対戦アクションゲームにおけるプレイヤーの挙動観察のためのシーン検索

三ツ井 慧太郎[1]　　岡部 誠[1]

概要.　我々は対戦型格闘ゲームのプレイヤーが強くなるための支援に取り組んでいる．強いプレイヤーになるための重要なアプローチの 1 つに，他の強いプレイヤーの挙動を観察し，強い行動パターンを学ぶことがある．そこで我々は YouTube などの動画共有サイトに上がっている多くの対戦動画から，強いプレイヤーの挙動を効率よく観察するために，ゲームシーンの検索手法を提案する．提案手法は，対戦中のある瞬間の画像をクエリーとして入力すると，その画像と似通った状況のシーンを対戦動画の中から複数検索して提示する．ユーザは検索結果を再生することで，強いプレイヤーならクエリーの状況からどういった行動をとるのかについて観察できる．このシーン検索を実現するために，対戦ステージや背景を無視し，キャラクターの位置関係のみを考慮した特徴ベクトルを生成できるディープニューラルネットワークを学習する．シーン検索の精度については主観評価を行い，約 8 割の精度でシーン検索ができることを確認した．また，ユーザスタディを行い，動画で上級プレイヤーの挙動を観察しながら CPU 対戦を繰り返したとき，提案手法を用いることでより高いレベルの CPU に勝てるようになることを確認した.

1　はじめに

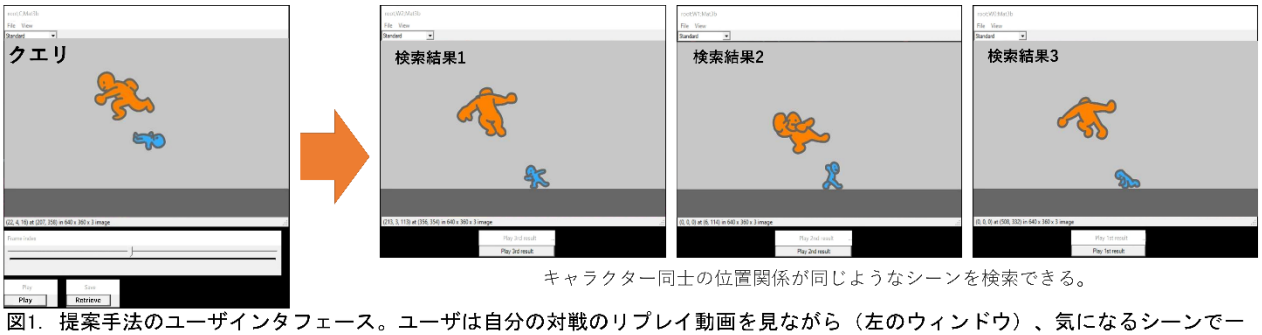
近年，競技性の高いビデオゲームをスポーツ競技として捉えて”eSports”と呼ぶほど，ビデオゲームに注目が集まっている．eSports はファースト・パーソン・シューティング(FPS)や，対戦型格闘ゲーム，サッカーやバスケットボールなどのスポーツゲーム，レーシングゲームなど，多種多様なジャンルがあり，それぞれのジャンルに複数のタイトルがある．eSports で活躍するためには，プレイヤーたちは各タイトルに特化してゲームを練習し強くなる必要がある．我々は毎年世界大会が開かれているニンテンドースイッチの大乱闘スマッシュブラザーズ SPECIAL（スマブラ）[1]を題材にプレイヤーが効率よく強くなるための支援に取り組んでいる．

スマブラのような対戦型格闘ゲームにおいて，強いプレイヤーになるために重要なことが 2 つある．1 つはゲームの操作技術を磨くことである．これはキャラクターの移動操作をミス無くこなしたり，技を出す際のコマンド入力を正確にできるようになるなどだが，これらはプレイヤー自身が時間をかけて練習して習得する必要がある．もう 1 つの重要なことは対戦相手を意識した練習をすることである．eSports はビデオゲームを媒体にしているが，要は人と人との試合であり，相手をだますようなフェイントを掛けたり，相手のミスを誘発するような行動を仕掛けたりする駆け引きに面白さがある．対戦相手を意識した練習には，YouTube などの動画共有サイトに上がっている対戦動画を見て，強いプレイヤーの挙動を観察し，強い行動パターンを学ぶことが有効である．しかしこれは簡単ではない．動画共有サイト上の動画数が多いため，自分の見たい対戦動画がなかなか見つからなかったり，また，対戦動画が見つかっても，その中のどこに自分の見るべきシーンがあるかが分からない，といった問題がある．

そこで我々は，多くの対戦動画から強いプレイヤーの挙動を効率よく観察するために，ゲームシーンの検索手法を提案する．提案手法は，対戦中のある瞬間の画像をクエリーとして入力すると，その画像と似通った状況のシーンを対戦動画の中から複数検索して提示する（図 1）．ユーザは検索結果を再生することで，強いプレイヤーならクエリーの状況からどういった行動をとるのかについて観察できる．

この目的を実現するために，既存の画像/動画検索手法について調査したが，我々の要求に対応するものは存在しなかった．既存の検索手法の大きな問題の 1 つは，我々はゲーム内のキャラクター同士の位置関係のみに興味がある．しかし，既存の画像/動画検索手法は画像全体の見た目が似たものを検索してしまうため，キャラクター同士の位置関係よりもむしろ，戦っているステージや背景が似たものを検索してしまう．そこで我々は入力画像をキャラクター

1 静岡大学大学院総合科学技術研究科工学専攻
数理システム工学コース



キャラクター同士の位置関係が同じようなシーンを検索できる。

図1. 提案手法のユーザインタフェース。ユーザは自分の対戦のリプレイ動画を見ながら（左のウィンドウ）、気になるシーンで一時停止し、「検索」ボタンを押す。すると、他の対戦動画の中から、類似した3つのシーンが検索される（右の3つのウィンドウ）。

同士の位置関係のみの情報を持った特徴ベクトルに変換できるようなディープニューラルネットワークを学習することで，シーン検索手法を実現する．

シーン検索の精度については主観評価を行い，約 8 割の精度でシーン検索ができることを確認した．また，ユーザスタディを行い，動画で上級プレイヤーの挙動を観察しながら CPU 対戦を繰り返したとき，提案手法を用いることで平均的に 0.714 回多く勝てることが示された．言い方を変えると，より高いレベルの CPU に勝てるようになることが確認できた．

2　関連研究

類似画像/動画の検索手法については数多くの関連研究が存在する[2]．基本的な考え方は，画像を大規模な畳み込みニューラルネットワーク(CNN)によってエンコードし，得られた特徴ベクトルの類似度によってデータベース内の画像にランキング付けをすることである[3]．同一クラス内の画像の特徴ベクトル同士は類似するように，逆に，異なるクラス間の画像の特徴ベクトル同士は異なるようにしたく，そのような特徴ベクトルを出力できる CNN を学習するために，ネットワークの構造や loss 関数に様々な工夫が提案されてきている．例として，Revaud らは画像検索学習モデルの評価指標の 1 つである mean Average Precision(mAP)を直接最適化できるランキング学習手法を提案した．その結果，画像集合に対し，そこに写った建物毎にラベル付けしたデータセットを学習することで，クエリー画像に写ったものと同じ建物が写っている画像を高い精度で検索できる手法を提案した[4]．こういった画像検索技術の発展形としては，Vo らは画像と文をクエリーとし，画像を文によって変化させたような画像を検索できる手法を提案した[5]．

しかし，上記の既存手法のいずれも，我々の目的であるキャラクター同士の位置関係のみに着目したシーン検索という問題にそのまま適用できる手法は存在しなかった．そこで，まず動画のセグメンテーション手法を適用し，動画からキャラクターのみを抽出してから上記の検索手法を適用するというアプローチを試みた．動画からキャラクターのみを抽出するために，SiamMask[6]と STM[7]という 2 つの手法を実験した．共に動画の最初のフレームで追跡したい物体を指定すると，2 フレーム目以降を自動的に追跡し，物体のマスクを生成する手法である．しかし，実際にスマブラのキャラクターを追跡させてみると，最初の数秒は上手く追跡できても，キャラクター同士の衝突やエフェクト（煙や炎など）が起こってキャラクターが一瞬隠れたり，キャラクターのポーズが大きく変化するなどすると，その後のフレームを追跡できなくなった．特に STM は動画のセグメンテーションの分野では最も精度の高い手法の 1 つであるため，STM が上手くキャラクターを抽出できないならば，このアプローチは不可能と判断した．

スマブラに特化した研究として，強化学習によって強い AI プレイヤーを育てる手法がある[7]．こういった AI プレイヤーとの対戦も，人がスマブラに強くなるための 1 つのアプローチかも知れない．

3　ユーザインタフェース

図 1 に提案手法のユーザインタフェースを示す．図 1-左のウィンドウでは，ユーザは自分の対戦のリプレイ動画を見ているとする．そして，図 1-左のウィンドウに表示されているようなシーンで「こんなふうにマリオの頭上にドンキーコングが浮いている状況のとき，強いプレイヤーならどういった行動をとるのだろう？」と気になったなら，ここでリプレイを一時停止し，「検索」ボタンを押す．すると，強いプレイヤーの対戦動画の中から，類似した 3 つのシーンが検索されて表示される（図 1-右の 3 つのウィンドウ）．検索結果はいずれもマリオの頭上にドンキーコングが浮いている状況である．ユーザはこれらの動画を再生することで，強いプレイヤーがここからどういった行動をとるのかを観察できる．

「検索」ボタンを押してから結果が表示されるま

での時間は，現在の我々の実装では 10 秒未満を要する．また，表示される検索結果の数は，現在は類似性の高いものから順に 4 つまで，としている．

尚、検索対象となる強いプレイヤーの動画は、ユーザが予め収集しておくものとする。動画の収集方法としては、①「世界戦闘力」という任天堂が公式に定めるプレイヤーの強さの指標に基づいて集める、②ゲーム大会におけるトーナメント上位の対戦動画を集める、③実績のある有名なプレイヤーの動画を集める、などが考えられる。

4　シーン検索手法

関数fは画像xを入力すると特徴ベクトル$y = f(x)$を出力するニューラルネットワークであるとする．入力画像x_iが図 2-a のような画像であって$y_i = f(x_i)$であるとする．このとき，背景は異なるがキャラクター同士の位置関係が図 2-a と同じような画像集合（図 2-b，正例と呼ばれる）の中の画像 1 つをx_pとすれば，その特徴ベクトル$y_p = f(x_p)$はy_iと似たものであってほしい．逆に，キャラクター同士の位置関係が図 2-a と異なるような画像集合（図 2-c，負例と呼ばれる）の中の画像 1 つをx_nとすれば，その特徴ベクトル$y_n = f(x_n)$はy_iと異なったものあってほしい．

このような関数fを実現できれば，特徴ベクトルyの類似度を使うことで，ステージや背景を無視しキャラクター同士の位置関係のみに着目したシーン検索が実現できる．類似度には内積を用い，検索対象の全対戦動画の各フレームに対し，特徴ベクトルの類似度に基づいたランキング付けをすることでシーン検索ができる．

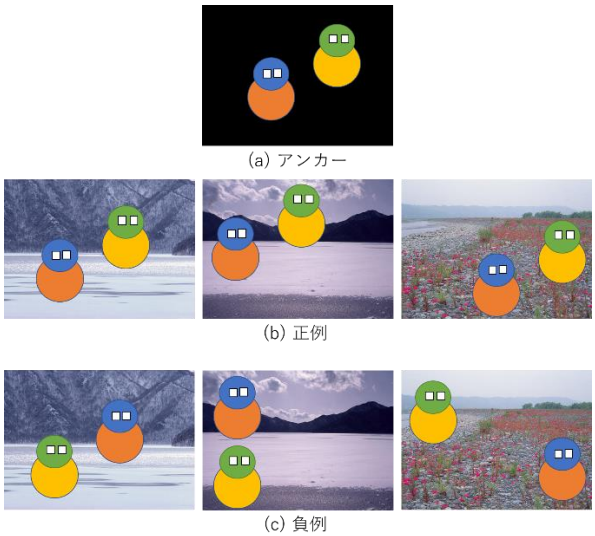


図 2.　Supervised Contrastive Learning におけるアンカー画像(a)とそれに対する正例(b)と負例(c).

関数fに ResNet-50[9]を用いた．画像xのサイズは 256×144 ピクセル×3 チャンネルであり，特徴ベクトルyのサイズは 2048 次元である．上記のような関数fを実現するために Supervised Contrastive Learning[10]を用いる．損失関数は

$$\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(y_i\cdot y_p/\tau)}{\sum_{a\in A(i)}\exp(y_i\cdot y_a/\tau)}$$

である．Iは全ての画像の添え字の集合，$P(i)$は正例の画像の添え字の集合，$A(i)$はアンカー画像以外の全ての画像の添え字の集合，τは温度と呼ばれるハイパーパラメータである．$a\cdot b$はベクトルaとベクトルbの内積である．上記の損失関数を最小化することによって，y_iとy_pの内積がy_iとy_nの内積よりも大きくなるように関数fを学習することができる．この関数fによってステージや背景を無視しキャラクター同士の位置関係のみの情報を持った特徴ベクトルを得ることができる．

4.1 学習

図 2 に示したように，関数fの学習にはアンカー画像に対する正例集合と負例集合が大量に必要となる．動画中のあるフレームをアンカーに選んだとき，それ以外のフレームは全て負例となり得る（キャラクターが移動しており位置関係がアンカーと異なるため）ので，負例集合はアンカー以外のフレームを選ぶことで簡単に用意できる．しかし，正例はキャラクターの位置関係がアンカーと同じで，かつ，背景が異なる画像であるため，通常は簡単に用意できない．そこで我々はスマブラの特徴を生かした正例の大量生成手法を以下に提案する．

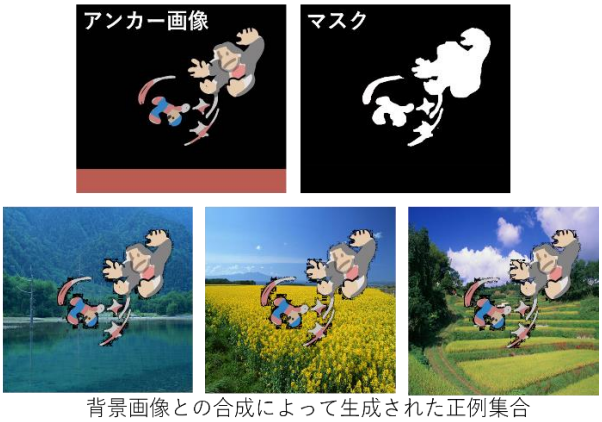


図 3.　ステージ「75m」を利用した正例の大量生成

データセットの作成にあたり，スマブラの対戦動画から 2000 フレームを抽出した．この動画は「75m」というステージで対戦されたものだが，このステージに関しては背景に現れるピクセルの色数が少ない（図 3-アンカー画像）．そのために容易に背景部分

のピクセルを特定することができる，即ち，キャラクター部分のマスクを作ることができる（図 3-マスク）. このマスクを用いて背景を別の画像に入れ替えることで，アンカー画像の正例を大量生成することができる（図 3-下段）.

このようにして背景を入れ替えたバージョンを 20 種類作ったため，合計で 2000×21＝42000 枚の画像からなるデータセットとなった.

学習は 2000 フレームから 5 フレームをランダムに抽出し，それらの背景を入れ替えたバージョン（合計で 5×21＝105 枚）を 1 つのバッチとして学習を行う．即ち，このバッチの中であるアンカー画像を 1 枚選んだ時，そのアンカー画像に対する正例がこのバッチには 20 枚，負例がこのバッチには 84 枚存在することになる．このような学習を 400 回繰り返すことで 2000 フレーム全てに対する学習を行ったとし，これを 1 エポックとする．400 回の損失関数の値の平均の推移（図 4）を見ると，学習が上手く進行していることがわかる.

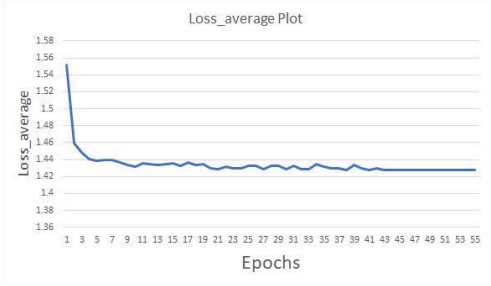


図 4．損失関数の値の変化.

4.2 シーン検索の精度評価

ステージや背景を無視しキャラクター同士の位置関係のみを考慮した検索が上手くできているかを調べるために主観評価を行った．今回の実験では 9 人の被験者（工学系の大学または大学院に所属する 20 代の男性）に協力してもらった.

各被験者には，クエリー画像 1 枚と，そのクエリー画像を入力として提案手法で検索を行って得られた上位 4 枚の検索結果の画像を提示する．クエリー画像と検索結果を見比べてもらい，以下のレベル 1 ～ 3 の質問に対し，Yes/No で答えてもらった.

- レベル 1 ：4 枚の検索結果の中に，クエリー画像とキャラクターの位置関係が似ているものが 1 枚でもあるかどうか
- レベル 2 ：4 枚の検索結果の中に，レベル 1 に加え，エフェクト（煙や炎など）までもが似ているものが 1 枚でもあるかどうか
- レベル 3 ：4 枚の検索結果の中に，レベル 1 & 2 に加え，キャラクターのポーズまでもが似ているものが 1 枚でもあるかどうか

各被験者には上記の評価を 97 枚のクエリー画像に対して行ってもらった.

97 枚のクエリー画像には，青のマリオと白のドンキーコングの対戦動画からランダムに選出したフレームを用いた．97 枚中の 49 枚は「75m」というステージでの対戦動画から選出されたフレームとなり，残りの 48 枚は「75m」以外のステージでの対戦動画から選出されたフレームとなった．検索対象は全て「75m」での対戦動画である．図 5 に主観評価の結果を示す.

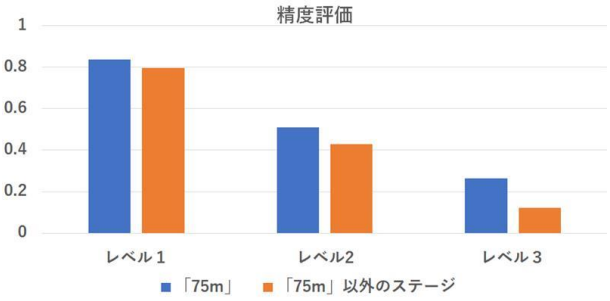


図 5. 主観評価の結果

図 5 を見るとレベル 1 の質問においては「75m」と「『75m』以外のステージ」の両方で 8 割の精度を達成できており，つまり，提案手法は上位 4 件の中にキャラクターの位置関係が似ているシーンを約 8 割の精度で検索できることが分かる．また，「75m」と「『75m』以外のステージ」の間でほとんど差が無いことから，ステージや背景を無視した検索に成功していることが分かる．レベル 2 の質問の結果からは，4 ～ 5 割の精度でエフェクトも似ているシーンが検索できていることが分かる．レベル 3 の質問の結果からは，キャラクターのポーズの類似性までを考慮した検索については精度が低いことが分かる．また，クエリー画像について，背景が単純な「75m」と背景が複雑な「『75m』以外のステージ」の間で精度に 2 倍程度の開きがあることから，キャラクターのポーズの特徴を捉える際に背景の複雑さがノイズとなって影響を与えていることが分かる.

5 ユーザスタディ

提案手法を用いて強いプレイヤーの挙動を観察したとき，それがプレイヤーの成長にどう影響を与えるかを調査するため，ユーザスタディを行った.

5.1 実験方法

今回の実験では 14 人の被験者（大学または大学院に所属する 20 代の男性 13 名と女性 1 名）に協力してもらった. 14 人のうち 7 人は提案手法を使って強いプレイヤーの挙動を観察するグループ，残りの

7 人は提案手法を使わずに強いプレイヤーの挙動を観察するグループとした.

各被験者にはスマブラの CPU 戦をプレイしてもらった.プレイするキャラクターは青のマリオとし,敵である CPU のキャラクターは白のドンキーコングとした. CPU のレベルは比較的弱い 3 から対戦を始め,勝った場合はレベルを 1 つ上げ,次のレベルの CPU と対戦してもらう.負けた場合は,負けたレベルの CPU と強いプレイヤー(任天堂公式の「世界戦闘力」の値がランキング上位 5〜6%に入っており、VIP の称号を持つ)が対戦している動画が予め用意してあるので,それを見て強いプレイヤーがどうやってその CPU を倒しているかを観察してもらい,その後,再び同レベルの CPU と対戦してもらった.以上を負けた回数が 5 回になるまで繰り返してもらい,最終的にどのレベルの CPU が倒せるかを調べた.尚, CPU のレベルの上限は 9 であり, レベル 9 に勝った場合はそれ以上レベルを上げることができないため,再びレベル 9 と対戦してもらうこととした.

強いプレイヤーの対戦動画を見るとき,提案手法を使うグループの被験者は,まず自分の負けた戦いのリプレイ動画を見る.任意のフレームで一時停止し,「検索」ボタンを押すことで.強いプレイヤーの対戦動画から 4 つの類似シーンが検索されるので,各シーンを再生してもらい,強いプレイヤーの挙動を観察してもらった.強いプレイヤーの対戦動画を見るとき,提案手法を使わないグループの被験者は,動画再生ソフト(Windows Media Player)で強いプレイヤーの対戦動画を自由に見るだけとした.どちらのグループも対戦動画を見て挙動観察をする時間は 5 分間とした.

各被験者の対戦は全て動画で記録するとともに,勝敗やゲーム内のスコアなども記録した.

5.2 実験結果

提案手法を使ったグループと使わなかったグループの対戦結果を図 6 と図 7 に示す.

	A	B	C	D	E	F	G	合計	平均
当初のレベル	3	6	8	5	8	4	9	43	6.142857
最後に負けたレベル	8	9	9	7	9	9	9	60	8.571429
勝った数	5	3	1	2	1	5	1	18	2.571429

図 6. 対戦結果:提案手法を使ったグループ

上図において,提案手法を使ったグループのメンバーは A さん〜G さんとし,提案手法を使わなかったグループのメンバーは H さん〜N さんとする.

「当初のレベル」とは,各被験者が最初に負けた CPU のレベルのことである.その平均値を見てみると,提案手法を使ったグループでは 6.14,提案手法を使わなかったグループでは 6.29 となっており,ほぼ同じ値となっている.つまり,片方のグループに初めから強い人または弱い人が集まった,ということはなかったことを示している.

	H	I	J	K	L	M	N	合計	平均
当初のレベル	6	8	5	7	9	6	3	44	6.285714
最後に負けたレベル	9	9	9	9	9	6	4	55	7.857143
勝った数	3	1	5	2	1	0	1	13	1.857143

図 7. 対戦結果:提案手法を使わなかったグループ

「勝った数」とは,「当初のレベル」が決まった後(つまり, CPU に 1 回負けた後),実験が終わるまでの間に CPU に勝った数を示している.即ち,ほとんどの被験者で「勝った数」は「最後に負けたレベル」から「当初のレベル」を引いた値になっている.ただし,被験者 G と L については,彼らが強いプレイヤーであって「当初のレベル」が上限の 9 であったため,異なる値となっている.

「勝った数」を比較すると,提案手法を使ったグループの平均値は 2.571,使わなかったグループの平均値は 1.857 となっており差がついた.これは,提案手法を使った場合の方が,使わなかった場合より,平均的に 2.571−1.857=0.714 回多く勝っていることを示している.言い方を変えれば, 0.714 分の高い CPU レベルに成長できたことを示している.

以上より,グループの間に当初の能力の差はなかったが,提案手法を使った場合では勝った数が多かったことより,提案手法の有用性が示せたといえる.

5.3 アンケート結果

上記の実験終了後に,各被験者に下記の 5 つの質問に 5 段階評価で答えてもらった(図 8).

「提案手法を使って,見たいシーンが検索できましたか?」という項目については,提案手法を使ったグループのみに答えてもらったが, 6 割程度の賛同が得られた.シーン検索の精度評価(4.2 章)においては約 8 割の精度で検索ができるという結果が得られていたにも関わらず,「見たいシーン」の検索精度となると 6 割程度の精度になってしまったということになる.この理由としては,キャラクターの位置関係が似ていても,再生してみると,その後の動画の展開が「見たいシーン」のものではなかった,という場合があるためである.今後は動画の展開も考慮した検索手法を模索していきたい.

「スマブラは面白いと感じましたか」という質問

について，提案手法を使ったグループの方がやや高い評価を得た．提案手法を使うことで，被験者にはやや複雑な検索操作を強いたが，それによってスマブラを面白いと感じなくなる，ということは無いことが分かった．

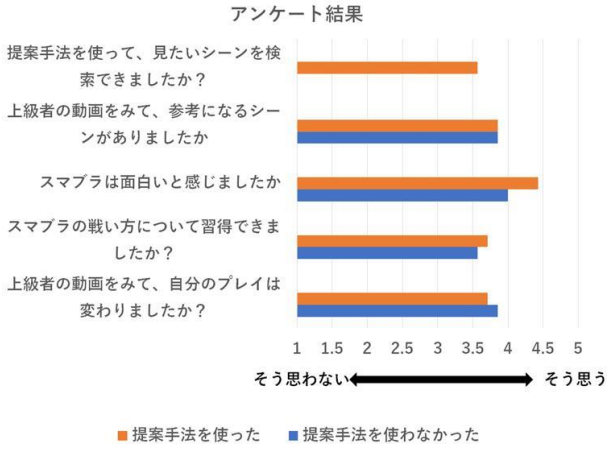


図 8．アンケート結果．棒グラフは平均点を示す．

提案手法を使ったグループに「システムの使いやすかった点，使いにくかった点，改善点」という項目で，文章を書いてもらうアンケートも実施した．「自分が見たい展開と検索されたシーンが異なっていたので展開まで指定したいと感じた．」「自分のみたいシーンが出てこないときがあった．」という意見から，画像レベルでの検索はできていても，ユーザの求める動画が出力できていないことを示すフィードバックが得られた．他には「どこで一時停止して『検索』ボタンを押せば参考になる動画を検索できるのかがわからないので教えてほしい．」や「動画で強いプレイヤーが入力しているコマンドを知りたい．」というような，より高度なサポートをシステムに希望するような指摘も頂いた．このような問題にも対応できるよう，今後も開発を進めたい．

6　今後の課題

今回のアンケートを受けて，初心者のプレイヤーに対しては操作方法の補助が有効であるように思われた．例えば，強いプレイヤーの動画に合わせて使っているコマンドを推定して表示するなどができれば，便利ではないかと思われる．また，対戦動画を見て挙動観察をするといっても，見る人のレベルによって，何を見たらいいのかがそもそも分からない，という問題があることが分かった．システム側からもっと積極的に「このシーンは真似する価値があるよ」というような提案ができるようになると面白いのではないかと思う．一方で，eSports に参加しているような上級プレイヤーに提案手法を使ってもらうようなユーザスタディも今後はやっていきたい．

今後は扱える検索対象を増やしていきたい。今回の実験は青のマリオと白のドンキーコングのみを対象に行ったが、スマブラでは他にも多種多様のキャラクターが存在し、更にそのカラーやスキンも変更可能である。それらも検索対象としたいため、より多くのデータセットを用意し、ニューラルネットワークを学習し直したい。検索対象を増やすと精度が低下するかどうかを調査したり、低下した場合にはニューラルネットワークのサイズを大きくしたり、状況毎に異なるニューラルネットワークを構築したりすることで対応していきたい。

参考文献

[1] 大乱闘スマッシュブラザーズ SPECIAL, https://www.smashbros.com/ja_JP/

[2] Shiv Ram Dubey, “A Decade Survey of Content Based Image Retrieval using Deep Learning”, IEEE Transactions on Circuits and Systems for Video Technology.

[3] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, “Neural codes for image retrieval,” in ECCV, 2014, pp. 584–599.

[4] Jerome Revaud, Jon Almazán, Rafael Sampaio de Rezende, César Roberto de Souza, “Learning with Average Precision: Training Image Retrieval with a Listwise Loss”, ICCV 2019.

[5] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, James Hays, “Composing Text and Image for Image Retrieval -An Empirical Odyssey”, CVPR 2019.

[6] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P. H. Torr, “Fastonline object tracking and segmentation: A unifying approach”, CVPR 2019.

[7] S. Wug Oh, J.-Y. Lee, N. Xu, and S. Joo Kim, “Video objectsegmentation using space-time memory networks”, ICCV 2019.

[8] Vlad Firoiu, William F. Whitney, Joshua B. Tenenbaum, “Beating the World's Best at Super Smash Bros. with Deep Reinforcement Learning”, arXiv:1702.06230 [cs.LG].

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”, CVPR 2016.

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan, “Supervised Contrastive Learning”, arXiv:2004.11362 [cs.LG].