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

ER16047 高田雅之　　指導教授：藤吉弘亘

1. はじめに

スキルの優劣を判定する Skill Assessment タスクの手法として，深層学習による Pairwise Deep Ranking (PDR) [1] が提案されている．スキル学習者は，PDR が選定した優良なデータを参考にすることで，技能スキルを効率的に習得可能となる．しかし，スキルの優劣判定の判断根拠が分からないと，PDR の信頼性が損なわれる．そこで，本研究では，判断根拠の視覚的説明が可能となる Attention Pairwise Ranking を提案する．提案手法は，視覚的説明となる注視領域を Attention branch により獲得し，注視領域の情報を活用することでの精度向上を目的とする．

2. Pairwise Deep Ranking

PDR は，2 つの動画像データを与えて，シーン内に映る動作の優劣を評価可能な深層学習ベースの Skill Assessment 手法である．2 ストリームの Temporal Segment Networks をベースとしており，動画内の動作を捉えることが可能である．PDR により，医療技術などの素人では優劣の判断が困難な専門的タスクの技能習得への適用が期待できる．

3. 提案手法

PDR は動作を評価した際の判断根拠を提示できない．そこで，本研究では PDR の出力結果に対する判断根拠を獲得できる Attention Pairwise Ranking を提案する．

提案手法は Attention branch および Attention 機構を導入し，学習中の注視領域を Ranking branch に与える．これにより，Attention Branch Network (ABN) [2] のように，精度向上が期待できる．また，優秀な部分を学習する Superior Network と，劣悪な部分を学習する Inferior Network に分割し，学習させることで，優と劣の注視領域を別々に獲得する．提案手法のネットワークを図 1 に示す．ここで，p_i，p_j は動画像データであり，優劣の関係は $p_j > p_i$ である．

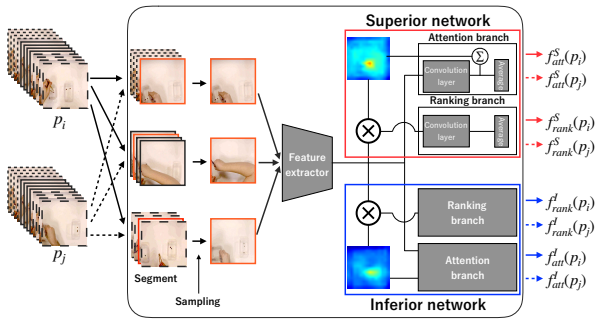

図 1：提案手法のネットワーク構造

Feature extractor は，3 つの Segment に等分した動画像データから，サンプリングした 1 フレームをベースネットワークに入力し，特徴マップを出力する．

Attention branch は，特徴マップと同サイズの重みフィルタを使用した Weighted Global Pooling(WGP) を用いて出力をする．また，総和したフレーム分の特徴マップを [0, 1] に正規化した Attention map を出力結果に対する注視領域として扱う．これにより，同時に全フレームを正規化することでフレーム毎の変化を表現する．

Ranking branch は，Feature extractor から抽出した特徴マップと Attention map を用いて最終的な出力をする．提案手法の損失 L は，Superior Network と Inferior Network の Attention branch，Ranking branch の誤差をそれぞれ，L^P_{att}，L^P_{rank}，L^N_{att}，L^N_{rank} とすると式 (1)〜(3) のようになる．

$$L = L^S_{att} + L^S_{rank} + L^I_{att} + L^I_{rank} \quad (1)$$

$$L^i_{att} = \log(1 + \exp(-f^i_{att}(p_i) + f^i_{att}(p_j))) \quad (2)$$

$$L^i_{rank} = \log(1 + \exp(-f^i_{rank}(p_i) + f^i_{rank}(p_j))) \quad (3)$$

4. 評価実験

本実験では，提案手法の有効性を評価するために従来手法との精度比較をする．

4.1. 実験概要

本実験では，EPIC-Skills 2018 Dataset の中から，Drawing (Dr)，Chopstick-Using (CU)，Suturing (Su)，Knot Tying (KT)，Needle Passing (NP) の 5 つのスキルデータを使用する．PDR のベースネットワークには ResNet34，ResNet101 を使用し，比較する．評価方法には，4-fold cross validation を用いる．

4.2. 実験結果

従来手法，提案手法の順位付けに対する精度を表 1 に示す．表 1 より提案手法は，従来手法と比較して，平均 15.2 ポイント精度向上した．

表 1：実験結果 [%]

手法	ResNet	Dr	CU	Su	NP	KT	平均
PDR	34	82.6	75.0	67.2	66.4	78.5	73.9
	101	86.3	75.0	66.2	76.1	76.4	76.0
ours	34	88.6	77.7	92.3	74.4	86.8	84.0
	101	**94.9**	**86.9**	**94.4**	**83.1**	**96.8**	**91.2**

4.3. 注視領域の可視化

提案手法の ResNet34 モデルによって獲得した Superior Network と，Inferior Network が注視したフレームを図 2，3 に示す．図 2，図 3 より，Superior Network と，Inferior Network で同フレームでも，異なる注視領域が獲得できていることが分かる．また，図中の赤く囲われた部分より，獲得した注視領域は動作部分や，動作結果が優劣に関係する場合はその部分にも反応することが確認できる．

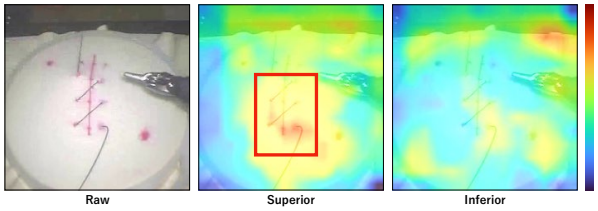

図 2：Superior Network が注視されるフレーム

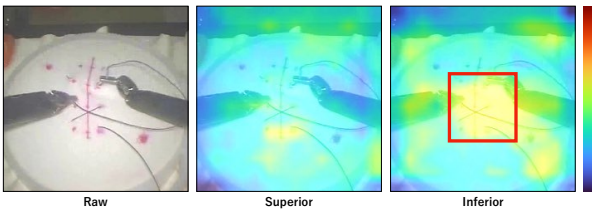

図 3：Inferior Network が注視されるフレーム

5. おわりに

本研究では，Attention branch を導入した Pairwise Deep Ranking を提案した．提案手法は全てのタスクで高精度であることを確認した．また，Superior Network，Inferior Network で異なる注視領域が獲得できることがわかった．また，動作結果にも反応することが確認できた．今後は，他のスキルデータに対して評価を行う．

参考文献

[1] H. Doughty, *et al.*, “Who ’s Better? Who ’s Best? Pairwise Deep Ranking for Skill Determination”, CVPR, 2018.

[2] H. Fukui, *et al.*, “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”, CVPR, 2019.