H-009

卓球動画における深層学習を用いたスローモーションシーン検出手法
A Method to Detect Slow Motion Scenes in Table Tennis Video Using Deep Learning Models

迫田　峻*　　大野　将樹**　　獅々堀　正幹**
Shun Sakoda　　Masaki Oono　　Masami Shishibori

1. はじめに

近年，インターネットによる動画配信サービスが普及している．映像およびネットワーク技術の発達により膨大の量の動画が視聴可能になっているが，全ての動画を視聴することは時間的に困難である．特にスポーツ番組を例に挙げると，野球やサッカー，バスケットボール，卓球などのスポーツでは，内容を理解するために 1 試合全てを視聴する必要があり，複数のスポーツ番組を視聴しようとすると多くの時間を費やしてしまう．そこで，スポーツ番組を興味があるシーンのみを取捨選択しながら視聴することで，動画視聴時間を短縮する技術が必要不可欠である．

本研究では，長時間の視聴が必要になるスポーツの中でも卓球に着目する．卓球のダイジェスト映像を生成するために，ラリーシーンや得点シーン，人物認識，リプレイ・スローシーンといった情報が必要となる．その中で，リプレイ・スローシーンの検出に焦点を当てて研究を行った．本稿では，各シーンの画像特徴量や時系列データを入力に，深層学習を用いて卓球の動画からスローシーンの検出する手法を提案する．本手法では，卓球の映像をフレーム単位で分割した後，画像の特徴量を取得するために，Convolutional Neural Network (CNN)[1][2]を用いる.次に，CNN で取得した画像特徴量を Long Short Term Memory(LSTM)[3]に入力し，各シーンがラリー等を行なっているプレイシーンかスローシーンか休憩中や観客等を写しているその他のシーンかを分類してスローシーンを検出する．

2. 関連研究

スポーツ映像のスローシーンを検出するための関連研究として，Vahid Kiani らの研究[4]がある.Vahid Kiani らは放送されているサッカー映像のスローモーションシーンを色や動きベクトルといった情報に着目して，シーンごとに SVM 分類器を用いて検出する手法を提案している.サッカーではスローシーンでも動きに関連した視覚的特徴が完全に異なっているシーンが存在するために，それぞれに対応したモデルを 1 つずつ作成して検出を行なっているため手間がかかる.本研究では，CNN を用いてシーンの画像特徴量を取得し，LSTM を用いて画像特徴量の流れを見ることでスローシーンを検出することで，スローシーンの中で複数のモデルを作成することなく検出する．

*徳島大学大学院先端技術科学教育部
Graduate School of Advanced Technology and Science Factory of Engineering,Tokushima University
**徳島大学大学院社会産業理工学研究部
Graduate School of Technology,Industrial and Social Science,Tokushima University

3. 提案手法

3.1 システムの概要

提案手法の流れを図 1 に示す．まず動画データをフレーム単位に分割する.得られたフレーム画像をプレイシーン，スローシーン，それ以外のシーンの 3 クラスで学習させた CNN に入力し，画像の特徴量を取得する．得られた画像特徴量を同じ 3 クラスで学習させた LSTM に入力し時系列を考慮しながらクラス分類を行い，スローシーンを検出する．

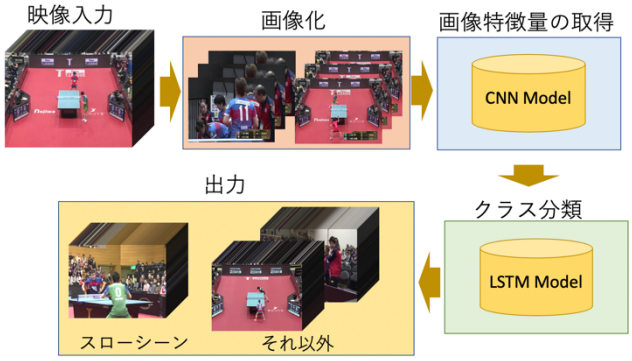


図 1:提案手法の流れ

3.2 画像特徴量の取得

映像をフレーム単位に分割し画像化したものから特徴量を取得する.まず特徴量を取得するためにクラス分類器を作成する.VGG-16[5]を Fine-tuning したモデルを用いて，プレイシーンとスローシーン，それ以外のシーンの 3 クラスで学習したモデルを作成する.学習に用いた画像データ数は表 1 に示す.そして作成したクラス分類器に映像から分割した画像を入力し，モデルの最終層のデータを取得する.そして取得したデータを LSTM の入力に用いる．

表 1:CNN の学習データ数

	分類クラス	学習数
学習データ	プレイ	700 枚
	スロー	700 枚
	その他	700 枚
バリデーションデータ	プレイ	200 枚
	スロー	200 枚
	その他	200 枚

3.3 クラス分類

クラス分類は CNN の最終層から取得した特徴量を時系列データとして入力し分類する.1 枚の画像から得られる特徴量の次元数は 256 次元で，分類クラスは CNN と同じプレイ，スロー，その他の 3 クラスで，学習に用いた動画は 3 クラスが含まれる約 40 秒の動画 5 本(総データ数 2000)である．

4. 実験

CNN のみでのクラス分類と CNN+LSTM でのクラス分類の 2 つの手法の結果を比較した.

4.1 実験データ

実験データには約 40 秒の卓球動画(T リーグより提供)を 3 本使用し，総画像数は 1197 枚(1 秒につき 10 枚のフレーム画像)である.動画内には，プレイシーンとスローシーン，その他シーンの 3 クラスが必ず含まれているものを使用した.

4.2 評価

システムの評価には再現率と適合率と F 値を用いる.再現率と適合率と F 値は以下の式で表す.

$$\text{再現率} = \frac{\textit{正しく検出できたスローシーン数}}{\textit{実際のスローシーン数}} \quad (1)$$

$$\text{適合率} = \frac{\textit{正しく検出できたスローシーン数}}{\textit{手法によって検出したスローシーン数}} \quad (2)$$

$$\text{F 値} = \frac{2 * \textit{再現率} * \textit{適合率}}{\textit{再現率} + \textit{適合率}} \quad (3)$$

4.3 結果

CNN のクラス分類の結果を表 2 に示す.また CNN+ LSTM のクラス分類の結果を表 3 に示す.

表 2:CNN 結果

	再現率(%)	適合率(%)	F 値
動画 A	100.00	100.00	100.00
動画 B	94.52	93.24	93.88
動画 C	81.42	94.85	87.62
平均	91.98	96.03	93.83

表 3:CNN+LSTM 結果

	再現率(%)	適合率(%)	F 値
動画 A	98.85	100.00	99.42
動画 B	94.52	93.24	93.88
動画 C	79.65	90.90	84.90
平均	91.01	94.71	92.73

また CNN+LSTM のクラス分類での各動画の変化を図 2, 図 3，図 4 に示す.

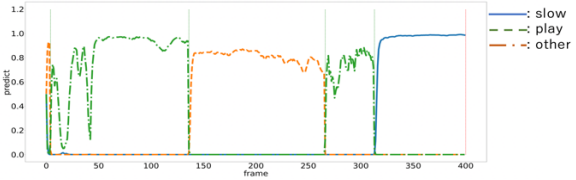


図 2:動画 A

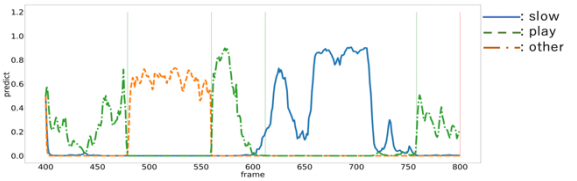


図 3:動画 B

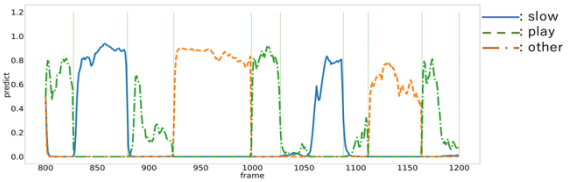
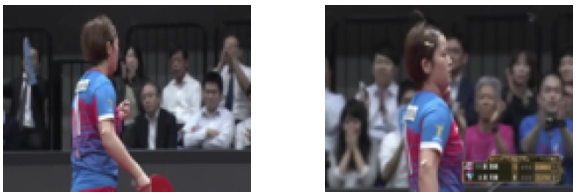


図 4:動画 C

4.4 考察

CNN のみと CNN+LSTM の結果を比べると CNN のみの方が再現率, 適合率ともに高かった.CNN では 1 フレームのみの情報を見ているため場面が切り替わってすぐに分類ができていたが, CNN+LSTM では前のフレームのデータを引き継いでいるために，場面が切り替わった直後に分類できていない部分が多かったために精度が下がった.CNN のみが精度が高くなったのは，今回使用した動画内にスローシーンと通常シーンとのカメラアングルが異なっていたことが原因と考えられる.一方 CNN のみは，図 5 のようにスローシーンとその他のシーンのカメラアングルが同じ時に分類精度が低くなっているが, CNN+LSTM では前後のフレームの情報を考慮しているために CNN で誤検出していたフレームを正しく検出している部分が見られた.



[a] スローシーン　[b] その他シーン

図 5:画像特徴が似ているシーン*

5. まとめと今後の課題

本論文では，深層学習を用いて卓球映像からスローモーションシーンの検出手法を提案した.CNN を用いて画像の特徴量を取得し，それを時系列データとして LSTM に入力し分類を行うことで前後のフレームを考慮しながら分類することを試みた.結果として CNN のみと CNN+LSTM では, CNN のみの方が精度は高かったが，前後のフレームを考慮することで CNN のみでは誤検出だった部分を正しく検出することができた部分が見られた.

今後の展望としては，LSTM を用いると前後のフレームを考慮してしまうことから，場面が切り替わった直後がうまく検出できないので LSTM に入力する前処理として動画をカットしてから入力することで精度向上を試みる.

7.参考文献

[1] 中山英樹,"深層畳み込みニューラルネットワークによる画像特徴抽出と転移学習",電子情報通信学会音声研究会 7 月研究会,pp.55-59,2015

[2] Karen Simonyan,Andrew Zisserman,"Verry Deep Convolutional Networks for Large-Scale Image Recognition",2014

[3] Sepp Hochreiter "LONG SHORT-TERM MEMORY" Technische Universitt Mnchen 1997

[4] Vahid Kiani,Hamid Reza Pourreza "An Effective Slow-Motion Detection Approach for Compressed Soccer Videos" ISRN Machine Vision,vol.2012 ,p8

[5] Karen Simonyan, Andrew Zisserman,"VERY DEEP CONVOLUTIONNAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION", arXiv:1409.1556v6 [cs.CV] 10 Apr 2015

*T リーグより提供