Vol.2017-CVIM-206 No.1
2017/3/9

Curriculum Learningを用いたネットワーク群による効率的な大規模動画像検索

松本 泰幸[1]　篠崎 隆志[2,3]　白浜 公章[4]　上原 邦昭[1]

概要：画像認識分野において、学習済みの畳込みニューラルネットワーク（CNN）を用いた特徴抽出器としての効率的な利用が知られている。一方、抽出した特徴量を用いた識別器の学習には計算コストがかかる．特に，識別概念が大規模な場合には問題となる。本稿では、抽出した特徴を、小規模なネットワークに学習させ、さらに Curriculm Learning (段階的な転移学習) を行えば、従来の識別器よりも高速かつ、認識する概念が大規模な場合に対しても柔軟な認識が可能となる手法を紹介する。

1. はじめに

近年，ネットワークの高速化や記憶装置の大容量化などに伴い，多種多様かつ大量の動画像が流通し，蓄積されている．同時に，動画共有サービスの普及により，多くの動画像が選択，視聴できる状況にある．このような動画共有サービスで，大量の動画像の中から欲しい情報に自在にアクセスするためには，動画像の内容に基づく検索やブラウジングが必要となる．こうした検索手法として，動画像に対して与えられた，アノテーションに基づく手法が考えられる．しかし，アノテーションの付与には，労力，作成する際の恣意性，主観性などの課題があり，適切に検索が実現されるのは困難である．

この問題を解決するためには，動画から自動でアノテーションを生成することが有効であり，これによって恣意性や主観性を排除しつつ，より大規模で時間粒度の細かい検索が可能となると期待されている．近年の深層学習の発展によって静止画像の自動認識の精度が飛躍的に向上し，現在その技術は動画へと浸透し，動画の自動アノテーションがまさに実用可能になりつつある．

深層学習とは，多層のニューラルネットワーク（Neural Network，以下 NN）を用いる機械学習の手法である．NN は，従来，NN の規模，特に層の数が大きくなれば過学習が避けられなかったが，Dropout による正則化で回避する手法が提案され，多層 NN の研究が加速している [1].

多層 NN のうち，画像認識の応用では畳み込みニューラルネットワーク（Convolutional Neural Network，以下 CNN）が有名である [2]．CNN は，畳み込み層とプーリング層と呼ばれる 2 種類の層を交互に積み重ねた，構造を持つ多層 NN である．CNN は，多層であっても事前学習（Pre-Training）を必要とせず，最初から教師なし学習を行えるという利点がある．

一方，CNN には学習時と異なるタスクに対しては，識別が難しいという問題がある．つまり，CNN はすべてのコンセプト（概念）[*1]について一般化されるわけではなく，学習時に含まれないコンセプトの識別には知識転移，すなわち，学習により獲得した情報を，別の学習につなげる工夫が必要となる．そこで，近年，注目されている手法の一つとして，学習済みネットワーク（Pre-trained network）を，特徴抽出器として用いる転移学習（Transfer learning）がある．これは，学習済みネットワークの中間層の出力を特徴ベクトルとして取り出し，サポートベクターマシン（Support Vector machine，以下 SVM）などの識別器に学習させて，目的とするコンセプトの識別を可能とする手法である．

しかし，従来の SVM などの学習では，計算コストが大きくなり，識別するコンセプトが大規模になれば，大量の学習データを取り扱うために，計算コストが障害となることが指摘されている．一方，NN は GPU を用いた最新のフレームワークが使用可能であるため、従来の SVM などに比べて数十倍以上の速度の学習が可能である．そこで本稿では，従来の識別器の代わりに，小規模な NN である micro Neural Network（microNN）を用いる手法を提案する．microNN は，GPU による高速化に加え，構造が単純

1　神戸大学大学院システム情報学研究科
2　情報通信研究機構 脳情報通信融合研究センター
3　大阪大学大学院情報科学研究科
4　ドイツ ジーゲン大学 パターン認識グループ

*1　物体や人，動作，風景など

ため，より高速な学習が実現できる．高速化によって，大規模な学習データセットに対しても柔軟な学習が可能となる．本稿では，この NN の柔軟性を利用して，Curriculum Learning (段階的な転移学習) を行い，多数のコンセプトや大規模なデータセットに対しても識別，検索が可能となる手法について述べる．

2. 動画像認識のための深層学習

動画像を静止画像の連続として扱えば，その時間的変化から，移動物体を見つけたり，カメラの動きを推定することが可能である．つまり，動画像認識は時系列情報を考慮した画像認識と考えられる．本節では，深層学習による動画像認識では，時系列情報の獲得に，どのような工夫や手法が提案されているかを紹介する．

2.1 時間的マックスプーリング

動画像における 1 ショット[*2]には，複数のフレームが含まれるために，CNN の構造から得られる特徴量ベクトルは，フレーム数分得られることになる．そこで，ショット内フレームの特徴量ベクトルの時系列に対して，時間的なマックスプーリング（Temporal Max-Pooling）を適用すれば，静止画像の場合と同様の，ショットのベクトル表現とみなすことができる．この時、特徴量ベクトルの空間が，もし充分にスパースならば，それぞれの特徴に対応するベクトルは独立に近い状態である．このため，フレーム間でのマックスプーリングは，ショット内のコンセプトに対する様々な特徴情報を集積することになる．例えば，“人物”の動画像は，横顔や正面の顔などの特徴情報が集積されれば，より “人物” と判定され，“飛行機” の映像であれば，遠くに小さく見える状態から近くで大きく見える画像などの特徴情報を集積されれば，より “飛行機” と判定されることになる．つまり，連続する静止画像の特徴量ベクトルを，時間方向に最大値をとれば，動画像でも 1 枚の静止画像に対する，画像認識と同様の手法で認識できることになる [3].

2.2 Long Short-Term Memory

動画像のような時系列データに対する手法として RNN（Recurrent Neural Network）[4] がある．RNN は，ノードの出力が次の時刻における自身のノードの入力となっており，時間的情報を考慮する学習が可能である．しかし，RNN の学習では，隠れ層を経るごとに勾配が小さくなってしまう問題がある．この問題で，長期の記憶を保持することが難しいという欠点が生じる．この欠点を解決するための手法として LSTM（Long-short term memory）がある．LSTM は，RNN の中間層のノードを，LSTM block に置き換えて実現されている [5]．LSTM block は，“入力ゲート”、“出力ゲート”、“忘却ゲート” と呼ばれる 3 つのゲートを持つ．入力ゲートは，前の層からの入力を通すか通さないかの判断を行い，出力ゲートは，次の層へ出力を通すか通さないかの判断を行う．忘却ゲートは，LSTM の内部状態を次の時刻に引き継ぐかどうかの判断をする役割を持っている．この構造によって，長期依存が学習可能となる [6].

2.3 Two-Stream Convolutional Networks

NN は，層のつなげ方，識別器を構成する方法の自由度が，従来の手法に比べて格段に高い．この性質を活用して，入力として複数の情報源を利用する，マルチストリーム学習（multi-stream learning）と呼ばれるアプローチがある．

例えば，動画像は空間的情報と時系列情報に分解することができる．空間的情報では，個々のフレームにシーンや物体に関する情報が表現されるが，時系列情報では，フレーム間を跨ぐ，動きの情報が獲得される．そこで，空間的情報と時系列情報を独立してマルチストリーム学習を行い，最後のソフトマックスの値で統合する手法がある．これを Two-Stream Convolutional Networks（Two-Stream CNN）と呼ぶ．具体的には，時系列上の異なるフレーム間での対象の移動量を，ベクトルで表現したオプティカルフロー（optical flow）から学習し，もう一方で CNN による空間的情報を獲得し，それらを統合して，両方の情報を考慮した識別手法が提案されている [7], [8].

3. TREC Video Retrieval Evaluation

大量の動画像から動画像検索を行うための動画像検索・解析技術に関する国際的なコンペティションとして，アメリカ国立標準技術研究所 (National Institute of Standards and Technology, NIST) が開催している TRECVID（Text REtrieval Conference Video Retrieval Evaluation）がある [9]．本研究では，TRECVID で実施される以下の 2 つの課題について Curriculum Learning に基づくアプローチの検証を行う．

3.1 Semantic indexing

2015 年の Semantic INdexing（SIN）タスクは，ショットから特徴を抽出し，その中に存在する意味のあるコンセプトを検出・認識して，動画像の内容に基づいたインデキシングを行うことを目的としている．このタスクは静止画像における一般物体検出を，動画像まで拡張したものと考えられる．具体的には，**図 1** の左のショットは “飛行機”，右のショットは “成人男性” のコンセプトがタグ付けされており，タスクではこれらの検出を行う．

[*2] 1 つのシーンを表す数秒程度の短い動画像

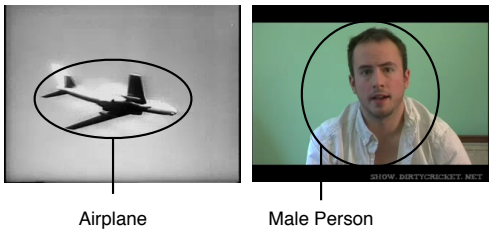

図 1　Semantic INDexing

図 2　Ad-hoc Video Search

3.2　Ad-hoc video search

2016 年の Ad-hoc Video Search（AVS）タスクでは，数十のクエリ*3が与えられ，クエリに最も適合する動画像を検出することを目的としている．クエリからコンセプトを抽出して，各コンセプトの検出・認識を行うことを考えると，AVS タスクは SIN タスクの拡張と考えられる．

検索例として，クエリ “屋外でギターを演奏している人” が与えられた場合を説明する．クエリからは，“人” コンセプト，“ギター演奏” コンセプト，“屋外” コンセプトを抽出が求められるため，3 つすべてのコンセプトが映り込むショットを検索する必要がある． 図 2 に示す 3 枚のフレーム例をみると，左の 2 フレームは，“人”，“ギター演奏”，“屋外” のすべてのコンセプトが写り込んでいるが，右のフレームは，“人”，“ギター演奏” コンセプトは写っているものの，“屋外” コンセプトが写っていない．よって，AVS タスクでは，クエリに適合する左の 2 フレームが正解例となる．

4.　提案手法

4.1　提案手法概要

本節では，我々の提案手法の概要を説明する．

CNN から獲得された特徴量を用いて，段階的な転移学習を行えば，異なるタスクやドメインにおいても識別が容易となる．このとき，抽出する特徴量の識別器には SVM などが用いられる．識別するコンセプトが大規模な場合，SVM の学習の計算コストが大きくなることが問題となる．TRECVID では，1 コンセプトの検索に何千の動画像を識別する必要があり，コンセプトが数十個になると，計算コストの問題は重大となる．

そこで本手法では，SVM の代わりに，小規模のニューラルネットワーク（micro Neural Network; microNN）の使用を提案する．NN では，GPU を用いた最新のフレームワークが使用可能であることから、SVM と比較して数十倍以上の学習速度が実現できる．さらに，microNN は極めて小規模の構造であるため，より高速な学習，識別が実現でき，1 度の転移学習に計算時間がかからないため，連続した段階的転移学習でも実現可能となる．次節以降では，まず microNN の高速かつ柔軟な段階的転移学習を概観する．続いて，Two-Stream CNN による，シーン認識と物体認識を行う手法を提案する．さらに，動画像の時系列情報の獲得に時間的マックスプーリングを用いた手法を示す．最後に microNN の中間層に LSTM を利用した手法についても提案する．

4.2　学習済みネットワークを用いた抽出

本手法では，初めに入力データへの Data Augmentation を行う．これは，CNN などで学習させる際，学習する画像をずらしたり，ぼかしたり，様々な変形を加えて学習データを増やして，認識を頑健にするというテクニックである [10]．本手法では，元の画像データを反転させることに加えて，小さなウィンドウを平行移動させて，1 枚の画像から 10 枚の画像を切り抜いたものを入力とするなどの Data Augmentation も行っている（ 図 4）．

次に，Data Augmentaion された入力データを用いて，学習済みネットワークから特徴量を抽出する．近年の著名な学習済みネットワークとしては，AlexNet[11] や VGGNet[12]，GoogLeNet[13] などが知られている．本手法では VGGNet を用いている．VGGNet は，安定性が高く，他の多くの研究でも使用されている．VGGNet は多層の CNN 構造で，16 層からなるモデルと 19 層からなるモデルが存在する．本手法では，16 層のモデルを利用し，全結合層の第 2 層目（fc7）の出力を抽出し，これを microNN への入力データとしている．

4.3　Micro Neural Network

4.3.1　基本構造

VGGNet から抽出した特徴量を入力データとして，識別器 microNN を学習させることを考える．microNN は 2 値分類を行う識別器で，ショット内にコンセプトが含まれるか含まれないかの分類のみを行う．miroNN の構造は，全結合の隠れ層が 1 層のみからなる構造で，入力層，隠れ層，出力層のノード数はそれぞれ，4096 次元，32 次元，2 次元としている．また，microNN の各層には Dropout を適用している．

4.3.2　学習

本研究では，microNN を用いて，以下の 3 段階の転移学

*3　いくつかのコンセプトを含む質問文

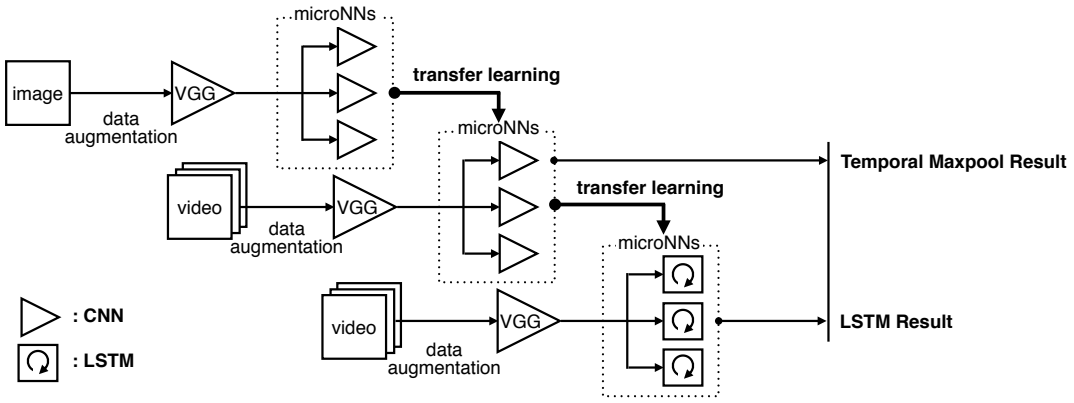

図 3　学習過程の全体像

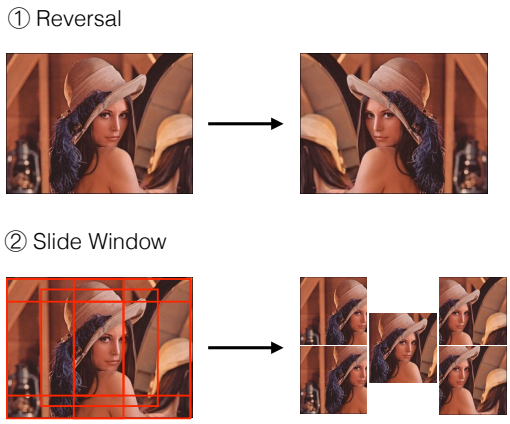

図 4　data augmention の例

習を行う．また，学習過程の全体像を 図 3 に示す．

(1) VGGNet の中間層の出力を用いて，静止画像データセットによる microNN の転移学習を行う（図の 1 段目）．
(2) 静止画像で学習済みの重みを初期値として，動画像データセットによる転移学習を行う（図の 2 段目）．
(3) 同様に，動画像データセットによる LSTM の転移学習を行う（図の 3 段目）．

一般に，CNN の学習は初期値によって強く影響を受けることが分かっている．特に，学習データセットが少ない場合，過学習を防ぐためにも適切な初期値の重みを獲得することが重要である．それゆえ，ゼロから学習する CNN と比較して，学習済みの重みを初期値として利用できれば，高精度な識別が可能になると考えられる．本手法の場合，動画像データセットによる学習を進める前に，同様の意味を持つコンセプトが含まれる静止画像データセットによって学習を行う．この過程は，図 3 で 1 段目から 2 段目に向かう矢印の transfer learning を表す．静止画像で学習した重みを初期値とすることで，動画像データセットのみでの学習に比べて，識別精度の向上が期待される．例えば，“飛行機” というコンセプトの動画像データセットが用意されているタスクで，いきなり動画像データセットで学習を進めるのではなく，別の “飛行機” コンセプトを含む静止画像のデータセットで学習を行う．2 段階目の転移学習時では，時間的マックスプーリングを適用する（ 図 3 の Temporal Maxpool Result）． 2.1 に述べるように，時間的マックスプーリングの適用で，動画像でも静止画像の認識と同様の手法で識別が可能となる．

さらに，3 段階目の転移学習として LSTM の学習を行う．2 段階目での時間的マックスプーリングとは異なり，対象コンセプトの時間経過による移動量や動き情報を，より厳密に学習を進めて，動画像として連続性を考慮した識別が可能となる．また，通常の RNN と異なり，長期記憶が可能となるため，時系列の中での，コンセプトの初期状態などを考慮した識別が考えられる．例えば，“走り幅跳び” コンセプトの場合，“走る”，“跳ぶ” と 2 つの状態が考えられる．長期記憶ができない識別では，“跳ぶ” 状態の短時間の動画像情報のみで識別しなければならない．しかし，長期記憶が可能となれば，“跳ぶ” 前の “走る” 情報を考慮した識別が可能となる．この結果，“走り幅跳び” と “立ち幅跳び” との差別化が可能となると期待される．本手法では，microNN における中間層の各ノードを，LSTM ブロックに置換して実現している．

4.3.3　Two-Stream CNN を利用した Scene 認識

本節では，miroNN による Two-Stream CNN で，シーン認識と物体認識を行う手法を提案する．シーン認識の Two-Stream CNN は，VGGNet による特徴量と，学習済みネットワークとして Place_CNDS（以下，Place）を用いた特徴量を組み合わせて，入力次元数を 2 倍とする microNN で学習を進めている．Place はシーンに特化して学習されたモデルである [14]．8 層の畳み込み層と 3 層のフル接続層から構成されており，VGGNet と同様に，識別層の一つ手前の中間層から，4096 次元の出力を抽出して利用している．Place と VGGNet により獲得された特徴量を組み合わせて，“海辺” や “室内”，“キッチン” といったシーンに特化したコンセプトと “人” や “飛行機” といった物体を意味

表 1　AVS タスクのクエリの例

ID	クエリ
501	a person playing guitar outdoors
502	a man indoors looking at camera where a bookcase is behind him
503	a person playing drums indoors
504	a diver wearing diving suit and swimming under water
505	a person holding a poster on the street at daytime

するコンセプトの識別精度向上が考えられる．

5. 評価実験

5.1 実験設定

提案手法の有効性について，TRECVID2015 の SIN タスク，及び TRECVID2016 の AVS タスクを用いて検証・評価する．また 3.2 節でも説明したように，AVS タスクでは文のクエリが与えられて，クエリに最も適合する動画像を検索することが要求される．クエリの例を **表 1** に記す[*4]．

5.1.1 コンセプト抽出

本手法を AVS タスクに適用するために，まずクエリから対象となるコンセプトの抽出を行わなければならない．ここでは，将来的に自動化が可能となるように，以下のような簡単なルールを用いてコンセプトを抽出している．

- 冠詞などを除いた動詞，名詞などの意味のある単語のみを選択
- 動名詞や複数形などは原型に戻して扱う
- 選択した単語の類義語についても抽出する

5.1.2 コンセプトの統合と検索

本手法の，クエリに対する適合率の算出方法を説明する．クエリに対する適合率を算出するために，まずクエリ内のコンセプトごとに識別結果の統合を行う．例えば，クエリ内に “人” コンセプトと “室内” コンセプトが抽出されたとき，それぞれのコンセプトについて，microNN により識別を行い，その出力値を組み合わせてクエリとしての適合率としている．本手法では，コンセプトごとの識別結果の値について，正規化を行い，和を取った値をクエリの適合率としている．正規化は，コンセプトごとの識別結果の値を $[-1, 1]$ の範囲の値に変形し，特定のコンセプトの結果に引きづられないようにしている．

統合による検索結果のサンプルを **図 5** に示す．これは AVS のクエリ *“502: a man indoors looking at camera where a bookcase is behind him”* に対する結果である．縦軸がコンセプトを表し，横軸が検索結果のランキングを表している．上段が “カメラに向かって話す人” に対する結果で，下段が “カメラに向かって話す人” に “本棚” の結果を結合したものである．どちらもカメラに向かって話している人が写っているが，本棚のコンセプトを統合すれば，よりクエリに適合した結果になっていることが見て取れる．

[*4] すべてのクエリの一覧は以下を参照．
http://www-nlpir.nist.gov/projects/tv2016/pastdata/tv16.avs.topics.txt

5.1.3 データセット

本実験における静止画像のデータセットには，ImageNet の画像を用いている．ImageNet は，自然言語処理の分野で有名な WordNet のオントロジーに従って，各単語に対応する画像を収集したものである．約 1400 万枚の画像データ（2 万 2 千カテゴリ）が用意されており，AVS タスクでは，39 コンセプトの画像データを用いている．

動画像データセットには，TRECVID のデータセットと UCF-101 データセット [15] を用いている．TRECVID データセットは，197,000 個 (400,238 ショット) の学習用映像と，8,263 個 (145,634 ショット) のテスト映像が用意されている．UCF-101 データセットは，13,320 個の Youtube の動画像データがもととなっており，1) 人や物，2) 人体の動き，3) 人と人，4) 楽器の演奏，5) スポーツなど 101 の行動クラスからなる．本実験では，そのうち 5 クラスを利用している．

5.1.4 評価指標

認識結果は，TRECVID の評価基準に従って，テスト用動画像のショットを microNN の出力値が高い順にランク付けしたときの，上位 2,000 ショットに対する “平均精度 (AP: Average Precision)” で評価している [16]．AP は，情報検索の分野で開発された評価尺度で，実際にコンセプトが映っているショットが上位にランク付けされているほど高くなる．また，AP の平均を MAP（Mean Average Precision）と呼ぶ．

5.2 実験結果

5.2.1 microNN の精度と学習時間評価

本節では，SIN タスクによる microNN の精度と学習速度の評価結果を示す．まずはじめに microNN の隠れ層のユニット数の影響についての確認を行った． **図 6** に各コンセプトに対する識別精度を示す．縦軸に認識対象の 30 種類のコンセプトを並べ，横軸に識別精度の AP を示す．図 6 よりユニット数の違いによる明瞭な傾向は必ずしも見られなかったが，32 次元が最も好成績であったことから以降の実験においても 32 次元を用いた．次に，従来の識別器である SVM との比較を行った． **図 7** に結果を示す．さらに， **表 2** に microNN と SVM のそれぞれの MAP と学習にかかった計算時間を記す．

表 2　Performance comparison between SVM and microNN

手法	MAP(%)	平均学習時間 (s/concept)
microNN	0.1626	0.1385
SVM	0.2148	110.44

図 7 から，幾つかのコンセプトでは microNN の AP が

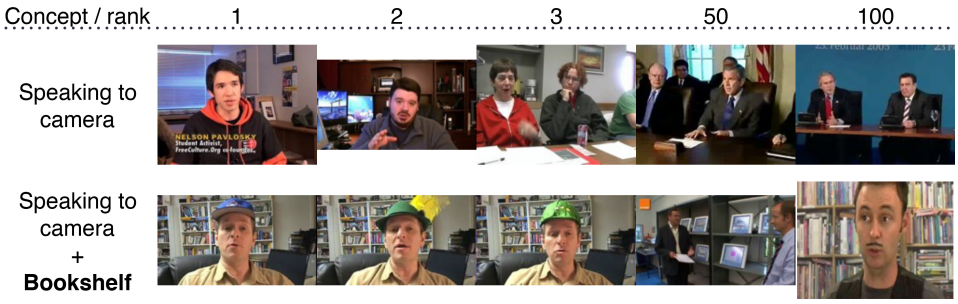

図 5　コンセプト統合結果の例

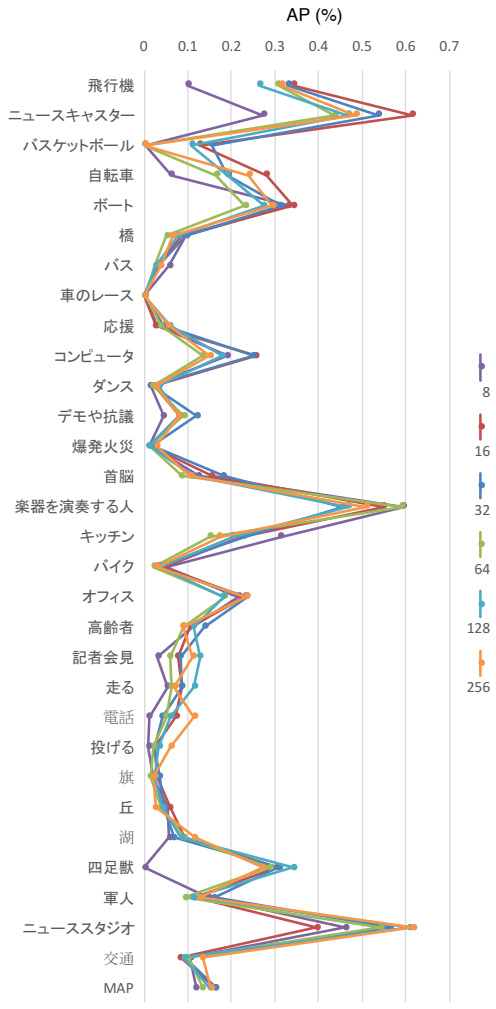

図 6　microNN の隠れ層のノード数変化による識別制度比較

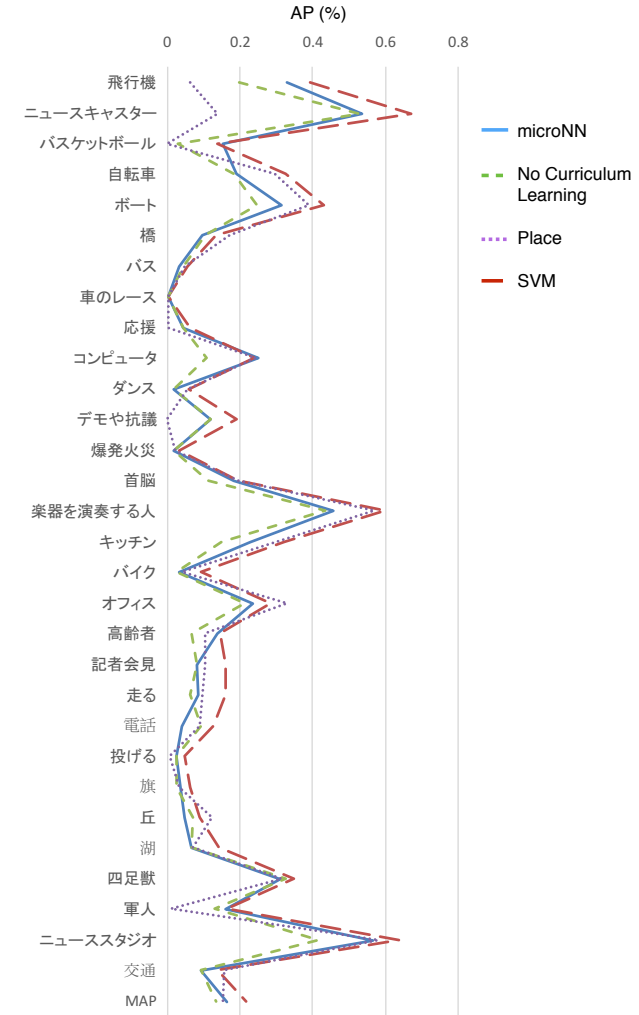

図 7　SIN タスクにおける各手法の識別精度比較

SVM の AP を，わずかに下回っていることが分かる．一方. 表 2 から，転移学習による計算コストの問題は，microNN を利用すれば解決することが分かる．具体的には，1 コンセプトに対する学習時間をみると，SVM の 1,000 倍のオーダーで高速化できることが分かる．学習データセットが大規模となる場合，microNN は高速な学習によって精度向上が見込まれるが，SVM は計算コストが増加する上に，これ以上の精度向上が見られないことが考えられる．

5.2.2　段階的な転移学習の有効性の評価

本節では，段階的な転移学習の有効性を評価するために，ImageNet のデータセットで転移学習を行った microNN に，TRECVID のデータセットを転移学習させた場合の精度と，ImageNet による転移学習を行わず，TRECVID のデータセットのみで学習した場合の精度について比較する．SIN タスクにおける結果を 図 7 に示す．図の microNN が段階的転移学習による結果を示し，No Curriculum Learning が TRECVID のみの学習結果を示す．

図から，転移学習により “Basketball” や “Computer”，“Studio_With_Anchorperon” などの，いくつかのコンセプトについて大きな精度の向上が見て取れる．これは，学習の初期値に大きな影響を受けていることを示しており，静止画像の学習を通した転移学習の有効性が示されている．AP が同程度，または下回るコンセプトについては，静止画像と動画像の同じコンセプトでの，データセットの差によって，学習に悪影響を及ぼされていることが考えられる．例えば，“飛行機” というコンセプトについて考えると，静止画像は機体の正面ばかりを捉えたデータに対し，動画像では側面ばかり捉えたデータである場合などがある．このような場合，同じコンセプトに対するデータセットに差がある場合には，転移学習のデータセットの選択が重要となる．

5.2.3 Two-Stream CNN によるシーン認識

AVS タスクを用いて，VGGNet から抽出された特徴量のみで学習させた microNN による精度と，Place から抽出された特徴量と VGGNet の特徴量を組み合わせた特徴量で学習させた microNN による精度を比較する．結果を 図 7 に示す．図の点線が Place による結果を示している．

“飛行機” や “ニュースキャスター” といったコンセプトでは，Place を用いたモデルの精度は VGGNet のみによるモデルの精度と比較して大きく減少している．一方，“橋” や “オフィス”，“丘” などシーンを意味するコンセプトは，大きく向上している．つまり，物体を表現するコンセプトの識別には，Place が悪影響を及ぼしているが，シーンコンセプトを含む動画像では，精度が向上していることがわかる．よって，アンサンブル学習などを利用した，学習時のモデルの使い分けにより，全体としての精度向上が期待できる．

5.2.4 学習データバランスによる評価

AVS タスクを用いて，正例，負例のサンプル数が不均衡な学習を行った場合と，サンプル数が同数で学習を行った場合の比較評価を行う．不均衡な場合の正例は，用意されている学習データはすべて用いて，負例は，合計で 30,000 個になるように用いている．均衡な場合は正例，負例それぞれ 15,000 個の動画像を用いている．正例の数が足りないコンセプトについては，オーバーサンプリングを行っている．結果を **図 8** に示す．図の microNN が不均衡データによる結果，Balance が均衡データによる結果を示している．

ほとんどのクエリでは，不均衡データによる学習の精度が，均衡データによる学習の精度を上回っていることがわかる．このことから，コンセプトが存在するかしないかの識別を行う場合，学習サンプル数の正例，負例のバランスを考えるよりも，より多くの負例を用いたほうが，精度の向上が見込まれることが考えられる．

5.2.5 時系列情報の有効性の評価

AVS タスクを用いて，時間的マックスプーリングを適用した microNN による精度と，さらに LSTM による転移学習を行ったときの精度と比較する．結果を 図 8 に示す．実線で示す microNN が時間的マックスプーリングによる結果，破線が LSTM を行ったときの結果を示している．LSTM を行う microNN が，AVS タスクで最高精度である，0.047 の MAP を達成している．特に，509 クエリにおける精度は，LSTM を用いていない microNN の精度に比べて 3 倍以上の精度が達成できている．これは，このクエリに対しては上手く時系列情報が獲得できたことを意味している．

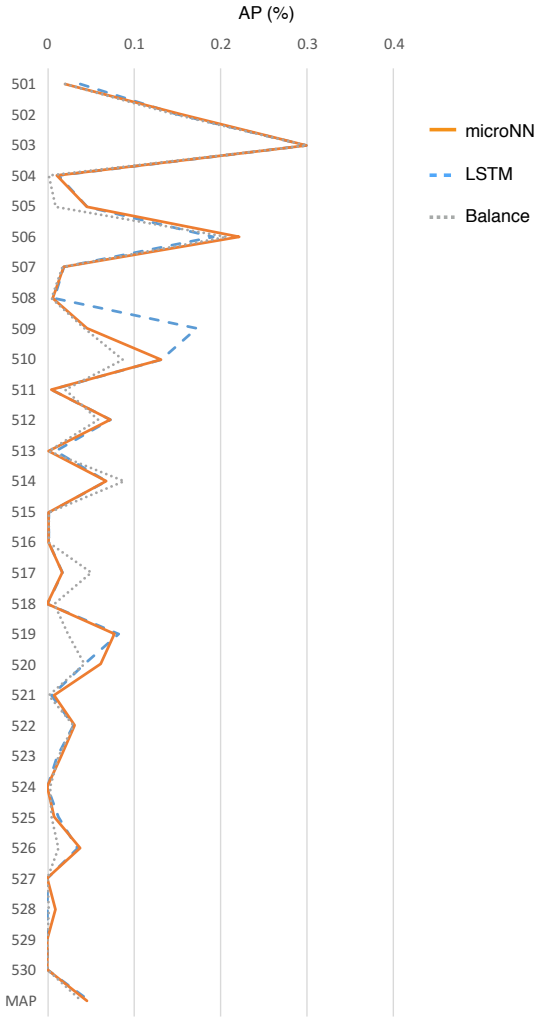

図 8 AVS タスクにおける各手法の識別精度比較

6. 結論

本研究では，動画像の内容に基づく検索手法として，識別器としてのニューラルネットワーク群と，段階的転移学習を提案した．また，microNN の利用では，動画像の時間的特徴に対する処理として，時間的マックスプーリングと LSTM による手法を提案した．評価実験には，学習済みネットワークとして VGGNet を利用し，静止画像に ImageNet データセット，動画像に TRECVID と UCF-101

のデータセットを用いた．TRECVID 2015 の SIN タスクと TRECVID 2016 の AVS タスクのテスト用映像のショットを SVM の出力値が高い順にランク付けしたときの，上位ショットに対する”平均精度 (AP: Average Precision)”で本手法を評価した．

本研究では，microNN の利用により大規模データセットに対する転移学習の問題点である，学習コストの問題を解決した．さらに，学習データのバランスについての評価を行い，不均衡データセットの優位性や，通常の学習と転移学習の比較による段階的転移学習の有効性を示した．この結果，空間的情報と時間的情報を考慮した柔軟な学習が可能となり，TRECVID2016 の AVS タスクで 2 位の精度を達成した．

今後の課題として，更なる識別精度の向上を目的とした，Two-Stream CNN によるアンサンブル学習が挙げられる．本稿での Place を用いた識別で，シーンコンセプトについての識別精度の向上が見られた．一方で，物体コンセプトについては，VGG Net のみを用いた識別の結果が上回った．そこで，コンセプトごとに Place を用いたモデルと，VGG Net のみを用いたモデルでアンサンブル学習を行い，MAP の最高値を目指す手法の導入を検討していく．

謝辞

本研究は科学研究費補助金 26280040 および 16K12487 の補助を受けて実施された．

参考文献

[1] Hinton, G. E., Osindero, S. and Teh, Y.-W.: A Fast Learning Algorithm for Deep Belief Nets, *Neural Comput.*, Vol. 18, No. 7, pp. 1527–1554 (online), DOI: 10.1162/neco.2006.18.7.1527 (2006).

[2] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D.: Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Comput.*, Vol. 1, No. 4, pp. 541–551 (online), DOI: 10.1162/neco.1989.1.4.541 (1989).

[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L.: Large-Scale Video Classification with Convolutional Neural Networks, *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, Washington, DC, USA, IEEE Computer Society, pp. 1725–1732 (online), DOI: 10.1109/CVPR.2014.223 (2014).

[4] Graves, A., Mohamed, A. and Hinton, G. E.: Speech Recognition with Deep Recurrent Neural Networks, *CoRR*, Vol. abs/1303.5778 (online), available from ⟨http://arxiv.org/abs/1303.5778⟩ (2013).

[5] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. and Darrell, T.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, *CoRR*, Vol. abs/1411.4389 (online), available from ⟨http://arxiv.org/abs/1411.4389⟩ (2014).

[6] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780 (online), DOI: 10.1162/neco.1997.9.8.1735 (1997).

[7] Simonyan, K. and Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos, *CoRR*, Vol. abs/1406.2199 (online), available from ⟨http://arxiv.org/abs/1406.2199⟩ (2014).

[8] Ng, J. Y., Hausknecht, M. J., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G.: Beyond Short Snippets: Deep Networks for Video Classification, *CoRR*, Vol. abs/1503.08909 (online), available from ⟨http://arxiv.org/abs/1503.08909⟩ (2015).

[9] Awad, G., Fiscus, J., Michel, M., Joy, D., Kraaij, W., Smeaton, A. F., Qunot, G., Eskevich, M., Aly, R., Jones, G. J. F., Ordelman, R., Huet, B. and Larson, M.: TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking, *Proceedings of TRECVID 2016*, NIST, USA (2016).

[10] Salamon, J. and Bello, J. P.: Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification, *CoRR*, Vol. abs/1608.04363 (online), available from ⟨http://arxiv.org/abs/1608.04363⟩ (2016).

[11] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25* (Pereira, F., Burges, C. J. C., Bottou, L. and Weinberger, K. Q., eds.), Curran Associates, Inc., pp. 1097–1105 (online), available from ⟨http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf⟩ (2012).

[12] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556 (2014).

[13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going Deeper with Convolutions, *Computer Vision and Pattern Recognition (CVPR)*, (online), available from ⟨http://arxiv.org/abs/1409.4842⟩ (2015).

[14] Wang, L., Lee, C., Tu, Z. and Lazebnik, S.: Training Deeper Convolutional Networks with Deep Supervision, *CoRR*, Vol. abs/1505.02496 (online), available from ⟨http://arxiv.org/abs/1505.02496⟩ (2015).

[15] Soomro, K., Zamir, A. R. and Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, *CoRR*, Vol. abs/1212.0402 (online), available from ⟨http://arxiv.org/abs/1212.0402⟩ (2012).

[16] Smeaton, A. F., Over, P. and Kraaij, W.: Evaluation campaigns and TRECVid, *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, ACM Press, pp. 321–330 (online), DOI: http://doi.acm.org/10.1145/1178677.1178722 (2006).