Context-based Recognition during Human Interactions: Automatic Feature Selection and Encoding Dictionary

Louis-Philippe Morency
USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292
morency@ict.usc.edu

Iwan de Kok
Human Media Interaction Group, University of Twente
P.O. Box 217, 7500AE
Enschede, The Netherlands
i.a.dekok@student.utwente.nl

Jonathan Gratch
USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292
gratch@ict.usc.edu

ABSTRACT

During face-to-face conversation, people use visual feedback such as head nods to communicate relevant information and to synchronize rhythm between participants. In this paper we describe how contextual information from other participants can be used to predict visual feedback and improve recognition of head gestures in human-human interactions. For example, in a dyadic interaction, the speaker contextual cues such as gaze shifts or changes in prosody will influence listener backchannel feedback (e.g., head nod). To automatically learn how to integrate this contextual information into the listener gesture recognition framework, this paper addresses two main challenges: optimal feature representation using an encoding dictionary and automatic selection of optimal feature-encoding pairs. Multimodal integration between context and visual observations is performed using a discriminative sequential model (Latent-Dynamic Conditional Random Fields) trained on previous interactions. In our experiments involving 38 storytelling dyads, our context-based recognizer significantly improved head gesture recognition performance over a vision-only recognizer.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motion*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*

General Terms

Algorithms

Keywords

Contextual information, visual gesture recognition, human-human interaction, head nod recognition

1. INTRODUCTION

Face-to-face communication is highly interactive. Even when only one person speaks at the time, other participants exchange information continuously amongst themselves and with the speaker through gesture, gaze, posture and facial expressions. Such feedback is an essential and predictable aspect of natural conversation and its absence can significantly disrupt participants ability to communicate [3, 23]. Recognizing these visual gestures is important for understanding all the information exchanged during a meeting or conversation, and can be particularly crucial for identifying more subtle factors such as the effectiveness of communication [17], points of confusion, status relationships between participants [18], or the diagnosis social disorders [16].

This paper argues that it is possible to significantly improve state-of-the art recognition techniques by exploiting regularities in how people communicate. People do not provide feedback at random. Rather they react to the current topic, previous utterances and the speaker's current verbal and nonverbal behavior [1]. For example, listeners are far more likely to nod or shake if the speaker has just asked them a question, and incorporating such dialogue context can improve recognition performance during human-robot interaction [14]. More generally, speakers and listeners co-produce a range of lexical, prosodic, and nonverbal patterns. Our goal is to automatically discover these patterns using only easily observable features of human face-to-face interaction (e.g. prosodic features and eye gaze), and exploit them to improve recognition accuracy.

In this paper, we show that we can improve the recognition of conversational gestures by considering the behaviors of other participants in the conversation. Specifically, we show that the multimodal context from the current speaker can improve the visual recognition of listener gestures. We introduce the idea of encoding dictionary, a technique for contextual feature representation inspired by the influence speaker context has on the listener feedback. We perform automatic selection of relevant contextual features by looking at individual and joint influences of context. The final contextual integration is done using a discriminative sequential model. We proof the importance of speaker context on a head nod recognition task using a large dyadic-storytelling dataset.

The following section describes previous work in visual gesture recognition and explains the differences between our context-based approach and other recognition models. Section 3 discusses the contextual information available during human-human interactions. Section 4 describes the details of our context-based recognition framework including the encoding dictionary and our feature selection algorithm.

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

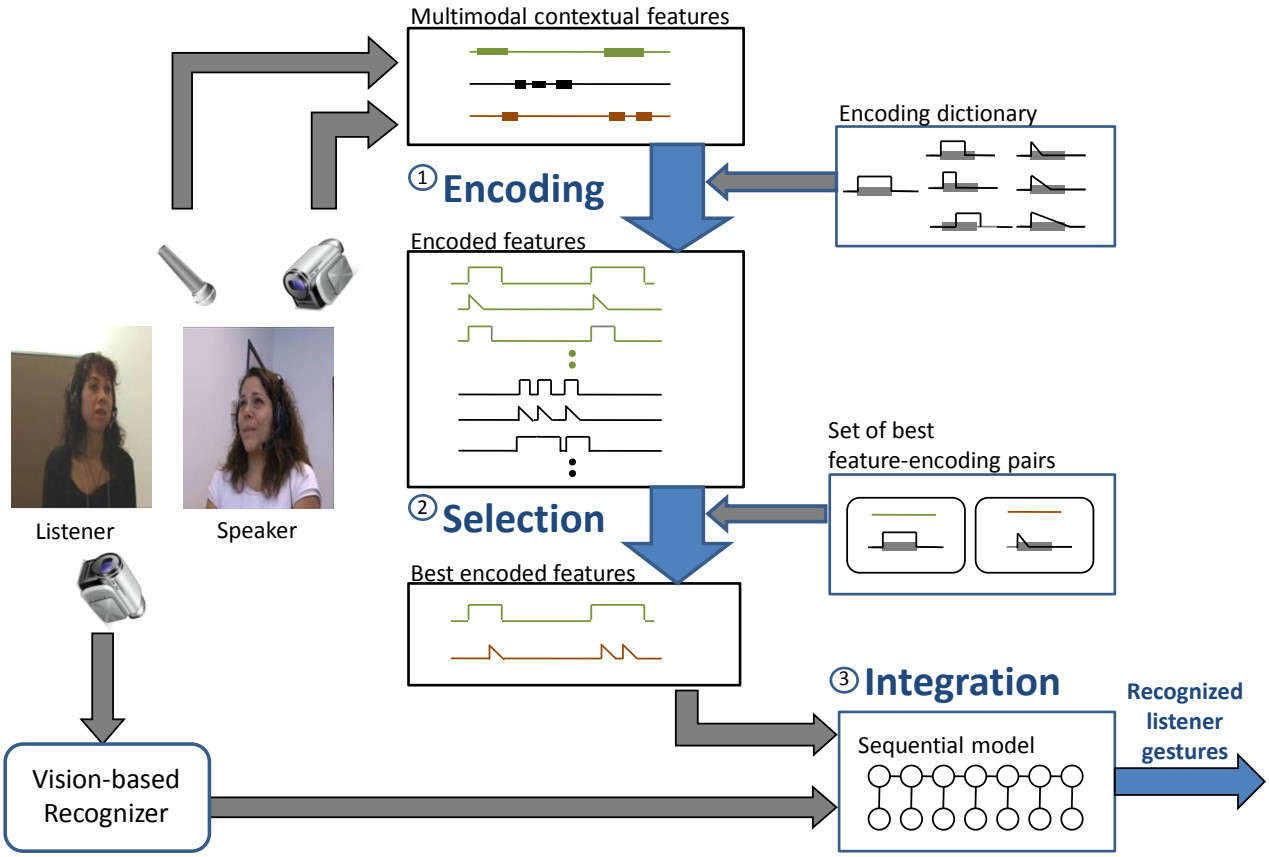

Figure 1: Overview of our context-based gesture recognition framework. To recognize online the visual gestures of a person (i.e., the listener), we integrate contextual information (e.g., prosodic cues, eye gaze,...) from other participants (i.e., the speaker) in three steps: (1) the contextual information is encoded using a template dictionary build based on co-occurrence of context and gestures, (2) select the feature-encoding pairs based on their individual and joint influences, and (3) integrate the contextual information to the visual observations using a probabilistic sequential model (e.g., Conditional Random Field).

Section 5 presents the way we collected the data used for training and evaluating our model as well as the methodology used to evaluate the performance of our approach. The results are discussed in Section 6.

2. PREVIOUS WORK

Our approach to recognize visual feedback is unique in that it incorporates contextual features of other human participants in an interaction beyond the person of interest. However it relates to and builds upon several approaches for audio-visual speech and multimodal recognition.

Recognition of head gestures has been demonstrated by tracking eye position over time. Kapoor and Picard presented a technique to recognize head nods and head shakes based on two Hidden Markov Models(HMMs) trained and tested using 2D coordinates from an eye gaze tracker [10]. Kawato and Ohya suggested a technique for head gesture recognition using between eye templates [11]. Fujie *et al.* also used HMMs to perform head nod recognition [7]. In their paper, they combined head gesture recognition with prosodic low-level features computed from Japanese spoken utterances to determine strongly positive, weak positive and negative responses to yes/no type utterances.

Several researchers have developed models to predict when backchannel should happen based mostly on unimodal inputs. Ward and Tsukahara [20] propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110ms during speech. Models were produced manually through an analysis of English and Japanese conversational data. Nishimura et al. [15] present a unimodal decision-tree approach for producing backchannels based on prosodic features. The system analyzes speech in 100ms intervals and generates backchannels as well as other paralinguistic cues (e.g., turn taking) as a function of pitch and power contours. Cathcart et al. [5] propose a unimodal model based on pause duration and trigram part-of-speech frequency. The model was constructed by identifying, from the HCRC Map Task Corpus [2], trigrams ending with a backchannel. In contrast to these gesture generation systems, our approach uses the contextual information from the speaker to improve recognition of listener gestures.

Context has been previously used in computer vision to disambiguate recognition of individual objects given the current overall scene category [19]. In contrast to the idea of fusing multiple modalities from the human participant to improve recognition (e.g., Kaiser *et al.* work on multimodal

interaction in augmented and virtual reality [9] and Yiong *et al.* Catchment Feature Model [22]), our approach takes its contextual information from the other participants. More closely related, Morency *et al.* [14] used dialogue state information to improve recognition accuracy in the context of a human-robot interaction. However, that application assumed privileged access to the mental state of one of the participants (i.e., the robot) that is not directly observable in human-to-human interaction. To our knowledge no previous work has explored the use of dialogue context from other human participants for visual recognition of interaction gestures.

3. CONTEXT IN HUMAN INTERACTIONS

Communication is a joint activity and social scientists have long argued that it cannot be properly recognized and understood by focusing on participants in isolation but rather one must see individual behaviors within the context of the group or dyad [4, 6]. Translating this proscription to the domain of gesture recognition, this argues that features outside of the person-of-interest should correlate with their behavior, and representing and exploiting these contextual features should improve recognition accuracy. Here, we explore this idea within the domain of dyadic conversations, specifically we consider whether adding contextual information about a speaker's behavior improves the ability to detect feedback gestures produced by a listener.

As our interest is in producing online (real-time) recognition systems, we focus on contextual features that would be readily available to a real-time system (i.e, surface behaviors rather than the privileged mental state of individual participants). Prior research into face-to-face conversation has identified a number of shallow features of a speakers behavior that correlate with listener feedback:

Prosody Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody [15]. For example, Ward and Tsukahara [20] show that short listener backchannels (listener utterances like "ok" or "uh-huh" given during a speaker's utterance) are associated with a lowering of pitch over some interval. We encode the following prosodic features, including standard linguistic annotations and the prosodic features suggested by Ward and Tsukhara [20]:

- Downslopes in pitch continuing for at least 40ms
- Regions of pitch lower than the 26th percentile continuing for at least 110ms (i.e., lowness)
- Utterances longer than 700ms
- Drop or rise in energy of speech (i.e., energy edge)
- Fast drop or rise in energy of speech (i.e., energy fast edge)
- Vowel volume (i.e., vowels are usually spoken softer)
- Lengthened words (e.g., "I li::ke it")
- Emphasized or slowly uttered words (e.g., "ex_a_c_tly")
- Words spoken with continuing intonation
- Words spoken with falling intonation (e.g., end of an utterance)
- Words spoken with rising intonation (i.e., question mark)

Pauses Listener feedback often follows speaker pauses or filled pauses such as "um" (see [5]). To capture these possible associations, we use the following contextual features:

- Pause in speech (i.e., no speech)
- Filled pause (e.g. "um")

Gesture display Gestures performed by the speaker are often correlated with listener feedback [4]. Eye gaze, in particular, has often been implicated as eliciting listener feedback. Thus, we encode the following contextual feature:

- Speaker looking at the listener

Lexical Finally, some studies have suggested an association between lexical features and listener feedback [5]. Although lexical features are not as easy to recognize in real time as the previous features, however there has been recent progress in real-time keyword spotting [8] and we include these for completeness (note: in our experiments none of the lexical features were selected):

- All individual words (i.e., unigrams)

4. CONTEXT-BASED RECOGNITION

The goal of our approach is to integrate contextual information from other human participants when recognizing visual feedback of a specific participant. During dyad interactions, the context is defined by the speaker verbal and nonverbal actions. Our approach integrates the speaker context with the visual observations from the listener to improve recognition of the listener gestures (see Figure 1). In an offline phase we learn based on previous recorded interactions which contextual features are important and how they should be encoded. During online recognition, the contextual features are encoded, selected and then integrated with the vision-based observations using a probabilistic sequential model, referred as multimodal integrator which outputs the final recognition results. Figure 1 shows an overview of our approach.

Two important challenges in training a probabilistic sequential model (e.g., Hidden Markov Model or Conditional Random Fields) for multimodal integration are (1) how to encode the contextual information in a format that will facilitate training of the sequential model and (2) how to select only the relevant contextual features out of the whole speaker context. To overcome these problems, we introduce two new approaches for feature representation and selection:

- **Encoding dictionary** Our encoding dictionary contains a series of templates designed to model different relationship between a contextual feature and visual gestures. The encoding dictionary and its usage are described in Section 4.1.
- **Automatic selection of feature-encoding pairs** We suggest two techniques for automatic feature and

encoding selection based on co-occurence statistics and performances evaluation on a validation dataset. Our feature selection algorithms are described in Section 4.2.

The following sub-sections describe the encoding, selection and integration stages depicted in Figure 1. The vision-based recognition algorithm used in our experiments is described in Section 5.3.

4.1 Encoding Dictionary

The goal of the encoding dictionary is to propose a series of encoding templates that potentially capture the relationship between contextual events and visual gestures. These templates were designed based on two main observations from our user study analysis:

- **Response delay** A delay is sometimes observed between the listener visual feedback and a specific contextual event from the speaker. This can be explained by the facts that the listener takes time to process the speaker information or that more than one contextual cue jointly triggered the listener feedback. For example, the listener may not head nod immediately after the speaker lowered his/her volume and pitch at the end of a sentence, waiting for the speaker to look back.
- **Lingering effect** The relationship between speaker contextual events and listener visual feedback may not always be constant over time.

It is important to note that a feature can have an *individual* influence on feedback and/or a *joint* influence. An *individual* influence means the input feature directly influences visual feedback. For example, a long pause can by itself trigger visual feedback from the listener. A *joint* influence means that more than one feature is involved in triggering the feedback. For example, saying the word "and" followed by a look back at the listener can trigger listener feedback. This also means that a feature may need to be encoded more than one way since it may have a *individual* influence as well as one or more *joint* influences.

Figure 2 shows the 13 encoding templates designed based on these observations. These encoding templates were selected to represent a wide range of ways that a contextual feature can influence the visual feedback. These encoding templates were also selected because they can easily be implemented in real-time since only the start time of the contextual feature is needed. Only the binary feature also uses the end time. In every case, no knowledge of the future is needed. The three main types of encoding templates are:

- **Binary encoding** This encoding is designed for contextual features which influence on visual feedback is constrained to the duration of the contextual feature. For example, it is unlikely that a listener will head nod if the speaker is not looking. The feature *Speaker looking at the listener* should then be encoded as binary so that it acts as a direct filter of listener visual feedback.
- **Step function** This encoding generalizes binary encoding by adding two parameters: width of the encoded feature and delay between the start of the feature and its encoded version. This encoding is useful if the feature influence on feedback is constant but with a certain delay and duration. For example, a listener may take more or less time before answering the speaker question with visual feedback (e.g., head nod or head shake) because they are thinking about it. After a certain time, a visual answer becomes unlikely (the person probably answered verbally).
- **Ramp function** This type of encoding linearly decreases for a set period of time (i.e., width parameter). This encoding is useful if the feature influence on feedback is changing over time.

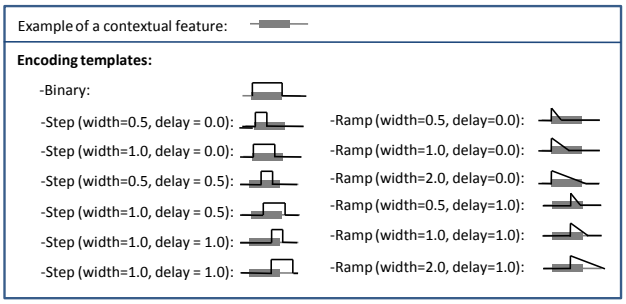

Figure 2: Encoding dictionary. This figure shows the different encoding templates used by our context-based approach. Each encoding template was selected to express a different relationship between contextual features and visual feedback. This encoding dictionary gives a more powerful set of input features to the sequential probabilistic model and improves the performance of our context-based recognizer.

4.2 Automatic Feature Selection

We perform the feature selection based on the same concepts of *individual* and *joint* influences described in the previous section. Individual feature selection is designed to assess the individual performance of each contextual feature while the joint feature selection looks at how features can complement each other to improve performance. In our experiments the original set of contextual features contained 1400+ features, including the lexical features (i.e., spoken words by the speaker). The automatic feature selection is important to reduce the chances of overfitting by our multimodal integrator (described in Section 4.3). Also, performing the *joint* feature selection on the original set of features would be too time consuming. For this reason we first perform *individual* feature selection, as described in the following section.

4.2.1 Individual Feature Selection

Individual feature selection is based on (1) the statistical co-occurrence of contextual features and visual feedback, and (2) the individual performance of each contextual feature when trained with any encoding template and evaluated on a validation set.

The first step of individual selection looks at statistics of co-occurrence between visual gestures and contextual features. The number of co-occurrence is equal to the number of times a visual feedback instance happened between the start time of the contextual feature and up to k seconds after it. In our experiments, k was set to 2 seconds after analysis of the average co-occurrence histogram for all contextual features. After this step the number of features will

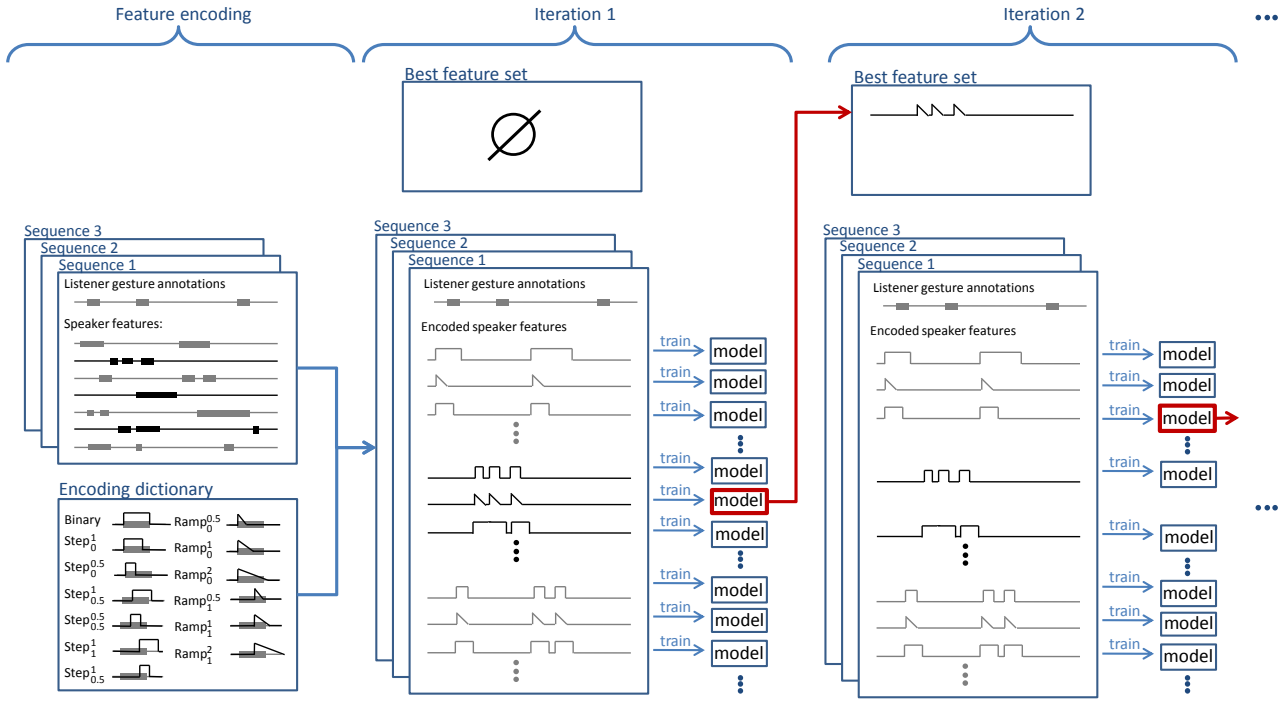

Figure 3: Joint feature selection. This figure illustrates the feature encoding process using our encoding dictionary as well as two iterations of our joint feature selection algorithm. The goal of joint selection is to find a subset of features that best complement each other for recognition of listener visual feedback.

be reduced to a more manageable size of 50 contextual features.

The second step is to look at the best performance an individual feature can reach when trained with any of the encoding templates in our dictionary. For each top-50 feature a sequential model (see Section 4.3) is trained for each encoding template and then evaluated. A ranking is made based on the recognition performance of each individual feature and a subset of 10 features is selected.

4.2.2 *Joint Feature Selection*

Given the subset of features that performed best when trained individually, we now build the complete set of feature hypothesis to be used by the joint feature selection process. This set represents each feature encoded with all possible encoding templates from our dictionary. The goal of joint selection is to find a subset of features that best complements each other for recognition of visual feedback. Figure 3 shows the first two iterations of our algorithm.

The algorithm starts with the complete set of feature hypothesis and an empty set of *best* features. At each iteration, the best feature hypothesis is selected and added to the best feature set. For each feature hypothesis, a sequential model is trained and evaluated using the feature hypothesis and all features previously selected in the best feature set. While the first iteration of this process is similar to the individual selection, every iteration afterward will select a feature that best complements the current best features set. Note that during the joint selection process, the same feature can be selected more than once with different encodings. The procedure stops when the validation performance starts decreasing.

4.3 Multimodal Integration

The multimodal integration step incorporates the speaker contextual features with visual observations to improve recognition of listener gestures. The contextual features are selected and encoded as described in Sections 4.1 and 4.2. An example of this multi-dimensional stream of information is labeled in Figure 1 as "Best encoded features".

While our framework supports multi-dimensional streams of information from the listener vision-based recognizer (e.g., outputs from multiple visual gesture classifiers), in our experiment we used a one-dimensional stream: the output from the vision-based head nod recognizer (see Section 5.3 for details). For each frame grabbed by the camera looking at the listener (see Figure 1), we get a visual measurement of how likely a head nod is happening. This suggested a sampling rate of 30Hz for both the contextual features and visual observations. For each time sample, the probability output from the vision recognizer is concatenated with the values of each contextual feature at that time. These concatenated feature vectors are used as input for the probabilistic sequential model.

While our approach generalizes to any sequential model, in our experiment we used a Latent-Dynamic Conditional Random Field (LDCRF) as it was shown to outperform Hidden Markov Models and Conditional Random Field for context-based gesture recognition during human-computer interactions [13]. The LDCRF model supports multiple output labels, making it possible to train a multimodal integrator for more than one listener gestures. This is an interesting avenue for future work.

5. EXPERIMENTAL SETUP

For training and evaluation of our prediction model, we used a corpus of 38 human-to-human interactions. This corpus is described in Section 5.1 while Section 5.2 describes the contextual features used in our experiments. Section 5.3 describes the vision system used for tracking and recognizing head gestures. Finally Section 5.4 discusses our methodology for training the probabilistic model and evaluate it.

5.1 Data Collection

Data is drawn from a study of face-to-face narrative discourse ('quasi-monologic' storytelling). 76 subjects from the general Los Angeles area participated in this study. They were recruited using Craigslist.com and were compensated $20 for one hour of their participation.

Participants in groups of two entered the laboratory and were told they were participating in a study to evaluate a communicative technology. Participants completed a pre-experiment questionnaire eliciting demographic and dispositional information. Subjects were randomly assigned the role of speaker and listener. The speaker remained in the computer room while the listener was led to a separate side room to wait. The speaker then viewed a short segment of a video clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video. Two video clips were selected and were merged into one video: The first, "Cyber-Stalker" is about a woman at work who receives unwanted instant messages from a colleague at work, and the second, "That's an Order!", is about a man at work who is confronted by a female business associate, who asks him for a foot massage in return for her business.

After the speaker finished viewing the video, the listener was led back into the computer room, where the speaker was instructed to retell the stories portrayed in the clips to the listener. The listener was asked to not talk during the story retelling. Elicited stories were approximately two minutes in length on average. Participants sat approximately 8 feet apart. Finally, the experimenter led the speaker to a separate side room. The speaker completed a post-questionnaire assessing their impressions of the interaction while the listener remained in the room and retold what s/he had been told by the speaker. Participants were debriefed individually and dismissed.

We collected synchronized multimodal data from each participant including voice and upper-body movements. Both the speaker and listener wore a lightweight headset with microphone. Three Panasonic PV-GS180 camcorders were used to videotape the experiment: one was placed in front the speaker, one in front of the listener, and one was attached to the ceiling to record both speaker and listener

5.2 Contextual features

From the video and audio recordings, 15 classes of contextual features were extracted (see Section 3 for details). Most prosodic features were extracted automatically. Pitch and intensity of the speech signal were computed from the speaker audio recordings using real-time signal processing software [20]. From this we automatically derived the first six prosodic features listed in Section 3, including downslope, lowness, long utterances, energy edge, energy fast edge, and vowel volume.

Human coders manually annotated the additional prosodic and the lexical features from the audio recordings. Every spoken word was used as a possible contextual feature. All elicited narratives were transcribed, including pauses, filled pauses (e.g. "um"), and prolonged words. These transcriptions were double-checked by a second transcriber.

Figure 4: Setup for training and evaluation corpus. This study of face-to-face narrative discourse (i.e., quasi-monologic storytelling) included 76 subjects. The speaker was instructed to retell the stories portrayed in two video clips to the listener.

Finally, from the speaker video the eye gaze of the speaker was annotated on whether he/she was looking at the listener. After a test on five sessions we decided not to have a second annotator go through all the sessions, since annotations were almost identical (less than 2 or 3 frames difference in segmentation).

Note that although some of the speaker features were manually annotated in this corpus, all of these features can be recognized automatically given the recent advances in real-time keyword spotting [8], eye gaze estimation [12] and prosody analysis [20].

5.3 Vision-based Gesture Recognition

For vision-based recognition, we used the Watson software [21] which tracks the head position and orientation in real-time with 6 degrees of freedom using a tracking framework called Adaptive View-Based Appearance Model. The library also recognizes two head gestures using a support vector machines (SVMs): head nods and head shakes. In our experiments we used the SVM classification margin returned by Watson as input for the multimodal integrator (see Section 4.3).

For ground truth comparison, listener video were annotated for head nods by two coders. These annotations form the labels we used in our context-based framework for training and evaluation. The dataset contained a total of 165 head nods.

5.4 Methodology

To train our context-based recognizer we split the 38 session into 3 sets, a training set, a validation set and a test set. This is done by doing a 10-fold testing approach. This means that 10 sessions are left out for test purposes only and the other 28 are used for training and validation. This process is repeated 4 times in order to be able to test our model on each session. Validation is done by using the hold-out cross-validation strategy. In this strategy a subset of 10 sessions is left out of the training set. This process is

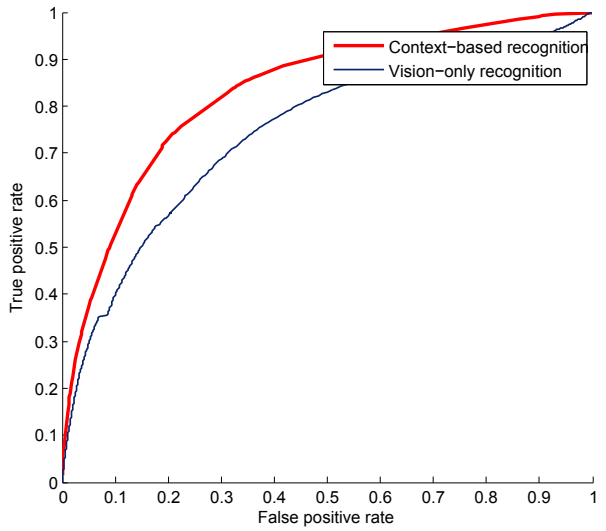

Figure 5: ROC curves of head nod recognition comparing our context-based approach to a vision-only approach.

Recognizer	Area	EER
Context-based	83.2%	76.5%
Vision-only	74.9%	69.4%

Table 1: Quantitative comparison between our context-based approach and a vision-only approach (same as Figure 5). The table shows both the area under the curve and the equal error rate (EER).

repeated 4 times and then the best setting for our model is selected based on the performance of our model.

The comparative evaluation of our context-based recognizer was performed at the time-sample level (i.e., frame level). A classification decision is made for each time sample and the true positive and false positive rates were computed based on these classifications. The true positive rate is computed by dividing the number of recognized frames by the total number of ground truth frames. Similarly, the false positive rate is computed by dividing the number of falsely recognized frames by the total number of `other-gesture` frames.

During validation and testing, our context-based recognition algorithm is applied on the unsegmented sequences meaning that no pre-segmentation of the gesture start and end times was done on these sequences. By evaluating the approach at the time-sample level, we are evaluating both the recognition performance as well as the segmentation performance. For this reason the expected error should be lower than an algorithm which only detect gestures (i.e., true positive and false positive rates computed at the gesture level [14]).

In our experiments we used Latent-Dynamic Conditional Random Field as sequential model used for contextual integration (see Section 4.3). The number of hidden states per label for the LDCRF model was set 2 states per label and the regularization term was validated with values $10^k, k = -1..3$.

6. RESULTS AND DISCUSSION

We designed our three main experiments to evaluate (1) the overall performance of our context-based recognition framework, (2) the gain from using joint feature selection and (3) the gain from using encoding dictionary.

Our first experiment compared the performance of our context-based recognition framework with a vision-only recognizer. Figure 5 shows the ROC curve for both approaches. The ROC curves present the detection performance for both recognition algorithms when varying the detection threshold. The two quantitative methods used to evaluate ROC curves are area under the curve and equal error rate. Table 1 shows the quantitative evaluation using both error criteria. The use of context improves recognition from 74.9% to 83.2%. Pairwise two-tailed t-test comparison show a significant difference for both error criteria, with $p = 0.021$ and $p = 0.012$ for the area under the curve and the equal error rate respectively.

As described in Section 4.2, our context-based recognition framework uses two types of feature selections: individual feature selection and joint feature selection (see Section 4.2.2 for details). It is very interesting to look at the features and encoding selected after both processes:

- *Vowel volume* using ramp encoding with a width 0.5 second and a delay of 0.5 seconds
- *Speaker looking at the listener* using a binary
- *Pause* using step encoding with a width 2.0 second and no delay
- *Low pitch* using ramp encoding with a width 0.5 second and no delay

These are the four features-encoding pairs selected after the joint feature selection process which stopped when validation performance started decreasing. We can see that only one feature was selected with binary encoding, suggesting that the use of the encoding dictionary was important. The first selected feature *Vowel volume* used an encoding with a ramp and a delay of 0.5 seconds, meaning that its influence on head nods is asynchronous and decreases over time. The second selected feature is related to the eye gaze of the speaker, confirming the importance of our multimodal context. The third and fourth features have also been reported by Ward and Tsukahara [20] as good predictive features for backchannel feedback. No lexical feature was selected by the *joint* selection algorithm. This result means that visual gesture recognition can be improved using only prosodic cues, pauses and speaker visual display.

The second and third experiments were designed to understand the influence of feature selection and encoding dictionary on the context-based recognition framework. Table 2 compares the recognition performance when using or not using the joint feature selection after the individual feature selection. Table 3 compares the recognition performance when using the complete encoding dictionary to using only binary encoding. This last comparison was done after the individual feature selection.

We can see from both Table 2 and 3 that the gain performance of our context-based recognition algorithm is directly related to the joint feature selection and the encoding dictionary. By using the encoding dictionary instead of the usual

Feature selection	Area	EER
Joint + Individual	83.2%	76.5%
Individual only	79.1%	72.0%

Table 2: Quantitative evaluation showing the gain in performance when using both individual and joint feature selection.

Feature encoding	Area	EER
Dictionary	79.1%	72.0%
Binary	76.1%	69.9%

Table 3: Quantitative evaluation showing the gain in performance when using the encoding dictionary for feature representation.

binary encoding, the performance improves from 76.1% to 79.1%. And by using the joint feature selection, the performance improves again from 79.1% to 83.2%.

Our experiments show that by using joint feature selection and an encoding dictionary, contextual information from other participant significantly improve the performance of vision-based gesture recognition.

7. CONCLUSIONS

Our results show that contextual information from other human participants can improve visual gesture recognition. We presented a context-based recognition framework that represents contextual features based on an encoding dictionary and automatically selects the optimal features based on *individual* and *joint* influence. By using simple prosodic, pauses and visual display contextual features available in real-time, we were able to improve the performance of the vision-only head gesture recognizer from 74.9% to 83.4%. Recognizing these visual gestures is important for understanding the full meaning of a meeting or conversation, and can be particularly crucial for identifying more subtle factors such as the effectiveness of communication or diagnosis social disorders. As future work, we plan to experiment with a richer set of encoding templates including gaussian density functions, and to apply our context-based approach on other visual feedback cues such as eye gaze patterns and body posture shifts.

Acknowledgements

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) and the National Science Foundation under grant # HS-0713603. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

8. REFERENCES

[1] J. Allwood, J. Nivre, and E. Ahlsén. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, pages 1–26, 1992.

[2] H. Anderson, M. Bader, E. Bard, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. The mcrc map task corpus. *Language and Speech*, 34(4):351–366, 1991.

[3] J. B. Bavelas, L. Coates, and T. Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

[4] J. K. Burgoon, L. A. Stern, and L. Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, Cambridge, 1995.

[5] N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. In *European ACL*, pages 51–58, 2003.

[6] H. H. Clark. *Using Language*. Cambridge University Press, 1996.

[7] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi. A conversation robot using head gesture recognition as para-linguistic information. In *RO-MAN*, pages 159–164, September 2004.

[8] S. Igor, S. Petr, M. Pavel, B. LukáŽ, F. Michal, K. Martin, and C. Jan. Comparison of keyword spotting approaches for informal continuous speech. In *MLMI*, 2005.

[9] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *ICMI*, pages 12–19, November 2003.

[10] A. Kapoor and R. Picard. A real-time head nod and shake detector. In *PUI*, November 2001.

[11] S. Kawato and J. Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the Şbetween-eyesŤ. In *FG*, pages 40–45, 2000.

[12] L.-P. Morency and T. Darrell. Recognizing gaze aversion gestures in embodied conversational discourse. In *ICMI*, Banff, Canada, November 2006.

[13] L.-P. Morency and T. Darrell. Conditional sequence model for context-based recognition of gaze aversion. In *MLMI*, 2007.

[14] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8–9):568–585, June 2007.

[15] R. Nishimura, N. Kitaoka, and S. Nakagawa. A spoken dialog system for chat-like conversations considering response timing. *LNCS*, 4629:599–606, 2007.

[16] A. Rizzo, D. Klimchuk, R. Mitura, T. Bowerly, J. Buckwalter, and T. Parsons. A virtual reality scenario for all seasons: The virtual classroom. *CNS Spectrums*, 11(1):35–44, 2006.

[17] L. Tickle-Degnen and R. Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293, 1990.

[18] L. Z. Tiedens and A. R. Fragale. Power moves: Complementarity in dominant and submissive nonverbal behavior. *Journal of Personality and Social Psychology*, 84(3):558–568, 2003.

[19] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *ICCV*, Nice, France, October 2003.

[20] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 23:1177–1207, 2000.

[21] *Watson: Head tracking and gesture recognition library*. http://projects.ict.usc.edu/vision/watson/.

[22] Y. Xiong, F. Quek, and D. McNeill. Hand motion gestural oscillations multimodal discourse. In *ICMI*, pages 132–139, Vancouver B. C., Canada, November 2003.

[23] V. H. Yngve. On getting a word in edgewise. In *Sixth regional Meeting of the Chicago Linguistic Society*, pages 567–577, 1970.