複数の深層学習モデルによるスキル優劣判定とその判断根拠に関する研究

TR20007 高田雅之　　指導教授：藤吉弘亘

1. はじめに

スキル優劣判定は，動画内の動作の優劣判定を行うタスクであり，これまでに Rank-aware Attention Networks (RAAN) [1] が提案されている．RAAN の学習には，優劣の 2 値ラベルが付与された 2 つの動画を入力して，Ranking Loss により重みの更新を行う．このとき，動画を K 個のセグメントに分割し，全てのセグメントが正解になるよう学習する．優劣の 2 値ラベルは動画全体を見てラベルが付与されており，動画内に含まれる全てのセグメントが該当するわけではない．しかし RAAN は，優劣の差が無いセグメントに対して全体に付与されたラベルを使用するため，誤った優劣判定を誘発するという問題がある．本研究では，この問題を解決するアプローチとして，Multiple Instance Learning (MIL) を用いて，優劣の差があるセグメントに着目したスキル優劣判定法を提案する．提案手法は，動画をバッグ，動画から切り出したセグメントをインスタンスとし，バッグのみにラベルを付与して優劣判定を行う．これにより，優劣の差があるインスタンスのみに着目した学習が可能となる．また，優れた動作と劣った動作を個別に判定することで，それぞれの時間情報に関する判断根拠の獲得を目指す．

2. 関連研究

スキル優劣判定は，スキルを評価する手法であるため，医療やスポーツ等を対象とした研究が行われている．動画共有サービスによって様々なスキルが含まれた動画が溢れる現在では，様々な動作に使用できるスキル優劣判定の研究が必要となる．様々な動作に使用できるスキル優劣判定手法の中で Rank-aware Attention Networks (RAAN) [1] は，比較的長い動画に対してスキル優劣判定を行う手法である．この手法では，Temporal Attention を導入し，動画内における優れたセグメントと劣ったセグメントを識別することができる．

3. 提案手法

本研究では，優劣の差があるセグメントに着目するためのスキル優劣判定手法である Skill MIL を提案する．

3.1 Skill MIL

提案手法の構造を図 1 に示す．まず，動画を K 個のセグメントに分割する．それぞれのセグメントから連続する 16 フレームをランダムに選択する．提案手法では MIL を導入するために，1 つのセグメントをインスタンス k とし，K 個のインスタンスの集合体をバッグ $\beta = \{k_x | x = 1, \ldots, K\}$ と定義する．インスタンスにはラベルを付与せず，バッグにのみラベルを付与して学習を行う．入力は，優れたスキルを表す動画を p_i と，劣ったスキルを表す動画を p_j のペアとする．動画は $p_i > p_j$ の関係であり，p_i から構成されるバッグをポジティブバッグ β^p，p_j から構成されるバッグをネガティブバッグ β^n とする．β^p，β^n に含まれるインスタンス k_x^p，k_x^n を Feature extractor に入力する．Feature extractor からの出力を，それぞれ Superior ranking module (SRM) と Inferior ranking module (IRM) に入力する．その後，SRM と IRM の出力を演算することで最終的なスコアを獲得し，優劣判定を行う．

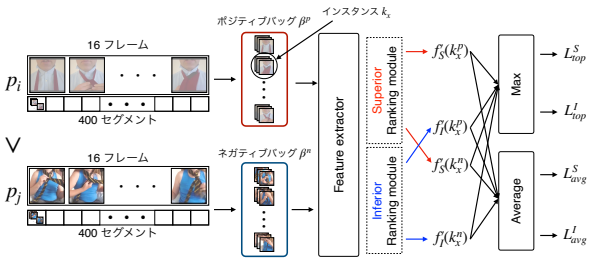


図 1：提案手法の構造

3.2 Ranking module

SRM と IRM の内部構造は同一であり，その構造を図 2 に示す．各モジュールは，3 層の全結合層と Attention sub module で構成する．Feature extractor の出力 $\mathbf{x}$ を入力としたとき，Attention sub module はインスタンスの重要度 $A_{\mathbf{x}} = \{a_k | k = 1, \ldots, K\}$ を出力する．各インスタンスの重要度 a_k の算出方法を式 (1) に示す．

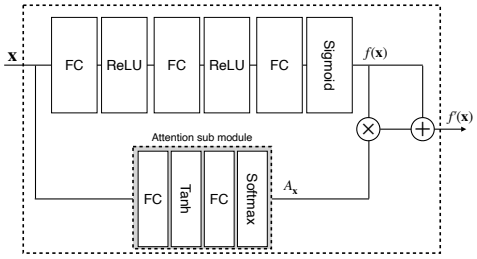


図 2：Ranking module の構造

$$a_k = \frac{\exp\{\mathbf{w}^{\mathrm{T}} \tanh(\mathbf{V}\mathbf{x}_k^{\mathrm{T}})\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^{\mathrm{T}} \tanh(\mathbf{V}\mathbf{x}_j^{\mathrm{T}})\}} \tag{1}$$

ここで，$\mathbf{w}, \mathbf{V}$ は学習するパラメータを表す．獲得したインスタンスの重要度 $A_{\mathbf{x}}$ は，式 (2) に示すように 3 層の全結合層からの出力である特徴マップ $f(\mathbf{x})$ に重み付けとして乗算し，重み付け前の特徴を加算する．これにより，特徴マップの消失を抑制し，インスタンスの重要度を優劣判定に反映させることができる．

$$f'(\mathbf{x}) = f(\mathbf{x}) + (A_{\mathbf{x}} \cdot f(\mathbf{x})) \tag{2}$$

3.3 損失関数と学習

Skill MIl で使用する損失関数について説明する．SRM の損失関数 L^S_{top} と L^S_{avg} は，$\beta^p > \beta^n$ の関係になるよう学習することで，動画内の優れた動作に焦点を当てた学習ができる．L_{top} は，バッグ内の最大値を比較する損失関数であり，L_{avg} は，バッグ内の平均値を比較する損失関数である．L^S_{top} と L^S_{avg} を，式 (3)，(4) に示す．

$$L^S_{top} = \max(0, 1 - \max_{k_x^p \in \beta^p} f'_S(k_x^p) + \max_{k_x^n \in \beta^n} f'_S(k_x^n)) \tag{3}$$

$$L^S_{avg} = \max(0, 1 - \frac{1}{\beta}((\sum_{\beta} f'_S(k_x^p)) + (\sum_{\beta} f'_S(k_x^n)))) \tag{4}$$

ここで，f'_S は SRM の出力を表す．L^S_{top} は，β^n より β^p の方が最大値が高くなるよう学習し，L^S_{avg} は，β^n より β^p の方が平均値が高くなるよう学習する．

一方で，IRM の損失関数 L^I_{top} と L^I_{avg} は，$\beta^n > \beta^p$ の関係になるよう学習することで，動画内の劣った動作に焦点を当てた学習が可能となる．L^I_{top} と $L^S_{avg}I$ を，式 (5)，(6) に示す．

$$L^I_{top} = \max(0, 1 - \max_{k_x^n \in \beta^n} f'_I(k_x^n) + \max_{k_x^p \in \beta^p} f'_I(k_x^p)) \tag{5}$$

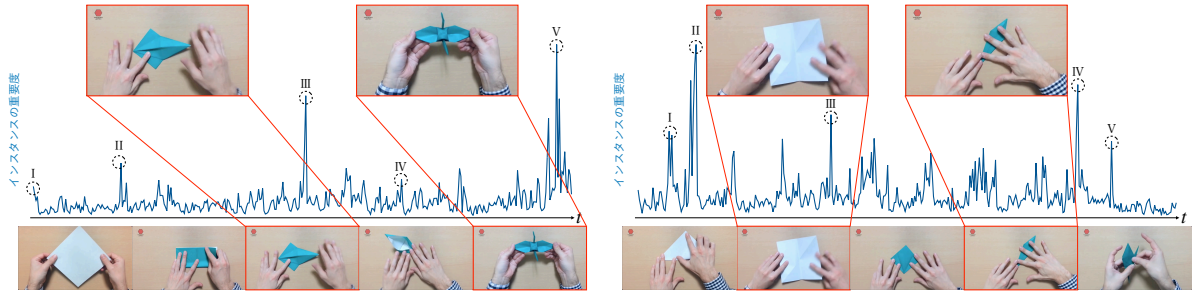
$$L^I_{avg} = \max(0, 1 - \frac{1}{\beta}((\sum_{\beta} f'_I(k_x^n)) + (\sum_{\beta} f'_I(k_x^p)))) \tag{6}$$

ここで，f'_I は IRM の出力を表す．L^I_{top} は，β^p より β^n の方が最大値が高くなるよう学習し，L^I_{avg} は，β^p より β^n の方が平均値が高くなるよう学習する．

Skill MIL 全体の損失関数を，式 (7) に示す．

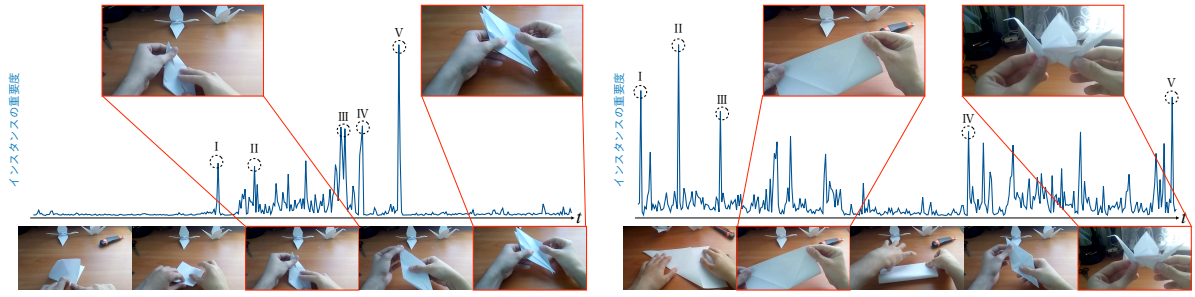
$$L = L^S_{top} + L^I_{top} + \lambda L^S_{avg} + \lambda L^I_{avg} \tag{7}$$

ここで，λ は寄与率を調整するための係数であり，$\lambda = 0.5$ を用いる．



(a) SRM (b) IRM

図 3：高いスキルレベルの動画に対するインスタンスの重要度 (Origami タスク)



(a) SRM (b) IRM

図 4：低いスキルレベルの動画に対するインスタンスの重要度 (Origami タスク)

3.4 優劣判定

本研究では．対応するインスタンスの優秀かどうかを表すスコア $score_{high}$ と，劣悪かどうかを表すスコア $score_{low}$ を比較することでスキル優劣判定を行う．それぞれ式 (8), (9) により求める．

$$score_{high} = \max_{k_x^p \in \beta^p} f'_S(k_x^p) + \max_{k_x^n \in \beta^n} f'_I(k_x^n) \tag{8}$$

$$score_{low} = \max_{k_x^n \in \beta^n} f'_S(k_x^n) + \max_{k_x^p \in \beta^p} f'_I(k_x^p) \tag{9}$$

$score_{high}$ と $score_{low}$ を比較し，スコアの高い方を判定結果とする．

4. 評価実験

提案手法の有効性調査を行う．比較する手法には RAAN を用いる．データセットには BEST Dataset を使用する．Best Dataset は，YouTube から取得した動画で構成され，Apply Eyeliner，Braid Hair，Origami，Scramble Eggs，Tie Tie の 5 つのサブセットで構成されるデータセットである．RAAN と実験条件を揃えるため，公開されている Best Dataset の I3D 特徴を使用する．本 I3D 特徴は，Kinetics 400 データセットで事前学習したモデルを使用して抽出されている．また，使用する際は過学習を抑制するためにノイズ $\{z|0 < z < 1\}$ を I3D 特徴に加算してから実験を行う．最適化手法には Adam を用い，学習率は 0.0001，バッチサイズを 128，$K = 400$ とし，試行回数を 2000 エポックとした．

4.1 実験結果

表 1：従来手法との比較 [%]

Method	Apply Eyeliner	Braid Hair	Origami	Scramble Eggs	Tie Tie	Average
RAAN	**84.3**	74.4	78.8	84.4	**87.1**	81.8
Ours	83.4	**75.2**	**80.7**	**90.3**	**87.1**	**83.3**

従来手法である RAAN と比較した結果を表 1 に示す．表 1 より，提案手法は平均して 1.5pt 従来手法より精度が向上した．特に Scramble Eggs タスクでは，5.9pt 向上した．次に，インスタンスの重要度の可視化例を図 3，4 に示す．図 3 (a) と図 4 (a) に示すように，SRM では，動画の中盤部分を強く重要視していることが確認できる．これは中盤部分の作業が折り鶴の完成度に大きく影響を与えている為だと考えられる．また，図 3 (a) では，中盤部分のみではなく動画の終盤に対しても強く重要視している．これはこの動画の終盤で，綺麗に折られた折り鶴をカメラに写している (図中 V) ためだと考えられる．一方で，IRM では，図 3 (b) と図 4 (b) に示すように動画の序盤や終盤を強く重要視していることが確認できる．スキルレベルの低い動画は，折り直す動作を複数回行うことが多い．序盤を重要視する原因は，複数回行う折り目を付ける動作を折り直す動作と判定しているからだと考えられる．また終盤を重要視する原因は，完成した折り鶴を見せている動画が多いからだと考えられる．

4.2 インスタンスの重要度の有効性

表 2：インスタンスの重要度 反転時の精度変化 [%]

Inversion	Apply Eyeliner	Braid Hair	Origami	Scramble Eggs	Tie Tie	Average
before	**83.4**	**75.2**	**80.7**	**90.3**	**87.1**	**83.3**
after	53.6	63.9	52.8	51.1	78.1	59.9

提案手法で獲得したインスタンスの重要度 $A_{\mathbf{x}}$ の有効性を評価する．評価方法として，インスタンスの重要度を反転させて推論を行う．反転する前と反転した後での精度を比較することで精度がどれだけ低下するかを調査する．表 2 にインスタンスの重要度の反転時の精度変化を示す．表 2 より，インスタンスの重要度の反転により大幅に精度が低下することが確認できた．特に，Scramble Eggs タスクでは 39.2pt の低下が確認できた．よって，スキル優劣判定に有効なインスタンスの重要度を得られる事が確認できる．

5. おわりに

本研究では Skill MIL を提案し，優劣の差があるセグメントに着目する学習を行い従来手法より高精度を獲得した．また，インスタンスの重要度を獲得し，ネットワークに応用することで，精度の向上と時間情報の判断根拠が獲得可能なこと示した．今後は，空間情報に対する重要度を獲得し，更なる精度向上を検討する．

参考文献

[1] H. Doughty. *et al.*: “The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos”, CVPR. 2019.

研究業績

[1] 高田雅之, 等,“Attention Pairwise Ranking によるスキル優劣判定における視覚的説明と高精度化”, 第 23 回 画像の認識・理解シンポジウム, 2020.