Vol.2019-CVIM-215 No.4
2019/1/17

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[2]　長原 一[3]　八木 康史[4]

概要：近年，監視カメラや車載カメラ等の IoT デバイスで撮影された映像の解析が盛んに行われている．これらのカメラで撮影された映像は，データセンタに集約され，解析等に用いられが，一般には通信路の容量を削減するために空間解像度や時間解像度 (フレームレート) を下げる等データの圧縮が行われている．空間解像度を下げれば細部が不鮮明になり，フレームレートを下げれば，動きに関する情報が失われてしまう．この空間解像度と時間解像度のトレードオフを解決する手段として符号化露光画像を用いた圧縮ビデオセンシング手法が提案されている．圧縮ビデオセンシングは，画像センサの各ピクセルをランダムに露光した符号化露光画像を撮影し，符号化露光画像に含まれる異なる時間の情報を用いることで，時間空間解像度の高いビデオを復元する．これを用いた解析として人物行動認識を考えた場合，そもそも撮影された符号化画像には時間情報が含まれているため，復元を介さなくても直接，符号化露光画像から行動認識を行えると考えた．本研究では，符号化露光カメラにより撮影される単一の画像から Deep Learning を用いて直接人物の行動認識を行う手法の提案を行う．比較実験として動画を用いた場合，単純な平均化した画像を用いた場合と比較し，本手法のデータ量に対する認識率の高さを確認した．

キーワード：符号化露光，圧縮センシング，行動認識

Human Action Recognition Using Coded Exposure Image

Tadashi Okawara[1,a)]　Michitaka Yoshida[2]　Hajime Nagahara[3]　Yasushi Yagi[4]

1. はじめに

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．近年のカメラ数の増加に伴い，人間のオペレータを支援，または人間のオペレータに取って代わる自動監視システムが注目され，監視カメラや車載カメラ等の IoT デバイスで撮影された映像の解析が盛んに行われている．これらカメラで撮影された映像は，データセンタ等の中央ホストに集約され，解析等

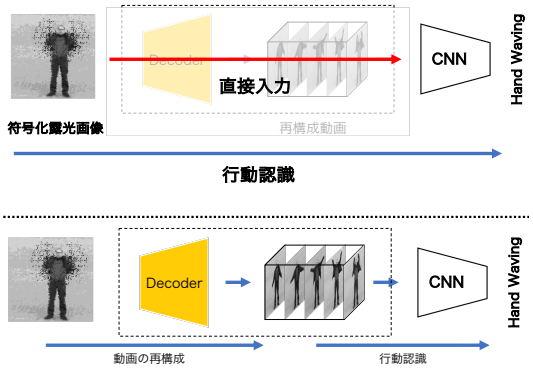


図 1: 提案手法の概要．符号化露光画像から直接，行動認識を行う提案手法 (上)．圧縮ビデオセンシングを用いた映像解析として行動認識を行った場合 (下)．

[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] 九州大学大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
[3] 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
[4] 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

に用いられる [1]．一般には通信路の容量やストレージの容量を削減するために空間解像度や時間解像度 (フレームレート) を下げる [2] 等データの圧縮が行われている．空間解像度を下げれば細部が不鮮明になり，フレームレートを下げれば，動きに関する情報が失われてしまうという問題ある．この問題は，カメラで動画を撮影する際のトレードオフな問題と類似する．スチルカメラでは，空間解像度の高い画像の撮影を行うことができるが，フレームレートを高くすることができない．一方で，ビデオカメラでは，フレームレートの高い撮影を行うことはできるが，空間解像度を高くすることができない．この空間解像度と時間解像度のトレードオフを解決する手段の一つとして，圧縮センシングの中でも動画の復元タスクを行う圧縮ビデオセンシングによる手法が提案されている [3], [4], [5], [6]．

映像解析におけるデータ圧縮のトレードオフな問題に対し，圧縮センシングによる解決を考える．単に圧縮ビデオセンシングを考えた場合，符号化露光画像からビデオの再構成を行うことで，通常のビデオと同様に映像解析を行うことが可能である．自動監視システムでは，視野内の人間の不審な行動を検出または予測し，オペレータに警告する必要がある．そのため本研究では，映像解析として人間の行動認識に焦点を当てる．人間の行動認識に圧縮ビデオセンシングの適用を考えた場合，符号化露光画像からビデオの再構成という高次元化を行った後，ビデオから行動ラベルの推定という低次元化を行うことになり非効率である．そこで，本研究では，符号化露光カメラにより撮影される単一の符号化露光画像から Deep Learning を用いて直接人物の行動認識を行う手法の提案を行う．

2. 関連研究

2.1 圧縮ビデオセンシングを用いた時空間解像度のトレードオフの解決

圧縮ビデオセンシングは，画像センサの各ピクセルの露光タイミングをランダムにずらした画像を撮影することで，時間情報を単一画像にサンプリングする．この符号化露光画像に含まれる異なる時間の情報をもいて再構成処理を行うことにより，映像センサの時空間サンプリングを超えたビデオの再構成を可能にする．この圧縮ビデオセンシングのモデルは，未知の動的シーン x，符号化露光画像 y，符号化露光パターン ϕ を用いて次のように表される．

$$y = \phi x \tag{1}$$

圧縮ビデオセンシングでは，符号化露光画像 y から未知のビデオ y の再構成を行うが，符号化露光画像のデータ量は，符号化露光画像のサイズ $W \times H$ であり，露光時間を T とすると，未知のビデオのデータ量は $W \times H \times T$ となる．これは，観測した情報よりも多くの情報を復元することとなるため，一意に定めることはできない．そこで，Hitomi ら [3] や Sonoda ら [4] は，動画は，基底となる動画とそのスパースな係数で表現できると仮定するスパース最適化による再構成手法を用いて，観測した情報よりも十分少ない数の係数を求めることで，ビデオの再構成を行った．Hitomi らはスパース最適化手法として，L_0 ノルム正則化を行う Orthogonal Matching Pursuit (OMP) アルゴリズムを用いた．一般に，スパース最適化は NP 困難な問題であることが知られている．したがって，スパース最適化を用いた再構成手法は，膨大な時間を要するものであり，実用的な手法であるとは言えない．Yang ら [5] は，動画は，Gaussian Mixture Model (GMM) で表現可能であると仮定し，符号化露光画像が与えられた事後確率の期待値から動画を再構成する，より高速な手法を提案した．また，Iliadis ら [6] は，Deep Learning を利用し，符号化露光をエンコーダとする AutoEncoder を学習することで，符号化露光画像から動画を再構成するデコーダを作成し，より高速な再構成手法を提案した．

符号化露光パターンにも様々なものが提案されてきた．圧縮ビデオセンシングでは，各ピクセルでランダムなタイミングで露光された画像を撮影する必要がある．しかし，一般的な CCD や CMOS センサは，すべてのピクセルが同時に露光するグローバルシャッタや画素の読み出し順に露光を行うローリングシャッタが一般的であり，圧縮ビデオセンシングで必要なセンサは一般には存在しない．そのため，理想的なランダムな露光を想定したものや，ハードウェアの実装上の制限を考慮した符号化露光パターンが用いられている．Iliadis ら [7] は，ランダムな露光が可能な理想的なセンサを想定し，各画素の露光時間を 16 分割し，$4 \times 4 \times 16$ のランダムなパターンを繰り返した $8 \times 8 \times 16$ の符号化露光パターンとしたシュミレーション実験を行った．Hitomi ら [3] は，一般的な非破壊読み出しができない CMOS センサの構造を想定して，1 回の露光で開始と修了を任意とする単一露光の符号化露光パターンを提案し，7×7 のパターンを用いて，シュミレーション実験と反射光学系と Liquid Crystal on Silicon (LCoS) を用いた疑似実装による実験を行った．Sonoda ら [4] は，ピクセル毎に露光を制御可能なプロトタイプの CMOS センサを用いて，擬似ランダム露光な符号化露光を実現した．ハードウェアの制約から縦列，横列で同時に露光する 8×8 の符号化露光パターンを用いた実証実験を行った．Yoshida ら [8] は，これらの符号化露光パターンをハードウェアの制約を考慮した上で，符号化露光パターンとデコーダを Deep Learning を用いて同時に最適化する手法を提案した．

2.2 行動認識における特徴表現

かつては，行動認識に 3D モデルを利用していた．しかし，映像から正確な 3D モデルを構築することは難しいため，多くの場合，代わりに全体的または局所的なアクション

(a) Bobick と Davis [9] の MEI と MHI. **(b)** Blank ら [10] の Space-Time Shapes.

図 2: 行動認識に用いられていた全体を考慮する動きの表現

の表現を利用する手法が取られている．全体を考慮する表現では，人体の構造や形状，動きのグローバルな表現を用いており，動きに関する情報を単一の画像にエンコードする二値画像を累積した Motion Energy Image (MEI) や 輝度で時間を表す Motion History Image (MHI) [9] (図 2a)，オブジェクトの輪郭を時間軸に沿って積み重ねた Space-Time Volume (STV) [10] (図 2b) が提案された．全体を考慮したこれらのアプローチは，視点や外観の変化を捕捉するのが難しく，STV では細部を捉えることができない問題点があった．一方，局所領域を考慮する表現では，一般的な画像認識と同様に，関心点の検出，局所記述子の抽出，局所記述子の集約という手順に従い，行動認識のための局所特徴を作成する．時空間領域における関心点を検出して，2 次元の Harris コーナー検出器を 3 次元に拡張する Space-Time Interest Points (STIP) [11] が提案された．時空間の局所記述子として，Histograms of Oriented Gradients (HOG) [12] をモーション記述子として利用することが提案され [13]，また，ビデオクリップ内のピクセルレベルの動きをエンコードする Histograms of Optical Flow (HOF) が提案された．記述子の集約では，画像認識と同様に Bag-of-Features (BoF) [14] が用いられた．特にカテゴリー分類では，テキスト分類で高い評価を受けていた Support Vector Machine (SVM) が BoF ベクトルに対しても用いられるようになった [15].

画像認識の分野で，畳み込みニューラルネットワーク (CNN) が注目されるようになると，映像認識の分野でも CNN を用いられるようになった．CNN は，関心点の検出，局所記述子の抽出，局所記述子の集約のいずれの段階でも使用でき，画像フレームを特徴化するだけでなく，オプティカルフローや HOG などと組み合わせても使用された．Simonyan と Zisserman [16] は，RGB の画像フレームとオプティカルフローを蓄積したものをそれぞれ外観とモーション情報として用いることを提案し，また，2 つのストリームを結合することでさらなる精度向上を示した．UCF101 や HMDB51 などのデータセットにおいて Deep Learning を使用しないかつての認識精度を大幅に改善し，2 ストリームネットワークに基づく数多くの研究がなされてきた．一方，Tran ら [17] は 3 次元で畳み込むことで外観とモーションを同時にモデル化するネットワーク (C3D) を提案した．2 ストリーム 2D CNN に劣るものの大規模動画データセットである Sports-1M を用いて良い精度を達成した．Kay ら [18] は，行動認識の大規模化かつ校正されたデータセットである Kinetics を提案した．比較的小規模な 3D CNN において，事前学習なしのモデルでありながら校正されたデータで学習することにより，ImageNet で事前学習した 2D CNN に迫る精度を達成することを示した．Carreira と Zisserman [19] は，22 層の 2D CNN である GoogLeNet (Inception v1) [20] を 3D に拡張した I3D を提案し，Kinetics データセットを用いて学習し最先端の精度を達成した．

表 1: 提案手法のネットワークアーキテクチャ．8 層の 2 次元畳み込みと 5 層の最大値プーリングと 2 層の全結合層から構成される単純な 2 次元の CNN.

layer	kernel size/stride	output size	params
Conv 2D (1a)	3×3/1	112×112×64	640
Max Pool (1)	2×2/2	56×56×64	
Conv 2D (2a)	3×3/1	56×56×128	73,856
Max Pool (2)	2×2/2	28×28×128	
Conv 2D (3a)	3×3/1	28×28×256	295,168
Conv 2D (3b)	3×3/1	28×28×256	590,080
Max Pool (3)	2×2/2	14×14×256	
Conv 2D (4a)	3×3/1	14×14×512	1,180,160
Conv 2D (4b)	3×3/1	14×14×512	2,359,808
Max Pool (4)	2×2/2	7×7×512	
Conv 2D (5a)	3×3/1	7×7×512	2,359,808
Conv 2D (5b)	3×3/1	7×7×512	2,359,808
Zero Padding		8×8×512	
Max Pool (5)	2×2/2	4×4×512	
FC (6)		4096	33,558,528
FC (7)		4096	16,781,312
Softmax		6	
Total			59,559,168

3. 提案手法

ビデオ監視システムにおける人間の行動認識でのデータ圧縮のトレードオフな問題に対し，圧縮センシングの適用を考える．単に圧縮ビデオセンシングの適用を考えた場合，第 2.1 節で述べた通り，符号化露光画像から動画の再構成を行うことが可能である．再構成した動画から通常のビデオと同様に行動認識を行うことができる．しかし，これは符号化露光画像という低次元のものから動画という高次元なものを再構成する NP 困難な問題を解き，不確実性の残る動画を元に行動認識を行うことになる．そもそも撮影された符号化露光画像には時間情報が含まれているため，動画の再構成を介さなくても直接，行動認識を行えると考えられる．そこで，本研究では，符号化露光カメラにより撮影される単一の画像から 2 次元の CNN を用いて直接，人

物の行動認識を行う手法を提案する．

本提案手法には，通常の行動認識に対して下記のような利点がある．

I. データ量が削減できる．

提案手法では，画像センサの各ピクセルをランダムに露光可能なセンサを用いて符号化露光画像を撮影する．この符号化露光の長さ分だけデータ量を圧縮することが可能である．これにより，通信量の削減や伝送にかかる消費電力の削減が期待される．

II. 撮影と同時に圧縮され効率的である．

通常の圧縮手法では，カメラで動画を撮影した後に圧縮処理を施す．一方，提案手法では画像センサの各ピクセルをランダムに露光し符号化露光画像を撮影し，動画の再構成に十分な情報を圧縮して取得するため非常に効率的であると考えられる．そのため，動画の圧縮処理にかかる電力などのコスト削減が期待される．

III. 3 次元の畳み込みが不要となる．

第 2.2 節で述べたように，近年，動画認識では 3 次元畳み込みによる時空間の特徴化で精度を改善している．これらのネットワークは，通常ネットワークの層が多く大規模なものが多く，ネットワークのパラメータも多くなる．また，ネットワークを十分に学習させるために必要なデータ数が増えるため，データセットも大規模なものが必要となる．したがって，大規模 GPU クラスタなどの演算資源が学習時に必要となり，学習にかかる時間は膨大なものとなる．一方，提案手法では符号化露光画像を入力とすることで，3 次元の畳み込みを必要とせず，2 次元畳み込みで時空間の特徴化が可能となる．3 次元畳み込みによるものに比べ，必要となるパラメータ数が減少しデータ数があまり大きくないものでも学習が可能となる．

提案手法のネットワークアーキテクチャとして，表 1 のような単純な 2 次元の CNN を考える．3×3 のストライド 1 の 8 層の 2 次元畳み込みと 2×2 のストライド 2 の 5 層の最大値プーリングと 2 層の全結合層から構成される．計算の簡略化のために bias 項を無視するとある畳み込み層のパラメータ数 P は，その層の入力チャネル数 C_{in} と出力チャネル数 C_{out} とカーネルサイズ K を用いて，

$$P = C_{in} \times C_{out} \times K \tag{2}$$

と表される．したがって，Tran ら [17] が最も良いとする 3 次元畳み込みのカーネル $3 \times 3 \times 3$ をすべての畳み込み層で用いた場合，本提案手法は 2 次元畳み込みのカーネル 3×3 であり，畳み込み層のパラメータ数に関しておよそ 1/3 となる．

提案手法のネットワークは次にように学習，評価を行う．K 種類の行動 $\boldsymbol{C} = \{C_1, C_2, \ldots, C_K\}$ についてのクラス分類を行うとする．ある行動 $a \in \boldsymbol{C}$ における長さ N の動画

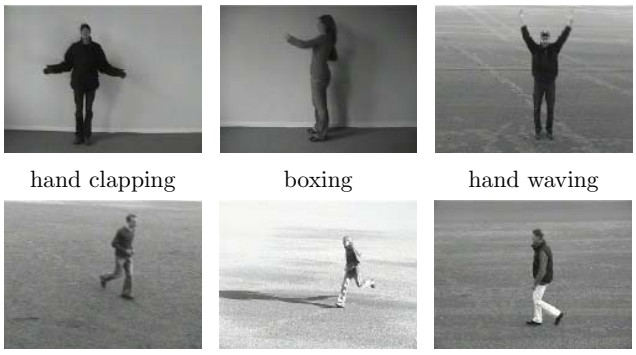


図 3: KTH Human Action データセット．各行動クラスの 1 例．

を $\boldsymbol{I} = \{I_1, I_2, \ldots, I_N\}$ とする．符号化露光パターンの長さを L とするとビデオクリップの長さは L であり，ビデオクリップは，

$$\begin{aligned} \boldsymbol{V} &= \{V_1, V_2, \ldots, V_{\lfloor \frac{N}{L} \rfloor}\} \\ V_i &= \{I_i, I_{i+1}, \ldots, I_{i+L-1}\} \end{aligned} \tag{3}$$

で表される．ビデオクリップに符号化露光パターンを適用し，

$$X_i = \phi V_i \tag{4}$$

とする．$\boldsymbol{I}$ に対して，$\{(X_i, a)\}$ のペアを用いてネットワークを学習する．各入力 X_i に対する出力 Y_i を動画全体で平均し，最大値を取ったものを動画における行動ラベルとして評価を行う．すなわち，ある時点での入力 X_i が行動 C_j に属する確率 $p(C_j|X_i)$ は，

$$p(C_j|X_i) = Y_{ji} \tag{5}$$

となるので，$\boldsymbol{I}$ に対して推定される行動ラベル a^* は，

$$a^* = \underset{c \in \boldsymbol{C}}{\operatorname{argmax}} \left(\frac{1}{\lfloor \frac{N}{L} \rfloor} \sum_{i=1}^{\lfloor \frac{N}{L} \rfloor} p(c|X_i) \right) \tag{6}$$

で表される．データセットの動画総数を M として，認識精度 (Accuracy) S は次のように計算される．

$$\begin{aligned} S &= \frac{1}{M} \sum_{j=1}^{M} f(a_j, a_j^*) \\ f(a, a^*) &= \begin{cases} 1 & (a = a^*) \\ 0 & (otherwise) \end{cases} \end{aligned} \tag{7}$$

4. 評価実験

符号化露光画像から直接、行動を認識するシュミレーション実験を行った．

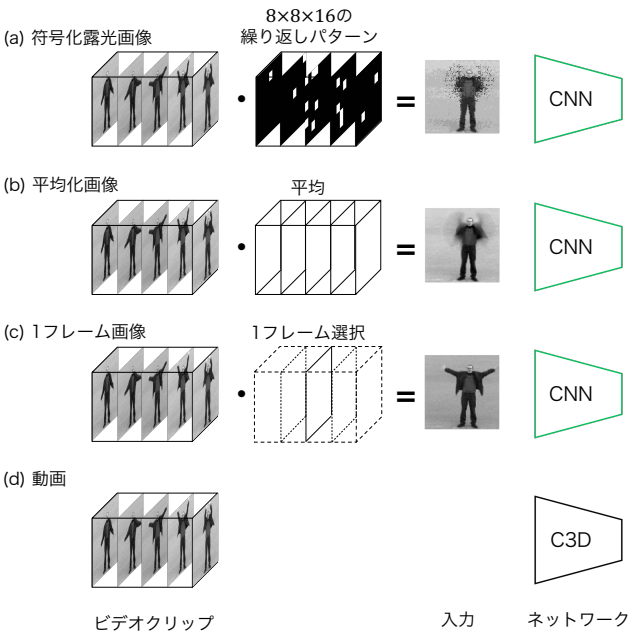


図 4: 比較手法の一覧．(a) ビデオクリップに符号化露光パターンを適用し符号化露光画像を生成し入力とする．(b) ビデオクリップを時間平均化し入力とする．(c) ビデオクリップから 1 フレーム選択し入力とする．(d) 動画を入力とし C3D で学習する．

4.1 データセット

シュミレーション実験には，KTH Human Action データセット [21] (図 3) を用いた．このデータセットは，固定されたカメラで，25 人の被験者による”walking”，”jogging”，”running”，”boxing”，”hand waving”，”hand clapping” の 6 種類の行動を 4 つのシナリオで撮影したもので，平均で 4 秒の 600 のグレースケールのビデオがある．25 fps で撮影され，160 × 120 の空間解像度にダウンサンプリングされている．Schuldt ら [21] の分割手法に従い，被験者を訓練で 8 人，検証で 8 人，テストで 9 人に分割し使用した．

4.2 比較手法

各動画は，学習時に重複なしの 16 フレームのビデオクリップに分割され，112 × 112 の空間解像度にランダムに切り抜きを行った．このビデオクリップもしくは，このビデオクリップに対して圧縮処理を施したものを入力として学習を行った．圧縮処理は，それぞれ 1/16 のデータ量に相当するよう圧縮した (図 4)．それぞれの手法は下記の通りである．

(a) 符号化露光画像
ビデオクリップに符号化露光パターンを適用し，符号化露光画像を生成し，これを入力とした．符号化露光パターンは，8 × 8 で各ピクセル 1/16 で露光するランダムなパターンを使用した．動画に対して 16 分の 1 のフレームレートで，各ピクセルの露光時間は符号化露光パターンによって変化する．この実験で用いた符号化露光パターンでは，露光時間は動画 1 フレームを撮影する露光時間と等しい (図 5(a))．

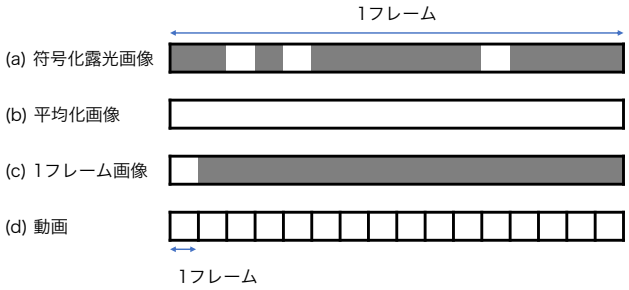


図 5: あるピクセルにおける露光の様子．格子は 1 フレームを意味し，白は露光，黒は露光していないことを意味する．(a) 符号化露光画像では，1 フレームを L 分割しそれぞれを露光するか否かランダムに決定する．(b) 平均化画像では，露光時間を動画 1 フレームの露光時間の L 倍として，フレームレートを $1/L$ で撮影する．(c) 1 フレーム画像は，L フレームのビデオクリップの 1 フレームである．

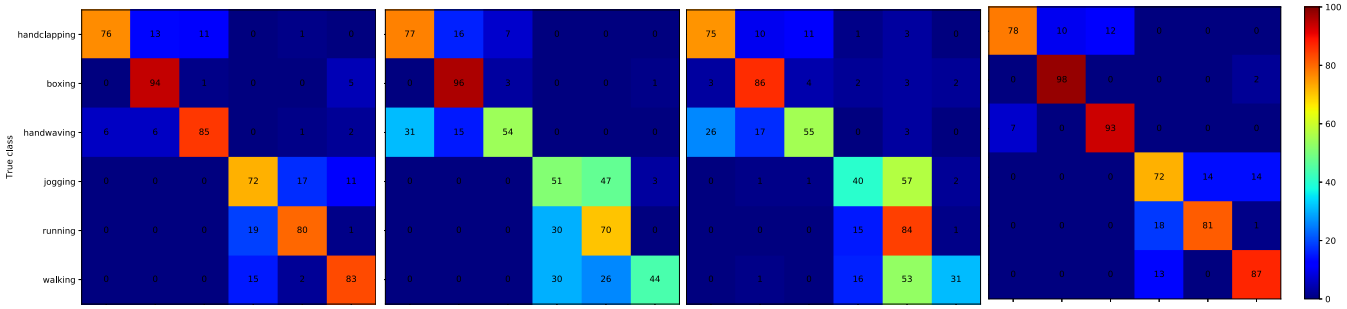
(b) 平均化画像
時間情報を 1 枚の画像に圧縮する単純な手法として．ビデオクリップを時間方向に平均化した平均化画像を用いた．平均化画像は，16 分の 1 のフレームレートで露光時間が 16 倍の動画の 1 フレームと等しい (図 5(b))．

(c) 1 フレーム画像
時間情報を持たない画像と比較するために，1 フレームの画像と比較した．ビデオクリップの 16 フレームのうち 1 フレーム選択し，これを入力とした．1 フレーム画像は，16 分の 1 のフレームレートで露光時間が等しい動画の 1 フレームと等しい (図 5(c))．

(d) 動画
ビデオクリップを入力とし，C3D [17] で学習した．C3D は，本来 RGB の 3 チャネルであるが，グレースケールの 1 チャネルに変更し，事前学習無しで学習した．

4.3 実験結果

実験結果を表 2 に，それぞれの混同行列を図 6 に示す．符号化露光画像以外を入力した場合は，動画を入力した場合の場合と比べ，著しく認識精度が低下した．図 6 の混同行列から，平均化画像を入力としたものは 1 フレーム画像を入力と同様に”hand waving”の認識精度の低下や”walking”，”jogging”，”running”の区別がついていないことが分かった．一方，符号化露光画像を入力としたもの



(a) 符号化露光画像　(b) 平均化画像　(c) 1 フレーム画像　(d) 動画

図 6: 混同行列．縦軸が真値，横軸が推定値を示している．数値はパーセンテージを表している．

表 2: シュミレーション実験の結果．

	認識精度 (Accuracy) [%]
(a) 符号化露光画像	81.57
(b) 平均化画像	65.38
(c) 1 フレーム画像	61.74
(d) 動画	84.94

は，動画を入力したものと同様の傾向を示し，認識精度についても動画を入力とした場合の認識精度に迫る高い精度を達成した．

ビデオクリップの長さ L を変化させ，認識精度がどのように変化するかシュミレーション実験を行った．その結果を図 7 に示す．C3D は 16 フレームのビデオを入力とするため，16 フレーム未満では $16/L$ 回同じフレームを繰り返すことで 16 フレームの動画にし入力した．また，16 フレームより多い場合は，C3D の入力フレーム数を L に変更した．そのため C3D を用いたものは，ネットワークの表現力の向上やデータセットの不足により，公正な比較ができないことに注意されたい．平均化画像を入力とするものは，ビデオクリップの長さが 4 フレームから 8 フレームまでは若干の精度改善が見られたが，ビデオクリップの長さを長くしていくと認識精度は低下した．これは，1 フレーム画像を入力としたものと同様の傾向を示しており，時間を平均化すると時間情報が失われていくと考えられる．一方，符号化露光画像を入力とするものは，ビデオクリップの長さが 16 フレームまでは認識精度が改善した．これは動画と同様の傾向を示しており，行動認識に必要な時間情報を十分に有していることが考えられる．しかし，符号化露光画像を入力としたものは，ビデオクリップの長さが 16 フレームより長くなると認識精度が低下した．これは特徴化しなければならない時間情報が増え，今回用いた符号化露光パターンでは時間情報を十分に表現しきれなくなったことが考えられる．

5. おわりに

本研究では，ビデオ監視システムにおける行動認識のトレードオフな問題に対し圧縮センシングを適用し，符号化露光カメラにより撮影される単一の画像から単純な 2 次元の CNN を用いて直接，人物の行動認識を行う手法の提案を行った．提案手法の有効性を評価するため，KTH Human Action データセットを用いたシュミレーション実験を行い，本提案手法は入力データを 1/16 に圧縮しているにもかかわらず，動画を入力とした 3 次元の CNN に迫る高い精度を達成した．

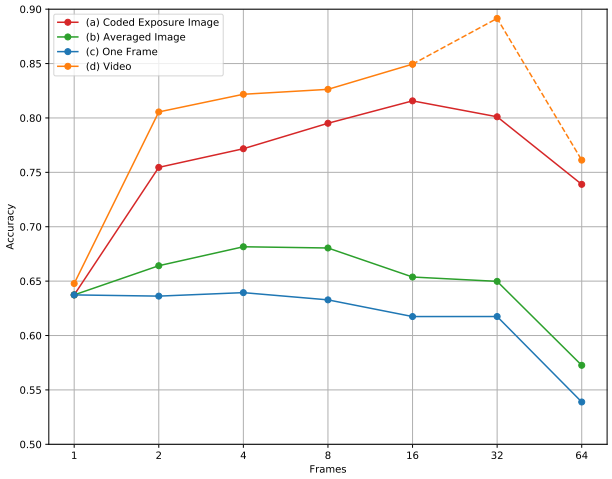


図 7: ビデオクリップのフレーム数に対する認識精度の変化．

今回用いた符号化露光パターンでは 16 フレームより長いフレームを圧縮する場合，時間情報を表現しきれなくなり認識精度の低下が見られた．多くのフレームを特徴化する符号化露光パターンや符号化露光パターンに特化した CNN のアーキテクチャの構築が今後の課題である．

参考文献

[1] Rty, T. D.: Survey on Contemporary Remote Surveillance Systems for Public Safety, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 40, No. 5, pp. 493–515 (2010).

[2] Li, Y., Ai, H., Yamashita, T., Lao, S. and Kawade, M.: Tracking in Low Frame Rate Video: A Cascade Particle Filter with Discriminative Observers of Different Life Spans, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 10, pp. 1728–1740 (2008).

[3] 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).

[4] Sonoda, T., Nagahara, H., Endo, K., Sugiyama, Y. and Taniguchi, R.: High-speed imaging using CMOS image sensor with quasi pixel-wise exposure, *Proceedings of IEEE International Conference on Computational Photography (ICCP)*, pp. 1–11 (2016).

[5] Yang, J., Yuan, X., Liao, X., Llull, P., Brady, D. J., Sapiro, G. and Carin, L.: Video Compressive Sensing Using Gaussian Mixture Models, *IEEE Transactions on Image Processing*, Vol. 23, No. 11, pp. 4863–4878 (2014).

[6] 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).

[7] Iliadis, M., Spinoulas, L. and Katsaggelos, A. K.: Deepbinarymask: Learning a binary mask for video compressive sensing, *arXiv preprint arXiv:1607.03343* (2016).

[8] 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)* (2018).

[9] Bobick, A. F. and Davis, J. W.: The recognition of human movement using temporal templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257–267 (2001).

[10] Blank, M., Gorelick, L., Shechtman, E., Irani, M. and Basri, R.: Actions as Space-Time Shapes, *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 1395–1402 (2005).

[11] Laptev, I.: On Space-Time Interest Points, *International Journal of Computer Vision*, Vol. 64, No. 2, pp. 107–123 (2005).

[12] Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1, pp. 886–893 vol. 1 (2005).

[13] Klaser, A., Marszalek, M. and Schmid, C.: A Spatio-Temporal Descriptor Based on 3D-Gradients, *Proceedings of British Machine Vision Conference (BMVC)* (Everingham, M., Needham, C. and Fraile, R., eds.), Leeds, United Kingdom, British Machine Vision Association, pp. 275:1–10 (2008).

[14] Csurka, G., Dance, C. R., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 1–22 (2004).

[15] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B.: Learning realistic human actions from movies, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8 (2008).

[16] Simonyan, K. and Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos, *Advances in Neural Information Processing Systems (NIPS)* (Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. and Weinberger, K. Q., eds.), Curran Associates, Inc., pp. 568–576 (2014).

[17] 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).

[18] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P. et al.: The kinetics human action video dataset, *arXiv preprint arXiv:1705.06950* (2017).

[19] 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. 4724–4733 (2017).

[20] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9 (2015).

[21] Schuldt, C., Laptev, I. and Caputo, B.: Recognizing Human Actions: A Local SVM Approach, *Proceedings of International Conference on Pattern Recognition (ICPR)*, Washington, DC, USA, IEEE Computer Society, pp. 32–36 (2004).