時空間情報を考慮した 動画からの複数動作の同時検出手法

杉山　裕哉† 　青野　雅樹†

† 豊橋技術科学大学工学科情報・知能工学　〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1 － 1
E-mail: †y-sugiyama@kde.cs.tut.ac.jp, ††aono@tut.jp

あらまし　バスケットボールにおけるシュート、ジャンプ、ブロックに代表されるなど複数の同時動作に対するマルチラベル問題としてのでの動作検出は、コンピュータビジョンにおいて最近着目されている重要な問題である。しかしながら、同時動作の正確な推定は難しい問題である。本研究では、高精度な動作検出法の開発を目的とする。従来手法では、RGB 画像からの空間情報とオプティカルフローからの時間情報を独立に学習し、同時動作検出法を与えていたが、我々は、これら二つの情報を併せ持つ時空間情報を用いたモデルを提案し加えることで、時間構造に強く、物体の判別に強いモデルを提案する。従来手法と提案手法の比較実験を行い、提案手法の有効性を確認した。本提案研究での典型的なアプリケーションとしては、防犯カメラでのセキュリティチェック、病院での患者のモニタリング、家庭での高齢者やペットなどのモニタリングなどがあげられる。

キーワード　行動検出、深層学習、オプティカルフロー

1 はじめに

動作検出は、防犯カメラでのセキュリティチェック、患者または高齢者のモニタリング、オンラインビデオ検索、ロボットの知覚など多くの社会的アプリケーションにおいて重要な要素であり、コンピュータビジョンにおいても重要な問題である。このタスクは連続的な動画を入力として与え、発生する全てのアクションに対応するフレームを検出することを目的としたタスクである。これはあらかじめトリムされた動画内の単一のアクションを分類する動画分類問題と比較して困難なタスクである。現実世界のビデオには図 1 のように複数のアクションが含まれているが、それらを動作が起きている時間毎にラベル化されているデータセットが少ないため、特にマルチアクションな動画を対象にしたものに関しては研究が進んでいない。

これまで畳み込みニューラルネットワーク (CNN) を使用するエンドツーエンドの学習方法は、動画解析において大いに貢献してきた。これらのアプローチは単一の RGB フレームや少数のフレームにわたるオプティカルフロー [1] など、動画のフレーム毎の情報を正常にモデル化している。最近では、I3D(Two-Stream Inflated 3D ConvNet) [2] のようなモデルが、より長期なアクションを補足するために開発されている。またこれに加え、それぞれのアクションには相関があるとして、それらを時間的構造から把握するためにフィルタを用いて学習を補助する Super-event 表現 [3] と呼ばれる手法も存在する。しかし、これらを用いても高い精度は得られていない。

本研究では、さらなる高精度な動作検出の手法を検討することを目的とする。動作の検出には RGB 画像から空間的特徴を抽出し、オプティカルフローから時間的特徴を抽出して学習を行う手法がある。我々はその 2 つの特徴を組み合わせることで

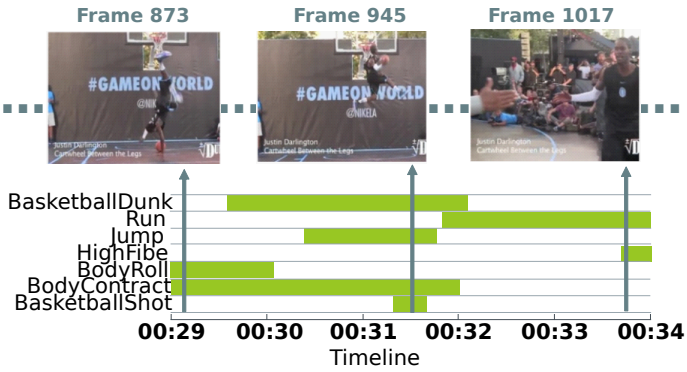

図 1: 動画とアクションの関係の例

時間構造と物体認識の両方に強いネットワークを構築することができると考える。すなわち時空間を考慮したネットワークを提案する。提案手法に対して実装、実験を行い、従来手法との比較をすることで、提案手法の有効性を確認する。2 節で行動検出に関する論文や、データセットに関する論文について紹介し、3 節で提案手法の構造を説明する。5 節でデータセットに関する説明をし、6 節で実験とその考察を行う。最後に 7 節でまとめと今後の課題について述べる。

2 関連研究

行動認識は、コンピュータビジョンにおいて一般的な研究課題である [4]。高密度な軌道特徴を用いたハンドクラフトな特徴量は過去に多くのベンチマークデータセットにおいて優れた結果を出している [5]。また近年では畳み込みモデルを用いた行動認識の学習機能に関する研究が増えている。そのひとつに Two-Stream ConvNet [6] と呼ばれるアプローチが存在する。これは動画の動きと画像の特徴をとらえるために、RGB 画

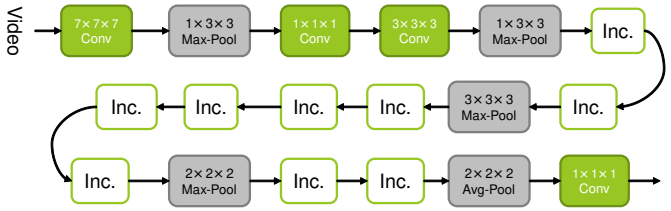

図 2: I3D の概要図

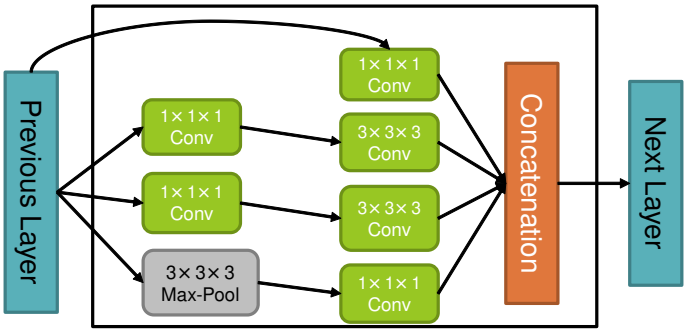

図 3: Inception Module

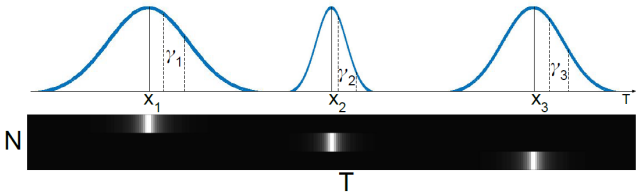

図 4: Temporal Structure filters の例 [3]

像とオプティカルフローを入力として使用する。またオプティカルフローにも動作の動きの差分を計算して生成する手法と、CNN を用いて最適なオプティカルフローを生成する手法がある (例　PWC-Net [7],FlowNet [8])。この CNN を用いたオプティカルフロー生成と動作検出のネットワークを用いて、動作検出に適したオプティカルフローを生成し動作検出の精度を上げる手法も研究されている。その他にも三次元 XYT(spatio-temporal) の畳み込みフィルタを学習する手法も多くの行動認識タスクに適用されている [9]。行動認識のデータセットとしては、THUMOS [10],ActivityNet [11],Charades [12] のような大規模なデータセットが存在する。

3　従 来 手 法

この節では本研究において従来手法の内、ベースとした手法について述べる。具体的には I3D(Two-Stream Inflated 3D ConvNet) [2] と、Temporal Structure filters の手法を利用した TGM(Temporal Gaussian Mixture Layer) [13] について述べる。

3.1　I3D

I3D とは Two-Stream Inflated 3D ConvNet のことで前述した Two-Stream ConvNet を応用した手法である。概要図を図 2 に示す。先ほども述べたように Two-Stream ConvNet は RGB 画像とオプティカルフローを入力として使用する。RGB 画像によりフレーム内の物体の位置 (空間要素) を学習し、オプティカルフローによりフレーム間の動き (時間要素) を学習する。通常の Two-Stream ConvNet はネットワーク部分に 2D CNN を用いている。しかし動画には空間要素と時間要素が含まれるため、3D CNN を用いた方が良い精度が得られると考えるのが自然である。この手法は、その 3D CNN に画像分類等に大きく貢献している ImageNet などの大規模なデータセットによる事前学習のパラメータを用いることができるようにするために、2D ConvNet を 3D ConvNet に膨張させた (Inflated) 手法である。膨張させる方法として入力の画像を複数枚重ねることにより、静止動画として扱うことで暗黙的に ImageNet 上で事前学習をすることを可能としている。これにより、ImageNet により事前学習されたパラメータを用いながら学習することができる。3D ConvNet のベースとして Inception Module(図 3) を用いた Inception-v1 をベースにしている。

3.2　Temporal Structure filters

Temporal Structure filters とは複数のアクションによって形成される時間的関係性を補足するように設計されたフィルタで、[14] で提案されたモデルの拡張/一般化したものである。これは「学習された」フレーム位置にのみ焦点を当て、固定長のベクトルを用いて可変長の動画を表現することである。各時間構造フィルタは、N 個のコーシー分布の集合としてモデル化される。これはガウス分布より訓練が容易であったため、用いている。各分布では幅を制御する中心 x_n 及び γ_n を学習する。動画の長さを T とすると、各フィルタは次のように構成される。

$$\hat{x_n} = \frac{(T-1)\cdot(\tanh(x_n+1))}{2} \tag{1}$$

$$\hat{\gamma_n} = \exp(1-2\cdot|\tanh(\gamma_n)|) \tag{2}$$

$$F[t,n] = \frac{1}{Z_n \pi \hat{\gamma_n}\left(\frac{(t-\hat{x_n})}{\hat{\gamma_n}}\right)^2} \tag{3}$$

ここで、Z_n は正規化定数であり、$t \in \{1,2.....,T\}$ と $n \in \{1,2.....,N\}$ である。図 4 に例を示す。これを訓練することで、モデルが動画全体においてどの時間間隔がフレーム毎の行動検出に関連するかを知ることができる。

3.3　TGM

Temporal Gaussian Mixture Layer（TGM） とは前述した Temporal Structure filters のようにガウス分布を用いて作成された畳み込みレイヤである (図 5)。ガウス分布によって動作の時間的間隔を学習し、「どの時間を見るべきか」を決定づける。この時に使用されたガウス分布を用いて、畳み込みカーネルを構築する。これにより各 Action クラスについて線形な時間構造をとらえることができる。$C_{in} \times D \times T$ 次元のベクトルを入力とする。この時 C_{in} は入力チャネル、D は各フレー

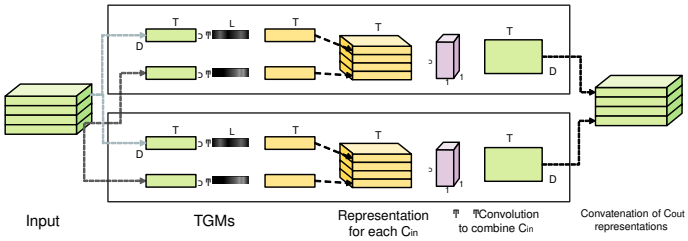

図 5: Temporal Gaussian Mixture Layer [13]

ム毎の CNN から抽出される特徴量の次元数 (本研究では 1024 次元)、T は動画の総時間である。TGM 層はそれを C_{out} 数の $1 \times L$ 次元に畳み込み、出力として $C_{out} \times D \times T$ 次元の特徴量を出力する。この時の L は Temporal Gaussian Mixture の長さである。各ガウス分布には中心 $\hat{\mu}$ と幅 $\hat{\sigma}$ の 2 つのパラメータがある。これらは以下の式で表される。

$$\mu = (L-1)\frac{\tanh(\hat{\mu})+1}{2}, \sigma^2 = \exp(\hat{\sigma}) \tag{4}$$

上記の μ と σ を用いて Temporal Gaussian Mixture を構築する。時間的ガウス分布畳み込みカーネルを以下の式のように構築する。

$$\hat{K_{m,l}} = \frac{1}{Z}\exp -\frac{(l-\mu_m)^2}{2{\sigma_m}^2} \tag{5}$$

この時 Z は $\sum_l^L \hat{K_{m,l}} = 1$ の正規化定数である。$\hat{K}$ は $M \times L$ 行列になる。各クラス毎に別々のガウス分布のセットを学習させる代わりに、クラス間で複数のガウス分布を共有することを維持するようなアプローチをとり、出力チャネル $i, \omega \in R^{C_{out} \times M}$ 当たりの一組の soft-attention の重みを学習する。その式を以下に示す。

$$a_{i,m} = \frac{\exp \omega_{i,m}}{\sum_j \exp \omega_{i,j}} \tag{6}$$

softmax 関数を用いて各入力チャネルの重みの合計を 1 にすることで、soft-attention の重みを作成する。最終的に時間的ガウス分布と soft-attention に基づいて Temporal Gaussian Mixture は以下のように計算される。

$$K_i = \sum_m a_{i,m}\hat{K}_i \tag{7}$$

この畳み込みカーネルを複数パターン用意し、チャネル方向に拡張、さらにそれを畳み込むことにより、より複雑で非線形な時間構造をとらえることを可能としている。またこのレイヤーを複数用いることで、抽象化レベルで動作をとらえることができる。本研究では先行研究と条件をそろえるためガウス分布ではなくガウス分布に類似しており、学習の収束が早いコーシー分布を用いる。実際生成される Temporal Gaussian Mixture のイメージを図 6 に示す。

4 提 案 手 法

前述の手法を用いて、時空間情報を考慮した新たなネットワークを構築する。我々の提案手法は主に空間情報と時間情報に加えることを提案とする。ネットワーク構成としては 3 つ提案する。それぞれを提案モデル 1、提案モデル 2、提案モデル 3 として以降説明する。一つ目は、空間情報を直接時間情報に加えた構成、二つ目は、取り出した特徴量を足し合わせた、時空間特徴量を入力としたネットワーク追加した構成、三つ目は二つ目のネットワークにおいて各 TGM ごとに特徴量を追加する構成である。

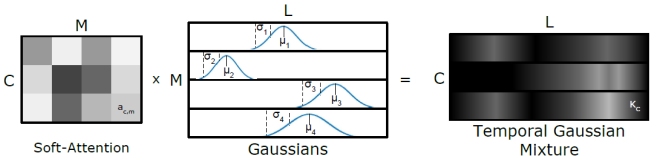

図 6: Temporal Gaussian Mixture

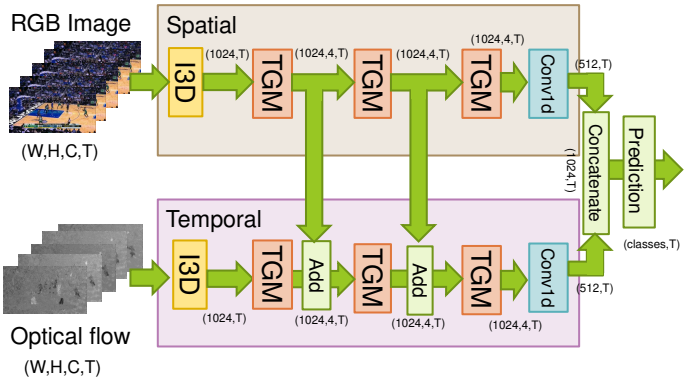

図 7: 提案モデル 1

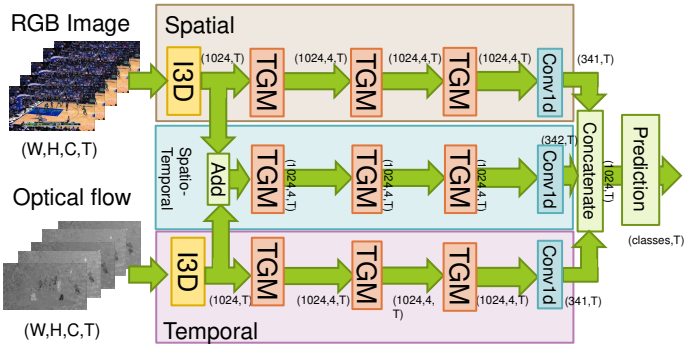

図 8: 提案モデル 2

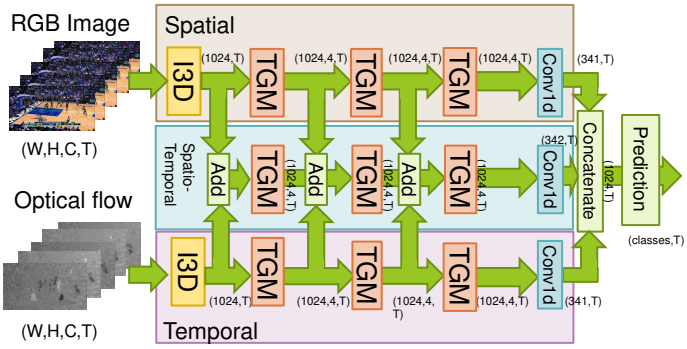

図 9: 提案モデル 3

4.1 提案モデル 1

提案モデル 1 は空間情報を直接時間情報に加えるようにしたモデルである。モデル図を図 7 に示す。I3D でそれぞれの特徴量を抽出する。そのあと各 TGM レイヤー一つ目と二つ目の TGM レイヤーをそれぞれ通した時、空間モデルの特徴量を時間モデルの特徴量に加える。オプティカルフローでは動作の動きを大まかなパーツのようにしてとらえるため、物体を投げる動作などの際に物体が何であるか (球体なのか、円盤なのかなど) をうまく判別できない場合がある。それを判別できるようにするために、RGB 画像の情報を加えることで物体情報を強調する。そしてそれぞれのネットワークから出力された特徴を連結して 1024 次元のベクトルにした後予測を行う。

4.2 提案モデル 2

提案モデル 2 は空間情報と時間情報を合わせた時空間情報を新たな入力とした時空間モデルを第 3 のネットワークとして追加したモデルである。モデル図を図 8 に示す。I3D から抽出した空間的特徴量と時間的特徴量を足し合わせる。その後、Spatial モデル、Temporal モデルと同様に 3 つの TGM レイヤーを通した後、各特徴量を連結させて推定する。これは空間、時間に特化した特徴量を合わせることで、物体に識別と時間軸に強い時空間を考慮した特徴量得られると考え構築した。この時空間情報を考慮したネットワークを Temporal-Spatial モデルとする。提案モデル 2 も提案モデル 1 と同様に各ネットワークから出力される特徴量を連結し 1024 次元のベクトルにして予測を行う。

4.3 提案モデル 3

提案モデル 3 は提案モデル 2 を拡張したモデルである。モデル図を図 9 に示す。提案モデル 2 では TGM レイヤーを 3 つ通すだけであったが、提案モデル 3 では各 TGM レイヤーを通すたびに、空間モデルと時間モデルから特徴量を取り出し、時空間モデルに加える。I3D による抽出直後では動作の関係性が学習されてない状態であるため、空間モデル、時間モデルから各 TGM を通したときに得られる、動作の関係性を学習した特徴量を足し合わせることで、動画の関係性に対して強い特徴量を得ることができると考える。

5 データセット

今回使用したデータセットは MultiTHUMOS [15] は、THUMOS データセットの拡張で、65 種類のアクションクラスがトリムされていない平均 3 分のビデオに高密度に付与されている (図 10)。ActivityNet や THUMOS と違って、MultiTHUMOS は、各ビデオに平均 10.5 種類のアクションクラス、1 フレームあたり 1.5 ラベル、25 種類のアクションインスタンスが付与されている。動画の内容としてバスケットボールやバレーボールの試合、ウェイトリフティングなど様々なスポーツの動画がある。これらはすべて Youtube から取得されたものである。我々は今回実験のために約 400 本の動画を 1:1 に分けて訓練とテストを行った。

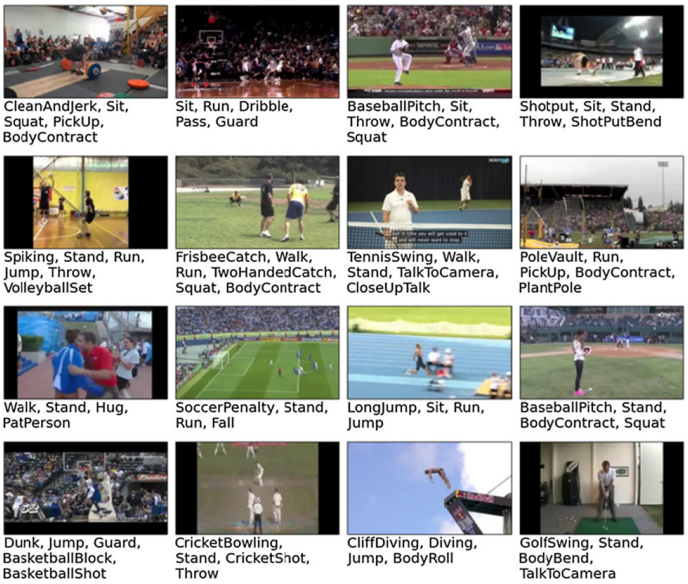

図 10: MultiTHUMOS データセット [15]

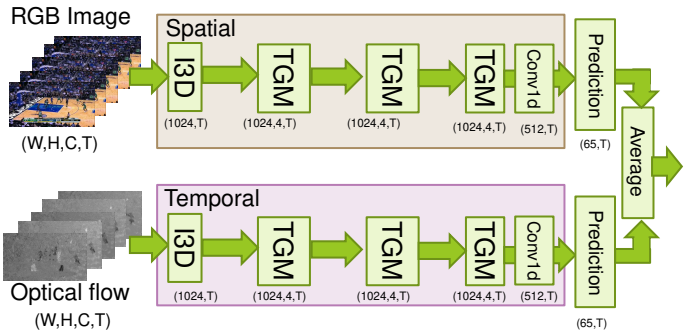

図 11: I3D+TGM

6 実　　験

5 節で述べたデータセットを用いて既存手法と提案手法の比較実験を行った。既存手法として、I3D のみ、I3D と TGM を組み合わせたものを用いる。I3D と TGM を組み合わせた概要図は図 11 に示す。評価指標として mAP(mean average precision) を用いる。これは総フレーム数に対してクラスごとの平均適合率 (AP) を求め、それをクラス数でさらに平均したものである。クラス数 C として計算式は以下に示す。

$$mAP = \frac{1}{C}\sum_{c}^{C} AP_c \tag{8}$$

エポック数は 60、バッチサイズ 2、Loss 関数は binary cross entropy を用いて実験を行う。その実験結果を表 1 に示す。

結果から、提案手法が既存手法より高精度を得ていることがわかる。また一番精度の向上が見られた提案モデル 3 と既存手法との各クラスごとの AP の差分値を図 12 に示す。この図からほとんどのクラスにおいて精度の向上が確認できており、全体として約 3%の精度の向上が確認できている。特に「WeightliftingJerk」に関しては約 16%精度の向上が確認できる。図 13 は予測された確率に対して、閾値を 0.5 として 2 値にして可視化したものである。動画はバレーボールの試合の動

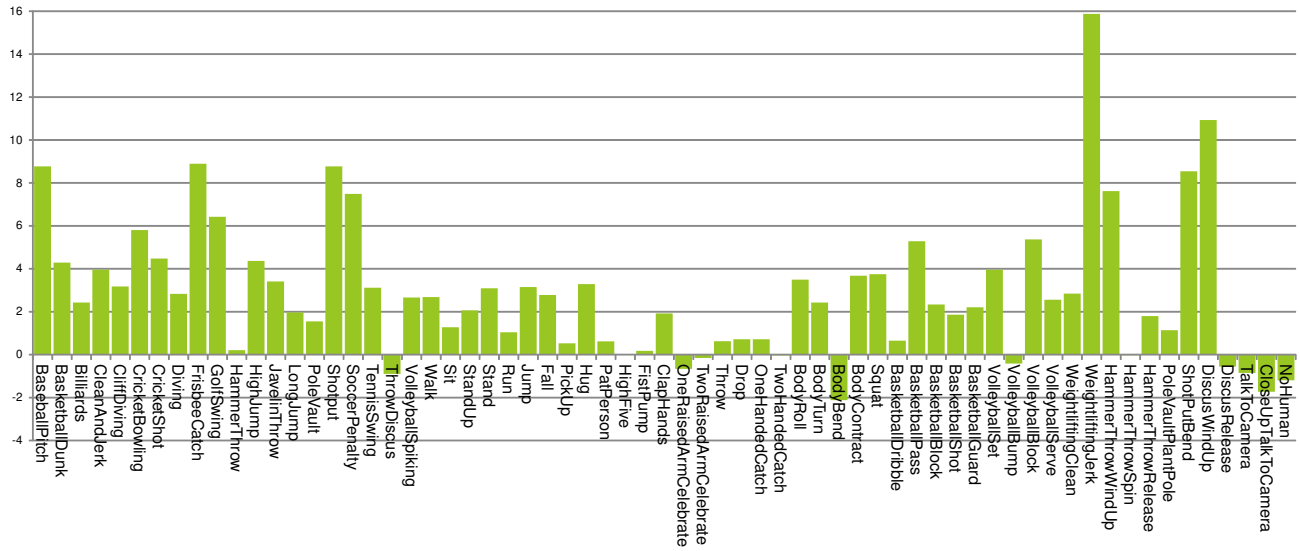

図 **12:** I3D+TGM と提案モデル 3 の比較 (各クラス (%))

表 1: 実験結果 (mAP(%))

手法	RGB	Flow	Two-Stream
I3D	23.6	18.1	26.0
I3D+TGM	32.5	24.8	34.2
提案モデル 1	-	-	34.8
提案モデル 2	-	-	35.7
提案モデル 3	-	-	37.1

画で成功例の一つである。従来手法と比較すると、バレーボールに関する動作 (スパイク、ブロック) や運動等において基本となる動作（走る、ジャンプ）共に提案手法のほうが、動作を継続的に検出できていることがわかる。これは時空間を考慮したモデル構成により、それぞれの動作が検出された際により強く検出できるようになったためであると考えられる。また「走る」動作においては従来手法では全く検出できていないのに対して、モデル 3 においては検出ができており、精度の向上が大きくみられる。これは「走る」という動作とほかの動作が同時に発生することが多く、従来手法ではほかの動作に強調されすぎてしまって、検出されなかった。しかし、提案手法では動作の関係性をより学習できていると考えられるため「走る」動作もしっかりと検出できている。

7　まとめと今後の課題

本研究では複数の同時動作に対するマルチラベル問題としての動作検出において高精度な手法を検討した。我々の提案として、空間的特徴と時間的特徴に特化したモデルを組み合わせ、新たに時空間特徴を考慮したモデルを複数提案した。提案手法を用いて、実験を行い既存手法と比較をした結果、動作の認識精度の向上を確認できた。今後の課題として、モデルの改良による精度向上、別のデータセットを用いての実験による提案手法の有効性の確認があげられる。

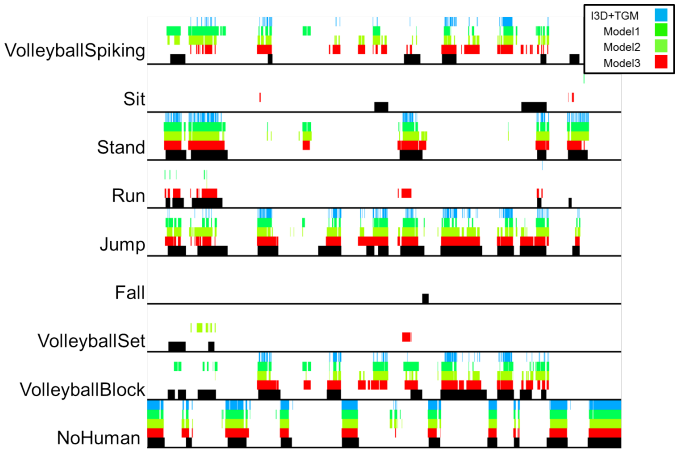

図 **13:** 成功例

謝　　辞

本研究の一部は、科研費基盤 (B) (課題番号 17H01746) の支援を受けて遂行した。

文　　献

[1] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. *CoRR*, abs/1604.06573, 2016.

[2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017.

[3] A. J. Piergiovanni and Michael S. Ryoo. Learning latent super-events to detect multiple activities in videos. *CoRR*, abs/1712.01938, 2017.

[4] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review.

[5] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, Washington, DC, USA, 2011. IEEE Computer Society.

[6] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

[7] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *CoRR*, abs/1709.02371, 2017.

[8] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *CoRR*, abs/1504.06852, 2015.

[9] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[10] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 510–526, Cham, 2016. Springer International Publishing.

[13] AJ Piergiovanni and Michael S. Ryoo. Temporal gaussian mixture layer for videos. In *International Conference on Machine Learning (ICML)*, 2019.

[14] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623, 2015.

[15] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Fei-Fei Li. Every moment counts: Dense detailed labeling of actions in complex videos. *CoRR*, abs/1507.05738, 2015.