The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

ミーティング中の頷きを対象とした ヒューマンインザループ動画アノテーションシステムの構築と評価

徳原 耕亮 *1 Kosuke Tokuhara　渡邉 洸 *2 Ko Watanabe　石丸 翔也 *3 Shoya Ishimaru　荒川 豊 *4 Yutaka Arakawa

*1*4九州大学　*2*3University of Kaiserslautern & DFKI GmbH

Manually labeling meeting videos is a burdensome task for annotators. In this study, we constructed a semi-automatic annotation system for meeting videos using a nodding recognition model. A comparison of the annotation of meeting videos using only manual annotation and human-in-the-loop annotation, which combines automatic annotation with human re-editing afterwards, showed that the time required for annotation was reduced by up to 17% compared to manual annotation, but increased by 28%. In some cases, the average reduction was -5%. The system is expected to improve accuracy and reduce work time through repeated use.

1. 背景

近年の働き方改革やコロナ禍の影響により，様々な企業でビデオ会議が広がっている．企業だけではなく，大学などの教育現場においてもビデオ会議が広く用いられるようになっている．そのためビデオ会議の解析や支援手法の研究が多く行われている．会議支援手法として言語情報や非言語情報を使用したものがあるが，会議の内容の機密性の高さによっては音声を記録することができない場合がある．そこで，頷きをはじめとする非言語情報を認識することでミーティングを解析する試みが提案されている [Watanabe 21]．認識モデルを構築するためには動画データ内に対応する行動ラベルデータが必要となる．会議動画のラベル付けはアノテーターにとって負荷のかかる作業である．そこで，これまでに構築した頷きの認識モデルを用いて，アノテーションコスト削減に向けたヒューマンインザループ型の半自動アノテーションシステムを提案する．本研究では，頷き認識モデルを用いたヒューマンインザループ型の半自動アノテーションによって，手動でアノテーションを行った場合とで作業内容の変化を評価する．

2. 実験

本章では，実験の手順，評価方法，評価結果について述べる．

2.1 実験手順

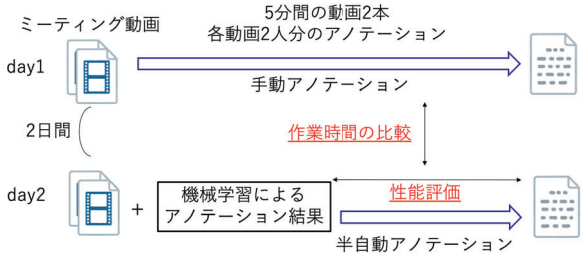

図 1: 実験の概要図

図 1 に実験の概要図を示す．提案システムの評価実験の協力者として，日頃からミーティング動画へのアノテーションを行っているアノテーター 1 名に比較実験を行ってもらった．

まず初めに，5 分間のミーティング動画 2 本に対しての頷きのアノテーションを行った．それぞれの動画は 2 名の参加者によるミーティングを記録しており，計 4 名に対するアノテーションを行った．この際のアノテーションでは，提案システムを使用せず，普段協力者がアノテーションを行っている手順に則って頷きのアノテーションを行った．次に同じ動画に対してモデルによる認識結果を抽出したのちに，頷きのアノテーションを行った．同じ動画に対してアノテーションを行うため，最初に手動でアノテーションを実施した日から間に 2 日間のインターバルを設けた．これらの手順によって得られたアノテーションの結果および作業にかかった時間を比較し，提案するシステムの有効性を評価する．

2.2 評価方法

本節では実験の評価方法について述べる．

時間短縮

提案する半自動アノテーションシステムを用いたことで短縮された作業時間．

T_{conv}：提案手法なしの場合の作業時間

T_{prop}：提案手法ありの場合の作業時間

$\frac{T_{conv}-T_{prop}}{T_{conv}}$ ：時間短縮

正解数

手動でのアノテーションとアノテーションシステムの両方で認識された頷きの数．

$n(A)$：アノテーターによる微調整後頷きとされたもの

$n(B)$：調整前頷きとされていたもの

$n(A \cap B)$：正解数

見落とし

実験初日の手動でのアノテーションでは「頷き」としていたにも関わらず，アノテーションシステムでは「頷き」と認識されなかった頷きの数．

$n(A \cap \overline{B})$：見落とし数

誤検知

モデル結果では「頷き」と認識されていたが，手動でのアノテーションでは「頷き」としていなかった頷きの数．

$n(\overline{A} \cap B)$：誤検知数

ずれの量

正解していたが頷きの時間 (長さ) が違い調整を行った際，どれだけ頷きの時間を調整したか．

連絡先: ko.watanabe@dfki.uni-kl.de

2.3 評価結果

本章では評価結果について述べる．本実験では 2 本の動画を用いた．以下に各評価基準についての結果を示す．

表 1: アノテーションに要した時間と削減率

	作業時間 (手動)	作業時間 (半自動)	削減率
参加者 1	9 分 3 秒	11 分 35 秒	-28%
参加者 2	44 分 12 秒	36 分 41 秒	17%
参加者 3	45 分 54 秒	47 分 12 秒	-3%
参加者 4	24 分 45 秒	22 分 35 秒	9%

表 2: 各ミーティング参加者に対するアノテーションを行った際の正解数，見落とし数，誤検知数．頷き認識モデルで出力されたものと，アノテーターによる修正後の比較．

	正解	見落とし	誤検知
参加者 1	7	0	28
参加者 2	22	21	8
参加者 3	49	10	59
参加者 4	12	16	17

表 3: 修正後の頷きの開始と終了時間のずれの量 (単位：秒)

	Start Time	End Time	Duration
max	3.192	3.784	4.88
min	0	0	0
ave	0.132	0.579	0.561

3. 考察

本章では，各評価基準に対する評価結果について考察を行う．

3.1 時間短縮について

4 人分全てのアノテーションに要した時間は手動の場合には 123 分 54 秒であったのに対して，提案システムを用いた場合では，118 分 3 秒と 5 分 51 秒の作業時間の削減に成功した．提案システムを利用すると，手動でのアノテーションと比較して作業時間が最大で 17%削減したが 28%増加したケースもあり，平均削減率は-5%となった．全体として十分なアノテーションコスト削減であり，さらにシステム利用を繰り返すことで更なる削減が期待できる．

しかし，アノテーションにかかった時間が伸びてしまったデータもあった．参加者 1 のデータに関しては実験で初めて触れたシステムの使用に手間取ってしまったことが原因として挙げられる．システム利用を繰り返すことでこれが原因で起こる作業時間の増加は軽減されることが予想される．また，原因として考えられるのは誤りの数である．作業時間が削減された参加者 2 では正解数 22 に対し見落とし，誤検知数が 29，参加者 4 のデータでは，それぞれ正解数 12 に対して見落とし，誤検知数が 33 であった．それに対して作業時間が増加してしまった参加者 1 では正解数 7 に対して見落とし，誤検知数が 28，参加者 3 では正解数 49 に対して見落とし，誤検知の数が 69 であった．誤りを修正する作業に時間を要した可能性が考えられる．更なる作業時間の削減に向けては認識精度の向上が必須である．また，作業時間が増加してしまった 2 つのデータセットに関して，参加者 1 では誤検知が 28 で作業時間削減率が-28%であったのに対して参加者 3 では誤検知が 59 と参加者 1 より大幅に多いにも関わらず，作業時間削減率は-3%であった．

この原因は 2 つ考えられる．1 つ目の原因は作業順によるものである．実験では表 1 の上から順に作業を行った．今回は初回の実験であり，被験者は参加者 1 のアノテーションを行う際に初めて提案システムを使用した．慣れない作業だったために作業時間が大幅に伸びてしまったのだと考えられる．2 つ目の原因は正解数による作業効率の上昇である．確かに参加者 3 では誤検知が多いが正解数も 49 と比較的多い．誤検知の修正作業による作業時間の増加と正解による作業時間削減が打ち消しあったことで大きな作業時間増加に繋がらなかったのだと考えられる．

3.2 正解，見落とし，誤検知について

表 2 にアノテーションを付与したミーティング参加者ごとの正解数，見落とし数，誤検知数を示す．全体的に見落としに比べて誤検知が多いことが読み取れる．見落としが少ないことから，比較的小さい動作による頷きも検知することができていることがわかる．それに対して誤検知が多いのは，「頷きでないが顔の動きを伴う動作」を頷きとして識別してしまっていることが原因として考えられる．例えば，座椅子に座っている際に少し腰を浮かせて座り直すような動作はカメラに対する顔の位置や角度が大きく変化するため，現在の頷き認識アルゴリズムではほとんど確実に頷きと認識してしまう．また「笑う」という動作に伴う体や顔の動きを頷きと認識してしまうことがあることも確認できた．

このような体全体の動きを頷きとして認識してしまうことを回避するためには，特徴量を追加し体の動きにロバストな特徴量を使用することが求められる．

3.3 ずれの量について

表 3 に正解していた頷き箇所の開始時間，終了時間，頷きの間隔のずれの大きさを示す．ずれの最大値は開始時間，終了時間では 3〜4 秒ほど，間隔はおよそ 5 秒のずれがあった．最小値はどれも 0 であり，調整を加える必要のない頷き区間もあった．平均値では開始時間のずれが 0.132，終了時間が 0.579，間隔が 0.561 であった．ずれの大きさの原因は誤検知数と同様に連続した頷きによるものだと考えられる．今回の実験を通じて，頷きに関して，動作の開始時間の認識よりも終了時間の方が大きくずれていることから，頷きの終了時間の判定が難しいことが分かった．今後の展望としては頷きの終了時に発生する特徴を組み込むことが大事になる．

4. 結論

本研究では，頷きの認識モデルを用いて動画の自動アノテーションシステムを構築した．提案するヒューマンインザループ型の半自動アノテーションシステムを利用すると，手動でのアノテーションと比較して作業時間が最大で 17%削減したが 28%増加したケースもあり，平均削減率は-5%となった．また，繰り返しシステムを利用することでさらに精度向上および作業時間削減が期待できることを示した．今後の展望として，再編集を実施した際の誤差情報をモデルに再学習させることによる継続的なモデル精度の向上を目指す．

謝辞本研究は，阪大 Society 5.0 実現化研究拠点支援事業（JPMXP0518071489）のもと実施するグランドチャレンジ研究，JSPS 科研費（JP18H03233 ）の支援により実施されている．

参考文献

[Watanabe 21] Watanabe, K., Soneda, Y., Matsuda, Y., Nakamura, Y., Arakawa, Y., Dengel, A., and Ishimaru, S.: Discaas: Micro behavior analysis on discussion by camera as a sensor, *Sensors*, Vol. 21, No. 17, p. 5719 (2021)