入門講座
インフォマティクス入門－9

コンピュータビジョンによる動画認識の最先端研究

Novel Video Recognition Researches in Computer Vision

片岡裕雄 Hirokatsu Kataoka
産業技術総合研究所
人工知能研究センター
コンピュータービジョン研究チーム　主任研究員

1 はじめに

2020年現在、深層学習の劇的な幕開けから早7年以上が経過しコンピュータビジョン分野の隆盛が続いている。深層学習の発展のきっかけとなった画像認識コンペティションであるILSVRC（ImageNet Large-Scale Visual Recognition Challenge）の2012年大会では、AlexNet[1]が従来型の局所特徴ベースの手法に大差をつけて勝利したことは機械学習分野においてはもはや周知の事実になりつつある[2]。ILSVRCで使用されていたデータセットであるImageNetによる学習済みモデルは画像識別（Image Classification）[3]や物体検出（Object Detection）[4-6]、領域分割（Semantic Segmentation）[7]など転移学習にも使用されコンピュータビジョン分野の発展に貢献してきた。ImageNetの転移学習に関する調査研究[8]にて、画像識別においてはImageNetにて学習したモデルが精度の面で有効に働くことを明らかにした。さらに「ImageNet事前学習の再考（Rethinking ImageNet Pre-training）」と名付けられた論文[9]において、物体検知や領域分割のための学習データが十分に揃えられている場合、ImageNetの事前学習は精度にあまり寄与しないが、学習の収束時間の早さに寄与しており、分野の発展に寄与していると位置付けている。

しかし一方、刻一刻と変化する時系列データである動画像の解析に関しては比較的進展が遅い状態である。動画像は時系列的に連続する画像の連続であるため、静止画を認識する際の難しさに加え、人物や情景の動きというものを定量化しなくてはならないこと、膨大な情報量をリアルタイムに処理しなくてはならないことなどから、より一層難しい問題設定であるということが言える。また、動画中において「何を認識するか？」という切り口についても「物体を検出する」「人物の領域を切り出す」「人物行動を理解する」「動画を説明する文章を自動作成する」など多数の問題設定が考えられるが、コンピュータビジョンにおける基本的かつ主要な問題設定としては1動画につき1つの人物行動ラベルを返却する問題設定がもっとも広く用いられるため、本稿においては1動画1ラベルを返却する問題設定を中心的に取り扱うこととする。図1は動画像に対する人物行動の認識である。図中では動画像の入力に対して、任意の手法で動画像からベクトルなど表現を獲得・識別を実行して、人物行動ラベルを返却するというフローが一般的である。図1ではテニスコートでテニスラケットを扱う様子が表示されており“Tennis Swing”を返却することが正解である。

本稿では、深層学習における動画認識手法の最先端研究や使用するデータセット等を紹介する。

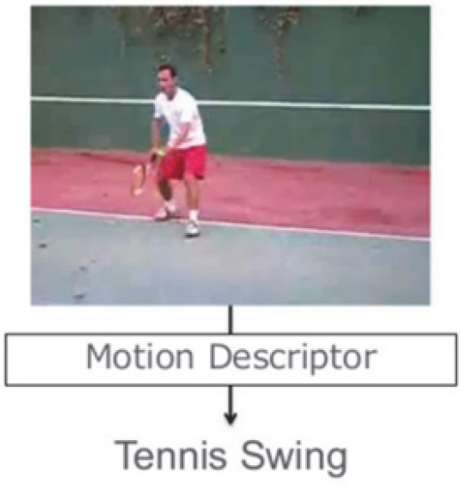

図1　動画に対する人物行動認識

2 動画像認識手法

まず動画認識手法の概要を示し、動画像認識手法における主な分類として、(i) 静止画認識において使用された2次元畳み込みネットワーク (2DCNN) や (ii) 時間と空間を同時に畳み込む3次元畳み込みネットワーク (3DCNN) が提案されている。図2に動画認識手法の変遷について示す。深層学習による動画認識手法が主流となったのは2014年のTwo-stream ConvNets以降である[10]。Two-stream ConvNetsはRGB動画像の入力と同時に、時間方向への変化を記録したオプティカルフロー画像を入力し、2種類の2DCNNを用いることからTwo-stream ConvNetsと命名されている。実は2013/2015年に時間と空間を同時に処理する3DCNNが提案されていたが、精度の面でTwo-stream ConvNetsには及んでいない。このような背景から深層学習による動画認識の初期は2DCNNによる改善が進められていた。転機となったのは2017年に提案されたKinetics-400[11] データセットによる転移学習である。Kinetics-400は30万動画に対して人物行動400カテゴリを収録したデータセットであり、転移学習に先立ち事前に学習を行うことで効果的に特徴表現を獲得する事前学習により転移学習の効果を高めるものである。3DCNNは2DCNNよりも次元数が多く、パラメータ数も膨大であったが、大規模動画データセットKinetics-400の登場により動画に対する特徴表現の獲得に成功した。Kinetics-400データセットを用いることで、20層程度の浅い層の最適化のみならず、152層までの深い層のパラメータの収束にも貢献できることが実験から明らかとなった[12]。2DCNNではImageNetによる事前学習が凡ゆるタスクに対する転移学習に有効であることが知られていたが、動画認識に対してはKinetics-400データセットが事前学習データセットにおけるスタンダードになったといえる。今日において最先端技術になったと言えるのは、DeepMind社の開発したI3D[13] である。I3DはKinetics-400の事前学習のほか、2Dで成功したパラメータを用いてネットワークを初期化している。アーキテクチャはそこまで深層ではないものの、2017年当時では主要なすべてのデータセットにおいて動画認識の最高性能を達成した。2020年現在でもI3Dやその特徴ベクトルを用いる研究は多く、スタンダードな手法として動画認識タスクに適用されている。

以降、本章では主要な動画認識手法であるTwo-stream ConvNet、C3D[14]、I3Dや3DCNNの派生として知られる(2+1) DCNN[15,16] について紹介する。

2.1 Two-stream ConvNets (2DCNN) [10]

Two-stream ConvNetsの構造を図3に示す。2つの畳み込みニューラルネットワークの確率分布をカテゴリごとに統合して平均値を取得することで最終的な行動カテゴリを推定する枠組みである。ストリームはRGB動画像 (Spatial Stream) とフロー動画像 (Temporal Stream) を入力とする2DCNNにより構成されている。ここで、RGB動画像はカメラなどのセンサにより撮影された時系列的に連続する画像群であり、フロー動画像はRGB動画像からオプティカルフローを計算して2D画像に投影した画像である。オプティカルフローの計算には、密なフロー計算の一種であるFarneback Optical Flow[17] が当初使用されていたが、最近では深層学習により計

図2　動画認識手法の変遷

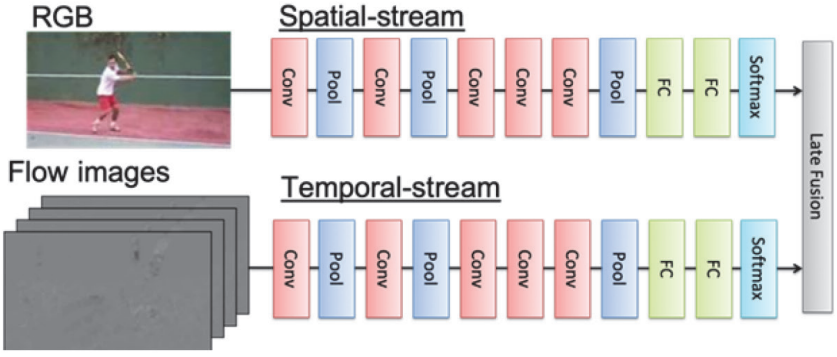

図3　Two-stream ConvNetsの構造

算されるFlowNet[18]などでも代用が可能である。RGB動画像とフロー動画像によるストリームはそれぞれ空間的な特徴と時間的な特徴を捉えていることに相当する。

ネットワークの入力について、Spatial Streamでは1枚のRGB画像（224 [pixel] ×224 [pixel] ×3 [channel]）である。最初の2つの数字は2D画像の（x, y）座標に対応し、3チャンネルは画像のRGB要素に起因する。Temporal Streamでは10枚の時系列的に連続するフロー画像（224 [pixel] ×224 [pixel] ×20 [channel]）の入力を準備する。2D画像のサイズはRGB画像と同様であるが、チャンネル数はそれぞれ（x, y）方向の勾配や強度を画像に投影したフロー画像を用い、[x1, y1, x2, y2, …, x10, y10]のようにチャンネル方向に蓄積したものである。

Two-stream ConvNetsのベースとなるネットワークは、オリジナルの論文ではVGG-M-2048をベースにしていたが、その後、より深い層を保有するVGG-16[19]やResNet-50[20]により改善されてきた。VGG-M-2048の構造は図3に示す通りであるが、VGG-16では畳み込み層が13層、全結合層が3層の計16層から構成される。ResNet-50では基本単位のモジュールにスキップコネクションを含み残差を計算しながら層を積み重ねていく構造を保有している。VGG-M-2048やVGG-16の構造では単純に畳み込み層を積み重ねていたが、ResNetによる残差の最適化により層を深くしても学習の過学習が発生せず、性能が高くなることが知られている。

また、Spatial StreamとTemporal Streamの特徴統合に関しては、当初確率分布の平均値を取る方法や識別器であるSupport Vector Machine（SVM）によりさらに学習する方法が用いられていた。SVMによる手法では出力層をさらに特徴ベクトルとみなして識別器により学習している。今日では確率分布の平均値を取る方法を用いることが多い。

Two-stream ConvNetsについて、RGBやフローの動画像を用いる方法は主に2DCNNにて用いられたが、現在では精度を向上させるためのテクニックとして3DCNNにおいても用いられることがある。外的にフロー動画像を計算して2DCNNに入力することにより、RGB動画像を用いた場合CNN内部では学習できない特徴表現の獲得に結びつくため精度が著しく向上（数％以上向上）する。しかし密なフロー計算を用いる場合、それ自体がリアルタイムでの処理が困難であるため、フロー動画像をできる限り用いることなく動画像認識を完結させる手法が考案され続けている。2020年現在においても、RGB動画像のみを用いてオプティカルフローの特徴表現をCNN内部で獲得できるようにする研究は行われており、計算量削減のためにも研究が続いている。

2.2 C3D（3DCNN）[14]

3DCNNのベースラインとなったのは3D Convolutional Networks（C3D）である。ここで、C3Dの詳細な説明の前に2D畳み込みと3D畳み込みの基本的な違いについて図4に示す。通常、2DCNNでは図4上にあるように画像の（x,y）平面に従ってフィルタを走査・処理することで畳み込みマップを獲得する。畳み込み処理や最大値プーリング（Max Pooling）を繰り返すことで最終的に識別結果を得ている。一方で3D畳み込みでは入力として時系列的に連続する画像を複数枚入力することで、画像平面（x, y）プラス時間tにより3次元のボリュームデータを構成する。すなわち、3Dでは画像平面（x,y）に加えて時間情報も含む（x, y, t）を畳み込み処理する。図4下の3DCNNでは畳み込みを実施するカーネル（畳み込みの重みがフィルタの基本単位に割り振られている）や最大値プーリングも3次元に拡張されている。単純に2次元の畳み込みカーネルや最大値プーリング処理が3次元空間にて行うのみであるが、時間方向に1枚ずつ処理をしていたものが、3DCNNでは時間方向のボリュームも含めて同時に処理可能である。

C3Dの入力は基本的にRGB動画像のみであり、時系列方向には16フレーム、画像サイズは112ピクセルを入力するため、112 [pixel] ×112 [pixel] ×3 [channel] ×16 [frame]となり、2DCNNからは次元がひとつ増えている。C3Dのベースとなるネットワークは VGGNetであり、（x, y, t）に対してそれぞれ3×3×3のサイズで畳み込み処理を行う。

2.3 I3D[13]

I3D（Inflated 3D Convolutional Network）も基本的には3DCNNにより構成されるネットワークである。I3Dが2017年時点で非常に高い精度を誇っていた2つの主な理由としてはKinetics-400事前学習モデルの利用や2DCNNの事前学習モデルで効果の高いImageNet事前学習を使用したことである。学習の手順としては、ImageNetの2D畳み込みカーネルの重みを3D畳み込みカーネルにコピーして、Kinetics-400に

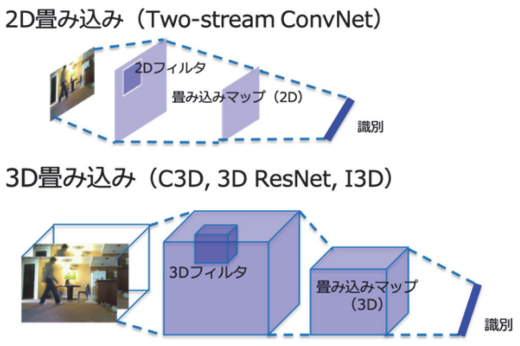

図4　2DCNNと3DCNNを構成する畳み込み機構

より学習を行う。2Dの畳み込みカーネルを膨張させて3D畳み込みカーネルとして扱うため、“Inflated”と言われている。静止画認識の文脈でスタンダードであるImageNetと動画認識の文脈でスタンダードであるKinetics-400のデータセットから、良好に特徴表現を獲得することに成功していることが、動画認識の分野で高精度を達成した理由である。

その他、I3Dに使用されているテクニックを紹介する。使用しているアーキテクチャとしてはGoogLeNet[21]を使用している。GoogLeNetはInceptionと呼称されるカーネルサイズなどのパラメータを変動させた複数種類の畳み込みにより構成される基本ユニットを数段階積み上げることで構成されている。ここで、もちろんGoogLeNetの畳み込みカーネルは3Dにより構成されている。GoogLeNetは22層構成であり、C3Dで提案されているようなVGGNetよりは深い構造となっていることからも、精度向上に寄与している。また、Carreiraらは論文中で実験設定についてグラフィックボードを64枚用いてできる限り学習時のミニバッチサイズを大きくしたと記載している。通常、64枚ものグラフィックボードを用いることは困難であるが、学習済みモデルは著者らにより公開されており、追加学習を行うことでより少ない計算機材の中でも精度の再現を可能としている。

2.4 3D ResNet[12]

著者らの研究では、より一般的に使いやすい動画認識のベースラインを作ると同時に「3DCNNは今後2DCNNと比較して発展するか？」を3D ResNet論文にて研究した。ベースラインという意味では、静止画認識でスタンダードとなったResNetの構造を用いて、畳み込みカーネルを3Dに置き換えた。また、調査の結果3DCNNの発展する起点となったのが大規模かつラベル付けの質が高いKinetics-400データセットを用いたことであると結論づけた。ImageNetは128万もの静止画像に対して良質なラベル付けを行ない転移学習に有効な特徴表現を獲得しているが、Kinetics-400は動画像認識の文脈でこれに匹敵するデータ量やラベルの質を確保できていることがわかった。ResNet + ImageNetでは152層まで層を積み重ねて精度向上が見られるが、3D ResNet + Kinetics-400も152層まで精度向上が見られることを明らかにした。また、Kinetics-400の部分はより小規模なデータセットであるUCF-101[22], HMDB-51[23], ActivityNet[24]など数千～数万規模のデータセットでは152層までは過学習をおこしてうまく精度が出ない中、Kinetics-400のように数十万動画を用いることで学習を成功に導くことができる。ただし、数量の面で多く用意してもSports-1M[25]やYouTube-8M[26]のように自動でラベリングしたような質の低いラベルでは精度があまり出ないことが知られている。

2.5 (2+1) DCNN[15,16]

3DCNNの派生形として、空間情報 (x, y) を畳み込んだのちに時間情報tを畳み込む (2+1) DCNNが最近提案されている。3DCNNでは (x, y, t) に対してそれぞれ3×3×3の畳み込みカーネルを用意していたが、(2+1) DCNNでは (x, y, t) に対して (3, 3, 1) のカーネルと (1, 1, 3) のカーネルを組み合わせながら順次畳み込みを行うことで空間と時間の情報を捉えつつもパラメータ数を減らすことに成功している。計算時間に関しては畳み込みを2回行うため3DCNNと比較すると (2+1) DCNNの方が若干処理時間を要する。

3 動画認識用データセット

ここで、動画認識に用いられる代表的なデータセットを図5に示す。図中1行目にはコンピュータビジョンによる動画認識の初期に用いられてきたデータセットであり、今日では用いられることが少なくなったが、深層学習以前の動画認識の発展に寄与した。図中2行目は現在では追加学習 (Fine-tuning) の文脈で用いられている。図中3行目で紹介されているデータセットは最近提案されたデータセットであり、数十万～数百万規模の動画数を誇る。深層学習時代にてよく用いられるデータセットであるが、2020年現在ではKineticsデータセットが転移学習用のデータとして使用されることが多い。

図5で紹介したデータは主に動画共有サイトであるYouTubeからダウンロードされたデータであることが多い。動画認識コミュニティではデータ取得の偏りを防ぐために最近ではユーザ自らが撮影したデータセットの共有 (Charades dataset[27]) や、キッチンにて撮影された料理行動や物体の行動・インタラクション認識 (Epic Kitchen dataset[28]) などのデータセットも収集・公開されている。異常なイベントを認識する取り組みや交通シーンにおける動画解析なども進んでおり、今後より動画認識は実利用化に向けて進んでいくと考えられる。

4 文脈による人物行動認識

これまでに紹介してきた動画認識研究にも最近弱点があることが報告されている。その中の一つが文脈による認識である。動画像認識では人物行動を認識することが多く、前景となる人物行動と背景に映り込む情景の認識を分離して考える必要があることが下記の研究により明らかとなった。

人を見ない人物行動認識[29]：図6に研究の概要を示す。通常の人物行動認識では左図のようにRGB動画などを入力として動画像中の人物が「何をしているか？」を明らかにする

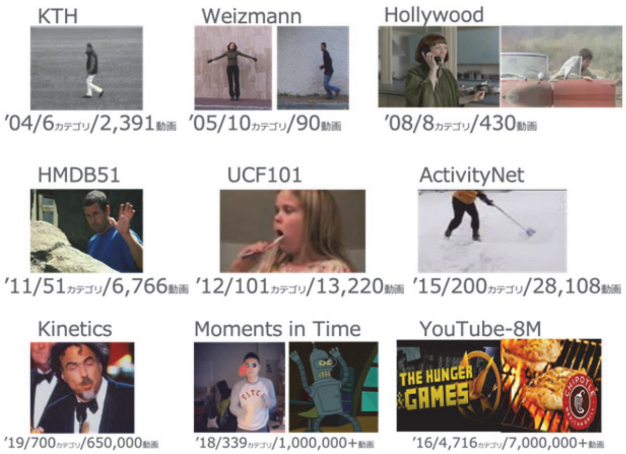

図5　動画認識用データセット

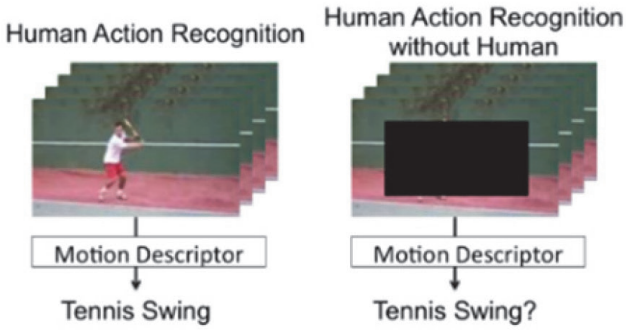

図6　人を見ない人物行動認識

技術であるが、人を見ない人物行動認識では右図のように人物領域を黒く塗りつぶしても人物行動が認識できるか調査する取り組みである。結果として、101カテゴリ認識のデータセットUCF-101（ランダムスコア：0.99％）において通常の人物行動認識が85〜90％認識できるところを、人を見ない人物行動認識では約50％もの精度で認識できてしまうことを明らかにした。これは、背景領域に現れる文脈を理解してしまうことを説明している。図中のシーンではテニスラケットをスイングしている人物行動が映っているが、人物を見ない状態でもテニスコートという文脈を見るだけでもテニススイングをしているとシステムが返却している。このように、人物行動認識において背景領域に現れる文脈は無視できないほど大きいということを明らかにした。

教師なし文脈外行動学習[30]：人を見ない人物行動認識に関連して、文脈を効果的に利用する取り組みも行われている。図7では教師なし文脈外行動学習の枠組みを示している。その文脈において大多数を占める人物行動においては、特徴空間内にてサンプルが集まりやすいが、文脈外の行動は特徴空間において外れ値に位置付けられる、という前提のもとで学習戦略が立てられている。例えばバットを振るという行動

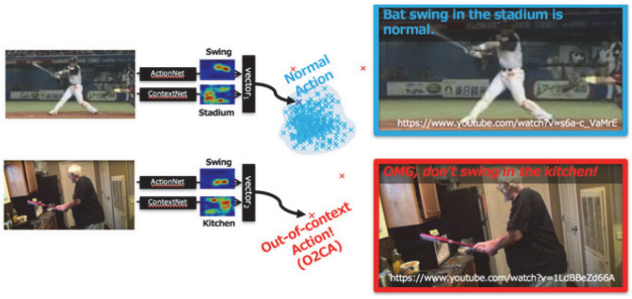

図7　教師なし文脈外行動学習

を扱った際、野球場でバットを振ることは正常だが、台所でバットを振ることは異常であると見なされる。ここから、自動で教師を作成できないか、ということが起点となっている。CNNにて実装する際にはミニバッチ内にて平均ベクトルを算出し、もっとも平均からの外れ値を排除するように学習することで文脈外行動を認識できるように学習でき、なおかつ人間による教師を必要としない。評価は合成データを用いて行なっているものの、高い精度で正常もしくは異常を識別することに成功している。

参考文献

1）A. Krizhevsky, I. Sutskever and G. E. Hinton：ImageNet Classification with Deep Convolutional Neural Networks, Neural Information Pro-cessing Systems (NIPS), (2012).
2）ImageNet Large Scale Visual Recognition Challenge 2012, (2012), http://www.image-net.org/challenges/LSVRC/2012/results.html, (accessed 2021-04-30).
3）J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell：International Conference on Machine Learning (ICML), (2014), 647.
4）R. Girshick, J. Donahue, T. Darrell and J. Malik：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2014), 580.
5）R. Girshick：IEEE International Conference on Comuter Vision (ICCV), (2015), 1440.
6）S. Ren, K. He, R. Girshick and J. Sun：Faster R-CNN, Neural Information Pro-cessing Systems (NIPS), (2015).
7）J. Long, E. Shelhamer and T. Darrell：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2015), 3431.
8）S. Kornblith, J. Shlens and Q. V. Le：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2019), 2661.

9) K.He, R.Girshick and P.Dollar：IEEE International Conference on Comuter Vision (ICCV), (2019), 4918.
10) K.Simonyan and A.Zisserman：Two-stream convolutionalnetworks for action recognition, Neural Information Pro-cessing Systems (NIPS), (2014).
11) W.Kay, J.Carreira, K.Simonyan, B.Zhang, C.Hillier, S.Vijayanarasimhan, F.Viola, T.Green, T.Back, P.Natsev, M.Su-leyman and A.Zisserman：The Kinetics Human ActionVideo Dataset. arXiv pre-print arXiv:1705.06950, (2017).
12) K.Hara, H.Kataoka and Y.Satoh：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2018), 6546.
13) J.Carreira and A.Zisserman：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2017), 6299.
14) D.Tran, L.Bourdev, R.Fergus, L.Torresani and M.Paluri：IEEE International Conference on Comuter Vision (ICCV), (2015), 4489.
15) Z.Qiu, T.Yao and T.Mei：IEEE International Conference on Comuter Vision (ICCV), (2017), 5533.
16) D.Tran, H.Wang, L.Torresani, J.Ray, Y.LeCun and M.Paluri：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2018), 6450.
17) G.Farneback：Proceedings of the Scandinavian Conference on Image Analysis, (2003), 363.
18) A.Dosovitskiy, P.Fischer, E.Ilg, P.Hausser, C.Hazırbas, V.Golkov, P.van der Smagt, D.Cremers and T.Brox：IEEE International Conference on Comuter Vision (ICCV), (2015), 2758.
19) K.Simonyan and A.Zisserman：Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representation (ICLR), (2015).
20) K.He, X.Zhang, S.Ren and J.Sun：IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), 770.
21) C.Szegedy, W.Liu, Y.Jia, P.Sermanet, S.Reed, D.Anguelov, D.Erhan, V.Vanhoucke and A.Rabinovich：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2015), 1.
22) K.Soomro, A.R.Zamir and M.Shah：UCF101：A Datasetof 101 Human Action Classes From Videos in The Wild, CRCV-TR-12-01, (2012).
23) H.Kuehne, H.Jhuang, E.Garrote, T.Poggio and T.Serre：IEEE International Conference on Computer Vision (ICCV), (2011), 2556.
24) F.C.Heilbron, V.Escorcia, B.Ghanem and J.C.Niebles：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2015), 961.
25) A.Karpathy, G.Toderici, S.Shetty, T.Leung, R.Sukthankar and L.Fei-Fei：IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (2014), 1725.
26) S.Abu-El-Haija, N.Kothari, J.Lee, P.Natsev, G.Toderici, B.Varadarajan and S.Vijayanarasimhan：YouTube-8M：A large-scale video classification benchmark, arXiv pre-print arXiv:1609.08675, (2016).
27) G.A.Sigurdsson, G.Varol, X.Wang, A.Farhadi, I.Laptev and A.Gupta：ECCV, (2016), 510.
28) D.Damen, H.Doughty, G.M.Farinella, S.Fidler, A.Furnari, E.Kazakos, D.Moltisanti, J.Munro, T.Perrett, W.Price and M.Wray：ECCV, (2018), 753.
29) Y.He, S.Shirakabe, Y.Satoh and H.Kataoka：ECCV Workshop, (2016), 11.
30) H.Kataoka and Y.Satoh：International Conference on Robotics and Automation (ICRA), (2019), 8227.

(2020年10月7日受付)