共生社会に向けた人間調和型情報技術の構築
平成21年度採択研究代表者

H25 年度
実績報告

河原達也

京都大学　学術情報メディアセンター
教授

マルチモーダルな場の認識に基づくセミナー・会議の多層的支援環境

§1. 研究実施体制

(1)京大グループ
① 研究代表者:河原 達也 （京都大学 学術情報メディアセンター・教授）
② 研究項目
マルチモーダルな場の認識に基づくセミナー・会議の多層的支援環境

(2)奈良先端大グループ
① 主たる共同研究者:猿渡 洋 （奈良先端科学技術大学院大学 情報科学研究科・准教授）
② 研究項目
セミナー・会議のための音響・音声処理

§2. 研究実施の概要

複数の人間による知的活動のマルチモーダルなインタラクションを長時間収録した音声・映像に対して、人間の直感に基づいて有用箇所を効率的に視覚化・提示できる人間調和型情報基盤を構築している。

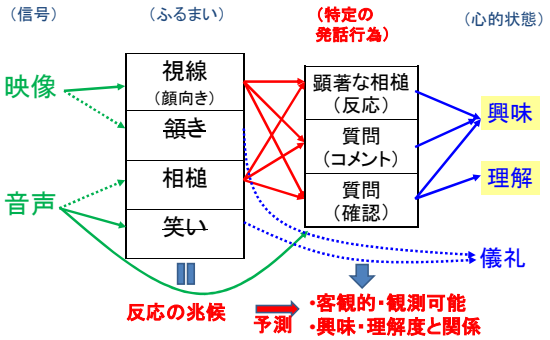
具体的には、学術的なイベント等で一般的になっているポスター形式のプレゼンテーションを対象に、多様なセンサを備えた大型ディスプレイによる「スマートポスターボード」の設計・実装を行っている。ポスター発表は今でも紙を用いる場合が多く、センサを備えたインタラクション環境は世界的にも例がない。また、このように長時間の複数人による自然な振る舞いを対象として、マルチモーダルな信号処理を行った例もほとんどない。本研究では、音声・音響・映像に関する研究者が結集して、視線や相槌・質問などの聴衆の反応をセンシングすることにより、興味・理解度を推定する枠組みを提案し、学会等でデモ展示が行えるレベルに達している。

本システム(スマートポスターボード;下左図)は、大型ディスプレイに設置したカメラやマイクロフォンアレイを用いて会話を記録し、誰がポスターに来て、どのような質問やコメントを行ったかを容易に検索できるようにすることを目指している。そのためにまず、視線(顔向き)検出と話者区間検出を統合したマルチモーダルな信号処理を実現している。さらに、視線配布や相槌などの聴衆のマルチモーダルな振る舞いに着目し、質問や顕著な相槌などの特定の発話行為を予測することで、興味・理解度の推定を行っている(下右図)。

また、講演形式のセミナーに対する字幕付与を目標として、音声認識の高度化にも取り組んでおり、京都大学 OCW(OpenCourseWare)への適用や、一般向けのシンポジウムで実演を行うレベルに達している。

スマートポスターボードの概観

スマートポスターボードの処理の流れ



§3. 成果発表等

（3－1） 原著論文発表

論文詳細情報（国内）

[1] 吉野幸一郎, 森信介, 河原達也.
述語項構造を介した文の選択に基づく音声対話用言語モデルの構築.
人工知能学会論文誌, Vol.29, No.1, pp.53--59, 2014.
http://dx.doi.org/10.1527/tjsai.29.53

[2] 村脇有吾.
階層的複数ラベル文書分類におけるラベル間依存の利用.
自然言語処理, Vol.21, No.1, pp. 41-60, 2014.

[3] 朝倉僚, 宮坂淳介, 近藤一晃, 中村裕一, 秋田純一, 戸田真志, 櫻沢繁.
筋電位計測と画像による姿勢計測を用いたリハビリテーション支援システムの設計.
電子情報通信学会論文誌, Vol.J97-D, No.1, pp.50-61, 2014.

論文詳細情報（国際）

[1] M.Ablimit, T.Kawahara, and A.Hamdulla.
Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language.
Speech Communication, Elsevier, Vol.56, 採録決定, 2014.
http://dx.doi.org/10.1016/j.specom.2013.09.011

[2] R.Miyazaki, H.Saruwatari, S.Nakamura, K.Shikano, K.Kondo, J.Blanchette, and M.Bouchard.
Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction.
Signal Processing, Elsevier, 採録決定, 2014.
http://dx.doi.org/10.1016/j.sigpro.2014.03.010

[3] T.Tung and T.Matsuyama.
Visual racking using multimodal particle filter.
International Journal of Natural Computing Research, IGI Global, 採録決定 2014.

[4] T.Tung and T.Matsuyama.
Invariant shape descriptor for 3D video encoding.
The Visual Computer, Springer, March, 2014.
http://dx.doi.org/10.1007/s00371-014-0925-6

[5] T.Mukasa, S.Nobuhara, T.Tung and T.Matsuyama.
Tree-structured Mesoscopic Surface Characterization for Kinematic Structure Estimation from 3D Video
IPSJ Transactions on Computer Vision and Applications (CVA), Vol. 6, pp. 12-24, March 2014.
http://dx.doi.org/10.2197/ipsjtcva.6.12

[6] *T.Tung and T.Matsuyama.
Geodesic Mapping for Dynamic Surface Alignment.
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), November, 2013.
http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.179

[proceedings（査読審査の入るものに限る）]

[7] T.Kawahara.
Smart posterboard: Multi-modal sensing and analysis of poster conversations.
In Proc. APSIPA ASC, (plenary overview talk), 2013.

[8] K.Yoshino, S.Mori, and T.Kawahara.
Predicate argument structure analysis using partially annotated corpora.
In Proc. IJCNLP, pp.957--961, 2013.

[9] T.Kawahara, S.Hayashi, and K.Takanashi.
Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations.
In Proc. INTERSPEECH, pp.1882--1885, 2013.

[10] K.Yoshino, S.Mori, and T.Kawahara.
Incorporating semantic information to selection of web texts for language model of spoken dialogue system.
In Proc. IEEE-ICASSP, pp.8252--8256, 2013.

[11] S.Nakai, R.Miyazaki, H.Saruwatari, and S.Nakamura.
Theoretical analysis of musical noise generation for blind speech extraction with generalized MMSE short-time spectral amplitude estimator.
In Proc. Intelligent Signal Processing (ISP) Conf., No.4.3, 2013.

[12] H.Saruwatari and R.Miyazaki.
Information-geometric optimization for nonlinear noise reduction systems.
In Proc. Int'l Sympo. Intelligent Signal Processing and Communication Systems (ISPACS), 2013.

[13] *R.Miyazaki, H.Saruwatari, S.Nakamura, K.Shikano, and K.Kondo, J.Blanchette, and M.Bouchard.
Toward musical-noise-free blind speech extraction: concept and its applications.
In Proc. APSIPA ASC, 2013.

[14] H.Saruwatari, S.Kanehara, R.Miyazaki, K.Shikano, K.Kondo.
Musical noise analysis for Bayesian minimum mean-square error speech amplitude estimators based on higher-order statistics.
In Proc. INTERSPEECH, pp.441-445, 2013.

[15] H.Yoshimoto and Y.Nakamura.
Cubistic Representation for Real-Time 3D Shape and Pose Estimation of Unknown Rigid Object.
In Proc. ICCV Workshop, pp.522-529, 2013.

[16] H.Yoshimoto and Y.Nakamura.
Free-Angle 3D Head Pose Tracking Based on Online Shape Acquisition.
In Proc. ACPR, pp.798-802, 2013.

[17] T.Tung, R. Gomez, T. Kawahara, and T.Matsuyama.
Multi-party Human-Machine Interaction Using a Smart Multimodal Digital Signage.
In Proc. HCI, LNCS, Vol. 8007, pp.408-415, 2013.

[18] T.Tung and T.Matsuyama.
Intrinsic Characterization of Dynamic Surfaces.
In Proc. IEEE-CVPR, 2013.

[19] Y.Murawaki.
Global Model for Hierarchical Multi-Label Text Classification.
In Proc. IJCNLP, pp. 46-54, 2013.