Vol.2018-AVM-101 No.10
2018/6/7

観客の挙動を用いた興奮・緊張度の推定にかかる一検討

松村 誠明†, 亀田 明男†, 磯貝 愛†, 能登 肇†, 木全 英明†

概要：スタジアムやアリーナではスポーツを観戦する観客に対し，より試合に熱中いただくため，会場では様々な演出が行われる．これら演出は，熱中度低い観客に対しては楽しみ方の気付きを与えるような演出が望ましく，熱中度が高い観客に対してはより会場全体と一体になれるような演出が望ましい．本研究では，観客の熱中状態を判断する材料として興奮・緊張状態に着目し，観客の挙動を波と捉え，周波数変換したデータを基に興奮度や緊張度を推定する技術を構築したので報告する．

A Study for Prediction of Heating and Strain using Audience Behavior

Masaaki Matsumura†, Akio, Kameda†, Megumi Isogai†,
Hajime Noto†, Hideaki Kimata†

1. はじめに

近年スタジアムやアリーナなど競技場の ICT 化が進み，チームや選手に対してプレイ内容に関する統計情報(スタッツデータ)をフィードバックするだけでなく，観客に対してもスポーツ観戦を行う際の満足感や高い理解を促すため，的確に情報を提示するサービスの検討が始まっている．また，ライブやコンサートなどベントの質を高めるため，AI を用いて観客の状態を推定する実証実験[1]も報告されており，今後は解析結果を活かしたサービス品質の向上が期待されている．

中でもスタジアムやアリーナ・ライブやコンサートなど，現地に赴きコンテンツを楽しむ際に得られる満足感の一つは会場全体との一体感[2]であり，卓越したプレイや会場の演出に対して，観客は興奮や緊張・喜びや悲しみ等の感情を一体となって共有できることに一定の価値を見出していることがわかる．そのため，観客の挙動には規則性が現れることが予想され，この規則性を用いて会場の状態を推定できるのではないかと仮説を立てる．

本研究では，上記仮説検証のための一検討として，観客の挙動から興奮度および緊張度を学習・推定する手法を提案する．作成した試験データに対して提案手法を用いた実験を行い，考察をまとめたので報告する．

2. 関連研究・課題抽出

音声から状態を推定する技術として，チョコパラ SSS[3]ではシーケンス内音声の振幅レベルに着目し，盛り上がるシーンを検出する機能を具備している．

興奮状態になると観客は歓声が大きくなる特徴が顕著に現れるため，音声による盛り上がり検知は安定して高いレベルで推測できる．しかし，マイクは周囲の音声も広くサンプリングしてしまうため，チーム別の応援エリアなど，領域毎に特徴を解析したい場合にはノイズが多くなる．そのため，観客の挙動を高精度に解析するためには音声に加えて映像も併用するほうが望ましいと考えられる．

画像から状態を推定する技術として，Microsoft Kinect[4]等に代表される顔の表情の変移を取得できる家庭用デバイスを用いて，表情から個人の感情推定を行う研究[5]も報告されている．また，前述の実証実験[1]の通り avex 社と Microsoft 社共同でライブ会場に設置した観客席向きのカメラから顔認識を行い，複数名の表情から会場の状態を推定する試みも行われている．

近年の顔認識精度は CNN(Convolutional Neural Network)ベースの AI 技術(Deep Learning)の発展により飛躍的に精度が向上しているものの，ある程度正面顔で撮影されている必要がある．実際の観客は試合展開に応じて見る方向を変えるため，多数の観客に対してロバストに表情を検出するには相応のカメラ台数が必要となり，システムを導入する会場の負担が増加する恐れがある．また，観客の表情が読み解けるほど近距離でユーザを撮影する場合，利用用途によっては肖像権の問題が発生する恐れがあるだけでなく，観客に不快感を与えてしまう可能性もあるため，これらはセンシティブな課題として取扱う必要がある．

3. アプローチ

本研究では，誰が映っているか判別できない程の遠方から俯瞰視点で撮影された客席映像のみを用いて，観客の興奮度および緊張度を学習・推定する．

観客の挙動は，表 1 に記すような特徴があると考えられ，遠方から撮影した観客の動きを周波数変換すると表内の右列に記載のような特徴が現れると思われる．そこで，本研究では入力映像に対して時系列を考慮した 3 次元フーリエ変換したデータを Deep Learning により学習させることで興奮度および緊張度の推定可否を検証する．

† 日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Labs. NTT Corp.

表 1　観客の挙動の特徴

状況	映像で見た観客の動き	周波数で見た観客の動き
平時	応援団のコールや音楽に合わせてリズミカルに応援 もしくは個別の雑多な動き	統一感のある低周波・低～中振幅の動き もしくは統一感のない低周波・低振幅の動き
興奮時	立ち上がり，プレイに対して激しく拍手をするなどして応援	統一感のない高周波・高振幅の動き
落胆時	動的な興奮状態から静的な動きへの移行	高周波・高振幅から低周波・低振幅への移行
緊張時	体を強張らせるなど，瞬間的な静止 (緊張が開放された後，顕著な興奮/落胆が現れる)	瞬間的な低周波・低振幅(後，高周波への移行もしくは低周波の継続)

表 2　ラベリングの評価基準

No.	プレイ内容	興奮/落胆度	緊張度	観客の動き
1	3 ポイント・シュートで点を取る	0.9 ~ 1.0	0.0 ~ 0.2	席から立ち上がるなどの大きな喜びが見られる
2	ファインプレーで点を取る			
3	2 ポイント・シュートで点を取る	0.6 ~ 0.8		激しく応援グッズを振る 手を上げて万歳する
4	フリースローで点を取る			
5	相手ファールやスティールによりオフェンスを止める		0.3 ~ 0.5	
6	ジャンプボールを取る			
7	審判がボールを持ち上げ，開始を告げる			
8	シュートを外すもファールを奪う(フリースローの機会を得る)			
9	シュートを放つ	0.6	0.6 ~ 1.0	シュートに集中し応援の動作が小さくなる
10	平常時(ドリブルやパス回ししている状態)	0.3 ~ 0.5	0.3 ~ 0.5	応援グッズをリズミカルに振っている
11	フリースローを放つ前	0.0 ~ 0.2	0.6 ~ 1.0	シュートに集中し応援の動作がなくなる
12	味方ファールやスティールによりオフェンスを止められる	-0.8 ~ -0.6	0.3 ~ 0.5	応援の動作がなくなる
13	ミスプレーにより，オフェンスが終わる			
14	2 ポイント・シュートで点を取られる		0.0 ~ 0.2	
15	フリースローで点を取られる			
16	ファインプレーで点を取られる	-1.0 ~ -0.9		口を大きく開ける，天井を仰ぐなどの明らかに落胆した様子が動作に現れる
17	3 ポイント・シュートで点を取られる			
18	大きくシュートを外す			
19	フリーでシュートを外す			

4. データ作成およびネットワーク構築

4.1 ラベル作成

本研究で対象とする入力データおよび正解データは一般公開されているデータが存在しないため，アリーナにて行われたバスケットボールの試合における客席の挙動を撮影した映像を用い，映像の試合展開に応じて興奮度・緊張度のラベル付与を行ったものを正解データとして作成した．なお，興奮度・緊張度は個別パラメータとして分解し，興奮/落胆度を[-1.0～1.0]，緊張度を[0.0～1.0]としてラベリングを行う(双方とも平時は 0.0 とする)．

ラベルの評価は表 2 を基準として設定し，客席の挙動を目視確認しながら適宜設定を行った値を 3 人で精査・検証・補正して当該フレームのラベルとして設定した．

試合全体の中でも興奮/落胆度・緊張度に大きな変化が見られた計 50 シーン(1 シーン約 10 秒)において，各シーン複数個所ラベリングした値を，スプラインで滑らかに補間したものを正解データとした(図 1 参照)．なお，入力データ(周波数データ)に対応する正解データは時間軸 64 フレームの平均値を適用した．

4.2 入力データ作成

映像から取得するデータサイズは解像度 : 64x64，フレーム数 : 64(29.97fps で撮影した映像場合は約 2 秒)とした．Deep Learning—に学習・推定させる入力データ(64x64x64x2(末尾の要素は sin, cos))は解像度 : 64x64 の範囲内に着席時 4～6 人の観客の全身が映るようリサイズし，同一のシーン内から複数個所を選出した 64x64x64 の各ブロックに対して以下の手順で作成した．

1. グレースケール変換
2. 信号の正規化[0.0～1.0]
3. 3 次元フーリエ変換(cuFFT[6]を使用)
4. sin/cos 成分別の分散正規化

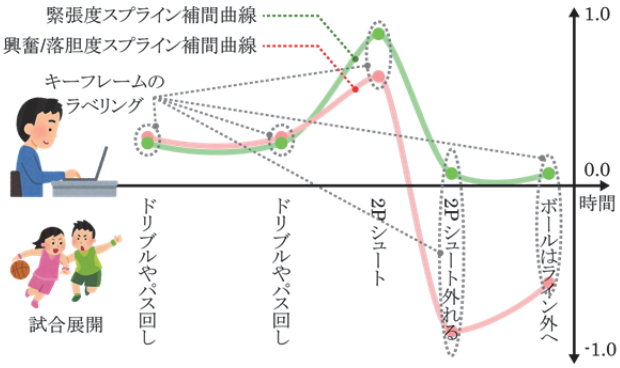

図 1: 正解データの生成

4.3 ネットワーク構築

Deep Learning を用いて動画像解析を行う際，CNN が広く利用されており，本研究においても CNN をベースとして図 2 のようにネットワークを構築する．

基本は Conv3D と BatchNormalization, MaxPooling の組合せで構成し，途中から分岐して興奮/落胆度・緊張度それぞれに対して解析するネットワークを設けた後，結合した 2 値を出力する構成をとる．過学習を防ぐ Dropout は出力層に近い MaxPooling 後(計 3 箇所)に挟み，設定値は全て 0.2 とした．

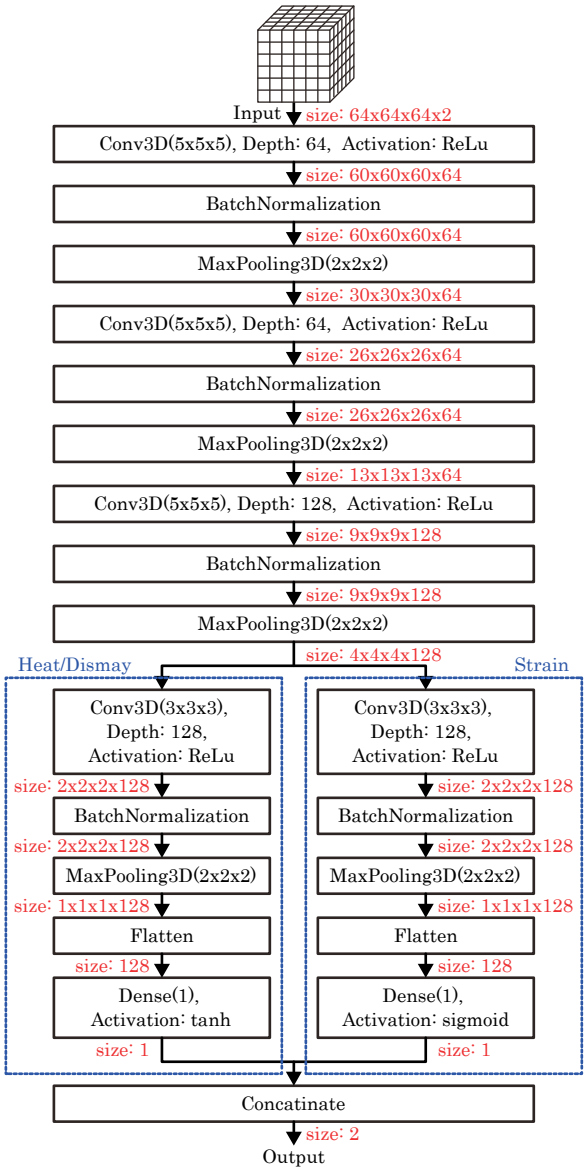

図 2: CNN の構造

5. 実験・考察

5.1 試験環境

試験環境および利用したフレームワークを表 3 に記す．なお，オプティマイザには学習率を 1.0e-5 に設定した Adam，誤差評価には平均絶対誤差，バッチサイズ 16 にて学習を行った．

表 3　試験環境

Index	Specification/Version
CPU	Intel Core i5-3750K (O.C. 4.2GHz)
Memory	DDR3-1600 32GB
GPU	NVIDIA GeForce GTX 1070 (8GB)
Framework	Anaconda 3 Python 3.6 keras 2.0.9 tensorflow-gpu 1.4 CUDA 8.0 (cuDNN 6.0)

5.2 試験結果

前述の入力データと対になる正解データを約 6000 サンプル生成し，中からランダムに 4096 サンプルを学習用に，1024 サンプルを検証用に抽出したデータセットを用いて学習を行った過程の loss の変遷グラフを図 3 に示す．

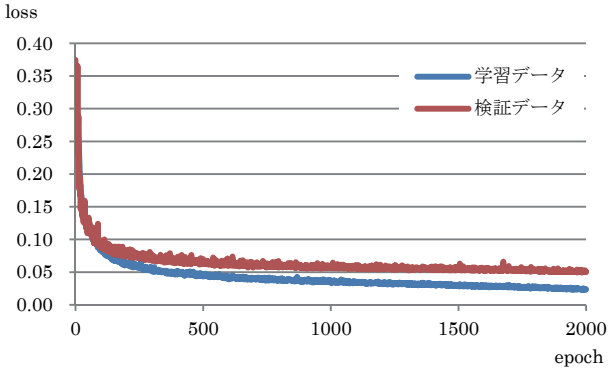

図 3: 学習過程における loss の変遷

loss は約 250 エポックにかけて大きく減少し，その後は緩やかに降下を続け，2000 エポックにて検証データで平均絶対値差 0.05 を下回る結果が得られた．

検証データについて興奮/落胆度・緊張度を個別に解析したところ，表 4 に示す結果が得られた．

表 4　検証データに対する
興奮/落胆度・緊張度別の解析結果

項目	平均絶対誤差	平均二乗誤差
興奮/落胆度	0.05626	0.04092
緊張度	0.04350	0.01060

興奮/落胆度は緊張度と比較して 2 倍のレンジがあるにも関わらず，緊張度とほぼ同等の平均絶対誤差を示しているため，興奮/落胆度の方が推定精度は高いと考えられる．しかし平均二乗誤差は大きな値を提示していることから，部分的に大きな誤差を生じていることが考えられる．

5.3 シーンへの適用結果

学習データ・検証データに含めなかった興奮/落胆度・緊張度の変動が大きい 2 つのシーン A, B に対し，連続したフレームで推定した解析結果を表 5 に，推定値のグラフを図 4～図 7 にそれぞれ示す．なお，各グラフはそれぞれ以下のデータを表す．

Original :　正解データ
Rect_0～4 :　同時間の異なる客席における興奮/落胆度・緊張度の推定結果
Average :　Rect_0～4 の平均

表 5　各シーンに対する
興奮/落胆度・緊張度別の解析結果

シーン	項目	平均絶対誤差	平均二乗誤差
A	興奮/落胆度	0.08171	0.01027
	緊張度	0.26072	0.10102
B	興奮/落胆度	0.14519	0.03093
	緊張度	0.12687	0.02217

興奮度/落胆度共に変異が大きな部分を上手く推定できていないことがわかる．表 4 と表 5 を比較すると，全てに共通して平均絶対誤差が大幅に増加していることが

分かる．これは学習データ・検証データ共に正解データの値域に偏りが大きかった可能性が挙げられる．

改善方法としては学習データ・検証データに対してランダムに選出するだけでなく，得られる正解データの値域(例えば興奮度：0.6～0.7)毎に一定数量のサンプルを用意して偏りを解消することで改善できると考えられる．また，誤差評価を平均二乗誤差に変更し，差分に対して大きなペナルティーを与えることで，変異が大きな部分の推定精度を向上させることが出来ると考えられる．

ネットワーク改善によって推定精度を向上させるにはCNN の深化・フィルタ数追加が考えられるが，3 次元の畳込みは膨大なメモリを消費するため，特に GPU を用いて計算する場合は慎重に設計する必要がある．

なお，推定のために要した計算時間は表 3 に記載の試験環境にて 1 サンプルあたり 16.7 msec(データ読込み等を除く)であり，4.2 節にて述べた入力データ作成も高速に動作するため，十分な推定精度が確保できれば会場内での演出用途などにも使用できると考えられる．

6. まとめ

本研究では，競技場にてスポーツ観戦する観客の興奮/落胆度・緊張度の予測を行うため，遠方から撮影された観客席の挙動を周波数変換し，観客の表情・音声に依存しない推定方法を考案・検証した．

結果として Deep Learning による学習の可能性は検証できたものの，実際のシーンに適用する場合には入力データ含め見直しが必要と考えられる．

様々なスポーツへの水平展開については見当の余地はあるものの，スポーツに応じて観客の挙動が大きく変わることが分かっているため，十分な検討が必要と思われる．バスケットボールの場合，攻守が激しく切り替わり，得点も短時間に大きく変化するが，例えばサッカーの場合は 1 点が勝敗を分けるケースも多く，1 点に対する重みがバスケットボールとは異なるため，チャンスに対する興奮度・緊張度の触れ幅が極めて大きい．野球では攻守がターン制のため，バスケットボールとは異なる観客の挙動が現れることが予想されるなど，バスケットボールの学習モデルを直接汎用化することは困難であると考える．

この問題に対しては一般的に採用されている手法と同様に，様々なスポーツのデータを基に汎用学習モデルを生成しておき，各スポーツに対してファインチューニングを適用することで，解決できると考える．

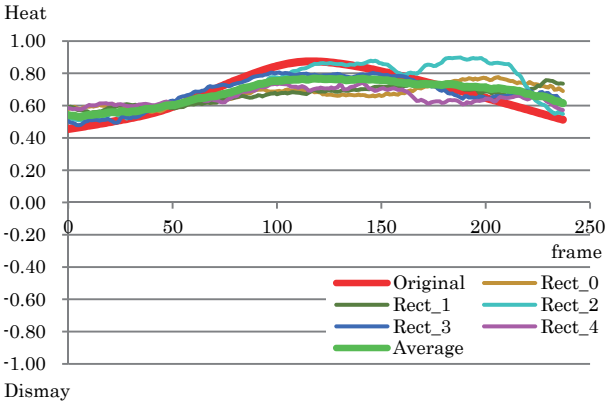

図 4: シーン A における興奮/落胆度の推定結果

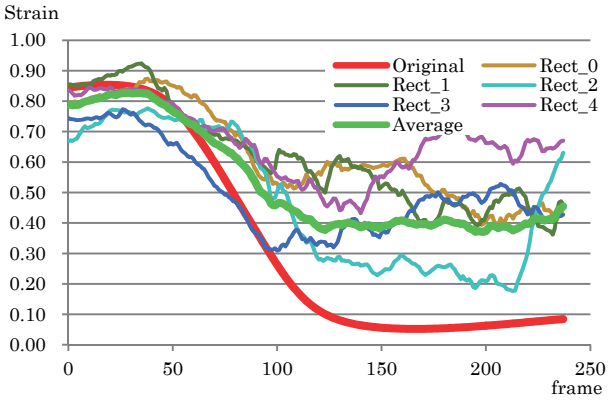

図 5: シーン A における緊張度の推定結果

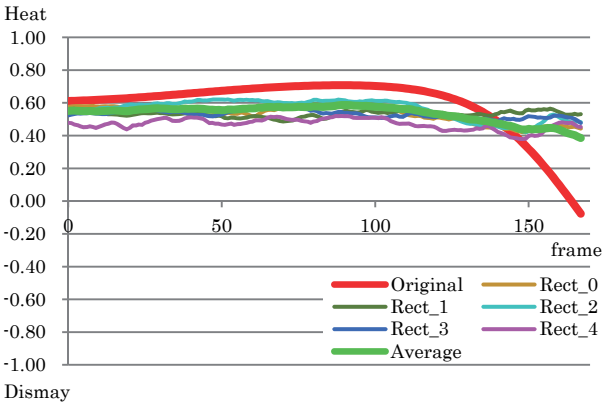

図 6: シーン B における興奮/落胆度の推定結果

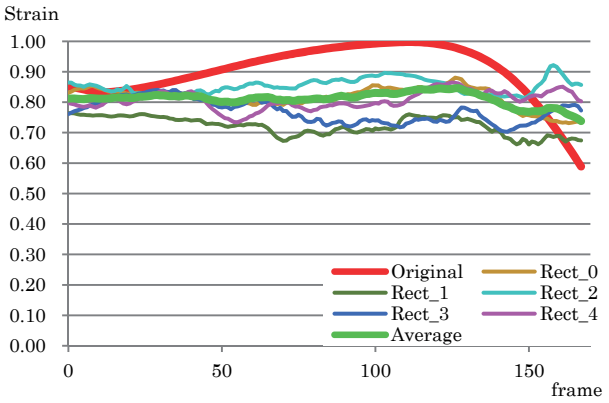

図 7: シーン B における緊張度の推定結果

参考文献

[1] より満足度の高いライヴ イベントの実現に向け、AI を活用した来場者分析システムを開発 (https://news.microsoft.com/ja-jp/2017/09/01/170901-avex-microsoft-faceapi/)
[2] 中塚 千恵, 小川 孔輔. なぜ,スタジアムに行ってしまうのか?--観戦型サービスにおける参加意図形成と顧客経験. マーケティングジャーナル vol.28 pp.43-62.
[3] 藤川 勝, 宮下 直也, 日高 浩太, 佐藤 隆, 下倉 健一朗. チョコパラ SSS：シーンに対する議論の盛り上がりを配信する映像コミュニティシステム. 画像電子学会第 34 会年次大会 06-31, pp.67-68 2006.
[4] Microsoft, USA. XBOX Kinect (http://www.xbox.com/Kinect)
[5] 高橋 正樹, 奥田 誠, Simon Clippingdale, 山内 結子, 苗村 昌秀. 表情強度と単純ベイズ推定を融合した顔表情認識. J-028, 第 12 回情報技術フォーラム FIT2013.
[6] cuFFT | NVIDIA Developer (https://developer.nvidia.com/cufft)