深層学習による動画データからの手元動作認識

Hand Motion Recognition from Movie Data by Deep Learning

東京大学	出町　和之	Kazuyuki	DEMACHI	Member
東京大学	陳　実	Shi	CHEN	Student-Member

A deep learning model has been proposed to recognize hand action for nuclear security. A system has been developed that can automatically recognize hand action from video data acquired by a single depth camera.

Keywords: Deep Learning, Convolutional Neural Network, Action Recognition, Behavior monitoring

1．はじめに

福島第一原子力発電所事故は、原子力施設が潜在的にテロの魅力的なターゲットとなる可能性があることを示唆した。IAEA(国際原子力機構)によると、核テロの脅威は Fig. 1 に示す 4 種類[1]：核兵器の盗取、核物質の盗取、ダーティーボムの製造、妨害破壊行為に分類されるが、その中でも原子力発電所にとっての最たる脅威は妨害破壊行為である。さらに、原子力施設関係者を意味する内部脅威者は、枢要区域へのアクセス権や専門知識を有し、妨害破壊行為者として一層の注意が必要である。

Fig. 1 Nuclear security defined by IAEA

内部脅威者の妨害破壊行為は通常の保全作業等に紛れて行われるが、原子力施設の膨大な数の作業員すべてを人の眼で監視することは現実的でなく、「技術の眼」による監視の補完が必要である。とくに、人の動作の大部分を占める手元動作の自動認識は開発が急がれる技術である。

手には関節が密集しているため手元動作は複雑となり、自動認識には難易度が高い。一般的な手元動作認識の手法の多くは RGB-D カメラという距離画像(D)を得られる特殊なカメラを使用している。しかし、実際に原子力施設やセキュリティの現場で主に使われているのは RGB（シングルデプス）カメラであり、これらに使用できる手元動作認識技術を新たに開発する必要がある。

そこで本研究では、シングルデプスカメラで取得した動画データから手元の動作を自動認識する手法を、深層学習の一種である畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) [2] を用いて開発した。

連絡先:出町和之、〒113-8654 東京都文京区本郷 7-3-1,
東京大学大学院工学系研究科原子力専攻,
E-mail: demachi@n.t.u-tokyo.ac.jp

2.　手法

2.1 提案する動作認識アルゴリズム

今回の認識対象とした手元動作は、手を握る、手を開く、親指を上げる、などを含む 6 種類である。Fig. 2 に、本研究で提案した手元動作認識のフローを示す。

まずは対象となる動画データの各フレームから手元を認識し、動画データを手の関節座標のみのデータに変換する。これによりカメラと手元の位置関係による影響を排除できる。このようにして得られた手の関節座標の連続データを 2 次元配列し、CNN による深層学習で、あらかじめ設定した 6 種の手元動作のどれかを識別する。

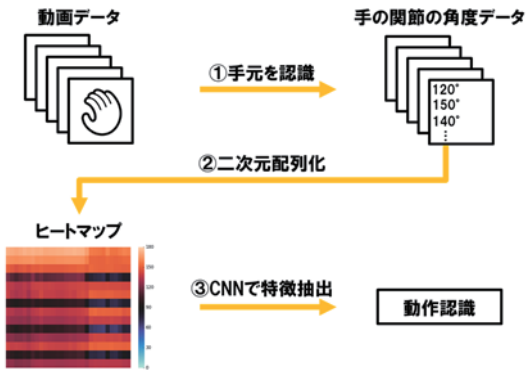

Fig. 2 Proposed Algorithm

2.2 手の関節角度のヒートマップ取得

動画データからの手元データ抽出には、オープンソースプログラムの hand3d[3]を用いた。これにより、シングルデプスカメラで得たデータから片手当たり関節など 26 箇所の三次元座標を取得できる。得られた三次元座標から、各指の第一関節、第二関節および付け根の片手当たり計 14 箇所の角度を解析的に計算した。

次に、撮影された動画のフレーム毎に 14 箇所の関節の角度を計算し、これを 1 次元配列とした。1 つの動画は 1 秒当たり 30 フレームで構成されるため、関節角度の 1 次元配列を横に結合することで、1 秒の動画から 14 行×30 列の行列データが作成できる。これを、動作識別のための CNN における学習データとした。なお、過学習による識別精度の低下を避けるため、各々の 14 行×30 列の行列データにランダムノイズを付与し、1 つの動画あたり 20 個の学習データを作成した。Fig. 3 は 1 秒間の関節角度行列データをヒートマップとして図示した例である。

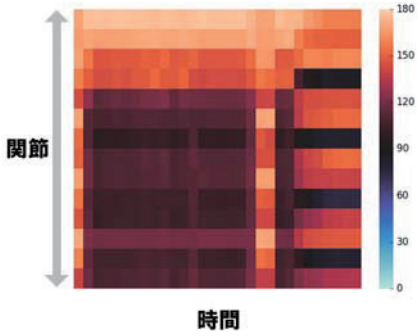

Fig. 3　Heat map of joint angle matrix for 1 second

2.3 畳み込みニューラルネットワーク (CNN)

時系列データ解析においては回帰ニューラルネットワーク(Recurrent Neural Network, RNN) [4]を用いるのが一般的である。しかし人間の動作は早さも動き方も一意ではなく、その差異に対するロバストな認識が求められる。CNN[2]ではデータの絶対値ではなく画像としての構造や分布から特徴抽出を行うため、要件となるロバスト性を満たすことが期待される。

Fig. 4 に、今回用いた CNN のアーキテクチャを示す。畳み込み層は二つあり、フィルタのサイズは3×3 とした。また、畳み込みにおけるストライドの幅は 1 とした。特徴抽出の後、分類のために全結合層を三層とした。活性化関数には ReLU 関数を用いた。

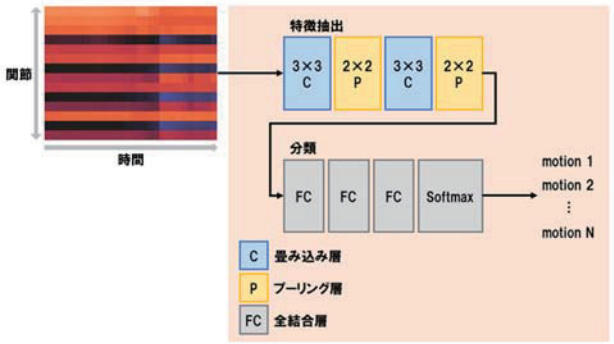

Fig. 4　Proposed CNN architecture

3.　手元動作推定結果

学習後の CNN を用いた手元動作認識結果の正解率を Table 1 に示す。全動作の平均は 0.899 であった。比較的高い正解率が得られたものの、更なる向上が必要である。

Table 1　Proposed CNN architecture

CloseThreeFingers	0.901
CloseThumb	0.895
OpenThumb	0.895
PushingWithOneFinger	0.904
ZoomingInWIthFullHand	0.902
ZoomingOutWithFullHand	0.897

4.　結論

原子力施設における内部脅威者の妨害破壊行為対策のための、動画データから手元動作を認識するための手法を開発した。今後は、学習用データの増加、手首や全身を組み合わせた認識技術の実装、アーキテクチャの改良による高速処理の実現などが課題となる。

References

[1] N. E. R. Headquarters, “Government of japan. 2011. Report of the japanese government to the iaea ministerial conference on nuclear safety the accident at tepcos fukushima nuclear power stations. Attachment xi-1,” 2011.

[2] A. Krizhevsky, I. Sutskever and G. E. Hinton “ImageNet Classification with Deep Convolutional Neural Networks”, Neural Information Processing Systems (NIPS2012)

[3] Christian Zimmermann and Thomas Brox, Learning to Estimate 3D Hand Pose from Single RGB Images. University of Freiburg, 2017.

[4] T. N. Sainath, O. Vinyals, A. Senior and H. Sak, “Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015)