マルチモーダルな認識に基づくポスター発表システム

林　宗一郎[†]　吉本廣雅[††]
平山高嗣[†††]　河原達也[†]

視線，音声などのマルチモーダルな情報をもとに，ユーザの興味・関心や理解の度合いを把握し，それに応じてユーザに適した情報を提供できるシステムを提案する．本稿では，説明内容とユーザの注視対象の同期関係に着目したユーザの心的状態の推定に基づいて，ポスターを用いたプレゼンテーションをオンラインで行なうシステムの概要を述べる．

Poster Presentation System based on Multi-modal sensing

Soichiro Hayashi,[†] Yoshimasa Yoshimoto,[††]
Takatsugu Hirayama[†††] and Tatsuya Kawahara[†]

We have investigated a poster presentation system present which can understand interest level of users by multi-modal sensing, especially the synchronization of the system's explanation and users' gaze. In this paper, we present the interaction flow,and the core techniques of our system.

1. はじめに

我々は，対話を通して，人の知的欲求をコンピュータに理解させ，必要な情報を提示したり，コミュニケーションを助けるシステムの実現を目指している．コンピュータから情報を得ようとする場合，自分の意思を明確化し，情報端末を駆使して検索ワードなどを用いてコンピュータに伝える必要がある．これに対して，マルチモーダルなインターフェースを用いてユーザが持つ潜在的な興味を探り，そしてユーザにそれを気づかせ，さらにはそれに応じてさりげなく，気の利いた情報提供を行う情報コンシェルジェシステムの実現を目指した研究を行なっている[5)6)]．情報コンシェルジェでは，エージェントを用いて積極的にユーザとのインタラクションを行なうことでユーザの意図を顕在化させ，情報提示を行なう（**図 1**）．

本稿では，学会やオープンハウスで一般的に行なわれているポスターを用いたプレゼンテーションに適用し，マルチモーダルな情報に基づいて聞き手の反応を分析することで，聞き手にとって有用な情報を簡潔に提示するシステムを提案する．

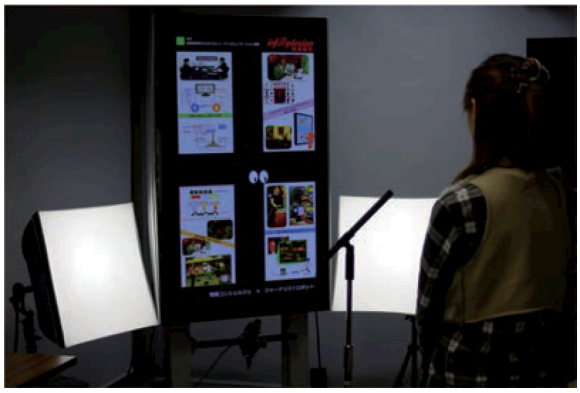

図 1　視線と発話に基づく Mind Probing を中核技術として構築された情報コンシェルジェ
Fig. 1　Info-concierge : using Mind Probing technology based on gaze and utterance

2. システムの概要

2.1 状況設定

本システムは，限られた時間で聞き手にとって有意義な情報を提示する必要があるポスター発表の場において，聞き手の興味や関心がどのコンテンツに向いて

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†† 京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University
††† 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University

いるかを推定し，聞き手に適した情報提示を行なう．

例えば，分野が多岐に渡っているプロジェクトの説明では聞き手の専門の違いによってどの要素に興味のあるか違ってくるうえ，説明で使用する用語や概念に対する理解にも個人差があると想定される．このような場合，聞き手の心理として自分が精通する分野に関してはできるだけ詳しく知りたいが，それ以外の分野は，概要を分かりやすく説明して欲しいといった要求が考えられる．しかし，既存の有人ポスター発表の場合，発表中にいつ自分が最も知りたい情報が説明されるか分からない．また，知らない用語が出てきても，発表者に自分の理解を逐一発表者に伝えるのは難しく，理解が曖昧な用語があってもそのまま説明を受けなければならず，結果的に発表者の意図したプレゼンテーションが実現しない．

本システムでは，電子掲示板を用いてシステム自身がポスター発表を行なう．1〜2 人の聞き手を想定し，非侵襲なデバイスによって聞き手に違和感を与えることなく聞き手の振る舞いを観測し，聞き手が関心のある部分をより詳しく説明し，分からない用語に関する情報は逐一提示する．なお，発表は聞き手の関心によって提示する情報を動的に変化できるよう大型のディスプレイを用いて行ない，音声はシステムが合成したものを利用する．

2.2 コンテンツ

本システムでの発表は，初めにいくつかのトピックについておおまかな内容を説明し，次にそれぞれのトピックについて詳しく説明するというトップダウン的な構成とする．そのために，ツリー状の構造のコンテンツを用意する．また，説明中に現れた用語の説明をするために利用する単語の辞書も用意する．

図 2 のように，ディスプレイは二分されメインにポスター内容を表示し，補助的に用語の説明を示す．ポスター部分は 2 × 2 の 4 枚のスライドを並べて提示する．

2.2.1 コンテンツの構造

発表内容は**図 3** のように階層構造にもち，1 つのノードに 4 つ以下のトピックを含む．各トピックは 1 枚のスライドで説明され，1 つもしくは 0 個の子ノードをもつ．ただし，子ノードのトピックは親トピックの内容を掘り下げたものとする．

ディスプレイのポスター描画部分に，1 ノードの含むトピックに関するスライドを表示する．

2.3 インタラクションフロー

本システムにおけるプレゼンテーションがどのような流れで行なわれるか，その際のユーザとシステムのインタラクションについて説明する．

図 2　本システムのディスプレイ画面
Fig. 2　The screen capture of our system

2.3.1 プレゼンテーションの流れ

(1)　ディスプレイ前方に聞き手が現れたのちすぐに説明を始めるのでなく，ユーザとシステムの心的距離を縮めるため，ディスプレイ内に視線がうつってから数秒後に説明を始めても良いかを音声を用いてユーザに確認する．

(2)　ユーザから説明開始の承諾をもらった後，根ノードのスライドを用いて説明を行なう．このとき，聞き手の振る舞いを観測する．

(3)　次に，聞き手が最も興味のあると推定されたトピックに関して詳しく説明するため，そのトピックの子ノードの含むトピックに関する説明を行なう．

(4)　(1) と同様に聞き手がどのトピックに興味があるかを推定し，

(a)　推定した興味度があるしきい値を超えるトピックがあれば，そのトピックの子ノードに関する説明を行なう．このとき，説明を突然切り替えるのではなく，(1) と同様にユーザに確認を行なう．このときの意図確認のアルゴリズム設計には，Misu らによって提案された推薦方略[3] を応用する．

(b)　もしなければ，1 つ上の階層に戻り，同条件を満たすトピックで，まだ説明していないものがあれば，それに関する説明を行なう．

(5)　(3)(4) を再帰的に繰り返す．条件を満たすコンテンツが無くなった時点でプレゼンテーションを終了する．

以上の流れを**図 4** に示す．

2.3.2 単語概念に関する説明

発表中，聞き手がはっきりと理解していないと考え

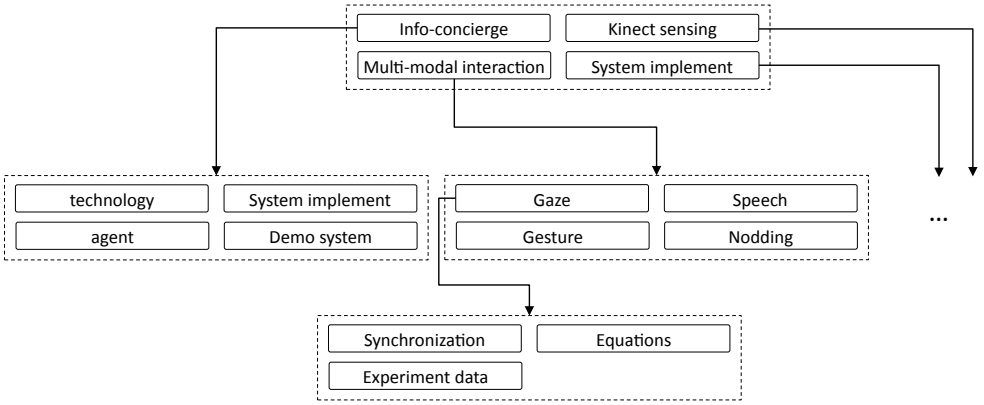

図 3 コンテンツ構造の一例
Fig. 3 An example of the content structure

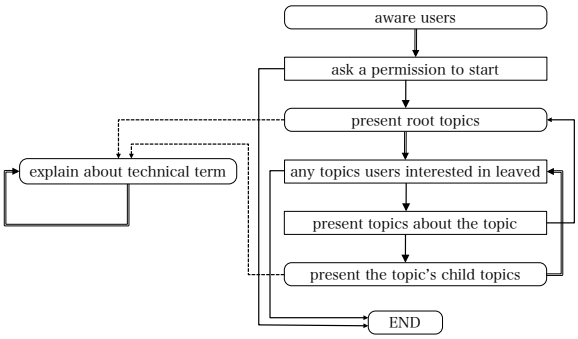

図 4 プレゼンテーションの流れ
Fig. 4 Presentation flow

られる用語に関しては，逐一説明するべきである．しかし，その度に音声で説明するのはプレゼンテーションの冗長化につながり，また複数人の聞き手が想定される中では現実的でない．そこで，出現した用語についての説明を画面の一部分に表示し，聞き手が読みたければ読むという情報提示法を考える．このとき，用語に関する説明を聞き手が見ていた場合，システムの音声の一文ごとの時間間隔を増加させる．

説明を表示する用語は，出現した用語ごとに聞き手の理解度を推定し，理解度が低いものだけにすべきであるが，本システムでは単語が一般的なものかどうか着目し，単語重要度（TF-IDF 値）を求めてその値が高いものの説明を表示する．

また，時間的制約によってシステムが説明を表示できる用語は限られるが，それ以外については聞き手がタッチインターフェースを利用して用語の一覧から説明を表示する用語を選択することも可能とする．

3. 興味・関心の推定

ユーザに有用な情報を提示するために，ユーザが何に興味をもっているのか，何に対する理解が足りていないかを推定する必要がある．本システムでは，聞き手の興味の度合いに合わせて，コンテンツを動的に提示する．以下，興味度の定義及び算出方法を示す．

3.1 興 味 度

興味度は，スライドに対応するトピックに対する興味の度合いを表すものと定義する．トピックは複数の概念を組み合わせた説明で構成され，概念は単語を組み合わせた説明で構成されるとする．本研究では，概念や単語といったトピックよりも小さい単位のものに対する興味は扱わない．また，本研究において “興味がある” とは，トピックに対して “より詳しい内容を知りたいという知的欲求を持つこと” とする．

これは，発表者が説明したいと考えている発表の本質に対する興味度だけを推定するためである．

3.2 興 味 推 定

聞き手は，システムが説明を行なっているトピックに常に関心を持っているとは限らない．より関心があるトピックが他にあればそちらに視線を送り，説明に対する反応が悪くなると考えられる．そこで本システムでは，合成音声での説明箇所と聞き手の注視対象が同期しているかを基に興味度を求める．

Nakano らは，HAI におけるユーザの会話への参加度を注視行動パターンに基づいて分析し[4]，会話を観察した第三者の主観評価との比較を含む分析結果として，会話参加度が高いシーンにおいては，他者との相互注視及びよそ見を行なわない傾向があることが示された．この知見に基づいて本システムでは，ポスターへの注視時間で各トピックにスコアリングを行い興味度を求める．このとき，注視対象が説明個所と同期している場合と同期していない場合で，注視のもつ意味が異なると考えられる．同期している場合はプレゼンテーションを聞くという意味合いが強く，同期していない場合は，そのトピックに注意を向けていると考え

られ，トピックに関する興味は後者の方がより顕著に表すと考えられるからである．そこで，説明箇所との同期の有無によって注視時間に重み付けを行なう．

また，対象に関心を持って注視を向けているかを評価するため，視覚コンテンツの切り替えによる Probing に基づいた興味推定[1] を導入する．人は興味がある視覚情報に注視を向けているときに，その周辺で起きた視覚情報の変化に注視を切り替える反応潜時が長くなるという分析結果が得られている．そこで，用語説明部を利用し，ポスターについての説明とは非同期に用語説明部の表示を切り替え反応潜時を計測することで，興味度の確信度を求める．

用語の数 N，i 番目の用語を表示した際の反応潜時 t_i，$i-1$ 番目の用語の表示から i 番目の用語の表示までの間にトピック X の説明中 $(p = X)$ に聞き手 u がトピック X のポスターを注視 $(g = X)$ していた時間 $c_{u,i(g=X,p=X)}$，重み α, β を用いると，聞き手 u のトピック A について興味度を表す式は，

$$I_{u,A} = \sum_{i=0}^{N} (1 - \exp(-\gamma t_i))(\alpha c_{u,i(g=A,p=A)} + \beta c_{u,i(g=A,p\neq A)}) \quad (1)$$

となる．

聞き手が複数人いた場合は，聞き手の興味度の総和をもとに次に説明するトピックを決定する．

3.3 Kinect を用いた注視対象推定

視線推定手法はすでに様々な方式が提案されているが，本稿ではポスターセッションに適した実用的な視線推定方式を実現した．まずセンサとしては Kinect を用い，撮影画像から聴衆の視線推定を行う．また画像の解像度及び眼鏡等の影響により眼球そのものを安定に撮影することは困難であるため，視線方向を頭部方向で代用する☆．

この方針に従い，本システムは下記処理で注視対象の特定を行う．

- 正面顔検出処理
- 頭部モデル獲得処理
- 頭部追跡処理
- 注視対象推定処理

以下各処理の概要を述べる．

3.3.1 正面顔検出処理

Kinect センサで撮影したカラー画像及び距離画像から正面顔探索を行う．アルゴリズムは Haar-like 特徴を利用した物体認識法を用いる．

3.3.2 頭部モデル獲得処理

正面顔の検出結果に従って，距離画像から頭部の 3 次元形状を，カラー画像からその色情報を計算する．計算結果はポリゴンとテクスチャ情報に変換し，頭部モデルとする．

3.3.3 頭部追跡処理

本処理では頭部方向の推定問題を画像への頭部モデルの当てはめ問題として計算する．具体的には頭部を剛体とみなし，頭部の 3 次元位置と姿勢を表す 6 変数をマルコフ連鎖モンテカルロ法に基づいた粒子フィルタによる追跡処理で逐次計算する．さらに得られた 6 変数を初期値とした最適化処理を行うことで，頭部の 3 次元位置と姿勢の高精度化を行う．

3.3.4 注視対象推定処理

頭部追跡処理で得た 6 変数から，3 次元空間で視線に対応する半直線を求める．この半直線と，電子ポスターや発表者との交差判定を行うことで，注視対象を決定する．

4. システムの構成

本システムは，**図 5** のように視線計測モジュール，音声認識モジュール，コンテンツ提示モジュール，インタラクション管理モジュール，コンテンツ制御モジュールの 5 つのモジュールで構成される．各モジュールはスレッド化され，並列的に動作する．1 人もしくは 2 人のユーザがシステムの前方 150cm の距離に現れ，視線計測モジュールがそれぞれの顔を検出した時点で，コンテンツ提示モジュールがプレゼンテーションを開始する．各モジュールはソケット通信によって情報をやりとりする．

4.1 コンテンツ制御モジュール

コンテンツ制御モジュールは，インタラクション管理モジュールから興味度を受け取り，それを基にコンテンツ提示モジュールに制御信号を送信する．このモジュールは，全モジュールのスレッドを管理する．

4.2 コンテンツ提示モジュール

本システムは大画面ディスプレイとステレオスピーカーを介して，Servlet コンテナを搭載したアプリケーションサーバ jetty を用いてブラウザ上で視聴覚情報を提示する．画面はサイズが高さ 80cm，幅 145cm のものを利用する．

コンテンツ提示モジュールは，コンテンツに関する情報をもち，コンテンツ制御モジュールからの制御信号に基づいてプレゼンテーション及び用語の説明部への情報表示を行なう．説明テキストは用意しておき，

☆ 文献[7] では，視線と頭部方向の角度のズレは平均 10 度程度であり，ポスターを注視する状況ではそのズレはさらに小さくなる傾向があることが報告されている．

HOYA VoiceText engine SDK を用いて音声出力する．一文一文の間隔は，その都度コンテンツ制御モジュールから送られる．どのトピックに関しての説明かを聞き手に意識させるため，説明中のトピックのポスターの背景色表示を明るくする．用語の説明の表示の間隔は最低でも 20 秒とする．表示は出現するときは即時に表示し，消滅するときはユーザに気づかれないようにゆっくりフェードアウトさせる．出現させるタイミングは，その用語が出てきた説明文が終わった時点とする．その時点で前回の用語を表示してから 20 秒経過していない場合，20 秒経過後に表示する．

ブラウザとの情報のやりとりは WebSocket を用いて行なう．

4.3 視線検出モジュール

視線検出モジュールは，コンテンツ制御モジュールから信号を受けて計測を開始する．

計測を開始後は，インタラクション管理モジュールに視線情報を送り続ける．インタラクション管理モジュールには，4 つのポスター，用語説明部，それ以外 のどれを注視しているかの情報を送信する．

本モジュールは，GPU を用いた実時間処理として実装しており，例えば GPU として NVIDIA 社の GeForce GTX 550 Ti を用いた場合 15fps 程度で注視対象を推定できる．

4.4 音声認識モジュール

本システムはディスプレイ前方にマイクを備え，音声認識モジュール Julius[2)] を用いてユーザ発話を認識する．Julius は，言語モデルや音響モデルなどの音声認識に必要なモジュールを組み替えることができる高い汎用性を持ち，低スペック PC でも音声ストリームに対してリアルタイム処理が可能である．言語モデルは，Web を利用して学習した語彙数 6 万のモデルに，「はい/いいえ」で答える質問に対する応答を追加したもの，音響モデルは不特定話者 PTM トライフォンモデルを使用した．

認識結果は，単語情報を含んだものをインタラクション管理モジュールに送信する．

4.5 インタラクション管理モジュール

インタラクション管理モジュールは，視線検出モジュールと音声認識モジュールからセンシング情報を受け取り，興味度の計算などの解釈を行なったのち，その結果をコンテンツ制御モジュールに送信する．

ユーザの発話についての推定は，音声認識モジュールから受信した音声認識結果からプレゼンテーションを先に進めるかどうかの判定を行う．

興味度の計算については，まずコンテンツ制御モジュールから用語説明の時刻に関する情報を取得する．コンテンツ制御モジュールからどのトピックに関する説明を行なっているかをトピックが変わる度に受け取り，興味度を計算する．

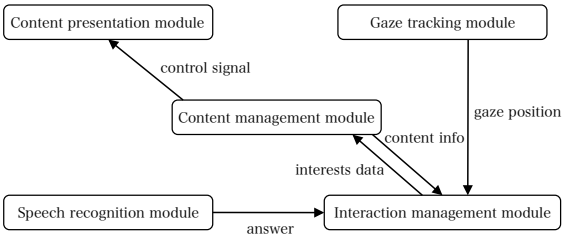

図 5 モジュール構成
Fig. 5 Structure of modules

コンテンツ制御モジュールから，4 つのトピックの説明が終了したことが伝えられた時点で，4 つ全てのトピックの興味度をコンテンツ制御モジュールに送信する．

5. おわりに

本稿では，聞き手の振る舞いを分析し最適な情報を提示する無人のポスター発表システムを提案した．しかし，本システムによるプレゼンテーションでは，学会でのポスター発表の利点である発表者への質問や他の聴衆とのやりとりについては考慮していない．また，興味の推定アルゴリズムを単純化するため，ポスター上にエージェントを出現させていない．今後は，聞き手同士のインタラクションを助長したり，質問を受付けることを可能にしていくべきだと考える．さらに，エージェントを用いた共同注視の効果についても検討し，興味や理解をより顕在化させることを検討していきたい．

今回は無人のポスター発表を想定したが，有人のポスター発表を支援することを想定したシステムも検討中である．

また，ポスター発表以外でも街中のデジタルサイネージ等大量の情報の中からユーザに適した情報を短時間で提示する必要がある状況が想定される．ポスター発表以外の状況でも利用できるより汎用的なシステムの構築を目指す．

参考文献

1) Hirayama, T., Dodane, J., Kawashima, H. and Matsuyama, T.: Estimates of User Interest Using Timing Structures between Proactive Content-Display Updates and Eye Movements, *IEICE TRANSACTIONS on Information and*

Systems, Vol.93, No.6, pp.1470–1478 (2010).
2) Lee, A. and Kawahara, T.: Recent development of open-source speech recognition engine julius, *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, pp.131–137 (2009).
3) Misu, T. and Kawahara, T.: Speech-based interactive information guidance system using question-answering technique, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol.4, IEEE, pp.IV–145 (2007).
4) Nakano, Y. and Ishii, R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations, *Proceedings of the 15th international conference on intelligent user interfaces*, ACM, pp.139–148 (2010).
5) 河原達也，川嶋宏彰，平山高嗣，松山隆司：対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ，情報処理，Vol.49, No.8, pp.912–918 (2008).
6) 平山高嗣，角　康之，河原達也，松山隆司：情報コンシェルジェ：Mind Probing に基づくマルチモーダルインタラクションシステム，信学技報，Vol.111, No.190, pp.55–60 (2011).
7) 矢野正治，中田篤志，福間良平，角　康之，西田豊明：非言語マルチモーダルデータを用いた会話構造の分析のための環境構築，情報処理学会研究報告 (ユビキタスコンピューティングシステム)，Vol.2009, No.22 (2009).