7U-05

スポーツ放送映像におけるマルチモーダル行動認識 -画像特徴量と実況音声テキスト特徴量の統合-

大久保 深, 秦野 亮, 西山 裕之 †
東京理科大学創域理工学部経営システム工学科 †

1 はじめに

行動認識とは動画から人間の動作を分類するタスクのことであり, スポーツ映像においては, 選手のパフォーマンス推定や自動ハイライト生成の目的で使用される.

スポーツ映像における行動認識の従来手法は, 画像特徴量のみを使用するものが多い [1, 2]. しかし, テレビやインターネットで公開されるスポーツ放送映像には, 実況者による実況音声が存在する. このような実況音声は, プレイヤーの動作を説明するものであることから, 実況音声から得られるテキストも有効な特徴量になることが考えられる.

そこで, 本研究では, 特に野球放送映像の行動認識において, 実況音声テキスト特徴量が有効となるかを検証する. 具体的には, 画像特徴量のみを用いた従来手法と, 画像特徴量と実況音声テキスト特徴量を用いたマルチモーダルな提案手法を作成し, 両者の精度を比較する. また, 追加実験として, 学習データに含まれていない実況者による実況音声や, 野球ではなくサッカー放送映像にも本研究の手法を適用し, 実況音声テキスト特徴量の汎用性を検証する.

2 関連研究

Piergiovanni ら [1] は野球放送映像データセットである MLB-Youtube を作成し, 画像特徴量のみを用いて行動認識ベンチマークを作成した. Chen ら [2] は MLB-Youtube において画像特徴量のみを用いて球種分類を行った. 上記のように, 野球放送映像の行動認識においては画像特徴量のみを用いた研究が多く, 実況音声テキストの有効性を検証する研究は十分にされていない.

3 提案手法

3.1 提案手法の概要

提案手法の概要を図 1 に示す. MLB-Youtube は, 20 本の MLB ポストシーズン試合放送映像に対して, 1 球ごとの投球時間区間と, {ball, swing, strike, hit, foul, in play, bunt, hit by pitch} の 8 クラスのマルチラベル付けされたデータを含むデータセットである. 本研究では検証に十分なデータ数が存在する {ball, swing, strike, hit, foul, in play} のみを使用する. 各データ数の内訳を表 1 に示す.

図 1 提案手法の概要図

表 1 各クラス数と全データ数

クラス	データ数
ball	2093
swing	2633
strike	3659
hit	2032
foul	1101
in play	937
全データ数	5752

画像特徴量の作成には, Piergiovanni ら [1] のベンチマーク手法を参考に, Inflated 3D ConvNet (I3D) を用いた. 動画を I3D に入力し, 画像特徴量を作成する. 実況音声テキスト特徴量の作成には, whisper-v2 large と text-emmbedding-ada-002 を用いる. まず, whisper-v2 large を用いて音声認識を行い実況音声テキストを作成する. 次に, 実況音声テキストをプロンプトとして, text-emmbedding-ada-002 に入力し, 実況音声テキスト特徴量を作成する.

モデルの構造を図 3,4 に示す. 従来手法モデルは Piergiovanni ら [1] のベンチマーク手法を参考に作成する. 提案手法モデルは従来手法モデルを拡張する形で作成する.

3.2 音声取得期間の遅延

MLB-Youtube に含まれる時間区間の終了点はおおよそ捕手の捕球までであり, その時点では実況者が十分に状況説明できていない可能性が考えられた. そこで実況者の十分な状況説明の音声を取得するために, 図 2 の通り, 音声取得期間の遅延を 2 種類実施した.

4 結果・考察

既存手法と提案手法の精度の比較を表 2 に示す. 表から提案手法が従来手法を上回っており, 実況音声テキスト特徴量が有効な特徴量となっていることが明らかになった. また, 提案手法において, 音声取得期間を 0 から 6 秒に遅延させることで, 精度向上

Multimodal Action Recognition in Sports Broadcast Videos - Integrating Image and Commentary Text Features -

† Shin Okubo, Ryo Hatano, Hiroyuki Nishiyama, Tokyo University of Science

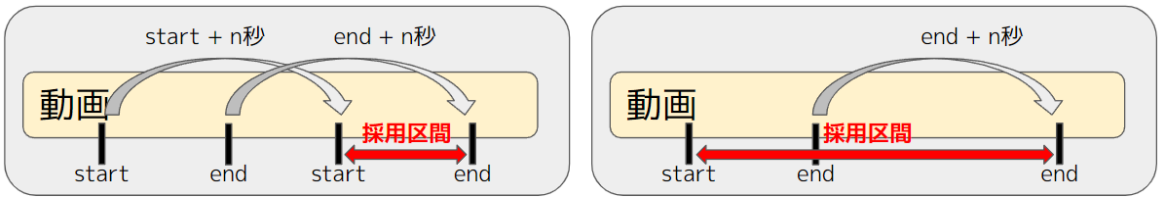

図 2　音声遅延パターン①（左）パターン②（右）概要図

表 2　各実験における提案手法の精度

	遅延秒数	音声不使用	0	1	2	3	4	5	6	7	8
本研究	既存手法	0.862									
	パターン①		0.866	0.874	0.883	0.888	0.891	0.873	**0.900**	0.887	0.876
	パターン②		0.866	0.871	0.880	0.885	0.881	0.884	0.885	0.879	0.879
追加実験 1	既存手法	0.812									
	パターン①		0.810	0.818	0.829	0.834	**0.839**	0.816	0.827	0.826	0.820
	パターン②		0.810	0.820	0.828	0.829	0.833	0.823	0.827	0.826	0.828
追加実験 2	既存手法	0.649									
	パターン①		0.648	0.651	0.649	0.662	**0.667**	0.660	0.653	0.659	0.655
	パターン②		0.648	0.655	0.655	0.657	0.656	0.665	0.657	0.656	0.654

本研究・追加実験 1 では mean Average Precision, 追加実験 2 では weight Average Precision を使用

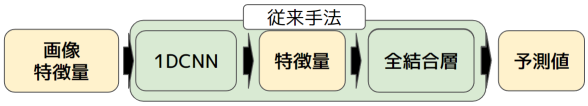

図 3　従来手法モデル構造図

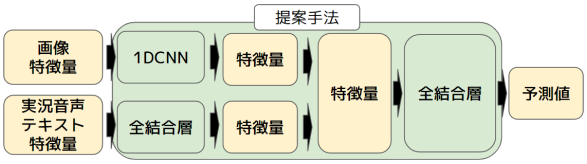

図 4　提案手法モデル構造図

の傾向がみられることから, 捕手の捕球後に行動認識において有効な実況音声テキストが存在すると考えられる.

5　追加実験

本研究では, 実況音声テキスト特徴量の汎用性を検証するため, 2 つの追加実験を行った.

5.1　学習データに含まれていない実況者による実況音声

実況音声テキスト特徴量には実況者ごとに異なる傾向が存在し, それが実況音声テキスト特徴量の汎用性低下を引き起こすことが考えられる. そこで本追加実験では, 学習データに含まれていない実況者による実況音声に対しても, 実況音声テキスト特徴量が有効に働くかを検証する.

本研究では, 学習・テストデータを試合動画毎に分割しており, テストデータの実況者が学習データにも含まれていたが, 本追加実験では, 実況者ごとに学習・テストデータを分割し, テストデータの実況者が学習データに含まれない状態で再度実験を行う.

既存手法と提案手法の精度の比較を表 2 に示す. 表から提案手法が従来手法を上回っており, 実況音声テキスト特徴量は, 学習データに含まれていない実況者による実況音声に対しても有効であると明らかになった.

5.2　サッカー放送映像

本研究では, 各プレー間に一時停止があり, その時間に実況者が十分な場面説明を行うので, スポーツの中でも特に実況音声テキストが有効な特徴量となると考え, 野球に着目した. そこで, 本追加実験ではサッカーのような一時停止の少ないスポーツに対しても, 実況音声テキスト特徴量が有効に働くかを検証する.

本追加実験ではサッカー放送映像データセットである SoccerNet で検証を行った.

既存手法と提案手法の精度の比較を表 2 に示す. 表から提案手法が従来手法を上回っており, サッカーのような一時停止の少ないスポーツに対しても, 実況音声テキスト特徴量が有効であることが明らかになった.

6　おわりに

本研究によって, スポーツ放送映像の行動認識において実況音声テキスト特徴量の有効性と汎用性の高さが明らかになった. 本提案手法を適用することで, より高精度な行動認識を行うことができ, 選手パフォーマンス評価や自動ハイライト生成の高精度化につながると考えられる.

参考文献

[1] Piergiovanni, AJ and Ryoo, Michael S. “Fine-Grained Activity Recognition in Baseball Videos“ 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 1821-18218. (2018), 10.1109/CVPRW.2018.00226

[2] Chen, R., Siegler, D., FaskoJr., M., Yang, S., Luo, X., Zhao, W. “Baseball Pitch Type Recognition Based on Broadcast Videos“ Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health, 328-344, (2019), https://doi.org/10.1007/978-981-15-1925-3_24