社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－GN－60 (1)
2006／5／31

キーワードを利用した映像音声遠隔コミュニケーション支援システム

浅井 紀久夫[†‡] 小林 秀明[‡] 斎藤 史彦[*]

† 独立行政法人 メディア教育開発センター 〒261-0014 千葉市美浜区若葉 2-12
‡ 国立大学法人 総合研究大学院大学 〒240-0193 神奈川県三浦郡葉山町
*（株）ソリッドレイ研究所 〒221-0835 横浜市神奈川区鶴屋町 2-20-1
E-mail: asai@nime.ac.jp

あらまし テレビ会議で発表者が提示したキーワードを使って，遠隔コミュニケーションを支援するシステムのプロトタイプを構築した．発表者はキーワード・マーカを持って話し，遠隔にいる視聴者はそのキーワードに重ねて付加情報を見る．キーワード・キャプションは外国語学習のためのビデオ学習環境に有効であると考えられるが，遠隔会議など双方向コミュニケーションにも利用できると着想した．プロトタイプ・システムでは，ビデオ映像に含まれる文字情報を抽出して語句として認識し，言語翻訳，三次元モデル提示，音声再生といった付加的機能を提供する．付加情報のコンテンツは，二層ディスプレイを使って現実のシーンに重畳される．提示ツールとしてのWebブラウザの利用は，マルチメディア・データの作成や編集を容易にする．また，情報提示機能を制御する文字マーカの検出にはオープン・ソース・ライブラリARToolkitが利用されるが，日本語の認識にはOCRミドルウエアが実装された．これは，利用するキーワードをマーカとして登録するという作業を軽減する．この支援システムは既存のTV会議システムの映像音声信号を効率的に利用するように設計され，映像音声による遠隔コミュニケーションにおいてキーワード提示によるコミュニケーションの円滑化を図る．

キーワード テレビ会議，キーワード，拡張現実感，Webブラウザ，二層ディスプレイ，OCR

Keyword-based System for Supporting Audiovisual Telecommunication

Kikuo ASAI[†‡] Hideaki KOBAYASHI[‡] and Fumihiko Saito[*]

† National Institute of Multimedia Education 2-12 Wakaba, Mihama-ku, Chiba 261-0014 Japan
‡ The Graduate University of Advanced Studies Hayama-cho, Miura-gun, Kanagawa 240-0193 Japan
*Solidray Co. Ltd. 2-20-1 Tsuruya-cho, Kanagawa-ku, Yokohama 221-0835, Japan
E-mail: asai@nime.ac.jp

Abstract We developed a prototype system to support telecommunications that uses keywords selected by the presenter in videoconference. The presenter holds to show keyword cards, and the listeners at remote sites can see additional information with the keywords. Although keyword captions are considered effective in video learning environments for learning foreign languages, we think they should also be available for interactive communications. Our prototype system recognizes letters and characters in a video image, and provides us with additional functions, such as language translation, 3D model visualization, and audio reproductions. The visual data are overlaid onto the real scene with a multilayered display, using a Web browser as a presentation tool to enable us to easily author/edit multimedia data. Optical character reader (OCR) middleware was implemented into the recognition function for Japanese, and an open-source library, ARToolkit was used to detect the character markers that controlled the functions of the information presentation. The support system is designed to efficiently exploit audiovisual signals of existing videoconferencing systems, so that the telecommunication can be fluent using keywords.

Keyword Videoconference, Keyword, Augmented Reality, Web Browser, Two-layer Display, OCR

1. はじめに

情報通信基盤が高等教育機関にも整備され，TV会議システムを利用した映像音声による遠隔会議や実時間遠隔講義などが頻繁に行われるようになった．遠隔会議や遠隔講義では，発表者あるいは講師が映像音声を利用したマルチメディア資料を提示しながら説明を行い，その後この説明に対す

る質疑応答が行われる形式が多い．授業や研究会，講演会などテーマや発表者が予め決まっている場合には，事前に資料を作成して十分準備しておくことができる．

一方，ディベートや異文化コミュニケーション，外国語会話などを対象とする場合，テーマはある程度決まっていようが，その内容は議論や討議内容に左右される．従って，必要な，あるいは関連する資料を予め全て用意することは難しい．

最近では，多くの TV 会議システムが商用として提供され（例えば，MediaPoint IP [1], ViewStation [2]），PC 上で動作するアプリケーション（例えば，NetMeeting [3], WebEx [4]）も存在するようになった．こうした遠隔コミュニケーション・ツールは遠隔地間で映像音声の交換を行うだけではなく，ファイル交換，チャット，画面共有，音声呼び出しといった多様な機能を提供するようになった．

しかし，これらは遠隔コミュニケーションの効率を向上させることには役立っているが，リアルタイムに内容の理解を助ける働きは少ない．つまり，遠隔コミュニケーション自体の質を改善することにはあまり貢献していないと考えられる．そこで，TV 会議の映像音声に含まれる文字情報に着目し，これをキーワードとして捉えて関連する付加情報を提供することにより遠隔コミュニケーションの円滑化を図る．図 1 に，キーワードを利用した遠隔コミュニケーション支援システムの概要を示した．

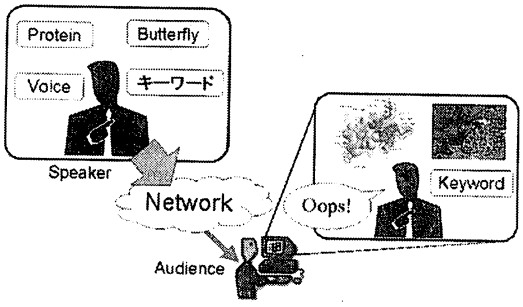


図 1: 支援の概要

キーワードに関連する付加情報はその関連性を視覚的に示すため，重畳提示することを考えた．現実のシーンに仮想物体を重畳させることにより，現実世界の認識を拡張させる技術として拡張現実感がある[5,6]．カメラ映像中の四角い枠のマーカを検出して位置と方向を計算する ARToolkit [7]が開発されて以来，様々なアプリケーションが作成されるようになった．しかし，アプリケーションの作成には CG の専門知識やプログラミングのスキルが必要だった．そこで，本システムでは，情報提示ツールとして Web ブラウザを利用することにして，既存のマルチメディア・コンテンツの有効利用，コンテンツ作成・編集の容易化を図った．

Web ブラウザには画面を重畳する機能はないため，二層ディスプレイを利用して物理的に画面を重ね合わせることにした．また，ARToolkit では利用するマーカを予め登録する必要があるが，様々な文字に対応するため，本システムには OCR（光学文字読み取り機能）ミドルウエアを実装した．

2. キーワード提示効果

キーワード提示の効果に関する研究は多数行われてきており，様々な知見が報告されている．しかし，映像音声情報に文字情報を付加する効果に関して見解が一致しているわけではなく，数多くの先行研究は文字情報が言語学習や内容理解に効果があるとしているが，状況によっては否定的になることも指摘されている．

ビデオ視聴時におけるキーワード提示効果の全体的な傾向として，キーワード提示は全文提示と同等の理解度が得られ，キャプションが全く提示されなかった場合に比べて理解度が良いというものである[8,9]．逆に，映像音声情報に文字情報を付加すると，注意が多言語メディアに分散するため，学習が阻害されるという報告もある[10]．

キーワードの提示の仕方によっては，キーワード提示の方が全文提示よりも良い理解度となる報告もある[11]．英語を母国語とする大学生がフランス語を学ぶ状況で，キーワード提示は全文提示と同様の理解度が得られたとしながらも，キーワードに関連した質問ではキーワード提示の方が全文提示よりも正答率が高いことから，各キーワードはビデオにおける特定の内容に注意を払わせると結論付けている．また，ビデオを利用した第二言語聴解練習では，学習者の聴解度を低下させることなく提示情報を読む負荷を減らすことができるため，キーワード提示が効果的であると述べられている．

この結果は，以下の 2 つの考えと矛盾しない．一つは一定時間内に処理できる情報量には限界がある[12]ため，全文提示による聴解促進には疑問もあることで，もう一つは映像内で文字情報が注

視されやすい[13]ことから，映像音声情報に付加された文字情報は注意を引きやすいことである．

キーワード提示の効果は，テレビのような一方向に情報を伝達するメディアだけではなく，双方向コミュニケーションにおいても有効であると予想される．映像音声を利用した遠隔講義や語学教育，国際コミュニケーションに応用するため，双方向通信が実施できる環境においてキーワードを利用するためのシステムを検討した．

3. システム

図1に，カメラ画像と関連情報を二層ディスプレイ上に重畳提示した例を示す．英語句が，日本語句上に重畳されている．利用者は単語マーカを二つ持ち，その翻訳語句が二つ別の Window に表示される．検出された“ひらがな”マーカはカメラ画像 Window 中でピンク色の四角で囲まれ，二つの英語句が提示 Window 中の対応する位置に表示される．“ひらがな”マーカを動かすと，英語句がそのマーカの位置と方向を追跡する．

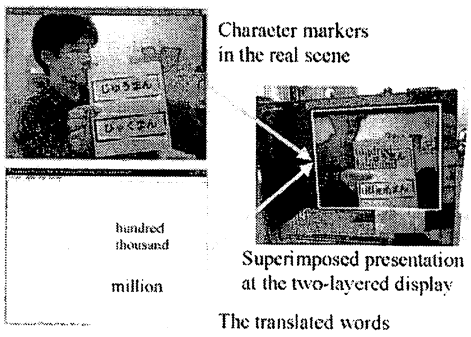


図 1: 二層ディスプレイ上の重畳提示例

3.1. 設計

システムは映像中の四角い枠を検出し，その四角枠中の文字群が予め登録した語句と一致するか検索する．設定ファイルの設定，あるいは提示された制御マーカに基づき，文字や画像，三次元モデルを提示，あるいは音の再生，合成音の発生を行う．1) Web ブラウザを組み入れる設計，2) 二層ディスプレイや 3) OCR の利用が，次のように本システムを特有にしている．

1) Web ブラウザ：Web ブラウザが，提示ツールとして使われた．Web ブラウザの利用には，幾つかの利点がある．まず，Web ブラウザはマルチメディア・データの様々な形式をサポートするし，plug-in を利用すればそのサポート形式は格段に増える．通常，AR アプリケーションを作成するには CG の知識やプログラミングのスキルが要求されるため，一般の利用者は AR コンテンツを作成したり編集したりする上で困難を抱えていた．Web ブラウザの利用は，コンテンツ作成・編集，既にあるコンテンツの再利用に対して柔軟性を与える．

2) 二層ディスプレイ：Web ブラウザを利用すれば，もはや AR ベースのシステムではなくなると言われるかも知れない．Web ブラウザは，仮想物体を現実のシーンに空間的に重畳する機能を持っていないからである．そこで，本システムでは，拡張現実感の特徴を維持するため，仮想物体が現実シーンに重畳するように，二層ディスプレイを実装した．また，同定したマーカの位置と方向を検出し，コンテンツを Web ブラウザのウィンドウ内に配置する際に反映させた．位置合わせは完全ではないものの，仮想物体は現実シーン中のマーカの近似的位置を追跡する（図 2a)．

3) OCR を利用した文字認識：大抵の AR アプリケーションでは，物体を同定して現実シーン中の位置と方向を検出するのにマーカを利用する．しかし，利用者は予めマーカを登録する必要があり，多くのマーカを利用する場合には煩雑な作業となる．OCR は，英数字や文字を読み取り語句として検出する．言語に依存して使われるため，特定の OCR ソフトが実装される必要があるが，利用できる語句の数を格段に増やすことができる．

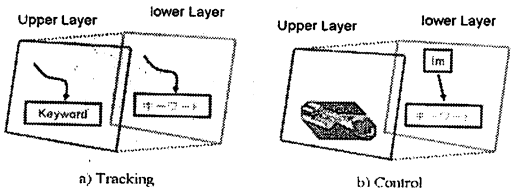


図 2: 位置追従と制御コマンド

3.2. 構成

図3は，プロトタイプ・システムの構成を示す．矢印は，処理データの流れの方向を示す．カメラは利用者が持つ語句マーカを含むビデオ映像を取得し，PC に送られる．

利用者は，二層ディスプレイで重畳された二つの Window を見る．一つの Window はカメラ画像を含み，もう一つはコンテンツである．カメラ画像 Window は，語句マーカを拭くんだ入力ビデオ

画像を表示する．コンテンツ Window は，認識した語句に関連する翻訳テキスト，画像，三次元モデルといった情報を提示する．

二層ディスプレイは，カメラ画像とコンテンツを重畳することができ，同時に提示できる．仮想物体は，語句マーカの位置と方向に基づいて現実シーンのカメラ画像に空間的に配置される．また，提示を制御するための制御コマンドとして，特別な文字マーカを用意した．利用者は制御コマンド・マーカを提示することにより，テキストから画像に提示形式を変更することができる．あるいは，テキスト・フォントのサイズを変更することができる（図 2b）．

プロトタイプ・システムは，ビデオ転送機能を持っていない．システムが遠隔通信に使われるとき，ビデオはインターネットや衛星通信などのネットワークを通して別途遠隔地に伝送されなければならない．

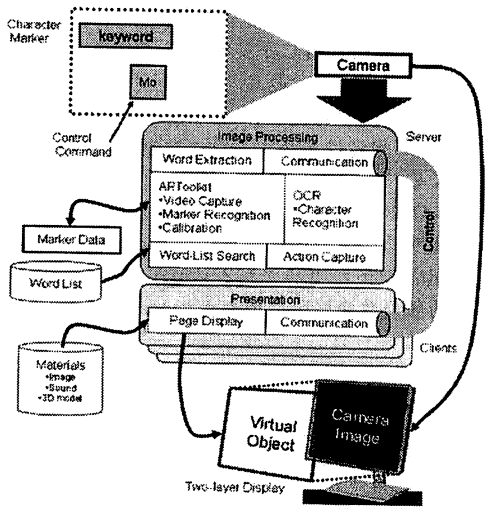


図 3: プロトタイプ・システムの構成

ソフトウエアは，主に画像処理部と提示部から構成される．二つの部分はサーバ・クライアントの関係になっていて，それらはソケット通信でデータを交換する．ネットワーク上でサーバから複数のクライアントに制御信号を送信するのに，UDP/IP が利用された．各クライアントは，サーバで認識された語句に基づいて同じ情報を取得する．この構成は，複数の視聴者に受信される同一情報に基づいてそれぞれ異なる表示を許容するような用途に利用できる．例えば，キーワード情報が視聴者に受信され，異なる言語で表示することも可能になる．

画像処理方法は，語句マーカ上の文字数に依存する．まず，OCR が語句マーカ中の語句を検出するために幾つかの文字を認識する．もし何も検出されないか，一文字のみが見つかれば，処理は ARToolkit に渡される．もし文字が事前登録されたマーカ・データ中の文字に一致すれば，その文字は制御コマンドとして扱われる．そうでなければ，“認識語句無し”というメッセージがディスプレイ上に表示される．もし OCR が何らかの語句を検出したら，それらは語句リスト中の語句に対して一致する語句があるかどうかが調べられる．

制御コマンドを同定するために，ARToolkit が使われた．制御コマンドの数が現在 30 程度で，これらのマーカを登録する作業はそんなに手間ではない．ARToolkit はまた，四角い枠を検出しその位置及び方向を決めるのにも利用された．

言語翻訳を実行するために商用の翻訳ソフトは利用せず，独自の語句リストを作成する設計とした．OCR から導出される文字認識結果は完璧ではなく，照明環境や語句マーカの動きなどによってときどき間違う．語句リストを使うことは，語句検出精度の向上に有効であると考えられる．語句リストは，利用者の興味がある領域から精選して作成される．プロトタイプ・システムでは，ひらがな，カタカナ，漢字を翻訳元の語句として指定し，日本語，英語あるいはタイ語の日常会話で使われる語句が翻訳先に指定された．表 1 に，語句リストの一部を示す．

表 1: 語句リスト

ひらがな	カタカナ	漢字	英語	タイ語
あい	アイ	愛	love	ความรัก
あいがんする	アイガンスル	哀願する	entreaty	วอน
あいこくしん	アイコクシン	愛国心	patriotism	ใจรักชาติ
あいことば	アイコトバ	合い言葉	password	รหัสผ่าน
あいさつ	アイサツ	挨拶	greeting	การทักทาย
あいしょう	アイショウ	愛称	nickname	ชื่อเล่น
あいじょう	アイジョウ	愛情	affection	ความรัก
あいしょう	アイショウ	相性	affinity	ชะตากัน
あいず	アイズ	合図	signal	สัญญาณ
あいすくりーむ	アイスクリーム	*	ice cream	ไอศครีม
あいすこーひー	アイスコーヒー	*	iced coffee	กาแฟเย็น
Hiragana	Katakana	Kanji	English	Thai
Japanese				

コンテンツは，利用者が持つマーカの位置と方向に基づいた配置で提示される．現実シーン中の語句マーカ位置に，テキスト，画像，あるいは三次元モデルを配置するように HTML ファイルが作

られる．位置及び方向が変化する度に，新しいHTMLファイルが作られ，ロードされる．しかし，ページ更新を頻繁に行うと提示が不安定に見えるため，ページ更新は最小限に抑えるべきである．そこで，仮想物体の位置変化はJavaScript機能を使って，ページ更新することなく円滑に制御できるようにした．マーカ枠の回転と大きさの変化に伴い仮想物体の配置を変化させる必要があるときだけ，新しいファイルが作られ再ロードされる．

この制御方法はまた，サーバとクライアントとの間のマルチキャスト信号のトラフィックを減少させるのにも役立つ．認識語句に関する情報の信号を送信するための三つのレベルを設定した．検出文字が同じ配置で前回認識したものと同一ならば，トラフックを発生しない（情報を送らない）．検出文字が前回の文字と同一だが，その配置が異なる場合，配置情報だけが転送される．検出語句と配置が変化したとき，語句と配置の両情報が転送される．

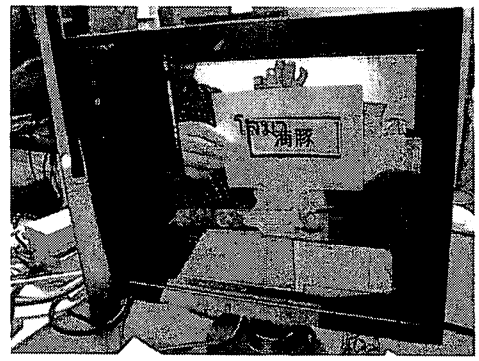
3.3. 実装

図4は，a) テキスト，b) 画像，c) 三次元モデルの提示のスナップショットを示す．システムは制御コマンド・マーカを挿入することにより，テキスト，画像，三次元モデルの提示形式を制御することができる．制御マーカ"Im"あるいは"Mo"が語句マーカに近づくと，対応する画像あるいは三次元モデルが語句マーカ位置に提示される．制御マーカがなければ，テキストだけが表示される．

VRML plug-inとしてCortonaを利用し，Webブラウザに三次元モデルを提示した．ただ，本システムはplug-inツールの制御に関して柔軟性をあまり持っていない．本報告では音は提示できないが，利用者が音声ファイルを出力形態として選択すれば，マーカ内の語句に関連付けられた音が再生される．

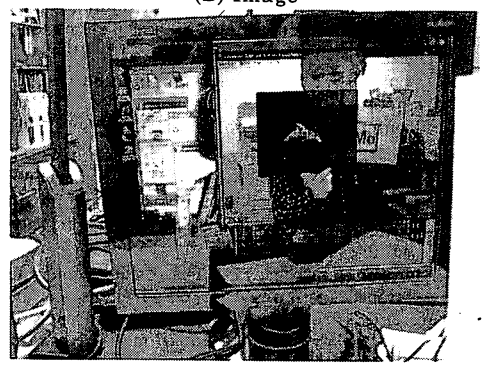
プロトタイプ・システムでは，画像処理（サーバ）が2.4 GHz Pentium IVのデスクトップPCに，コンテンツ提示が1 GHz Pentium IIIデスクトップPCに実装された．映像を取得するためのDVカメラ（DCR-HC1000, SONY）が，IEEE1394を通して画像処理PCに接続された．二層ディスプレイとして利用されたPureDepth MLD 3000は，17 inchサイズのモニタで，1280*1024*2の解像度を持つ仕様である．日本語文字認識には，OCRミドルウエア“Yonde!!KOKO”(A.I.Soft)が利用された．本OCRは，JIS第一レベルの漢字，ひらがな，カタカナ，英数字を認識する．

フレーム・レートは約20f/sだったが，語句マーカを提示してからテキストが適時されるまで約一秒の遅延が観測された．マーカがカメラ画像のなかで素早く移動すると，OCRによる認識は不安定になった．語句認識の精度を計測したところ，照明環境が良好な場合，静止した8つのサンプル・マーカ（日本語ひらがな及び漢字）に対してカメラから1mくらいまでほぼ100%の認識率が得られた．



(A) Text (Japanese to Thai)

(B) Image



(C) 3D Model

図4: スナップショット

4. むすび

TV 会議で映像音声による遠隔コミュニケーションを支援する仕組みとして，映像に含まれる文字情報を抽出して付加的な情報を同時提示するシステムを開発した．拡張現実感技術を応用し，オリジナル情報と付加情報の関連性を向上させる設計とした．また，情報提示ツールとして Web ブラウザを利用することにより既存マルチメディア・コンテンツに適用し易くした．一方，拡張現実感機能は二層ディスプレイを用いて維持した．さらに，OCR を利用することにより事前マーカ登録の手間を軽減した．

本システムでは，キーワードに付随した情報が提示される利点を強調したが，キーワードの提示自体の効果も期待される．今後，ディベートや三次元情報が必要な状況，英語を母国語としない人たち同士の英語によるコミュニケーションなどへの実践を検討したい．

謝辞

本研究の一部は，科学研究費補助金(17300283, 18500754)の支援を受けました．

文　献

[1] MediaPoint IP, NEC エンジニアリング, http://www.nec-eng.com/kaigi/index.html

[2] ViewStation, Polycom, http://www.polycom.co.jp/

[3] NetMeeting,Microsoft, http://www.microsoft.com/japan/windowsxp/using/windowsmessenger/default.mspx

[4] WebEx, サイバネットシステム, http://www.cybernet.co.jp/webex/

[5] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, Augmented reality: a class of displays on the reality-virtuality continuum, SPIE proceedings: Telemanipulator and Telepresence Technologies, vol. 2351, pp.282–292, 1994.

[6] R. Azuma, A survey of augmented reality, Presence: Teleoperators and Virtual Environments, vol. 6, no. 4, pp. 355–385, 1997.

[7] Kato, H., Billinghurst, M., Poupyrev, I., Imamoto, K., and Tachibana, K. Virtual object manipulation on a table-top AR environment. Proc International Symposium on Augmented Reality, pp.111-119, 2000.

[8] C. E. Kirkland, et al., The effectiveness of television captioning on comprehension and preference, Proceedings of the Annual Meeting of the American Educational Research Association, 1995.

[9] Toshikazu Kikuchi, A preliminary study of the effectiveness of key-word English captions in listening comprehension, Research Reports of Numazu Technical College, pp.135—146, 1999.

[10] S. D. Reese, Visual-verbal redundancy effects on television news learning, Journal of Broadcasting, vol.28, 79—87, 1984.

[11] Helen Gant Guillory, The effects of keyword captions to authentic French video on learner comprehension, Journal of the Computer Assisted Language Instruction consortium, vol.15, pp.89—108, 1998.

[12] 高野陽太郎（編），認知心理学(2)，東京大学出版会，1995.

[13] Hideko Itoh, An analysis of eye movements while watching educational TV programs, Bulletin of the National Institute of Multimedia Education, no.5, p.147—162, 1991.