F-001

ブレインストーミングの動画を対象とした深層学習による認識結果統合システムの開発

An Integrating System for Recognition Results from Deep Learning for Brainstorming Videos

永井 隆介[1)]　藤田 茂[2)]
Ryusuke Nagai　Shigeru Fujita

1　序論

ビデオ会議で行われるブレインストーミングにおいて，参加者の振る舞いを把握しコーチの支援を図るビデオ会議支援エージェント [1] の研究を行っている．

ブレインストーミングの映像を新たに深層学習することは学習時間，計算リソースの負荷が高く困難である．これに対し，複数の一般的な目的のための既存学習済み分類器の認識結果を記号処理的に組み合わせることで状況把握を行う設計が示されている [2].

[2] によるシステムは図 1 のような階層構造をとる．まず会議状況の映像から複数の深層学習分類機による認識を行い，それら独立した認識結果を統合し関連付けて会議状況を把握する上で意味のある情報とする．そしてそれらの統合情報から，Atomic Action[2] と定義された「歩く」,「話す」,「見る」などといったブレインストーミングにおける行動の基本単位を取得する．システムはこのように得られた Atomic Action を記号処理することで会議状況を把握してブレインストーミングのルールに基づいた支援行動を行う．

本研究では図ブレインストーミングの動画を対象とした深層学習による認識結果統合システムの開発を行う．その手法としてフレーム間での人物情報の教師なし統合手法と，個人識別，物体検出及び視線検出情報からの各人物の注目物体情報への認識結果統合手法を提案し実装と評価を行った．

2　関連研究

Person Re-Identification (re-ID) とは異なる地点における人物の画像データに対して，人物の同一性を評価することで個人を識別する手法である [3]．従来はパターンマッチング等を用いて行われていたが，2019 年には Zhou らによって Omni-Scale Network (OS-Net) を利用する手法が提案された [4]．OS-Net とは大局的な特徴と局所的な特徴の組み合わせの特徴量である Omni-Scale 特徴量を学習する CNN アーキテクスチャである．この OS-Net をバックボーンとしたモデルである person-reidentification-retail-0277[5] が Intel により OpenVINO Model Zoo で公開されている．

また本稿では分類器に用いる既存の深層学習モデルとして，物体のバウンディングボックスとその種類を確信度とともに取得できる Detectron2[6] と，人物の注目物体情報を，顔のバウンディングボックスと注目点の 2 次元画像上での座標の組合せとして検出できる attention-target-detection[7] を利用している．

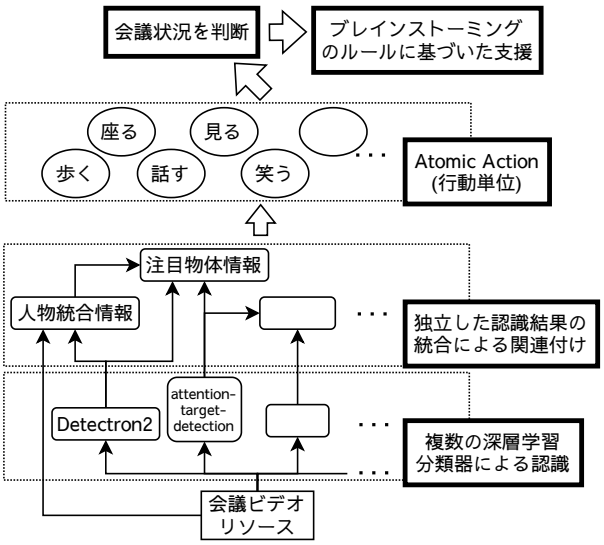

図 1　ビデオ会議支援システムの構造

3　提案手法

異なるフレーム間での人物情報の統合手法と，個人識別，物体検出及び視線検出情報からの各人物の注目物体情報への認識結果統合手法を提案する．

3.1　異なるフレーム間での人物情報の統合手法

下層のシステムから得られる入力情報は同一の分類器を用いたものであっても時間的に独立しているため，それらを関連付ける必要がある．

Atomic Action を得るためには人物の連続性が最も重要であるため，Detectron2[6] により得られる物体検出情報と入力映像のフレーム画像から人物追跡を行うことでフレーム間での人物情報の統合を行う手法を開発した．

入力情報は下層のシステムの処理の関係上フレーム間の開きが大きく離散的であり，またブレインストーミングは狭い会議室で積極的に行動する人物が存在しうるというシチュエーションであるため画像処理によるパターン認識を用いた追跡は困難である．

そこで本手法ではフレーム画像と物体検出情報から得られる人物画像から person-reidentification-retail-0277[5] で個人特徴量を取得してデータベースに格納し，累積した特徴量に対して k-means 法でのクラスタリングを行ってそのクラスタラベルにもとづいて人物を識別する．

なお処理間におけるクラスタの同一性については，前回実行時のクラスタと Jaccard 係数に基づいて比較すること再類似クラスタを対応させることで維持している．

1)　千葉工業大学 大学院 情報科学研究科.
Graduate School of Information and Computer Science, Chiba Institute of Technology
2)　千葉工業大学 情報科学部.
Department of Computer Science, Faculty of Information and Computer Science, Chiba Institute of Technology

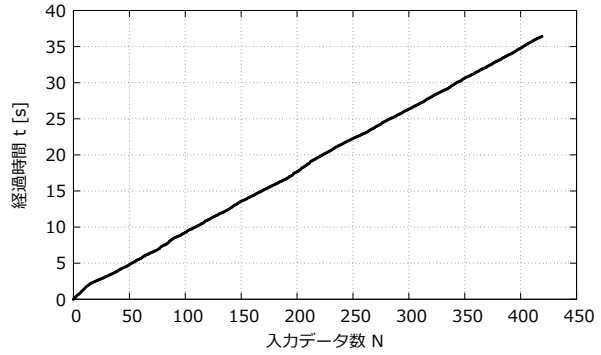

図 2　人物情報統合実験結果

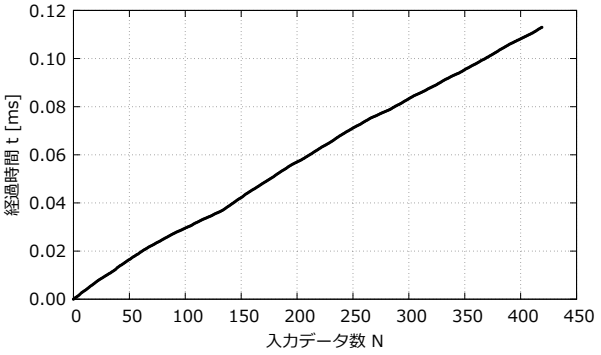

図 3　視線情報統合実験結果

3.2　個人識別，物体検出及び視線検出情報からの各人物の注目物体情報への認識結果統合手法

Look at や Talk to，Listen to, Write といったカテゴリの Atomic Action を得るうえで，誰が何を見ているかという情報は重要な要素である．

そこで本手法では，前項手法にて得られる個人識別情報と，Detectron2[6] による物体検出情報および attention-target-detection[7] によって得られる視線検出情報を統合し、個人ごとの注目情報を確信度として得る．

確信度の定義は視線元候補集合 A，視線先候補集合 B の要素 $a \in A, b \in B$ に対してある視線情報が誰から何に向けられたものであるのかについての確信度 $P(a,b)$ を式 1 としており，これはある視線情報に対して視線元と注目対象候補の組合せから自分自身同士の組合せを除いた値の逆数である．

$$P(a,b) := \begin{cases} \dfrac{1}{|A-B||B| + |A \cap B|(|B|-1)} & (a \neq b) \\ 0 & (a = b) \end{cases} \quad (1)$$

4　評価実験

前述した 2 つの手法について実装を行い動作とパフォーマンスを評価した．

対象とした入力は 419 秒間のブレインストーミングの動画と，この動画の 1 秒おきのフレーム画像から Detectron2 及び attention-target-detection を用いて得られた認識結果情報である．

また実験に使用した計算機環境は CPU: Intel Core i3-9100，メモリ: 32 GB である．

実験の結果，人物情報統合手法については，フレーム毎の実際に存在する人物数に対する検出できた人物数の割合としての人物検出率の平均は 0.9699 であり，フレーム毎の検出情報に対する正しい検出結果の割合としての検出正解率の平均は 0.9981 となった．

各フレーム処理時点のプログラム経過時間についてはそれぞれ図 2,3 のようになった．線形関係が見られることから安定した動作をしていると言える．また 1 秒おきという入力間隔に対して十分高速に動作することが確認できた．

5　結論

本研究では，遠隔地におけるビデオ会議支援エージェントについて，ブレインストーミングにおける会議状況を把握するシステムのために，既存の深層学習による学習済みモデルで認識された情報を統合する手法を提案，実装しその動作について評価を行った．

提案した手法は異なるフレーム間での人物情報の統合手法と個人識別，物体検出及び視線検出情報からの各人物の注目物体情報への認識結果統合手法の 2 つである．

また評価実験の結果として，どちらの手法も特別な前処理や訓練データを必要とせずに，一般的な計算機環境で十分高速かつ安定的に深層学習による認識結果の統合ができることを示した．

今後より多くの種類の認識情報を得てそれらの統合手法を作成し組み合わせていくことで，目標の Atomic Action を得られるようになると期待できる．

参考文献

[1] Gidel Thierry, Tucker Andrea, Fujita Shigeru, Moulin Claude, Sugawara Kenji, Suganuma Takuo, Kaeri Yuki, and Shiratori Norio, “Interaction Model and Respect of Rules to Enhance Collaborative Brain- storming Results,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 2, pp. 484–493, 2020.

[2] Shigeru Fujita, Thierry Gidel, Yuki Kaeri, Andrea Tucker, Kenji Sugawara, and Claude Moulin, “Ai-based automatic activity recognition of single persons and groups during brainstorming*,” In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3782–3787, 2020.

[3] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C.H. Hoi, “Deep learning for person re-identification: A survey and outlook,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[4] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang, “Omni-scale feature learning for person re-identification,” In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3701–3711, 2019.

[5] “Openvino toolkit - open model zoo repository.” https://github.com/openvinotoolkit/open_model_zoo/. (Accessed on 12/21/2021).

[6] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, “Detectron2.” https://github.com/facebookresearch/detectron2, 2019.

[7] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg, “Detecting attended visual targets in video,” In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.