Vol.2020-CVIM-220 No.31
2020/1/24

符号化露光画像を用いた人物の行動認識

大河原 忠[1,a)]　吉田 道隆[1]　長原 一[2]　八木 康史[3]

概要：現在，社会では監視カメラやカメラを備えたスマートフォンなどの様々なカメラが普及しており，こうしたカメラで異常な動作を検出したり，IoT デバイスを用いたマンマシンインターフェースなど，人間の行動を分析する需要が高まっている．カメラには，センサの読み出しの制限やカメラと処理サーバのネットワーク帯域の制限のため，空間解像度とフレームレートにはトレードオフがあり高解像度かつ高フレームレートでの撮影やデータ送信は困難である．低解像度の動画はオブジェクトの詳細が失われ，低フレームレートの動画はモーションの詳細が失われるため，低解像度や低フレームレートの動画は行動認識には適していない．この問題を解決するひとつのアプローチとして，符号化画像から動画を復元する，圧縮ビデオセンシングによる手法が考えられる．圧縮ビデオセンシングでは，ランダムなタイミングで露光可能なセンサを用いて撮影された単一の符号化露光画像から，センサの読み出しよりも高いフレームレートの動画を再構成することが可能である．符号化露光画像は動画が再構成できるというように時間情報を有しているので，単一の符号化露光画像から行動を認識できるのではないかと考えた．本研究では，行動認識のための符号化露光画像を利用を提案する．Deep Learning を使用して，分類モデルと同時に符号化露光パターンを最適化する．提案手法は，単一の符号化露光画像のみから人間の行動認識できることを実証した．同一データ量に圧縮する他の手法と比較し，提案手法の利点を示した．

キーワード：符号化露光，圧縮センシング，行動認識

Action Recognition from a Single Coded Image

Tadashi Okawara[1,a)]　Michitaka Yoshida[1]　Hajime Nagahara[2]　Yasushi Yagi[3]

1. はじめに

近年，社会の安全性や交通監視のため，人間のオペレータによってカメラの監視が行われてきた．しかし，カメラの普及により人間による監視は限界に達しつつあり，こうしたカメラから人間の行動を分析する需要が高まっている．

従来の行動認識では，シーンの動画を入力として行動ラベルを予測しており，シーンに依存せず動きに固有の特徴

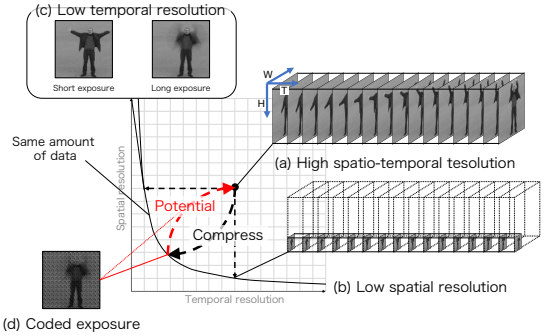


図 1: データ圧縮におけるトレードオフな問題．(a) 高時空間解像度の動画．(b) 低解像度かつ高フレームレートの動画．(c) 高解像度かつ低フレームレートの動画．(d) 符号化露光動画，データ量は (b) や (c) と同じだが、(a) に匹敵する情報を持つ．

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University, Osaka, Japan
[2] 大阪大学データビリティフロンティア機構
Institute for Datability Science, Osaka University, Osaka, Japan
[3] 大阪大学産業科学研究所
Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan
[a)] okawara@am.sanken.osaka-u.ac.jp

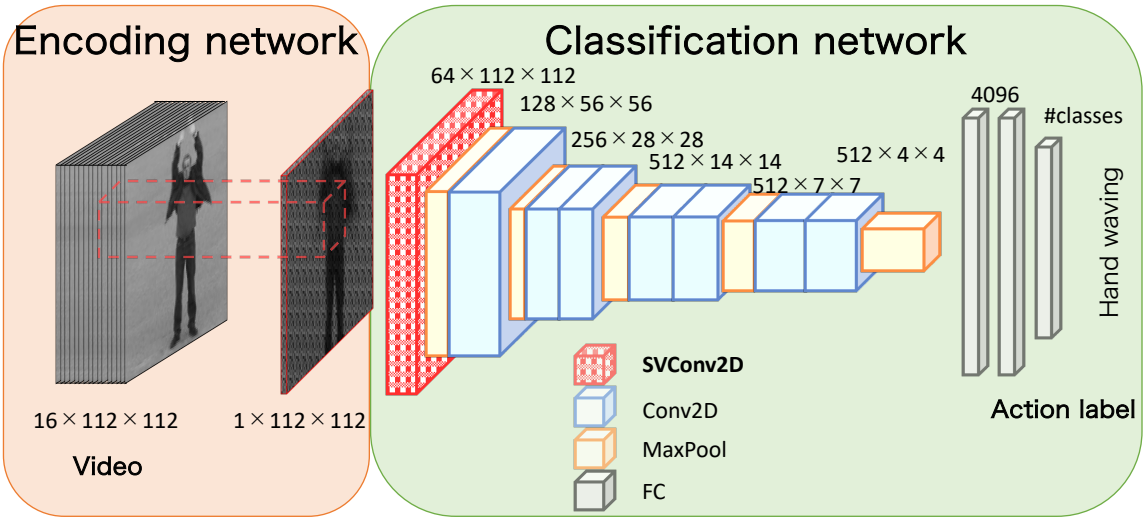


図 2: 符号化露光パターンの最適化. 符号化露光パターンはネットワークの重みとして学習され，実環境では学習されたパターンがプロトタイプの符号化露光カメラに実装された．Shift-variant convolution(SVConv2D) については 2.2 節で説明する．

を記述するため，動画認識の動き特徴としてオプティカルフローがよく利用されている [1], [2]．行動認識では動きに加えて物体見かけの情報が重要であるので，空間と時間の２つのネットワークからなる２ストリームネットワークが提案されている [3]．このネットワークでは，空間ネットワークでは RGB 画像を，時間ネットワークではオプティカルフローを処理し，最終的にそれらを統合することで，行動認識を行っている．しかし，畳み込みニューラルネットワーク (CNN) は，事前に空間情報と時間情報に明示的に分離せずとも，3 次元 (3D) の動画を学習することで時空間特徴を取得することができる．そのため，Kinetics などの大規模データセットを用いて学習することにより，行動認識に最適化された時空間特徴を抽出する 様々な 3D CNN が提案されている [4], [5], [6]．一方で，He ら [7] は動画中の人間を隠しても人間の行動を認識できると報告しており，Sevilla-Lara ら [2] は，動画の 1 フレームのみから人間の行動を認識できると報告した．さらに Huang ら [8] によって，3D CNN が認識しやすいフレームを選択していると報告されており，CNN は従来の行動認識のように動き特徴ではなくシーン内の物体を使用して行動を認識しているように思われる．我々は，行動認識の学習において学習データには多くの冗長性が含まれており，時空間情報を効率的に利用できていないと考えている．

シーンには密なテクスチャや動きの情報が含まれるため，時空間解像度の高い動画はイベント分析に有利である (**図 1-(a)**)．しかし，カメラにはハードウェアの制約のため空間解像度とフレームレートの間にトレードオフがあり，高解像度かつ高フレームレートな動画を撮影することは容易ではない．さらに，監視カメラや IoT デバイスなどのクライアントカメラが動画をシーン内のイベントや行動を分析するサーバに送信するといったクラウドコンピューティングの場面を考えると，撮影された動画は圧縮され一部の空間、時間解像度が失われる．したがって、高フレームレートで動画を撮影する場合 (図 1-(b)) は動画の空間解像度を下げる必要があり，物体の詳細な特徴は失われる．また逆に，高空間解像度で撮影する場合 (図 1-(c)) はフレームレートを下げる必要があり，詳細な動きをとらえることはできない．つまり，従来のビデオ撮影において解像度とフレームレートにはトレードオフの関係にある．

このトレードオフを解決する手法として，圧縮ビデオセンシング [9], [10], [11] がある．圧縮ビデオセンシングは図 1-(d) に示すように，画素ごとに露光を制御可能なセンサを用いて隣接する画素に異なる時間情報を持つ符号化露光画像を撮影し，単一の圧縮画像から複数のサブフレームを再構成することでセンサのサンプリング能力を超えた動画を得る手法である．高時空間解像度の動画を行動認識に用いるために動画の時空間トレードオフを解決する単純な手法として，圧縮ビデオセンシングを使用して単一の符号化露光画像から高解像度かつ高フレームレートな動画を再構成し，再構成された動画用いて行動認識を行うことが考えられる．しかし，認識アルゴリズムへの入力は必ずしも人間が認識できる動画である必要はなく，符号化露光画像には既に動画を再構成するのに十分な情報を含んでいるため，動画を再構成することなく，符号化露光画像から直接 CNN を用いてシーン内の行動を認識することが可能である．この単一の符号化露光画像から直接行動を認識する手法は，動画を再構成する必要がなく，また大規模なネットワークと計算コストが必要な 3 次元の情報を用いずに済むため，計算コストとモデルのサイズを削減することができる効率的なアプローチである．

本研究では，単一の符号化露光画像から直接行動を認識する手法を提案する．CNN のアーキテクチャを用いて符

号化露光画像の撮影から行動認識までのパイプライン全体をモデル化する．この CNN を end-to-end で学習することにより，行動認識のネットワークに加えて，符号化に最適な露光パターンを同時に決定する．シュミレーションと実証実験により，単一の符号化露光画像と短時間露光画像と長時間露光画像による認識精度を比較した．また，提案手法と 3D CNN を使用した通常の動画による認識精度を比較した．提案手法は他の単一画像よりも高い認識精度を達成し，単一の符号化露光画像の 16 倍のデータを含む動画を用いた 3D CNN に匹敵する認識精度を示した．

2. 提案手法

この章では，提案手法である単一の符号化露光画像から行動を認識する手法について述べる．提案モデルは図 2 に示すように，主に符号化ネットワークと分類ネットワークの 2 つの部分で構成されている．

符号化ネットワークは，符号化露光カメラによる符号化露光画像の撮影を表現する．このネットワークは長さ L の $p \times p$ のブロックごとの 2 値化 1D CNN として記述される．詳細な構造については 2.1 節で述べる．この 2 値化 1D CNN のパラメータを学習することで最適化された符号化露光パターンを得ることができる．

分類ネットワークは，符号化露光画像から行動ラベルを推定する第 1 層に Shift-variant convolution を用いた 12 層からなるニューラルネットワークである．通常の畳み込みカーネルは，隣接する画素は空間的に滑らかであると仮定してシフト不変 (shift-invariant) であるのに対し，Shift-variant convolution は，符号化画像の隣接する画素が異なるタイミングで露光される特徴に合わせて，画素位置毎にシフト変位 (shift-variant) カーネルでのコンボリューションを実現する．この Shift-variant convolution の詳細については 2.2 節で述べる．

このネットワーク全体を通常の CNN の学習と同様に動画と行動ラベルのセットを入力として end-to-end で学習する．これにより，行動認識に最適な符号化露光と，その符号化露光画像から行動を分類するモデルを同時に学習することができる．獲得された符号化露光パターンは分類モデルとタスクに最適化され，分類モデルも同様に符号化露光画像の露光パターンに最適化される．したがって，このフレームワークの下，カメラの符号化露光と行動認識の分類モデルという 2 つを同時に最適化することが可能である．

実実験では学習により最適化された符号化露光パターンを符号化露光カメラに実装することで実シーンの撮影を行い，撮影画像を行動認識の分類モデルに入力することで行動認識を行う．

2.1 符号化露光のための符号化ネットワーク

符号化露光カメラは各画素で露光の ON，OFF をブロックごとに切り替えながら画像を撮影するため，撮影プロセスは次のように表される．

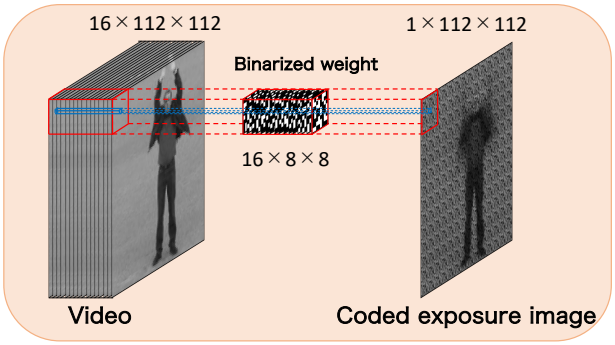


図 3: 符号化ネットワーク．このネットワークは符号化露光画像の撮影プロセスを 0 または 1 の 2 値の重みを使用して表現する．ここで，1 は露光を表し，0 は露光しないことを表す．図のように入力画素は時間方向に 2 値の重みで畳み込まれる．この図は $p = 8$ かつ $L = 16$ の場合を示している．

$$
\begin{aligned}
y_{i,j} &= \sum_{t=0}^{L-1} \phi_{t,i_p,j_p}\, \chi_{t,i,j}, \\
s.t. \quad i_p &\equiv i \bmod p, \\
j_p &\equiv j \bmod p,
\end{aligned}
\tag{1}
$$

χ は圧縮されるシーン，$\phi(\in \{0,1\})$ は符号化露光パターン，L は符号化露光パターンの長さ，y は単一の符号化露光画像の画素 (i,j)，mod はモジュロ演算を表している．提案モデルでは，この撮影プロセスを文献 [10], [11] と同様に，CNN でモデル化する．

図 3 に，$p \times p$ で長さ L の 1D CNN として符号化露光画像の撮影プロセスを表現する符号化ネットワークを示す．図 3 は $p = 8$ かつ $L = 16$ の場合である．符号化露光は ON，OFF の切り替えで実現するため 2 値の重みで表現する必要があり，通常の CNN のように連続値の重みを用いることはできない．そこで我々は Binarized neural network [12] を符号化ネットワークのモデルとその学習に使用した．Binary neural network [12] では重みを-1 または+1 に 2 値化するために Sign 関数を使用する．我々はそれに 1 を加えて 2 で割ることによって 0 または 1 の 2 値化された重みを獲得する．ここで，1 は露光を表し，0 は露光しないことを表す．

$$
\begin{aligned}
\psi_{i,j} &= \frac{1}{L} \sum_{t=0}^{L-1} w^b_{t,i_p,j_p}\, \chi_{t,i,j}, \\
s.t. \quad w^b &= \frac{\mathrm{Sign}(w)+1}{2}, \\
i_p &\equiv i \bmod p, \\
j_p &\equiv j \bmod p,
\end{aligned}
\tag{2}
$$

$\psi_{i,j}$ はネットワークによって生成された符号化露光の画素

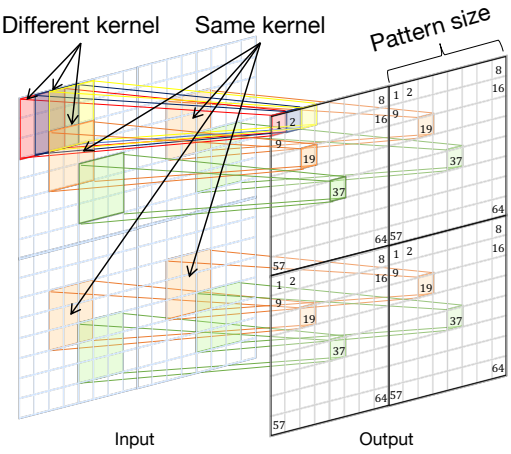


図 4: Shift-variant convolution (SVConv2D). ブロック内の異なる位置の画素は異なるカーネルで畳み込まれる．この図は $p = 8$ の場合を示している．

(i, j)， w は連続値の重み，$w^b (\in \{0, 1\})$ は 2 値化された重み (符号化露光パターンに相当する) を表す．また，Sign 関数は次の式で表される．

$$\omega^b = \mathrm{Sign}(\omega^c) = \begin{cases} +1 & \text{if } \omega^c \geq 0, \\ -1 & \text{otherwise}, \end{cases} \tag{3}$$

$\omega^b (\in \{-1, +1\})$ は 2 値化された重み，ω^c は連続値の重みを表す．

符号化露光パターンを行動認識に最適化するために符号化露光と分類ネットワークから構成される提案手法のネットワークを end-to-end で学習する．この学習によって符号化露光パターンが行動認識に最適化され，認識精度が改善される．実環境での行動認識を行う際には，学習により最適化された符号化露光パターンを符号化露光カメラに実装し撮影する．

2.2 Shift-variant convolution を使用した分類ネットワーク

均一露光画像はすべての画素が一様に露光され，隣接する画素に時間的な違いはない．そのため，通常の畳み込みは空間的な滑らかさを仮定し，シフト不変なカーネルが画像のすべての画素を畳み込む．一方，符号化露光画像では，時空間情報が符号化露光によって隣接画素にスパースに符号化されている．符号化露光画像はブロック内の各画素は異なる時間に露光されれるため空間的な滑らかさはなく，通常のシフト不変な畳み込みを符号化露光画像に適用することはできない．

そこで，Shift-variant convolution (SVConv2D) を第 1 層とした分類ネットワーク (SVC2D) を提案する (**図 4**)．図 4 では，同一カーネルは同じ色で異なるカーネルは異なる色で示している．符号化露光画像がブロック単位の周期的な性質を示すため，符号化露光パターン上の同一ブロック内は異なるカーネルにより畳み込まれ，ブロック間で同じピクセル位置にある画素は同一のカーネルにより畳み込まれる．

表 1: 提案する SVC2D を用いた分類ネットワークの構造.

Layer	(#Filters)	**SVC2D**	(Kernel/stride)
conv1	(64)	**SVConv2D**	
pool1		MaxPool2D	$(2^2/2)$
conv2	(128)	Conv2D	$(3^2/1)$
pool2		MaxPool2D	$(2^2/2)$
conv3a, 3b	(256)	Conv2D	$(3^2/1)$
pool3		MaxPool2D	$(2^2/2)$
conv4a, 4b	(512)	Conv2D	$(3^2/1)$
pool4		MaxPool2D	$(2^2/2)$
conv5a, 5b	(512)	Conv2D	$(3^2/1)$
pool5		MaxPool2D	$(2^2/2)$
fc6, fc7	(4096)	FC	

通常のシフト不変な畳み込みは次の式で表される．

$$z_{i,j} = \sum_{s=0}^{K_h-1} \sum_{t=0}^{K_w-1} w_{s,t}\, x_{i+s,j+t} + b_{i,j}, \tag{4}$$

x は入力，z は出力，$K_h \times K_w$ はカーネルサイズ，w は重み，b はバイアスを表す．これに対して，次のようにピクセル位置に応じて畳み込みカーネルを変更する．

$$\begin{aligned} z_{i,j} &= \sum_{s=0}^{K_h-1} \sum_{t=0}^{K_w-1} w_{pi_p+j_p,s,t}\, x_{i+s,j+t} + b_{i,j}, \\ s.t. \quad & i_p \equiv i \bmod p, \\ & j_p \equiv j \bmod p. \end{aligned} \tag{5}$$

提案する Shift-variant convolution を符号化露光画像に適用し，時空間特徴を考慮する．提案する分類ネットワークは第 1 層に SVConv2D を持つ，Maxpooling と通常の畳み込み，全結合層からなる 12 層のネットワークである (**表 1**)．これは，第 1 層に SVConv2D を使用し，残りの 11 層を C3D [4] の 3D コンボリューションを 2 D コンボリューションに置き換えたネットワークである．

3. 評価実験

3.1 実験設定

Tran ら [4] と同様に，学習ではオリジナルの動画像を 171×128 にリサイズし，$L \times 112 \times 112$ にランダムに切り取った動画をグレースケールへ変換して学習とテストで使用した．ここで，L は符号化露光パターンや入力動画の長さである．

図 5 に示す 3 つの手法の比較を行った．グローバルシャッターカメラを想定する例として，長時間露光画像と短時間露光画像を使用した．また，認識精度の理論的上限として元の動画を使用した．動画には単一画像の L 倍の情

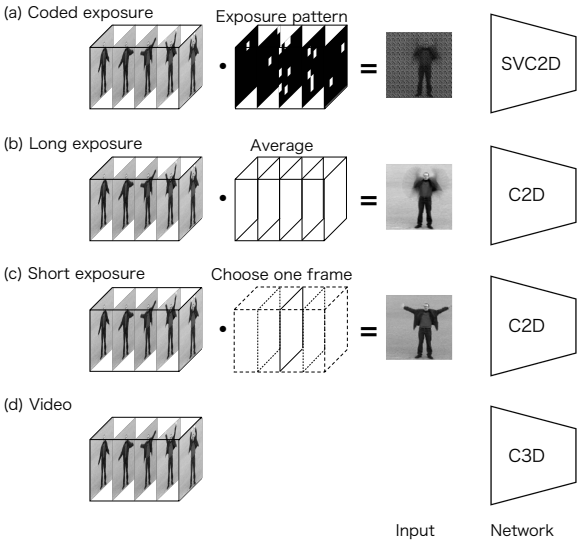


図 5: 露光の種類による圧縮手法の比較. (a) 符号化露光画像は，符号化露光パターンによって生成され，SVC2D を用いて学習される．(b) 長時間露光画像は，動画の L フレームを平均化して生成され，C2D によって学習される．(c) 短時間露光画像は，動画の L フレームから 1 フレームを選択し，長時間露光画像と同様に C2D によって学習される．(d) 動画は圧縮することなく C3D を用いて学習される．これはすべての情報を保持した理想的なケースである．

報が含まれている．比較実験では，動画の行動認識に C3D [4] を使用し，長時間露光画像や短時間露光画像の単一画像に対して C3D の 2 次元に変換したものである C2D を使用した．提案する分類モデルは，2D の符号化露光画像を入力とするため 2D の畳み込み層で構成されているが，C3D は動画の 3D ボリュームを入力とするため 3D の畳み込み層で構成されている．したがって，提案手法には C3D よりもパラメータが少なく，計算コストも小さいという利点がある (**表 2**).

(a) 符号化露光画像:

符号化露光画像は式 (1) を用いて生成した．ランダムな符号化露光パターン $\widetilde{\phi}$ は次の式で生成した．

$$\widetilde{\phi}_{t,i,j} = \delta_{t,u_{i,j}} \quad s.t. \quad u \sim \mathcal{U}[0, L-1] \tag{6}$$

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$

u は 0 から $L-1$ の一様乱数，δ はクロネッカーのデルタを表す．最適化された符号化露光パターンは 2.1 節で述べたように SVC2D によって同時に最適化した．このようにして生成された符号化露光画像を SVC2D に入力し学習した．

表 2: 各ネットワークにおけるパラメータ数と計算コスト. SVC2D は C3D に対してパラメータ数を 23%削減し，計算コストを 95%削減する．

Model	#Params	FLOPs
SVC2D	60.62 M	1.91 G
C2D	60.58 M	1.91 G
C3D	79.01 M	37.88 G

(b) 長時間露光画像:

長時間露光画像は次の式で生成した．

$$y_{i,j} = \frac{1}{L} \sum_{t=0}^{L-1} x_{t,i,j}, \tag{8}$$

x は圧縮されるシーン, L は動画の長さ, y は長時間露光画像を表す．この長時間露光画像は C2D で学習した．

(c) 短時間露光画像:

短時間露光画像は次の式で生成した．

$$y_{i,j} = \sum_{t=0}^{L-1} \delta_{t,\lfloor \frac{L}{2} \rfloor} x_{t,i,j}, \tag{9}$$

y は短時間露光画像を表す．短時間露光画像も長時間露光画像と同様に C2D で学習した．

(d) 動画:

動画は圧縮することなく C3D で学習した．これはすべての情報を保持した理想的なケースである．

これらのモデルは学習率 0.01 で SGD をオプティマイザとして 200epoch 学習した．

3.2 シミュレーション実験

3.2.1 Something-Something データセットでの評価

Something-Something [13] は人間と物体のやり取りに関する大規模なデータセットであり，時間的な関係性が必要とされる曖昧なカテゴリを含む 174 クラスの行動ラベルがある．このデータセットはテストセットが公開されていないため，検証セットでの結果を**表 3** に示す．

時間情報が含まれていないため短時間露光画像での認識精度は低かった．長時間露光画像では，時間の経過とともに明るさの変化が積分され，カメラや物体の動きによりモーションブラーが発生する [14], [15]．モーションブラーは，ブラーの形状としてモーション情報を提供するが，物体の形状を不鮮明にする．したがって，モーションブラーにより物体の認識が困難になるため，長時間露光画像の認識精度は低かった．ランダムな符号化露光パターンを使用した C2D での認識精度は前の 2 つの結果よりも高い精度となり，通常のシフト不変の畳み込みを使用した C2D においても符号化露光パターンを最適化することで認識精度を大幅に改善した．さらに，我々の提案手法である Shift-variant convolution を使用した SVC2D を用いた場合は，他の単一

表 3: Something-Something の検証セットでの平均精度.

Input	Pattern	Model	Top-1	Top-3	Top-5
(a) Coded	optimize	SVC2D	29.37	47.39	56.33
	optimize	C2D	26.77	45.66	54.37
	random		12.75	25.63	32.69
(b) Long		C2D	10.82	22.83	30.20
(c) Short		C2D	10.32	21.85	28.56
(d) Video		C3D	**39.31**	**61.97**	**70.05**

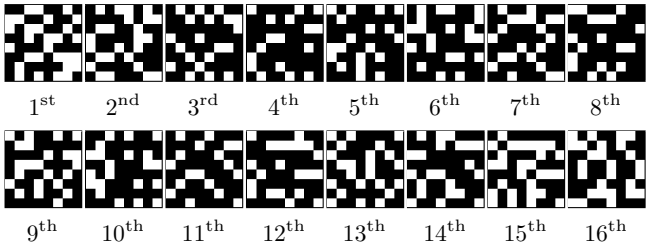


図 6: 行動認識に最適化された符号化露光パターン．ここでは，白い画素は露光を表し，黒い画素は露光しないことを表す．これは，均一な露光よりもスパースな露光が行動認識により適していることを示している．

図 7: XY-T 図．ここでは縦軸に時間次元を横軸に空間次元を示す．これは，短時間露光や長時間露光よりも時空間的にスパースな露光が行動認識により適していることを示している．

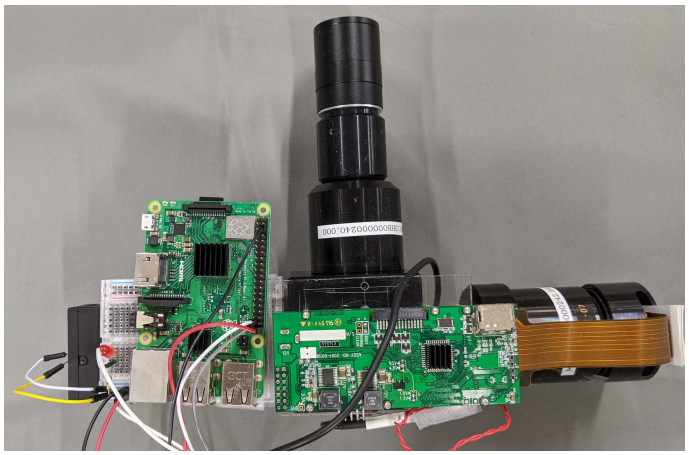
画像からの行動認識の結果よりも高い認識精度を示した．

通常畳み込みでは，符号化露光の固有な時空間パターンは画像の一部から決定されるため，シーンによって誤対応が発生する．我々は符号化露光パターン上の位置に応じて畳み込みカーネルを変更するというシンプルなアプローチでこの問題を解決した．

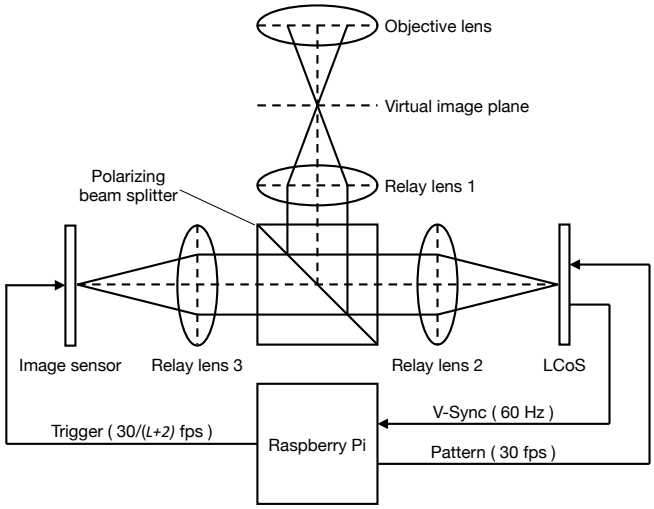
SVC2D と学習し最適化した符号化露光パターンを図 **6** に示す．最適化された符号化露光パターンは，グローバル露光でなくスパースな露光であることに注目されたい．縦軸に時間次元を横軸に (高さや幅が 1 次元に均らされた) 空間次元を示した XY-T 図を図 **7** に示す．これは，行動認識においては短時間露光や長時間露光よりも時空間でスパースな露光が適していることを示している．

3.3 プロトタイプ符号化露光カメラを用いた実環境での評価

商用の符号化露光カメラが存在しないため，実環境で符号化露光画像を撮影するために画素ごとに露光を制御可能なプロトタイプの符号化露光カメラを作成した (図 **8**-(a))．



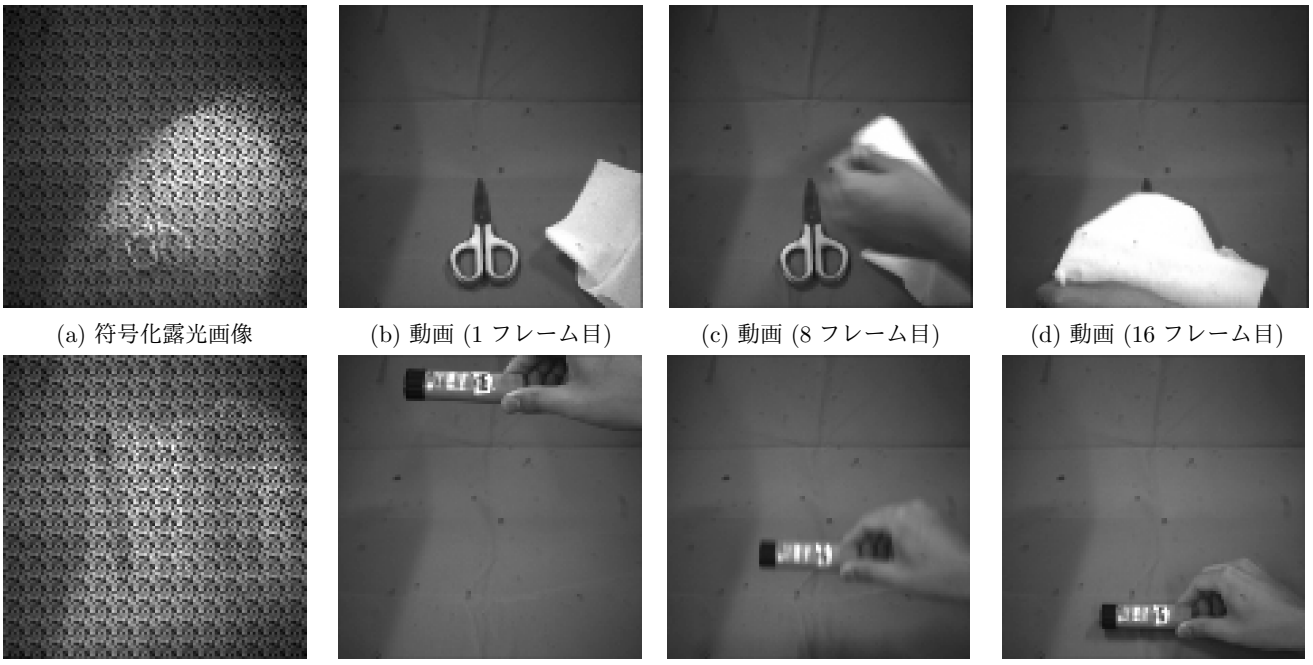
(a) プロトタイプカメラ．LCoS (右下)，LCoS 制御基板 (中央下)，Raspberry Pi (左下)，カメラ (Raspberry Pi の裏) で構成されている．このカメラは LCoS を用いて露光を画素ごとに制御可能である．



(b) プロトタイプカメラの光学図.

図 8: Liquid Crystal on Silicon (LCoS) を使用したプロトタイプの符号化露光カメラ．

プロトタイプカメラでは，リレーレンズを介して光学的にセンサに対応する位置に配置された Liquid Crystal on Silicon (LCoS) を使用して露光制御を光学的にシュミレートした (図 8-(b))．この LCoS に白黒の 2 値パターンを表示することにより，画素ごとに入射光をブロックすることで符号化露光を実現した．Raspberry Pi は LCoS からの垂直同期信号 (V-Sync) を受信し，同期して 30 fps で LCoS に符号化露光パターンを表示し，$30/(L+2)$ fps で同期してカメラのトリガーを入力す．ここで開始と終了のフレームの調整のために，L フレームの符号化露光パターンの前後に 2 つのブラックパターンを表示している．3.2.1 節で学習した符号化露光パターン (図 6) をこのプロトタイプカメラに実装した．Something-Something データセットか



(a) 符号化露光画像 (b) 動画 (1 フレーム目) (c) 動画 (8 フレーム目) (d) 動画 (16 フレーム目)

(e) 符号化露光画像 (f) 動画 (1 フレーム目) (g) 動画 (8 フレーム目) (h) 動画 (16 フレーム目)

図 9: 同一の行動 (上: **covering something with something**，下: **moving something down**) について撮影された符号化露光画像 (a，e) と動画 (b-d，f-h).

表 4: Something-Something データセットのサブセット (25 クラス) の実環境とシュミレーション環境での認識精度.

		実環境			シュミレーション環境		
Input	Model	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
(a) Coded	SVC2D	**72.0**	84.0	**88.0**	41.6	58.9	67.2
(b) Long	C2D	20.0	40.0	52.0	13.8	30.4	39.4
(c) Short	C2D	21.0	47.0	60.0	14.6	32.5	40.5
(d) Video	C3D	71.0	**88.0**	**88.0**	**47.1**	**69.4**	**76.9**

ら含まれる動画数の多い方から 25 クラス選択し，通常の動画と符号化露光画像を撮影した (**図 9**)．1 つの行動につき 4 つの物体を撮影し，合計で 100 の動画と 100 の符号化露光画像を撮影しました．

3.1 節と同様に，撮影した動画から長時間露光画像と短時間露光画像を生成した．3.2.1 節で学習したモデルと Something-Something データセットのサブセットを用いて実環境とシュミレーション環境での比較を行った．実環境で撮影した動画と符号化露光画像に対する、シュミレーション環境で事前に学習したモデルで推定した結果を**表 4** (左) に，実環境で撮影したサブセットと同一の 25 クラスの行動についてシュミレーション環境で実験した結果を表 4 (右) に示す．この結果から実環境においても，符号化露光画像は動画と同様に高い精度で認識できることがわかった．

4. おわりに

本研究では，単一の符号化露光画像から行動を認識する手法を提案した．符号化露光画像は圧縮ビデオセンシングにおいて高時空間解像度の動画の再構成に利用されてきたが，行動認識においても十分な時間情報が含まれていると考え，動画を再構成することなく単一の符号化露光画像からシーン内の行動を認識する手法を提案した．符号化露光カメラにより符号化露光画像の撮影を行う段階から行動を認識するまでの過程を，符号化ネットワークと分類ネットワークで表現し，同時に最適化することで行動認識に最適化された符号化露光パターンと符号化露光に最適化された分類ネットワークを獲得した．Shift-variant convolution を提案し，符号化露光画像という空間的に滑らかでない画像に対する効率的な畳込みを実現し，さらなる精度の改善を示した．提案手法では単一の符号化露光画像を使用したにも関わらず，16 倍の情報を有する動画に匹敵する高い認識精度を達成した．LCoS を使用したプロトタイプ符号化露光カメラを作成し，Something-Something のサブセットを撮影し評価を行い，提案手法が実際の環境でも機能することを確認した．

謝辞

本研究の一部は日本学術振興会基盤 (S)17H06102 および

挑戦的研究 (萌芽)18K19818 の支援を受けた.

参考文献

[1] Guney, F., Sevilla-Lara, L., Sun, D. and Wulff, J.: "What Is Optical Flow for?": Workshop Results and Summary, *Proceedings of European Conference on Computer Vision (ECCV) Workshops* (2018).

[2] Sevilla-Lara, L., Liao, Y., Güney, F., Jampani, V., Geiger, A. and Black, M. J.: On the Integration of Optical Flow and Action Recognition, *Proceedings of German Conference on Pattern Recognition (GCPR)*, Vol. LNCS 11269, pp. 281–297 (2018).

[3] Simonyan, K. and Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos, *Advances in Neural Information Processing Systems (NIPS)*, pp. 568–576 (2014).

[4] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks, *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 4489–4497 (2015).

[5] Carreira, J. and Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6299–6308 (2017).

[6] Hara, K., Kataoka, H. and Satoh, Y.: Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[7] He, Y., Shirakabe, S., Satoh, Y. and Kataoka, H.: Human Action Recognition Without Human, *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 11–17 (2016).

[8] Huang, D., Ramanathan, V., Mahajan, D., Torresani, L., Paluri, M., Fei-Fei, L. and Niebles, J. C.: What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7366–7375 (2018).

[9] Hitomi, Y., Gu, J., Gupta, M., Mitsunaga, T. and Nayar, S. K.: Video from a single coded exposure photograph using a learned over-complete dictionary, *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 287–294 (2011).

[10] Iliadis, M., Spinoulas, L. and Katsaggelos, A. K.: Deep fully-connected networks for video compressive sensing, *Digital Signal Processing*, Vol. 72, pp. 9–18 (2018).

[11] Yoshida, M., Torii, A., Okutomi, M., Endo, K., Sugiyama, Y., Taniguchi, R.-i. and Nagahara, H.: Joint optimization for compressive video sensing and reconstruction under hardware constraints, *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 634–649 (2018).

[12] Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R. and Bengio, Y.: Binarized Neural Networks, *Advances in Neural Information Processing Systems 29*, pp. 4107–4115 (2016).

[13] Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M. et al.: The" Something Something" Video Database for Learning and Evaluating Visual Common Sense., *Proceedings of International Conference on Computer Vision (ICCV)*, Vol. 1, No. 2, p. 3 (2017).

[14] Shengyang Dai and Ying Wu: Motion from blur, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8 (2008).

[15] Agrawal, A. and Yi Xu: Coded exposure deblurring: Optimized codes for PSF estimation and invertibility, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2066–2073 (2009).