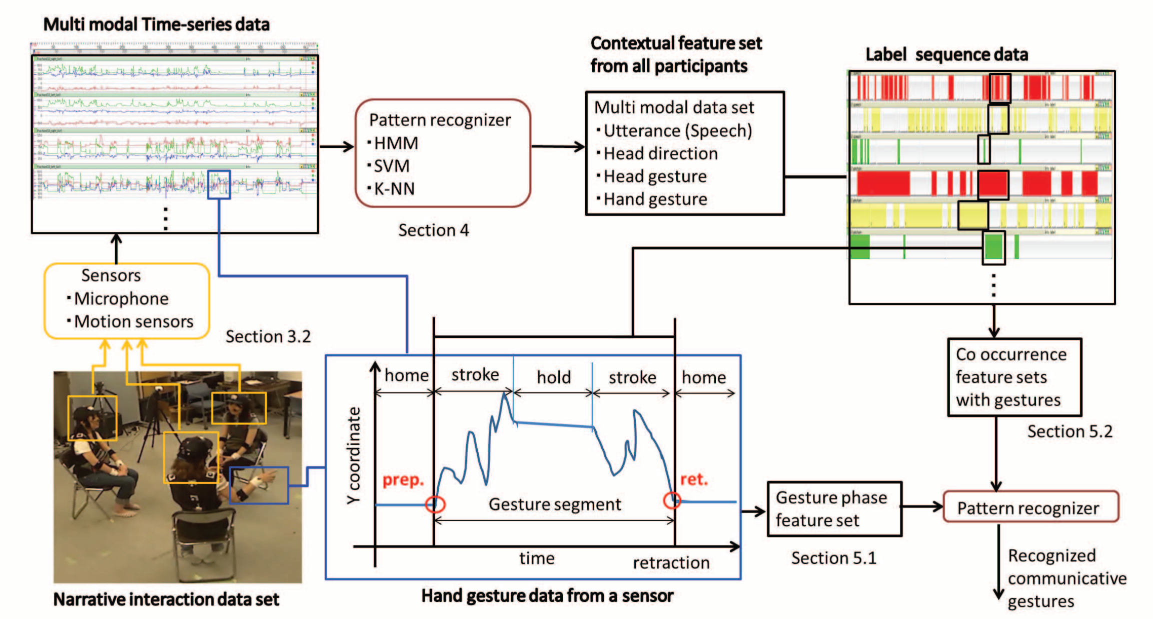

Figure 1: Overview of proposed frameworks: We use gesture phase features and contextual feature set from all participants for gesture modeling. Lower figure shows example of gesture phase features. “prep.”, “ret.” denote preparation phase and retraction phase, respectively.

spontaneous self-directed gestures or hand motions may also be used when participants do not gaze at the speaker. In this paper, we propose an autonomous framework to model hand gesture patterns used in conversations by using contextual information from other participants. First, we annotate primitive nonverbal patterns (speech, gesture, head gesture, and head direction) as binary on/off or three variables by using pattern recognition techniques.

Figure 1 shows an overview of the proposed framework. In this setting, the frequency of narrative gestures and their types are different among groups. This is because different speakers using gestures for explanations depend on specific individuals and may differ even if the speakers narrate the same episode. For this research, we collect a narrative interaction dataset where two people narrate a story from memory to one person. With this data, we can observe the gestures accompanying the narrative speech (narrative gesture), the self-directed gestures used when the storyteller does not remember the words, and the gestures of the non-speaker for interrupting a speaker’s turn. We set a task to classify the observed gesture patterns as gestures accompanying the narrative speech and other gestures.

Our paper has two contributions. First, the role classification of communicative gestures used in natural small group conversations is an unexplored problem. To our knowledge, using contextual features for hand gesture classification has not been addressed. We extract gesture features using the concepts of gesture phase and contextual features. We show that the classification of conversational hand gestures can be improved by considering the behaviors of other participants in the conversation. After the classification, we discover all participants’ nonverbal patterns coexisting with gestures using a probabilistic topic model to analyze how contextual information from other participants can be used to classify gesture roles.

2. PREVIOUS WORKS

Our challenge is to classify the communication gestures used in conversations. Our approach is different from other gesture recognition approaches in that we use a combination of the features of three motion phases, coexisting features, and contextual features of modality patterns from other participants. In other words, our research focuses on gesture recognition and multimodal recognition.

2.1 Psycholinguistic Research for Gestures

There have been many studies on human gestures in psycholinguistic research. McNeill argues that thought is related to the relationships between language and gestures that accompany a discourse and that gestures plays a role in representing a part of language [16]. Kendon describes a philology of gesture, consisting of gesticulation, language-like gestures, pantomimes, emblems, and sign language [11]. Kendon also defines gesture phases (preparation, stroke, hold, and retraction) for composition of communicative gestures. Communicative gestures are arbitrary and change depending on the conversational situation and the type of discourse. In such cases, it is not easy to extract common gesture features, such as hand shape, and measure the similarity between gestures in order to determine a position for classification. On the other hand, gesture phase features are composed from the structure of hand movements and are available for any communicative gestures. Psycholinguistic studies show that a stroke may be distinguished from other gesture phases, since a stroke contains maximum information. We also show that the hold phase is also an important feature that discriminates narrative gestures.

2.2 Gesture Recognition Techniques

There are numerous researches on autonomous hand gesture recognition, as summarized in [17][27]. According to [17], Hand gestures are often the most expressive and frequently used. They involve the following: (1) a posture: static finger configuration without hand movement and (2) a gesture: dynamic hand movement, with or without finger motion. Recognizing the start and end points of a meaningful gesture pattern from a continuous stream of input signals and subsequently segmenting the relevant gesture is a very difficult task due to the segmentation ambiguity and the spatial-temporal variability involved. To sense hand gestures, wearable sensors such as accelerometer sensors, computer vision devices, and motion capture devices are available.

To develop the recognition model from gesture patterns, Hidden Markov Models (HMM) [22], Conditional Random Fields (CRF) [13], Hidden Conditional Random Fields [26], Latent Dynamic Conditional Random Fields [19] are used. The main purposes of these systems are to provide a natural interface for interactive systems, to control robots, and to recognize sign language.

On the other hand, there are some existing researches for conversational gesture modeling and analysis of gestures by computational approaches. [14] proposed a vision-based sensing method for analyzing the upper body communicative patterns. In this research, manipulative gestures such as touching self, and beat gestures used in the job interview data are recognized automatically. [23] presents a multimodal framework for improving the gesture recognition accuracy. In this framework, alignment of speech act and hand kinematics is modeled by a Bayesian network and has been applied to the recognition of gestures of the forecasters in The Weather Channel broadcasts. [28] address oscillatory gesture recognition in natural conversations. In this research, oscillatory gestures are detected by using the frequency features of hand motion trajectory signals. They analyze relationships between the phases of speech and oscillatory gestures. [23] and [28] focus on communicative gesture recognition using multimodal signals (visual and acoustical features). Our approach uses not only multimodal signal from a participant making a gesture but also context features from other participants.

2.3 Context-based Multimodal Approach

Özyürek shows that speakers change the orientation of their gestures depending on the location of shared space [25]. We are inspired by the theory and set hypotheses to design gesture features in conversations. If speakers wish to impart meaning to a listener or other participants by using gestures, speakers will use gestures after other participant focus their gaze on them, interrupt floor gazing, or make other participants gaze at them. From this hypothesis, we decide to use context features from other participants (group gaze cue, speaking cue) for gesture classification.

[18] shows that using contextual cues (prosodic and lexical features) related to the speaker improves the recognition accuracy of the listener's head nods in a dyadic interaction setting. The finding is important for multimodal recognition. This research focuses on listener's head gesture recognition in a dyadic interaction. On the other hand, We focuses on hand gesture recognition in small group narrative interactions.

3. DATA SETTING

We collect small group narrative interaction datasets to evaluate our methods. We change the setting of a dyadic narrative interaction designed by McNeill [15] to that of a small group narrative interaction.

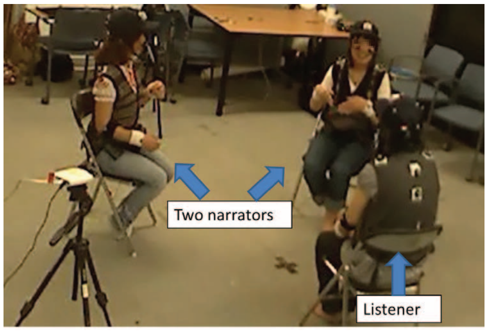

Figure 2: Interaction scene in small group narrative conversation: Two participants are asked to narrate a participant from memory a cartoon story (Canary Row)

3.1 Sensing Devices

We used a wearable motion capture system, Motion Analysis MAC 3D, and wearable accelerometer sensors, ATR promotion WAA-010 to accurately track hand and head motions stably (Figure 3). A passive optical system is used as a motion capture system. Several markers are placed to allow sensors to triangulate the three-dimensional (3D) position of a subject among ten cameras. The markers are placed at head and are listed to sense hand motions. A wearable accelerometer sensor is attached to the back of the head to measure head vertical motion including nodding. The motion capture system captures 120 frame samples per second. Shure wireless microphones (head set type) are used to collect voice data. We collected synchronized multimodal data from each participant, such as voice, hand and head movements.

3.2 Narrative Interaction Dataset

In this task, a participant is asked to narrate from memory a cartoon story to a participant. The name of the cartoon story is "Canary Row", and this story has been used for gesture analysis in narative tasks [15] ; In our setting, a group is composed of three participants, where two participants, who have watched the video, explain it to the other participant. Various gestures accompany the narrative speech provided by the two narrators. By increasing the number of narrators, we can observe that one narrator helps another to explain the story and observe the overlapping of gestures between participants. In this setting, we often observe meaningless hand motions that are not related to narrative speech; these are actually self-directed gestures (e.g., a gesture made when the narrator hesitates during speech).

Figure 2 shows an interaction scene where three participants sit on chairs without armrests, facing each other. The three participants were not acquainted with each other and have never watched the video or listened to the story. We recruited 24 women participants aged between 20 and 25 years through a temporary employee company to collect the dataset. We collected 8 sessions dataset in cooperation with these participants. Average time length of recorded datasets is 11 minutes (total is 700 minutes). Hand motion data including gesture has total 668214 frames.

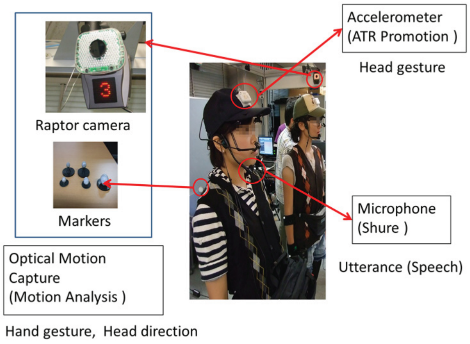

Figure 3: Sensors used in this recherches

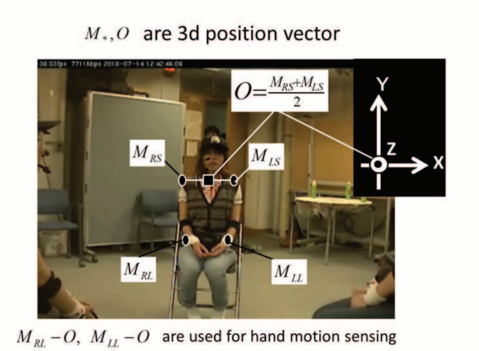
M_*, O are 3d position vector
$O=\frac{M_{RS}+M_{LS}}{2}$
M_{RS}
M_{LS}
M_{RL}
M_{LL}
Y
Z
X
$M_{RL}-O,\ M_{LL}-O$ are used for hand motion sensing

Figure 4: Coordinates relative to the center of participants

4. MULTIMODAL PATTERN ANNOTATION

Autonomous annotation is performed on the multimodal nonverbal dataset. We use the features: utterance, hand gestures, head gesture and head direction. Our final goal is to produce an autonomous recognition system. Therefore, we try to annotate each label by using heuristic combinations of the pattern recognition techniques. Manual annotation for the dataset is also performed to evaluate the accuracy of the autonomous annotation by using iCorpusStudio [24], software environment for browsing and analyzing the interaction corpus.

4.1 Speech Segment Annotation

Manual annotation setting : Utterance features are important for recognizing the gestures that accompany speech. We use standard speech signal processing techniques to detect speaking status as a binary variable to indicate whether a participant is speaking. Labels were annotated from manual transcriptions by an annotator. We define valid speech segments as utterances longer than 700 ms.

Autonomous annotation procedure : First, we apply heuristic rules by using the zero-crossing rate and amplitude to produce speech segment candidates. Second, two pre-trained Gaussian mixture models (for speech and non-speech) are used to recognize the speech segments. We use the following parameters for audio signal processing.

- Signal sampling: 16 kHz
- Frame length: 25 ms
- Frame shift: 10 ms
- Feature vector: 13 Mel-Frequency Cepstrum Coefficients (MFCCs), 13 ΔMFCCs

We use Julius software for this implementation[1] and

4.2 Hand Gesture Segment Annotation

The 3D positions of each marker attached to both lists are used as gesture features. Therefore, we do not sense finger movements in this research.

Manual annotation setting : First, we detect the gesture status as a binary variable indicating motion and nonmotion. We define nonmotion segments as the Ąghome positionĄh in the gesture phase. Here we define a position where both hands are placed on the thighs as home positions and annotate nonmotion segments when hands are set in the home position. All participants set their hands to this position when they do not make gestures because tables and armrests are not available. Thus, participants rest their hands on their thighs when not gesturing. On the other hand, we define segments that do not meet the criterion for nonmotion segments as motion segments.

We manually annotate training data to evaluate the hand gesture recognition algorithm and use it for gesture classification tasks in experiments. Labels were annotated by three individuals including the authors. The gesture annotation procedure is as follows.

Step1. Annotate nonmotion segments label.

Step2. Annotate start and end points of movements and labels motion segments between these points.

Step3. Annotate gesture phase labels: stroke, hold to motion segments.

Step4. Annotate preparation and retraction. The preparation and retraction phases are very shorter than hold and stroke phases, so we approximate the former two phases as a point (The lower figure in Figure 1).

Step5. Combine labels of right hand gesture and that of left hand one. When both hand gesture labels are equal at a frame (ex. both label are stroke or hold), the label is annotated. When both hand gesture labels are not equal at the frame and one label is annotated as stroke, The label is annotated as stroke. When these labels are annotated as home position and hold, the label is annotated as hold.

Autonomous annotation procedure : For preprocessing, the 3D positions of each marker are transformed into participant-centered coordinates (Figure 4). The 3D origin positions are set to the median value between the left and right shoulders. The original signals from markers are smoothed by a Gaussian filter of window size 50 ms; any missing value segments of less than 500 ms are interpolated by the linear interpolation method.

Hidden Markov Model (HMM) is used to recognize gesture segments and annotate on gesture phases. The HMM is the ergodic model based on Gaussian probabilities, where a full-covariance matrix is used for Gaussian function. We segment 2-s time-series data slices from the dataset of Session 1 by shifting one frame, and then collect the training data. The time-series data of 3D positions and their differentials are used as features to train the HMM. Here we prepare three HMMs for motion segments, nonmotion segments (home position), and hold phases. Preparation and retraction phases are recognized by HMMs as the change point from home position and the change point to home position, respectively. Finally, Three labels (home position (nonmotion), stroke, hold) are annotated on the gesture dataset.

[1] Julius: http://julius.sourceforge.jp

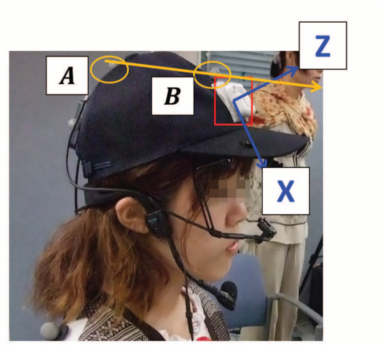

Figure 5: Features to detect head direction and head gesture from sensors

4.3 Head Gesture Segment Annotation

[8] classifies the roles of head gestures. [8] describes that the head gesture of not only the listener but also the speaker plays an important role. For example, a speaker nods to control and organize the interaction. The listener's nods are effective features to detect the backchannel context. We also annotate the head gesture by both a speaker and a listener.

Manual annotation setting : In the proposed frameworks, we annotate head nodding (vertical head motion) patterns as binary variables indicating a gesture or nongesture for each participant. In this narrative task, we do not record head shake patterns and annotate only vertical head movements including nodding. Labels were annotated by using video data by an author.

Autonomous annotation procedure : We detect head gestures by an approach [21] using the discrete wavelet transform (DWT). We use a wearable accelerometer and gyro sensor (A red rectangle in Figure 5) to extract the features of head movements. DWT features are calculated for the X rotation component of a gyro sensor and the X,Z components of an accelerometer. For wavelet analysis, we use the Daubechies wavelet of order 2 (db2) and a decomposition scale of 1-3. The maximum, mean, and standard deviations are calculated from the coefficients of levels 1 and 2. These values are used to define the feature vector of gestures. An SVM is trained to classify the feature vector into two categories: gesture or nongesture. This paper employs a Gaussian kernel and a soft margin criterion.

4.4 Head Direction Annotation

Detecting gaze patterns is useful to identify addressees. To detect gaze direction, head pose and head direction are used for the approximation of gaze direction, and visual focus of attention in group conversations [7] can be applied toward the automatic identification of addressees in multiparty cases. In this research, we also extract head direction features to detect the visual focus of attention. Although head tracking from vision remains a challenging task, we can observe head positions from motion capture sensors and avoid tracking difficulties.

Manual annotation setting : We annotate head direction patterns as binary variables indicating whether a participant is facing the right or left. Labels were annotated from manual transcriptions by an annotator.

Autonomous annotation procedure : The difference vectors in 3D position B of the head-front marker and S of the head-top marker are used to define feature vectors (Figure5). We use template matching by the k-nearest neighbor method for the classification of head direction patterns. We set $k = 2$ in this approach.

4.5 Summary of Autonomous Annotation

We have applied the standard pattern recognition techniques to annotate utterance, head gesture, hand gesture, and head direction.

We evaluated the performance of these algorithms. Training and classification was done by using dataset from each participant in one session. Annotation of that of head direction and that of hand gesture were done in session 1. Annotation of head gesture was done in session 9. Manual annotation data of one narrator in narrative interaction was used as a labeled training dataset for making the templates in KNN, training HMM, SVM. GMMs which have been trained from unspecified speakers in Julius are used for utterance annotation.

We summarize models an features, which were used for annotation in Table 1. Performance (recall and precision) of each recognizer are also shown in the table. Recall and precision ratio are calculated by using number of frames overlapped with manual annotated frames. Table 1 indicates that the annotation of utterance (on/off), hand gesture (on/off) and head direction (right/left) annotation were moderately successful. On the other hand, it indicates that we need to improve the performance to annotate head gesture (on/off) and hand gesture phase (rest/stroke/hold). Here, these trained models or templates are used for annotation of whole dataset.

5. CONTEXT BASED GESTURE FEATURES

After preparing label segment datasets annotated from each nonverbal signal, we define feature vectors of each gesture pattern from these label sets. The gesture feature sets are composed of two feature sets: gesture phase features and context features obtained from all participants.

5.1 Hand Gesture Features

Gesture segments (Section 4.2) include stroke segments or (and) hold segments. The duration of stroke and hold is different between gestures. Hold gestures are categorized by prestroke and poststroke hold phases [16]. In particular, a prestroke hold is an important feature related to stroke gesture. In this research, we focus on function of stroke and hold gesture parts in gesture segments and extract features related to the use of such gestures. The gesture feature set is defined as follows.

Total duration of gesture segment : The features denote the time length of gesture segment: TD

Mean and variation of 3d positions sequence : The features are calculated from the 3D positions of markers attached to both lists of a participant. The 3D position sequences within a timespan of the subjective gesture segment is defined as follows: $M = \{m_1, \ldots, m_T\}$, where m_l denotes the six-dimensional coordinate vectors of 3D positions in both lists. We calculate mean and variation statistic values: $|\triangle m_t| = |m_t - m_{t-1}|, (2 \leq t \leq T)$. These statistics denote the amount of hand movement variation. We define mean and variation from 3D positions as : Ms, Vs, respectively.

Frequencies of hold and stroke segment : In narrative gestures, hold and stroke gestures are repeated by turns. To characterize this property, repetition frequencies are important features. Here we count number of hold and stroke segments including in each gesture segment and this number is defined as the frequency. We define the features included in a gesture segment as $F = \{F_H, F_S\}$. In the gesture data

Table 1: Summary of autonomous annotation techniques: Performance denotes the result of experimental evaluation. r, p denote recall and precision.

	Features	Recognizer	Performance
Utterance	MFCCs, △MFCCs Zero cross ratio, Amplitude	GMM + Rule base	r=0.82, p=0.92
Head Direction	3d positions of head makers	KNN (k=2)	r=0.86, p=0.82
Head Gesture	Features extracted using discrete wavelet transform (DWT)	SVM	r=0.75, p=0.69

Hand Gesture	3d positions of list makers	HMM	r=0.80, p=0.75 (Gesture phase annotation)
			r=0.99, p=0.82 (Gesture segment annotation)

example of Figure 1, two hold segments and one stroke segment are observed in the segment, therefore $F_H = 2$, $F_S = 1$.

Sum of duration of hold and stroke segment : The sum of duration of hold Sd_H is calculated by the following equation: $Sd_H = \sum_{i=1}^{F_H} T_i^H/TD$, where T_i^H denotes the time length of the ith hold segment in a gesture segment. Sd_S is also calculated similarly. Finally, we define this feature as $Sd = \{Sd_H, Sd_S\}$

5.2 Conversational Context Features

Conversational context feature sets are composed of self-context features from participant making subjective gestures and context features from other participants. All context feature sets are calculated using co-occurrence relations between gesture segments and each nonverbal segments (utterance, head gesture, and head direction). We define the annotated gesture segment set as
$\mathcal{G}_i = \{g_1, \ldots, g_x, \ldots, g_{N_g}\}$, where $g_x = \{st^g_x, et^g_x\}$. N_g denotes the total number of gesture segments, st^g_x, et^g_x are the start and end time of gesture segments g_x, respectively. We define a participant who makes subjective gestures as $S1$, the other narrator as $S2$, and the listener who has not watched the animation as $S3$. Here, $\mathcal{G}_{i,j}$ are composed of all gestures made by the two narrators in 8 sessions. Index i denotes the session number $(1 \leq i \leq 8)$.

Next, we explore how looking, speaking, and head gesture cues influence gesture recognition accuracy. We define the other nonverbal segment dataset in a manner similar to the gesture segment dataset. The nonverbal dataset is defined as follows.

Speech utterance (speech) segment dataset is $\mathcal{SP}_i$, where $\mathcal{SP}_i = \{SP1_i, SP2_i, SP3_i\}$. $SP1_i$ denotes the speech segments from $S1$.

Head gesture : head gesture segment dataset is $\mathcal{HG}_i$ where $\mathcal{HG}_i$ is defined to be the same as that in $\mathcal{SP}_i$.

Head direction : head direction segment dataset is $\mathcal{H}_i$ We define gaze conditions of the three participants as six states. $\mathcal{H}_i$ is defined below.
$\mathcal{H}_i = \{H12_i, H13_i, H21_i, H23_i, H31_i, H32_i\}$, where $H12_i$ denotes the segments where $S1$ faces $S2$, $H12_i$ is defined to be the same as that in $\mathcal{G}_i$

We calculate the co-occurrence joint features as context features by using the overlap between two segments. Overlap ratio between gesture segments and other segments is calculated by the following equation:

$$P(g_x, np_y) = \frac{min(et^g_x, et_y) - max(st^g_x, st_y)}{l_x},$$
$$P(g_x, np) = \max_y P(g_x, np_y), \quad (1)$$

where l_x is the time length of g_x, $max(a, b)$, $min(a, b)$ denotes the maximum and minimum values of frame a and frame b, respectively. When $P(g_x, np) < 0$, $P(g_x, np) = 0$, because np_y which overlaps with the gesture does not exist. We input each segment in $SP1 - SP3, HG1 - HG3, H12 - H32$ into segments np_y and calculate the ratio between gesture segments and other context feature segments from all participants. We calculate this ratio in all sessions. Finally, we complete the context features for xth gesture segment in the session i as bellow.

$$\mathcal{P}_{i,x} = \{P(g_x, sp1), \ldots, P(g_x, hg1), \ldots, P(g_x, h32)\}, \quad (2)$$

where $sp1, hg1, h32$ denotes the segment of $SP1, HG1, H32$, respectively . The context feature is a 12-dimensional vector.

6. EXPERIMENT SETTING

In this section, we evaluate the proposed framework by the classification task of multimodal gesture patterns into narrative and other gestures. Annotations are performed on all datasets by applying the pattern recognition techniques described in Section 4. Head gesture detection, estimation of head direction, and utterance segment detection are performed automatically. On the other hand, we use manual annotation data for hand gestures to evaluate the effectiveness of the gesture phase feature in combination with other features. First, 547 gesture segments are annotated manually, which included 162 narrative gestures. We define the 162 narrative gesture segments as a positive class and the other 385 gesture segments as a negative class.

To train the proposed context-based classifier, we balance the number of positive and negative samples, so we choose 162 negative gesture samples from the dataset randomly. In this experiment, a 10-fold testing approach is used. In each class, 17 samples are used for test purposes only, and 145 samples are used for validation and training. This process is repeated 10 times, and a total of 100 experiments are performed.

We calculate the accuracy (number of true positives/total test data) to evaluate the classification performance. In our experiments nonlinear SVM and AdaBoost [6] algorithms are used for the classification of multimodal context features. We use a Gaussian kernel for nonlinear SVM and validate the bandwidth parameter γ of the kernel $(\gamma = 10^k, k = 1.. - 3)$. In AdaBoost the number of tree splits for weak learning is validated with the value $T = 5, ..10$.

We explore the feature set to discriminate gestures that accompany narrative speech and other gestures. To explore the feature set, we prepare four combinations of multimodal context features.

- $F1$: Gesture signal feature set $(TD + Ms + Vs)$
- $F2$: All gesture feature set ($F1 + F + Sd$: This feature set includes gesture signals and phase features.

Table 2: Classification accuracy [%] of gesture patterns by SVM and AdaBoost

[%]	F1 Motion features	F2 F1 + Gesture phase features	F3 F2 + Self context	F4 (Proposed) All features
SVM	59.1 ± 7.3	71.3 ± 7.0	71.7 ± 7.5	**75.6 ± 7.0**
AdaBoost	70.8 ± 8.5	71.7 ± 7.9	72.8 ± 7.7	73.7 ± 7.3

- $F3$: Self-context-based gesture feature set: $(F2 + P(SP1) + P(HG1) + P(H12) + P(H13))$

- $F4$: conversational context-based gesture features and gesture phase feature set (complete feature combination) $(F2 + \mathcal{P})$

On the case that different kind of features are combined into 1 vector, we have to normalize the scale of sample value. We use the min-max data normalization method for this purpose.

7. RESULTS AND DISCUSSION

Table 2 shows the classification results of SVM and AdaBoost. In the experiments, the classification accuracy is calculated by dividing the total number of true positives and true negatives by total number of test data.

In both algorithms, increasing the number of features improves the classification accuracy. In particular, the proposed feature fusion strategy results in better classification accuracy in the SVM. First, using gesture phase features improves the accuracy from 59.1% (F1) to 71.3% (F2). Second, using context features from other participants improves the SVM accuracy from 71.7% (F3) to 75.6% (F4). Classification using the complete feature set F4 produces the best accuracy in both algorithms. We compare the accuracies from both F4 and other groups using the Student's t-test in SVM. A Student's t-test shows significant difference ($p < 0.05$) between F4 and the other groups (F1-F3). On the other hand, the t-test shows the significant difference between F1 and F4 in AdaBoost. Totally these results show that gesture phase features and context features from all participants improve classification accuracy.

7.1 Context Feature Analysis using Topic Model

To identify the context features and gesture phase features which contribute to recognize the narrative gesture patterns, we use one of topic models: Latent Dirichlet Allocation (LDA) [2] to analyze the effective features. By using LDA, we extract the typical co-occurrence pattern set (joint features) with gestures accompany to narrative speech. We explore features which characterize the narrative gestures by relating the each typical co-occurrence pattern to each conversation scene,

LDA is a probabilistic generative model for documents composed from word set (Bag of Words feature set). In LDA, a text document d is modeled as a distribution $p(z = t|d) = \theta_t^{(d)}$ over topics t, and a topic is as a multinomial distribution $p(w|z = t) = \phi_w^{(t)}$ over words w. It is also used for nonverbal behavior analysis [9]. Here we set a document in the topic model as a gesture segment and words as the gesture features (gesture phase and context features).

Procedure of analysis by using LDA is as follows.

- Discretize each feature value by using a clustering method. Hierarchical clustering based a ward method is used for discretization of it and number of discretization levels (clusters) is set to 4. Number of topics T is set from 5 to 15. In this paper, we show the analysis result in $T = 10$.

- Estimate parameters of LDA from all gesture dataset ($N = 547$) by using Gibbs sampling.

- Identify topics which are likely to generate narrative gesture segment. Let narrative gesture segment set is $\mathcal{G}_p$ and the other gesture segment set is $\mathcal{G}_n$, we calculate $\theta_t^{(d_i)}, d_i \in \mathcal{G}_p$ and $\theta_t^{(d_j)}, d_j \in \mathcal{G}_n$. Next, we compute $M_t^p = p(z = t|\mathcal{G}_p)$ by averaging over $p(z = t|d_i)$ and M_n^p in same manner. If a topic t^* is generated from gesture segments d_i in $\mathcal{G}_p$ frequently, M_t^p is larger than M_t^n. We compare $\theta_t^{(d_i)}$ and $\theta_t^{(d_j)}$ by using Student's t-test. If a Student's t-test shows significant difference ($p < 0.05$) between them, we defined the topic as t^*.

- Analyze the word distribution for topic t^*.

Table 3 shows set of topic t^* and words (gesture features) which generate t^*. Probability P_f of features is calculated as $P_f = \phi_w^{(t^*)}/(\sum_w(\phi_w^{(t^*)}))$. We report P_f which is larger than 0.2 in the table. The upper table shows topics which generate the narrative gestures frequently.

Topic 2 shows that sum of duration of stroke Sd_S, frequency of hold segments F_H, total duration TD contribute to generate the topic. This means that participants are more likely to use hold gesture and long stroke gesture accompany to narrative speech. Topic 4 shows that the other participants are likely to make head gestures (nodding and backchannel) when participants make narrative gestures.

Topic 5 shows that participants gaze to the other participants (P_f of $HG2, 3$ is high) and also use long hold gesture. Topic 7 shows that narrative gesture segments has long time length (TD) and are made accompany to speech ($SP1$). This is natural conversation context. Topic 10 shows that the other gestures are used frequently, when another narrator $SP2$ speaks. We observed this scene from video data, and it is found that hands of a narrator $S1$ move when a narrator $S2$ speaks. This gestures are not be paid attention and will be self directed.

Totally, when participants use communicative gesture accompany to narrative speech, (1) the other participants are likely to make head gestures, (2) hold gestures are used frequently and (3) long stroke and long hold gestures. This analysis also shows that the context features and the gesture phase features are effective to identify these gesture.

7.2 Gesture's Role in Multimodal Scenario

We showed possibility of the gesture-role-classification via linking of contextual features with it. Seeing from the other view point, these results also showed a possibility that communicative gestures can be used as features to recognize the conversation scene, such as turn taking, floor transition and addressing toward participants; and behaviors of speaker: explanation and hesitate of saying.

Table 3: Topics which generate gesture segments of each class frequently

Narrative gesture								The others	
Topic 2		Topic 4		Topic 5		Topic 7		Topic 10	
word	P_f	word	P_f	word	P_f	word	P_f	word	P_f
Sd_S	0.52	$HG2$	0.40	H_{12}	0.41	TD	0.45	$SP2$	0.52
F_H	0.27	$HG3$	0.37	H_{13}	0.32	$SP1$	0.30	Ms	0.48
TD	0.21					Sd_H	0.25		

Toward understanding the roles of gestures in various conversational scene, we need to analyze the role of gestures in richer multimodal scenario with lexical and prosodic features and link the gestures with each conversation scene or social attitude of the participant. These features help us to characterize and model communicative gestures. [12][3] show that prosodic features such the rhythm, pitch and intonation of speech are also important features which relates to communicative gestures. Therefore using prosodic features and lexical features are future works.

8. CONCLUSION

This paper proposed a feature extraction method for the classification of communicative gestures used in small group interactions. As effective features to discriminate conversational gestures accompanying narrative speech, we extracted gesture phase features and co-occurring nonverbal patterns with the gestures, i.e., utterance, head gesture, and head direction of each participant. Experimental results showed that using a combination of these features improved the classification accuracy related to conversational gestures. This research focuses on classifying communicative gestures made with narrative speech. As future work, we plan to apply this feature extraction method to multiclass communicative gesture role classification.

Acknowledgment

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for B, 22700146, 25730132, 2010-2013.

9. REFERENCES

[1] J. M. Atkinson and J. Heritage. Structures of Social Action: Studies in Conversation Analysis. Cambridge University Press., 1984.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[3] N. Campbell. On the use of nonverbal speech sounds in human communication. In Proc. of the COST 2102 Workshop on Verbal andNonverbal Communication Behaviours, 2012.

[4] J. Cassell, S. Prevost, J. Sullivan, and E. Churchill. Embodied Conversational Agents. MIT Press, 2000.

[5] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang, and F. Quek. A multimodal analysis of floor control in meetings. In Proc. of international conference on Machine Learning for Multimodal Interaction, pages 36–49, 2006.

[6] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In Proc. of European Conference on Computational Learning Theory, EuroCOLT '95, pages 23–37. Springer-Verlag, 1995.

[7] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. Image Vision Computing, 27(12):1775–1787, nov 2009.

[8] D. Heylen. Challenges ahead head movements and other social acts in conversation. In In Artificial Intelligence and Simulation of Behaviour (AISB 2005), Social Presence Cues for Virtual Humanoids Symposium, pages 45–52, 2005.

[9] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In Proc. of ACM ICMI, pages 433–440, 2012.

[10] S. Kawato and J. Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". In Proc. of IEEE FG, pages 40–45, 2000.

[11] A. Kendon. Gesture: Visible Action as Utterance. Cambridge University Press, 2004.

[12] S. Kettebekov. Exploiting prosodic structuring of coverbal gesticulation. In Proc. of ACM ICMI, ICMI '04, pages 105–112, 2004.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of ICML, pages 282–289, 2001.

[14] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. S. Nguyen, and D. Gatica-Perez. Body communicative cue extraction for conversational analysis. In Proc. of IEEE FG, pages 1–8, 2013.

[15] D. McNeill. Hand and Mind: What Gestures Reveal about Thought. Psychology/cognitive science. University of Chicago Press, 1996.

[16] D. McNeill. Gesture and Thought. University of Chicago Press, 2008.

[17] S. Mitra and T. Acharya. Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 37(3):311–324, 2007.

[18] L.-P. Morency, I. K. de, and J. Gratch. Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In Proc. of ACM ICMI, pages 181–188, 2008.

[19] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In Proc. of IEEE CVPR, pages 1–8, 2007.

[20] T. Nishida. Conversational informatics: an engineering approach, volume 9. Wiley. com, 2008.

[21] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "who responds to whom, when, and how?" from gaze, head gestures, and utterances. In Proc. of ACM ICMI, pages 255–262, 2007.

[22] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proc. of IEEE, pages 257–286, 1989.

[23] R. Sharma, J. Cai, S. Chakravarthy, I. Poddar, and Y. Sethi. Exploiting speech/gesture co-occurrence for improving continuous gesture recognition in weather narration. In Proc. of IEEE FG, pages 422–427, 2000.

[24] Y. Sumi, M. Yano, and T. Nishida. Analysis environment of conversational structure with nonverbal multimodal data. In Proc. of ACM ICMI-MLMI, pages 44–47, 2010.

[25] Özyürek, A. Do speakers design their cospeech gestures for their addressees ? Jounal of Memory and Language, pages 688–704, 2002.

[26] S. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In Proc.of IEEE CVPR, volume 2, pages 1521–1527, 2006.

[27] Y. Wu and T. S. Huang. Vision-based gesture recognition: A review. In Proc. of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction, GW '99, pages 103–115. Springer-Verlag, 1999.

[28] Y. Xiong, F. Quek, and D. McNeill. Hand motion gestural oscillations and multimodal discourse. In Proc. of ACM ICMI, pages 132–139, 2003.