F-012

高精度シーン認識と高安全知能システムへの応用
High-Accuracy Scene Recognition and Its Application to Highly-Safe Intelligent Systems

高田 健一† 亀山 充隆‡
Kenichi Takada Michitaka Kameyama

1. はじめに

人間の生活環境での危険を事前に検出し，警報・危険回避支援などを自動で行う高安全知能システムの実現が望まれている[1]．この高安全知能システムは，人間が生活する実世界環境の情報を入力し，環境の認識，危険要素の検出・予測を行い，危険回避行動計画を警報等で出力するシステムであり，それを実現するためには入力情報から危険事象等を正確に認識する技術，すなわち，シーンの高精度な認識が必要不可欠となる．本稿では，畳み込みニューラルネットワーク（CNN）による画像の特徴抽出と種々の識別手法を最適に組み合わせることにより，このシーン認識が高精度に実現可能となることを明らかにする．

2. CNN と他の識別手法を組合せたシーン認識

人間の生活環境には，様々な危険事象が存在しており，高安全知能システムを構築するにあたっては，目の前で発生している事象（シーン）をいかにして高精度に認識できるかが重要となる．

2.1 シーン認識の方法

強盗シーンの認識を例にとると一般的には凶器の検出など難しい問題をクリアしなければならないが，人間は画像を一見して「強盗」であると認識できる．これをシーン認識として機械学習により実現できれば、高安全知能システムへの応用も可能となる．

その手法であるが、現在、物体認識の精度が人間を上回るレベルに達している CNN を利用する．CNN は多層構造をしたニューラルネットワークで，その学習には膨大なラベル付き学習用画像，時間，高性能なマシンなどが必要となり，手軽にはシーン認識に利用することはできない．しかし，既に大量のデータを学習済みの CNN が公開されており利用できる．学習済み CNN は物体カテゴリ認識用の識別器であるためそのまま画像を入力してもシーン認識はできないが，CNN の出力を画像の特徴量として利用することは可能であり[2]， その特徴量を図 1 のように他の機械学習で学習することによりシーン認識を行う組合せモデル(シーン認識モデル)を検討する．

2.1.1 シーン認識モデルの構成

本稿で使用する学習済み CNN は，ディープラーニングのフレームワークである Caffe の bvlc_googlenet.caffemodel である．この学習済み CNN は、GoogLeNet の複製で，1000 種類の物体カテゴリ認識が可能である[3]．これを特徴抽出器として使用し，出力された特徴をベイジアンネットワーク(BayesNet)やサポートベクトルマシン(SVM)などの他の識別手法により学習することでシーン認識を行う．

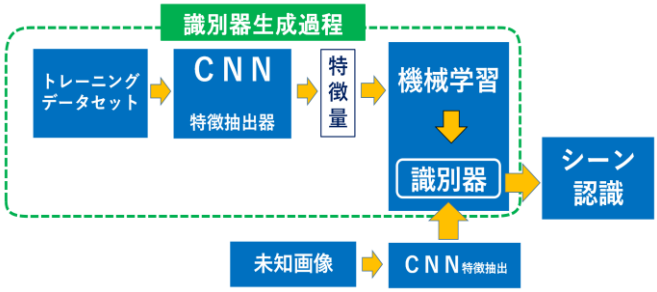

図 1　シーン認識モデルの概念図

2.1.2 シーン認識モデルの課題

このモデルにより，高精度のシーン認識を実現するための課題は次のとおりである．

(1) 特徴量の選択
　多階層ネットワーク構造である CNN のどの層の値が特徴量として有効なのか．

(2) CNN と組合せる識別手法の選択
　最も CNN の出力値（特徴量）と親和性の高い識別手法は何か．

(3) トレーニングデータ数と学習度合い
　どの程度の規模のトレーニングデータ数を用意すれば十分な学習ができるのか．

これらを評価実験により検討する．

3. シーン認識の評価実験と高安全知能システムへの応用

本実験では，高安全知能システムへの応用を念頭に，以下の 3 シーンを取り上げる．

(1) 強盗シーン
　コンビニエンスストアや金融機関などにおいて発生した強盗を認識し，関係機関への通報を行うことを想定している．
　データは，正例としてコンビニエンスストアや金融機関などでの強盗シーン 175 枚．負例としてレジ等で会計しているシーン 175 枚である．

(2) 笑顔のシーン
　表情を知ることで，その人の生理的・精神的状態を推定し，健康管理支援などへの応用を想定している．
　データは，正例として老若男女問わず笑顔 120 枚．負例として同じく泣顔・無表情などのシーン 120 枚である．

(3) 人物転倒シーン
　路上，駅のホーム，室内などで転倒・転落・急病などが発生していることを検知し，関係機関への通報を行うなどの緊急時対応を想定している．

† 石巻専修大学理工学研究科　Graduate School of Science and Technology, Ishinomaki Senshu University
‡ 石巻専修大学理工部　Faculty of Science and Technology, Ishinomaki Senshu University

データは，正例として屋内・屋外などで人が倒れている・苦しんでいるなどのシーン 600 枚．負例として危険・緊急性を感じさせない単に横になっているなどのシーン 600 枚である．

これらの画像データを用いて評価実験を行い，それにより得られた結果からモデルの有効性及び高安全知能システムへの応用に対するシーン認識の有用性を検証する．

なお，本実験で使用した画像は，全て Web 上から無作為に関連するシーンを収集し，教師信号（クラス）を付与したものである．

図 2　強盗・転倒シーンのサンプル画像

3.1 CNN から抽出する特徴量

抽出する特徴量は，入力画像が十分に抽象化され有効な特徴が出現していると考えられるネットワークの最終出力に近い層の値を使用するのが最も妥当である．本実験では GoogLeNet のカテゴリ分類結果(Softmax 関数の出力) 及び“AveragePool 7x7+1(V)”層の出力値を特徴量としたが，図 3 に示すとおり，結果としては，前者の特徴(以下「特徴量 F1」という)に比べて後者の特徴(以下「特徴量 F2」という)を使用した方が高い正解率を得られた．

3.2 CNN と組合せる識別手法

CNN と組合せる識別手法（機械学習）について，選定した識別手法は一般的に画像認識などで使われている以下の 4 種類とした．

(1) ベイジアンネットワーク（BayesNet）
(2) アダブースト（AdaBoost）
(3) サポートベクトルマシン（SVM）
(4) ニューラルネットワーク（NN）

3.3 評価実験の結果

3.3.1 強盗シーン

特徴量及び CNN と組合せる識別手法の選定について明らかにするため，強盗シーンを用い実験を行った．

なお，識別能力の評価方法は，トレーニングデータセットが比較的小規模なことから，10 分割の交差検証法による正解率により行った．

実験結果を図 3 に示す．SVM が特徴量 F2 で最も正解率が高く 98.0%であった．

また，どの程度の規模のトレーニングデータ数であれば十分な学習が可能であるかを明らかにするため，データ数の増加による正解率の推移から，十分な学習ができたと考えられるデータ数を調べた．その結果，強盗シーンの認識おいてはトレーニングデータ数が 400 枚程度で正解率が収束した．

3.3.2 笑顔のシーン

特徴量は F2 とし，組合せる識別手法は，強盗シーンの認識評価と同様の条件により実験を行った．結果は表 1 のとおりで，SVM が最も正解率が高く 99.6%であった．

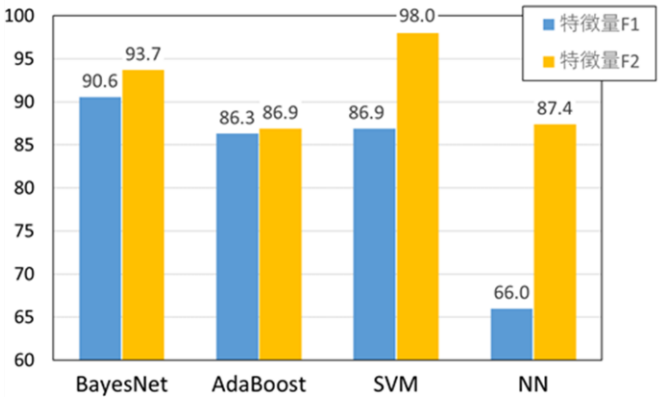

図 3　強盗シーン認識の評価結果

3.3.3 人物転倒シーン

特徴量は F2 とし，組合せる識別手法はこれまでの評価結果を考慮し SVM のみにより実験を行った．また，本実験では，テストデータセット（正例 40，負例 40）による評価も行った．結果は表 2 のとおりで，トレーニングデータセットによる正解率が 98.2%，テストデータセットによる正解率は 96.7%であった．

表 1　笑顔のシーンの認識評価実験結果

	BayesNet	AdaBoost	SVM	NN
正解率(%)	96.3	90.0	99.6	91.3

表 2　人物転倒シーンの認識評価実験結果

	正解率(%)
トレーニングデータセット	98.2
テストデータセット	96.7

4. おわりに

本稿では，強盗シーン等の評価実験において CNN の最終プーリング層の出力を特徴量とし，それを SVM で学習したとき最も高い正解率が得られるという結果から，シーン認識モデルの構成を明らかにした．また，このモデルは，比較的小規模なトレーニングデータセットでも十分な学習が可能である．

以上のことから，高安全知能システムへの応用が期待できる高精度シーン認識が可能であることが明らかとなった．

なお，本稿でのシーン認識は，正例・負例を識別するという 2 値分類であったが，今後は，多クラス分類への拡張も必要となる．また，動画などリアルタイムでの処理が要求されるような場面でのシーン認識の可能性についても興味ある課題である．

参考文献

[1] 大河原茂樹，亀山充隆,"危険要素抽出に基づく高安全システムの高品質化と VLSI プラットフォーム"，FIT 情報科学技術フォーラム,C-017(2015).
[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database”, In Neural Information Processing Systems(NIPS), pp.487–495 (2014).
[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D.Anguelov, D. Erhan, V.Vanhoucke, and A. Rabinovich, “Going deeper with convolutions”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1-9 (2015).