4Q1-12in2

個人に紐づくメディア情報を用いた感情解析プラットフォームの開発

内橋 堅志 *1　高濱 隆輔 *2　宮戸 岳 *3
Kenshi Uchihashi　Ryusuke Takahama　Takeru Miyato

*1京都大学大学院 情報学研究科 システム科学専攻
Department of Systems Science, Graduate School of Informatics, Kyoto University

*2京都大学大学院 情報学研究科 知能情報学専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

*3Preferred Networks

We focus on the invisibility of psychological effects caused by communication and develop a platform to analyze this. Specifically, content of dialogue and facial expressions are extracted from a moving picture, data is secured by associating with persons, and the change in emotion is analyzed from them. The effectiveness of the platform was demonstrated by analysis with real data.

1. 序論

近年、ネットワークサイエンスと呼ばれる新たな分野の研究が盛んに行われている [1]。ネットワークサイエンスとは、人間を孤立した個人であると考えるのではなく、社会的なネットワークの文脈の中で相互に影響を及ぼし合っていると考え、それを理解することを目指す学問である。この社会的なネットワークは SNS や会社をはじめ様々なコミュニティで発生しており、それぞれ異なる多様なモデリング方法が研究されている。本研究では特に、対面での会話によって構築されるネットワークを扱う。

対面で行う会話には、言葉ではなく、社会的な関係性を軸とする非言語的なコミュニケーションチャンネルが存在し、コミュニケーションにおいて非常に重要な役割を担っていることが知られている [2]。これまでは、この非言語的な作用を理解するために、個人や小さな集団におけるごく限られた観察しか行われなかった。しかし、近年のワイヤレス通信技術とセンサー技術の発展により、人間の自然な日常行動を詳細に捉えることができるようになったことで、大規模かつ正確な日常行動データの収集が技術的に可能となった [3]。

具体的には、対面での会話を撮影することによって得られる表情や発話の時系列的な変化のような話者の内部状態が反映される情報（これを非言語シグナルと呼ぶ）の解析を行う。ここで、既存の非言語シグナル情報の解析を行うシステムやサービスは、そのほとんどが音声情報を主な解析対象としていた [3]。しかし、本来非言語的なシグナルは表情に現れるものも非常に多い。これについては、認知心理学の観点からの裏付けも存在し、人間の内部状態を知る主要な手がかりとなることも分かっている。

本研究ではこの点に着目し、音声特徴量抽出と画像からの表情特徴量抽出がデバイス上でリアルタイムで行えるアプリケーション「SignaLog」を開発した。サーバでの計算をデバイス側が部分的に肩代わりできるようにしたことも，副次的な利点である。また、本研究では Android アプリケーションとして開発を行ったが、ライブラリの形式でまとめている機能も多く、ラッパーさえ用意すれば iOS や FPGA 上など、異なる環境でも動作させることが可能である点もプラットフォームとして優れている点である。

Contact: 〒 606-8501 京都市左京区吉田本町 36-1 京都大学 大学院情報学研究科 システム科学専攻 <ishii-secr@sys.i.kyoto-u.ac.jp>

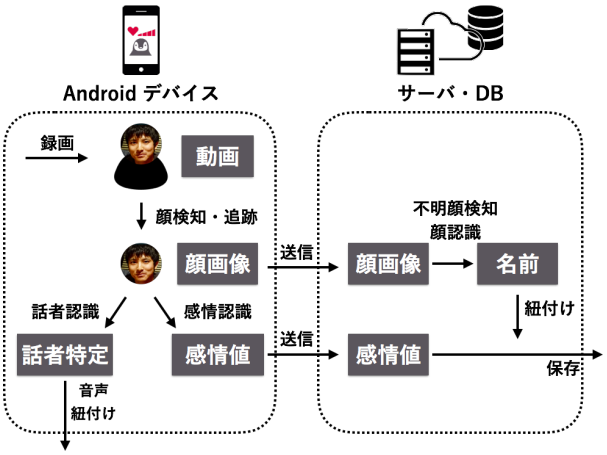

図 1: 動画情報処理の流れ。デバイス側で行う処理を左に、サーバ側で行う処理を右にそれぞれ示している。ここで、デバイス側で行う処理は全て撮影中リアルタイムに行われる。

2. 手法

Android 上で動作する会話情報収集アプリケーション SignaLog を開発し、得られたデータの解析を行った。SignaLog は、動画情報と音声情報をリアルタイムで処理する機能を持っており、これらについて順に説明する。

2.1 動画情報処理

まず、動画情報処理について詳説する。処理の流れを図 1 に示す。動画情報処理では、Android デバイスによって撮影された動画に対して顔検知と顔追跡を行い、動画に顔が入っていれ

ば顔画像情報を毎フレーム抽出する。顔画像はサーバに送信され、多量の顔画像から学習されたモデルを介して顔認識を行い、会話相手の名前を特定する。顔画像の系列から、口の動きが大きければ相手が話者であると判断することによって話者認識を行う。同時に、顔画像を用いた感情認識システムによって感情値を計算する。顔画像のデータは名前や感情値と紐づき、データベースに保存する。

感情認識を行うため、画像を入力として 7 値（怒、嫌、恐、幸、悲、驚、無）の分類を行うシステムを実装した。データは Challenges in Representation Learning: Facial Expression Recognition Challenge Dataset[*1] を用いて、モデルは多層 Convolutional Neural Network [4] を採用し、TensorFlow[*2] によってモデルを記述した。ハイパーパラメータについては複数のパターンを実験的に試しながら、 GPU による大規模計算で学習を行った。

さらに、学習を行ったモデルのパラメータとグラフ構造を圧縮し、アプリケーション側で展開して、 C++ で記述した I/O を Java から読み込むことによってデバイス側で学習結果を利用した。リアルタイムで動作するよう、パラメータの最適化に基づくモデルの圧縮やデバイス側でのスレッド分割を適切に行った。

2.2 音声情報処理

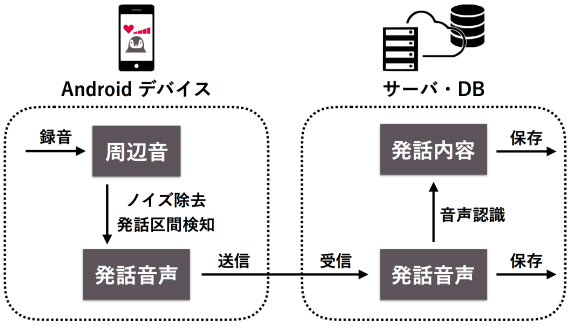

図 2: 音声情報処理の流れを示す。デバイス側で行う処理を左に、サーバ側で行う処理を右にそれぞれ示している。ここで、デバイス側で行う処理は全て撮影中リアルタイムで行われる。

次に、音声情報処理について詳説する。処理の流れを図 2 に示す。音声情報処理では、Android デバイスに入力された音声に対してノイズ除去と発話区間検知を行い、発話区間とみなされた部分については、音声から音量特徴量を抽出し、特徴量と音声そのものの両方をサーバに送信する。サーバに送信された発話音声は逐次的に音声認識処理にかけられ、発話内容のスクリプトを獲得する。ノイズ除去、発話区間検知、音声認識には、フェアリーデバイセズ株式会社によって提供されている聴覚プラットフォーム mimi®[*3] を利用した。

3. 実験

SignaLog によって収集された実際の会話データを用いて、人間関係を示唆する情報を抽出する 2 件の実験を行い、限られた状況下ではあるが、その有効性を検証した。

∗1 https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
∗2 https://www.tensorflow.org/
∗3 http://www.fairydevices.jp/mimi/

表 1: 上下関係を仮定した会話に関する特徴量を示した表。

	A → B	B → C	A → C
発言率 (%)	87.0	78.9	78.9
自分の音量の平均 (RMS)	472.1	413.8	598.6
相手の音量の平均 (RMS)	267.9	311.7	404.3
自分の音量の分散 (RMS)	307.1	523.6	296.0
相手の音量の分散 (RMS)	112.1	210.9	199.6

表 2: 彼女及び女友達と行った、日常生活における会話に関する特徴量を示した表。

	A → 彼女	A → 女友達
発言率 (%)	75.6	70.4
自分の音量の平均 (RMS)	584.1	821.5
自分の音量の分散 (RMS)	137.8	192.2
自分の話速 (文字数 / s)	4.4	3.7
笑顔率 (%)	55	41

3.1 人間関係を仮定した会話実験

本実験では、実験参加者 A, B, C の 3 人の間に A > B > C という上下関係を仮定した。この仮定のもとで A, B, C で相互に会話を行い、SignaLog を用いた会話データの収集を行った。実験の結果、表 1 に示すように、収集した特徴量のうち、会話における発言率、音量の平均と分散について、より立場が上の人間のほうがいずれの特徴量においても大きな値を示し、他の特徴量についてはほぼ同じ値が得られることを確認した。この結果は、仮定した上下関係が「上司は部下との会話で数倍発言している」「上司は部下に対して威圧的に（大きな声で、抑揚をつけて）話している」といった非言語シグナルに反映されていることを示唆するものである。本実験により、非言語シグナルが関係性を反映することおよび SignaLog が会話の非言語シグナルを正常に取得できていることを確かめた。

3.2 異性との会話実験

本実験では、実験参加者 A が自身の交際相手（彼女）及び女友達と行った、日常生活における会話データを SignaLog によって収集した。A が彼女と会話したデータと女友達と会話したデータの比較の結果、表 2 に示す通り、A の音量の平均と分散、話速、会話相手の笑顔率に違いが現れた。これらの結果は、非言語シグナルを解析することで「親密さ」と呼べるような抽象的な関係を抽出できる可能性を示唆している。

4. 結論

非言語的なシグナルが解析対象となることによって、人間関係がより本質的な意味で明らかになる可能性が示された。これまで行われてきたコミュニケーションデータからの関係性抽出は、職場や病院といった狭いコミュニティ内で収集したデータが中心であったが、本アプリが示した顔認識や話者認識をによる自動データ収集によって、一般に広く利用されるライフログ収集アプリとなる可能性がある。

さらに、上記機能を実装する過程で開発したリアルタイムで映像解析を行うプラットフォームとしての機能は、今後広がっていくであろう IoT プラットフォームと親和する形で提供していくことによって、非常に高度な映像解析を IoT デバイス上で行うことも可能になる。

本研究で行った実験は非言語シグナルによる関係性の抽出可能性を示唆するものではあるが、実験の規模、期間等について

は不十分な部分があるため、SignaLog を用いた大規模なデータの収集と解析を行うことが今後の課題であるといえる。

参考文献

[1] Alex Pentland. Socially aware, computation and communication. *Computer*, 38(3):33–40, 2005.

[2] Alex Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, volume 5, 2004.

[3] D Olguin Olguin, Joseph A Paradiso, and Alex Pentland. Wearable communicator badge: Designing a new platform for revealing organizational dynamics. In *Proceedings of the 10th International Symposium on Wearable Computers (Student Colloquium)*, pages 4–6, 2006.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.