修 士 論 文

会話動画中の表情認識に関する研究

平成 30 年度修了
三重大学大学院　工学研究科 情報工学専攻
ヒューマンインターフェース研究室

吹上雅樹

はじめに

人間は，絶え間なく他者とコミュニケーションをとっている．人がコミュニケーションをとるときには，言語情報だけでなく視覚情報と聴覚情報を含む非言語情報から相手の感情を推定している．その中で最も重要な視覚情報の代表は表情であり，相手の感情を推定する重要な手がかりとなる．また近年，コミュニケーションロボットの開発・普及により，家庭内や店頭で人とロボットがコミュニケーションをとる機会が増えている．ロボットが人の様々な表情を認識することにより，言語情報には含まれていない話者の感情を推定することができ，人とロボットのコミュニケーションを含むヒューマン・コンピュータ・インタラクションの高度化につながる．そのため，多種多様な表情を認識できる自動表情認識システムの実現と精度向上は重要な課題となっている．

表情認識に関するこれまでの多くの研究では，基本的な感情を表す 6 表情 (喜び，悲しみ，怒り，驚き，軽蔑，恐怖) を認識対象とした手法が提案されている．ところが，本来表情はコミュニケーション時にやり取りされる情報であるにもかかわらず，従来研究が対象としている表情にはコミュニケーション時によく見られる表情が含まれていない．その問題を解決するために，コミュニケーション時の表情を再現した表情動画データベース”The lerge MPI facial expresion database”が公開されている．

しかし，このデータセットには 2 つの大きな問題がある．1 つ目は少サンプル多クラスで，かつ表情変化が複雑であるため，汎化能力が高く表現力のある学習モデルが必要となること，2 つ目は表情だけでなく頭部が上下左右に回転する動きがあるために，従来の表情認識手法が適用できない点である．そのため本研究では，少サンプル多クラスかつ変化に富んだ表情を含む動画データに対してどのような表情認識手法が有効かを検証すること，東部の動きを考慮した会話動画中の表情認識に有効な手法を提案することを目的とする．

大量のデータを必要としないハンドクラフトの手法と表現力が高い深層学習を組み合わせることで，少サンプル多クラスかつ表現力の必要なデータに対してどのような手法が有効かを検証する．様々な手法を組み合わせた比較実験の結果，少サンプル多クラスでかつ変化に富んだ表情を含む動画データに対して，ハンドクラフト手法である HOG 特

徴と深層学習である LSTM の組み合わせがもっとも有効であるとわかった．また，会話動画中の表情変化に対して有効な手法として，表情の動きと頭部の動きの両方を学習できる 2-stream Recurrent Convolutional Neural Network アーキテクチャを提案した．このアーキテクチャは画像内から認識に必要な特徴を自動で抽出できる Convolutional Neural Network と時系列データを学習できる Recurrent Neural Network で構成される 2 つのネットワークを組み合わせたアーキテクチャである．それぞれ表情の動きを学習するネットワークと頭部の動きを学習するネットワークであり，2 つのネットワークを組み合わせることで単一のネットワークと比較し，認識精度向上を実現した．

本論文では，1 章では表情認識研究の背景や課題，2 章では本論文に関する技術，3 章では提案手法，4 章では検証実験，5 章では精度向上のために行った予備実験，6 章では考察と全体のまとめについて述べる．

目次

第 1 章

緒言

1.1　研究背景

人間は日常生活の中で，絶え間なく他他者とコミュニケーションをとっている．コミュニケーション時には，意識的に，あるいは無意識的に相手がどう感じているかを会話の内容や視覚情報，聴覚情報など様々な情報から推定している．コミュニケーション時に言語情報から推定される感情と非言語情報から推定される感情に矛盾が生じる場合，非言語情報から推定される感情が重視されることがわかっている [1]．非言語情報の中でも，表情は最も重視される情報である．

近年，コミュニケーションロボットなどの普及により，家庭や店頭で人とコンピュータがコミュニケーションをとる機会が増えている．人とコンピュータが高度なコミュニケーションをとるためには，コンピュータがその時々のユーザの感情を正確に推定する必要がある．日常生活をサポートするサービスである Google Home[2] や Apple Siri[3] は，ユーザの音声を認識することで言語コミュニケーションを行っている．しかし，これらのサービスはコミュニケーションに非言語情報を用いることができないため，ユーザの正確な感情を推定することができない．また，高齢化と介護従事者不足が深刻化する介護の現場でコミュニケーションロボットの活躍が期待されている．コミュニケーションロボットを導入することで，被介護者の 34.2%が活動の質と量の改善効果がみられたという報告がある [4]．現段階のコミュニケーションロボットは被介護者の状態を検知し反応を返すが，同時に感情を認識することでその時々の感情にあった反応を返すことができれば，より高い改善効果が期待できる．非言語情報を用いた感情推定を行うことで，多彩なコミュニケーションが可能となり，より高度なヒューマンコンピュータインタラクションが実現できる．そのため多種多様な感情の推定が可能な自動感情認識システムの実現と精度向上は重要な課題である．

1.2 感情推定手法

身体表現や音声，表情など様々な非言語情報を用いた感情推定手法が提案されている [5]．この章ではこれらの手法について説明する．

- 表情認識
  表情は感情推定のために広く用いられており，静止画や動画を用いた表情認識に関する研究はおよそ 30 年以上前から盛んに行われている [6, 7]. 従来の表情認識の手法は，顔画像から抽出された特徴を識別機に入力する統計的手法と表情筋の動きを定量化し表情を認識する商法に分けられる．
  顔の外観の動きを認識する研究は，顔画像から表情の動きを特徴ベクトルとして抽出し，特徴ベクトルを分類器に入力し認識する手法を提案している．顔の動きを定量化し表情を認識する研究は，Ekman らによって開発された Facial Action Coding System[8] に基づいた手法を提案している．このシステムは顔の動きの最小単位を Action Unit (AU) として定義し，表情を定量的に表す (例:AU12 は”口角を上げる”) ことを可能としている．表情認識を行う多くの研究が，Ekman らが定義する基本の 6 感情 (悲しみ，怒り，喜び，驚き，軽蔑，恐怖) を認識対象とした手法を提案している．また近年，上記の基本表情以外の表情を認識対象とする手法 [9] が提案されている．表情から様々な動き特徴を抽出することが可能なため，多種多様な感情を認識することに適しているといえる．
- 音声認識
  会話はコミュニケーションにおいて最も確実で自然な情報伝達手段であるため，人間とコンピュータの主な情報伝達手段として会話を用いることが期待されている [10]. 音声には声のピッチやトーンなどの情報が含まれているため，音声情報による話者の感情推定は可能である．しかし，話者の感情以外にも話す内容，話す速度，話し方により音声情報が変動する欠点がある．
- 身体表現認識
  身体表現を用いた感情推定手法を用いた研究の多くは，表情と体の動きの情報を合わせることで精度向上を目指している．体の動きはその人の感情状態を表すことがあり，感情を推定するための情報として役立つ．Castellano らの研究 [11] では，怒り，喜び，楽しみ，悲しみの 4 クラスを身体表現のみで感情推定を行う手法を提案しており，4 クラス感情推定精度が 61%となっていることから，身体表現から簡単な感情を推定可能であることがわかる．Gunes らの研究 [12] では，表情のみを用いた手法と表情と身体表現を合わせた手法による比較実験が行われており，表情と

身体表現を合わせた方がより感情推定精度が高くなることが示されている．身体表現と表情の両方を用いることにより精度向上が期待できるが，身体表現のみの手法では多くの種類の感情推定は困難だといえる．

- 生体信号認識
  心拍数の変動や呼吸頻度のような生体信号による感情認識が試みられ，表情などの情報と組み合わせることで感情推定が可能である [13]. しかし，認識精度は高くなく，複雑な感情を認識することは困難である．また生体信号を計測するためには，計測機器が必要となるため生体信号を取得できる場面が限られる欠点がある．

感情を推定するために様々な情報を用いた認識手法が提案されているが，多種多様な感情を高精度で認識するためには表情認識手法を用いるのがもっとも適しているといえる．

1.3　表情認識に関する主な従来研究

従来の表情の動きを認識する自動表情認識手法は，前処理，特徴抽出，分類の三つの段階からなる [14].

- 前処理
  前処理は顔画像に対して目鼻の位置合わせなどを行う処理であり，表情認識手法における基本的な処理である．前処理を行う利点として，個人間の変動と顔の向きの変動を低減できる点が挙げられる．多くの手法が顔画像全体を用いるが，顔の部分画像を前処理に使用する手法がある. 顔画像全体が対象となっている手法では Active Appearance Model (AAM) [15] が用いられている．図 1.1 で示す AAM は顔の形状および外観の統計モデルを顔画像に一致させるためのコンピュータビジョンアルゴリズムである．AAM を用いて入力顔を顔のプロトタイプモデルと一致させるように画像を変形させる．多くの手法で顔特徴点は両目，または両目と鼻から取得されている [16, 17].

図 1.1: Active Appearance Model のプロトタイプモデル

  顔部分画像が対象となっている手法では，顔特徴点から顔部分画像を生成する．大きな 2 つの部分画像に分ける手法や小さな 36 の部分画像に分ける手法など，分割される顔部分画像は手法により様々である [18, 19]．顔部分画像を用いることで，顔の認識で必要としない部分を削除できる利点がある．

- 特徴抽出
  従来の表情認識における有効な特徴として，Local Binary Patterns (LBP) [20], Histogram of Oriented Gradients (HOG) [21], Local Phase Quantization (LPQ) [22], Histogram of Optical Flow [23], 顔特徴点座標 [15, 24], and PCA-based methods [25] が挙げられる. これらの特徴の多くは特定の課題を解決するために作成された特徴であるため，照明や見かけの変動が大きい場合に汎化性が損なわれる欠点がある．

- 分類
  時系列データを分類する上で効果的なのは，系列間の変動の学習である．時系列データに有効な分類器として，Hidden Markov Models (HMM) [26, 27, 28], Spatio Temporal Hidden Markov Models (ST-HMM) [29], Dynamic Bayesian Networks (DBN) [30], Conditional Random Fields (CRFs) [31, 32, 33, 34] を拡張させた Latent-Dynamic CRFs や Hidden CRFs [35] が挙げられる．これらの手法を用いることで表情変化の推移を学習することが可能になり，静止画のみを用いるより複雑な表情の認識が可能になる．

近年，深層学習の一種である Convolutional Neural Network (CNN) があらゆる分野で注目を集めている．物体認識の分野において高い認識率でを誇る AlexNet[36] や GoogLeNet[37] が CNN を用いた手法として挙げられる．Mollahosseini らは CNN を用いた表情認識手法を提案しており，従来のハンドクラフト手法と同程度か, それより高い認識精度を得ている [38, 39]．CNN の問題として時間の経過による表情変化を考慮できない点が挙げられるが，時間方向にも畳み込む 3DCNN が提案されている．Hasani らは時系列に対応した 3DCNN を用いることで通常の CNN より精度が向上することを示した [40].

同じくディープニューラルネットワークの一つである Recurrent Neural Network[41] は，時系列データに対応している．RNN は音声認識や自然言語処理などの分野で用いられているが，勾配消失・発散問題から長時間の系列データを扱うことができなかった．Long Short Term Memory (LSTM) [42] は，RNN にはない 3 つの制御を採用することで，長時間の系列データを扱うことを可能にした．動画を入力とした表情認識において，LSTM を用いることにより，認識精度が向上することが示されている [43].

1.3.1 従来研究の課題

従来手法により基本的な 6 感情を対象とした認識精度は大きく向上したが，基本表情はコミュニケーションにおいてやり取りされる表情を対象としていない．ヒューマンコンピュータインタラクションの高度化のため，より多くの種類の感情を対象とした表情認識システムの実現と精度向上は重要な課題である．基本表情以外の表情を認識可能とするため，日常的な会話の中で現れる表情を再現した表情動画データセット”The MPI large facial expression database”[44] が提供されている．このデータセットを用いることで基本表情以外の表情を対象とした認識手法の研究が可能となる．しかし，基本的でない表情を再現しているため以下の問題点が挙げられる．

- 多くの感情ラベルがあるが，各感情に対する動画数が少ない
- 変化に富んだ表情を認識するため，表現力のあるモデルを用いる必要がある

以上の問題点に対して，学習データ数が少ない場合には大量のデータを必要としないハンドメイドの手法，表現力が必要な場合には大量のパラメータを持つ Convolutional Neural Network (CNN) や LSTM (Long Short Term Memory Recurrent Neural Network) を用いることが有効である．

また，コミュニケーション表情には「口角を上げる」といった表情の動きだけでなく，「うなずく」といった頭部の上下左右の動きも必要となるため，従来の表情の動き特徴のみを利用した手法は適用できない場合もある．

1.4 本研究の取り組み

1.4.1 研究目的

前述の課題を解決するため，本研究では，多クラス少サンプルかつ変化にとんだ表情を含む動画データに対してどのような表情認識手法が有効かを検証すること，頭部の動きを考慮した会話動画中の表情認識に有効な手法を提案することを目的とする．

1.4.2 研究内容

会話動画中の表情に対して有効な表情認識手法として，表情の動きを学習するネットワークと頭部の動きを学習するネットワークを組み合わせた 2-stream RCNN アーキテクチャを提案する．多クラス少サンプルでかつ表現力の必要なデータに対してどのような手法が有効かを検証するため，様々な手法を用いた比較実験を行う．前述のとおり，表情認

識は前処理，特徴抽出，分類の三段階で行う．比較実験では，以下の特徴抽出手法と分類手法の組み合わせを変更し有効性を検証する．

特徴抽出

ハンドメイド手法として顔認識で用いられている画像中の輝度の勾配を特徴化した Histograms of Oriented Gradients 特徴 (HOG 特徴) を用いる．また深層学習の手法として，Convolutional Neural Network (CNN) と約 200 万枚からなる顔画像を学習した CNN(VGG16) を用いる．

分類

分類では，長期時系列データに対応した深層学習の手法である Long Short Term Memory Recurrent Neural Network(LSTM) を用いる．

第 2 章

関連技術

2.1　特徴量

画像認識において，入力された画像から認識に必要な情報を数値ベクトルとして取得することを特徴抽出という．特徴抽出アルゴリズムは多種多様であり，それぞれ異なった特徴を画像から抽出できる．顔画像認識においては，画像のテクスチャ情報を記述する LBP や，輝度勾配を記述する Histograms of Oriented Gradients (HOG)，Scale-Invariant Feature Transform (SIFT) が用いられている．ここでは本稿で用いる HOG 特徴の特徴抽出アルゴリズムとその性質について説明する．

2.1.1 HOG 特徴 (Histograms of Oriented Gradients)

局所領域 (セル) の輝度の勾配方向をヒストグラム化し，そのヒストグラムを特徴量としたものが HOG 特徴量 [21] である．輝度勾配を用いているため幾何学的変換に頑健であることが利点である．

抽出アルゴリズム

1. 入力画像 $I = I(x, y)$ の各画素に対して，勾配強度 $m(x, y)$ と勾配方向 $\theta(x, y)$ を求める．

$$I_x(x, y) = I(x + 1, y) - I(x, y) \tag{2.1}$$
$$I_y(x, y) = I(x, y + 1) - I(x, y) \tag{2.2}$$
$$m(x, y) = \sqrt{I_x(x, y)^2 + I_y(x, y)^2} \tag{2.3}$$
$$\theta(x, y) = tan^{-1} \frac{I_y(x, y)}{I_x(x, y)} \tag{2.4}$$

2. 勾配方向を 0°〜180° の範囲を 20° ずつ 9 方向に量子化し，$N(\text{pixel}) \times N(\text{pixel})$ のセルごとに輝度勾配ヒストグラムを作成する．
3. $M(\text{cell}) \times M(\text{cell})$ の n 番目のブロックを輝度勾配ヒストグラムを式 (2.5) により正規化する．

$$v(n) = \frac{v(n)}{\sqrt{\sum_{k=1}^{M \times M \times K} v(k)^2 + \epsilon}} \tag{2.5}$$

ここで，K は勾配方向数であり, ϵ の値は 1 と設定する．

手順 3 の正規化により，ヒストグラムの形状が整うので照明の変動が低減できる．また，生成される特徴ベクトルの次元数 dim は以下の式により求めることができる．

$$dim = \left(\frac{W}{N} - M + 1\right) \times \left(\frac{H}{N} - M + 1\right) \times M^2 \times K \tag{2.6}$$

ここで，W, H はそれぞれ入力画像の縦画素数と横画素数を表す．また，画像サイズ，セルサイズ，ブロックサイズを次の図で定義する．

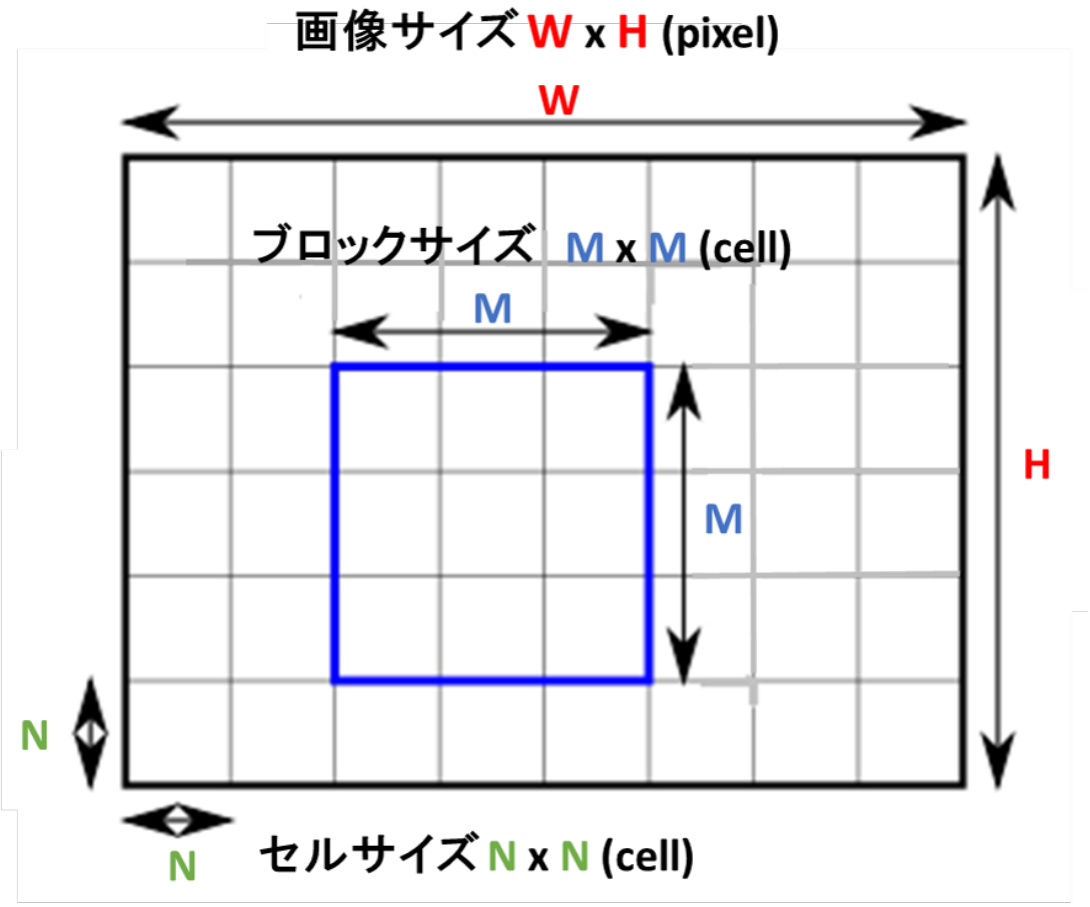


図 2.1: 画像サイズ，セルサイズ，ブロックサイズの関係図

2.2　主成分分析 (Principle Component Analysis : PCA)

高次元の原特徴ベクトルは，計算コストを指数関数的に増加させるだけでなく，次元の呪いと呼ばれる学習を妨げる状況を引き起こす．そのため，高次元の原特徴を低次元の特徴空間に射影する必要がある．本稿では，PCA を用いて次元削減を行う．PCA は多次元空間上のデータの分布の分散が最大となる直交部分空間に線形射影する手法である．直交部分空間に射影するため，主な情報を保持したまま原特徴ベクトルの次元を削減できる．

2.2.1　アルゴリズム

N 個の d 次元特徴ベクトル $\boldsymbol{x}_i \in R^d (i = 1, \ldots, N)$ の平均ベクトル $\overline{\boldsymbol{x}}$ と共分散行列 $\boldsymbol{S}$ は以下の式で表される．

$$\overline{\boldsymbol{x}} = \frac{1}{N} \sum_{i}^{N} \boldsymbol{x}_i \tag{2.7}$$

$$\boldsymbol{S} = \frac{1}{N} \sum_{i}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}} \tag{2.8}$$

共分散行列は固有値行列 Λ, 固有値ベクトル行列 Φ によって以下のように表される．

$$\boldsymbol{S}\Phi = \Phi\Lambda \tag{2.9}$$

Λ は共分散行列 $\boldsymbol{S}$ の固有値を要素とする対角行列である．また，Φ は Λ に対応する固有ベクトルの行列である．上位 k 個の固有値に対応する固有ベクトルの行列 Φ_k を用いて，以下の式により d 次元特徴ベクトル $\boldsymbol{x}$ を k 次元のベクトル $\boldsymbol{y}$ に次元削減する．

$$\boldsymbol{y} = \Phi_k \boldsymbol{x} \tag{2.10}$$

2.3 Deep Neural Nework : DNN

近年，多数のネットワークを結合したディープニューラルネットワークが，パターン認識において高い認識精度を得るための手法として，様々な分野で注目を集めている．ディープニューラルネットワークがハンドクラフトの手法と大きく異なる点は，設計することなく目的に最適化された特徴記述が可能になる点である．しかし，ネットワーク構造の設計やパラメータの最適化が難しく，大量の学習データが必要である欠点を持つ．ここでは本稿で用いる時系列データに対応した Long Short Term Memory Recurrent Neural Network と画像処理で用いる Convolutional Neural Network について説明する．

2.3.1 Long Short Term Memory Recurrent Neural Network : LSTM

多層パーセプトロン，RNN，LSTM について順を追って説明する．

単純パーセプトロン 図 2.2 で示す単純パーセプトロンはデータが入力される入力層，結果が出力される出力層，入力層と出力層の間にある中間層から構成されるニューラルネットワークである．誤差逆伝播法により与えられた正解データに近づくよう重みを更新し，任意の関数に近似できる．中間層の数を増加させた多層パーセプトロンを用いることで表現力は増すが，学習計算の過程で微分の乗算を中間層の数だけ繰り返すため勾配が消失する問題が発生する．

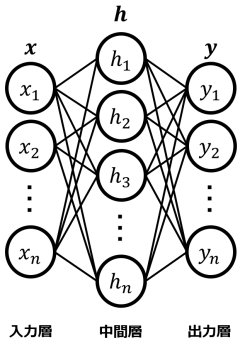


図 2.2: 単純パーセプトロン

RNN 図 2.3 で示す RNN はパーセプトロンの中間層が再帰構造を持つことによって，データ間の関係を学習することが可能になったニューラルネットワークである．時系列データを扱うことが可能なので，音声認識，自然言語処理の分野で活用されている．図 2.4 で示すように，時間軸で見ると RNN は非常に深いニューラルネットワークとなり勾配消失問題が発生するため，長期間の系列データを扱えない欠点がある．

図 2.3: RNN

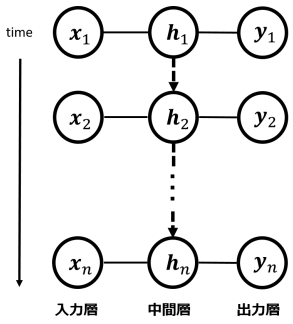


図 2.4: 時間軸方向に展開した RNN

LSTM 図 2.5 で示す LSTM は，RNN では勾配消失問題により学習できなかった長期間の系列データを学習可能とした RNN の拡張である．入力制御，出力制御，忘却制御，Constant Error Carousel (CEC) を導入することで短期間と長期間の両方の系列データに対応した．入力制御と出力制御はそれぞれ重み更新を同時に受ける入力重み衝突と出力重み衝突の問題を解消した．忘却制御は保持している状態を保持し続けるか更新するかを決定する．そのため入力系列のパターンが突然変更された場合にも対応できる．CEC は前の系列の状態を線形和として保持しておくことで勾配消失問題を解決している．

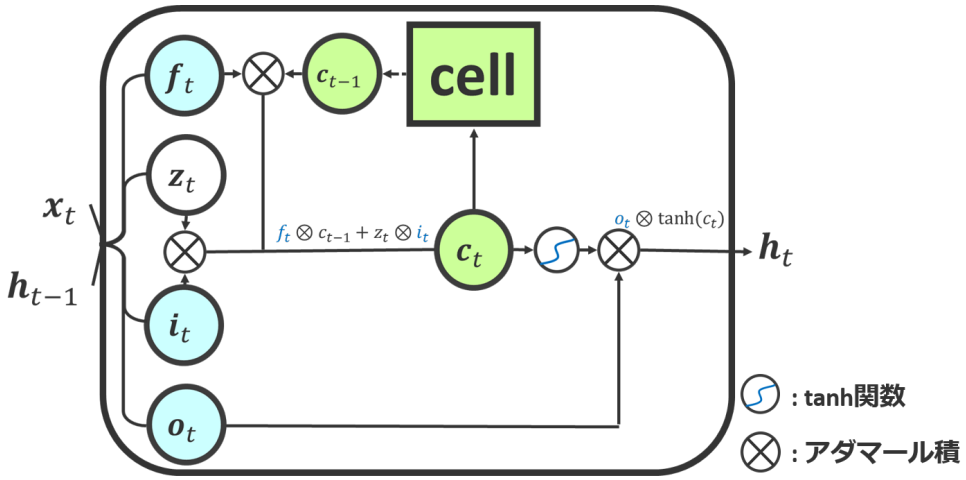


図 2.5: LSTM ブロックの構造

図 2.5 中の $\boldsymbol{x}_t$ と $\boldsymbol{h}_t$ はそれぞれ t フレーム目の入力ベクトルと t フレーム目の LSTM ブロックの出力であり，$\boldsymbol{f}_t, \boldsymbol{z}_t, \boldsymbol{i}_t, \boldsymbol{o}_t$ は以下の式で表せられる．

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{R}_f \boldsymbol{h}_{t-1} + b_f) \tag{2.11}$$
$$\boldsymbol{z}_t = tanh(\boldsymbol{W}_z \boldsymbol{x}_t + \boldsymbol{R}_z \boldsymbol{h}_{t-1} + b_z) \tag{2.12}$$
$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{R}_i \boldsymbol{h}_{t-1} + b_i) \tag{2.13}$$
$$\boldsymbol{o}_t = \sigma \boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{R}_o \boldsymbol{h}_{t-1} + b_o) \tag{2.14}$$

ここで σ はシグモイド関数，W, R, b はそれぞれ $\boldsymbol{x}$ の重み係数，$\boldsymbol{h}$ の重み係数，バイアス重み変数である．

2.3.2 Convolutional Neural Network : CNN

CNN は画像処理分野において高い認識精度をほこる，畳み込み層とプーリング層と全結合層から構成されるネットワークである．CNN は入力画像に対して畳み込み処理を行い，特徴マップを生成する．また，画像中の対象物の位置不変性を維持するためにプーリング処理を行う．これらの処理を繰り返し入力画像に対して行うことで，画像から特徴を抽出できる．

畳み込み処理は，画素値と任意の大きさのフィルターの内積の計算を繰り返し行う処理である．畳み込み処理の概略図を図 2.6 に示す．

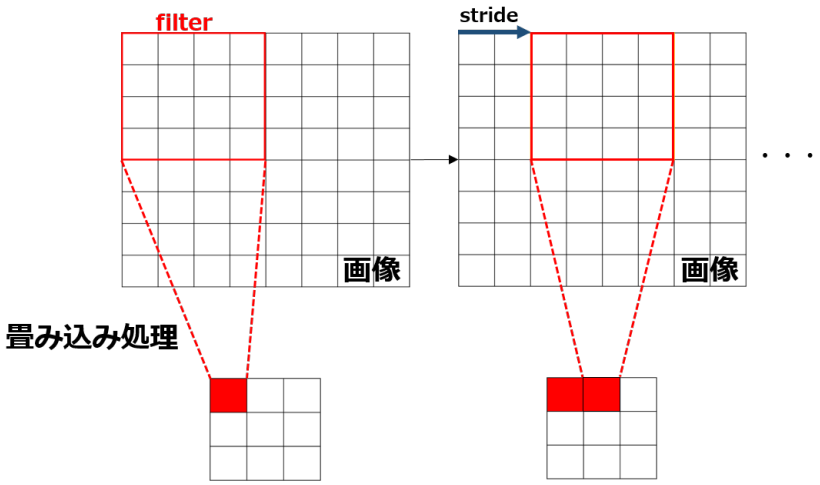


図 2.6: 畳み込み処理の概略図

プーリング処理は，任意の大きさの領域内の最大値をとり画像の圧縮を行う処理である．本研究では Max Pooling 処理を用いる．Max Pooling 処理の例を図 2.7 に示す．

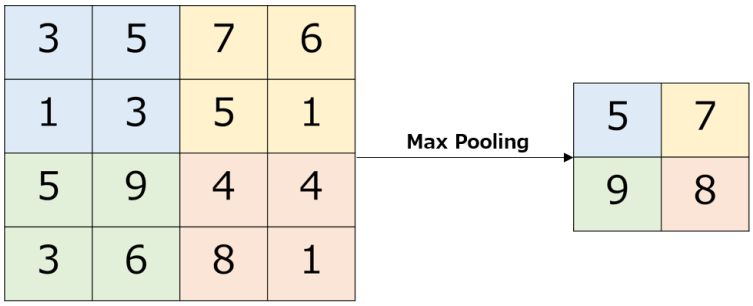


図 2.7: Max pooling の例

第 3 章

提案手法

3.1　概要

会話動画中の表情から表情の動きを学習する Facial ストリームと頭部の動きを学習する landmark ストリームを組み合わせた 2 ストリーム Recurrent Convolutional Neural Network (RCNN) アーキテクチャ [45] に基づく手法を提案する．このアーキテクチャの構造を図 3.1 に示す．入力は表情動画であり，出力は表情ラベルである．それぞれのストリームでは各入力画像に対して，CNN と LSTM を用いて事後確率値を算出する．それぞれネットワークの事後確率値の平均値から推定ラベルを決定する．

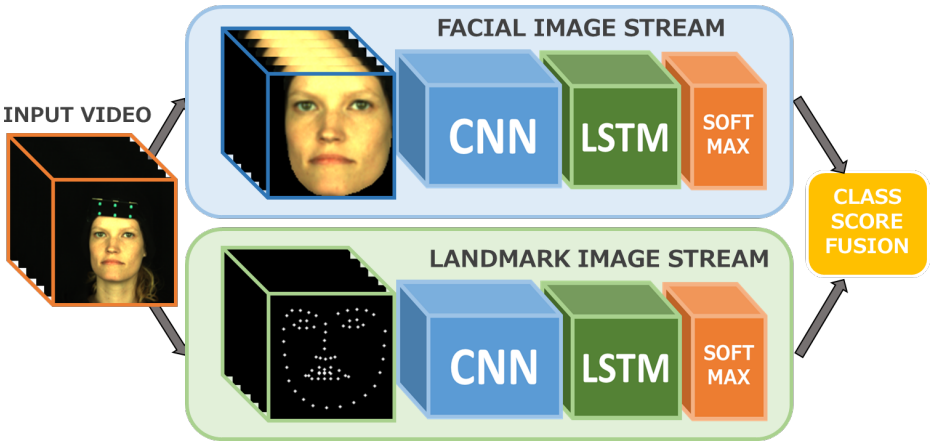


図 3.1: 2-Stream RCNN の構造

3.2　データセット

ここでは実験で用いるデータセットについて説明する．

3.2.1 The MPI large facial expression database

本実験では，会話中の表情を再現したデータセット”The MPI large facial expression database”[44] を使用する．ドイツ人男性 5 人，女性 5 人の計 10 人が 51 種類の表情を再現した 510 の表情動画が含まれている．このデータセットに含まれている 51 種類の感情ラベルを表 3.1 に示す．各ラベルは Emotion, Subordinate Emotion, Conversation,

表 3.1: データセットに含まれるすべての感情ラベル

51 Expressions			
Agree considerd	Agree continue	Agree pure	Agree relctant
Aha light bulb moment	Annoyed botherd	Annoyed rolling eyes	Arrogant
Bored	Compassion	Confused	Contempt
I didn’t hear	I didn’t care	I didn’t know	I didn’t understand
Disagree considered	Disagree pure	Disagree reluctant	Disblief
Disgust	Embarrassment	Fear oops	Fear terror
Happy achievement	Happy laughing	Happy satiated	Happy schadenfreude
Imagine negative	Imagine positive	Impressed	Insecurity
Not convinced	Pain felt	Pain seen	Sad
Remember negative	Remember positive	Thinking considering	Thinking ploblem solving
Smilling encouraging	Smilling endearment	Smilling flirting	Smilling sad nostalgia
Smilling triumphant	Smilling uncertain	winning	yeah right
Smilling sardonic	Tired	Bambi-eyes	

Subordinate Conversation のいずれかに分類されている．動画長は様々であり，これらの動画では自然な表情から徐々に表情が表出され，また徐々に自然な表情に戻る過程が録画されている．各動画の解像度は 768×576 ピクセルであり，フレームレートは 60fps である．

基本的な感情ラベルが含まれていることから Conversation, Subordinate Conversation に分類されている感情ラベルのみを用いる．また，それらに分類されている”Thinking consider”と”Thinking problem solving”など異なるラベルでも表情の変化が大きく変わらないラベルを統合し再編した 16 種類の感情ラベルを認識対象とする．16 種類の感情ラベルを表 3.2 に示す．

3.2.2 The Extended Cohn-Kanade Dataset (CK+)

提案したアーキテクチャが従来の表情認識手法と比べて有効なのかを調査するために，従来の表情認識で用いられている基本的な表情を再現した The Extended Cohn-Kanade+ Dataset (CK+)[46] を用いて比較実験を行う．このデータセットには全部で 123 人分の

表 3.2: 認識対象となる感情ラベル

16 Expressions			
Agree	Aha light bulb moment	Annoyed	Bored
Compassion	Confused	I don’t know	Disagree
Disbelief	Imagine negative	Imagine positive	Insecurity
Not convinced	Thinking	Tired	Bambi-eyes

593 つの表情動画が含まれているが，その中で感情ラベルが与えられている動画は 118 人分の 327 つである．感情ラベルとして基本的な 6 感情 (怒り，軽蔑，恐怖，喜び，悲しみ，驚き) に加え，困惑が含まれた計 7 種類の感情ラベルがつけられている．これらの動画は自然な表情から，徐々に表情表出される過程が録画されている．

3.3　2-stream アーキテクチャ

会話中の表情に有効な表情の動きと頭部の動きの両方を学習する 2-stream アーキテクチャを提案する．各ネットワークは，前処理，CNN による特徴抽出，LSTM による分類を行っている．それぞれネットワークから出力された事後確率値から推定感情ラベルを決定する．それぞれの処理について説明する．

3.3.1　前処理

各ネットワークに入力動画に対して行う前処理について説明する．

Facial Image Stream

このネットワークは表情の動きを学習する．そのための前処理として動画に含まれるフレームに対して，顔検出処理，テンプレートマッチングによる顔領域追跡を行う．顔領域追跡の処理の流れを以下に示す．

1. 0 フレーム目の顔画像から顔検出処理により，顔領域を抽出する．
2. 抽出された顔領域をテンプレート画像とする．
3. 次のフレームの顔画像からテンプレートマッチングにより顔領域を抽出する．
4. 2. の処理に戻る．

顔検出処理は画像処理用のオープンライブラリである OpenCV[47] を用いて行う．これらの処理の目的は横を向くことによって顔の特徴が取得できなくなり検出不可能になった

顔領域に対してテンプレテートマッチングを行うことで抽出可能にすることである．正面顔画像と側面顔画像の顔領域抽出の例をそれぞれ図 3.2 と図 3.3 に示す．

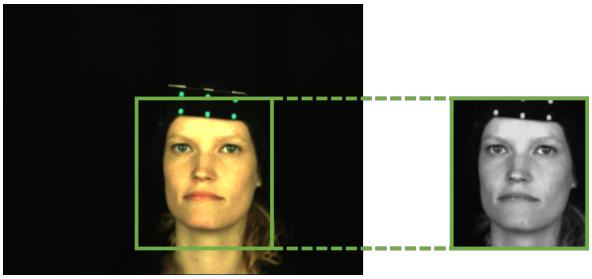


図 3.2: 正面顔の顔領域の抽出例

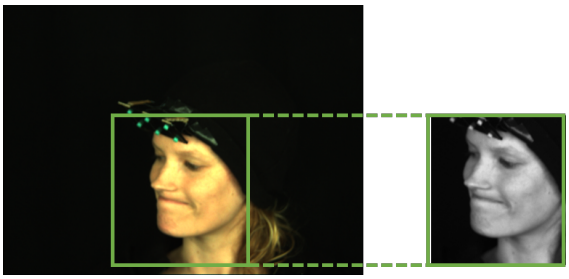


図 3.3: 横顔の顔領域の抽出例

Landmark Image Stream

このネットワークは頭部の動きを学習する．前処理として各フレームに対して，顔領域抽出，顔特徴点抽出，特徴点膨張処理を行う．顔領域抽出は前述の処理と同様である．顔特徴点抽出処理では顔画像処理ライブラリである Dlib[48] を用いて行った．抽出される特徴点は 68 個ある．特徴点膨張処理では，特徴点を中心とする 7×7 画素の範囲内にある 49 画素に対して以下の重み関数 (式 3.1) を用いて画素値を決定する．

$$\omega(L, P) = 1 - 0.1 \cdot d_{M(L,P)} \tag{3.1}$$

ここで，L は顔特徴点座標，P は注目画素，d_M はマンハッタン距離を示す．特徴点原画像と膨張処理後の特徴点画像をそれぞれ図 3.4，図 3.5 に示す．

3.3.2　CNN

本研究で用いる CNN の構成を図 3.6 に示す．ここで Input は入力画像であり，チャンネル数×縦サイズ×横サイズで表す．また， Conv は畳み込み層，Pool はプーリング層，

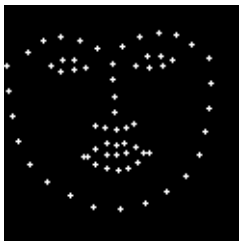


図 3.4: 特徴点原画像

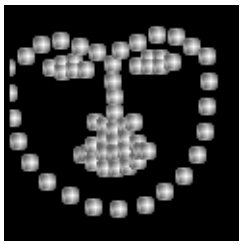


図 3.5: 膨張処理後の特徴点画像

fc は全結合を示す．活性化関数には Relu 関数を用いる．出力された 1024 次元の特徴ベクトルを LSTM-RNN に入力する．

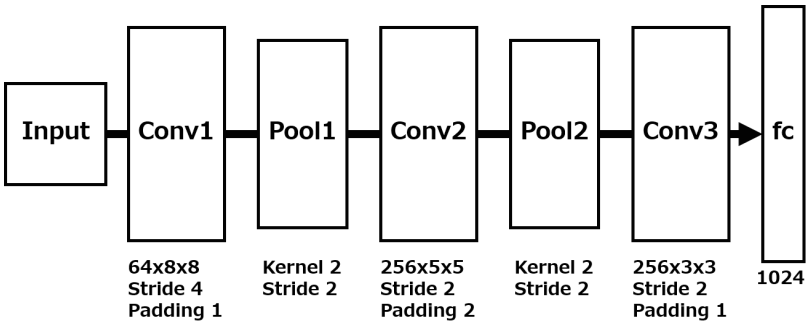


図 3.6: CNN のネットワーク構成

3.3.3　LSTM

CNN から抽出された特徴ベクトルを入力し，出力された事後確率値を平均したものから推定ラベルを決定する．LSTM のネットワーク構成は中間層が 6 層，出力層が 1 層で構成されており，それぞれ中間層のユニット数は 50，出力層のユニット数は 16 である．損失関数として softmax cross entropy を用いた．softmax cross entropy は全結合層の出力を softmax 関数を用いて正規化を行った後，cross entropy 関数により誤差を算出する．softmax 関数．cross entropy 関数はそれぞれ式 (3.2) と式 (3.3) に示す．

$$S(\boldsymbol{x}) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \tag{3.2}$$

$$H(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{i=0}^{N} y_i log x_i \tag{3.3}$$

ここで，$\boldsymbol{x}$ は全結合層の出力ベクトル，$\boldsymbol{y}$ は正解ラベルを示す One-hot ベクトル，N は感情ラベル数を表す．

3.3.4 Class Score Fusion

各ネットワークの最終フレームの各ラベルの事後確率値から平均値を求め，最も事後確率値が高いラベルを推定感情ラベルとする．

第 4 章

評価実験

4.1 実験概要

本研究では，提案手法の評価実験と少サンプル多クラスかつ変化に富んだ表情を含む動画データに対して有効な手法を検証する比較実験を行った．

4.2 実験条件

評価実験と比較実験の両方に共通する実験条件を以下に示す．

- 実験データセット
  本実験では，会話中の表情を再現した MPI データセットと従来の研究で用いられた CK+データセットを用いる．
- 交差検証法
  MPI データセットの場合は，1 人分のデータを評価用，1 人分のデータを検証用，8 人分のデータを学習用に分割した Leave One Person Out 交差検証により学習，評価を行う．CK+データセットの場合は，被験者を考慮した 6 分割交差検証により学習，評価を行う．
- 評価方法
  動画中の最終フレームの出力のみを用いて，その動画の感情推定可否を決定する．表情認識成功率を以下の式から求める．

$$\text{成功率} \quad = \frac{\text{推定成功動画数}}{\text{全動画数}} \times 100\ [\%] \tag{4.1}$$

4.3　提案手法の評価実験

表情と頭部の動き両方を学習するアーキテクチャがどれほど有効なのかを評価する実験を行う．表情の動きを学習するネットワークのみ，頭部の動きを学習するネットワークのみ，表情と頭部両方の動きを学習するネットワークを組み合わせたアーキテクチャのそれぞれで実験を行う．

4.3.1　実験結果

MPI データセットでの認識精度と推定結果の混合行列を以下の表と図に示す．

表 4.1: MPI データセットの認識正解率

手法	認識正解率 (%)
2-Stream (提案手法)	42.9
Facial Image Stream	35.7
Landmark Image Stream	21.4

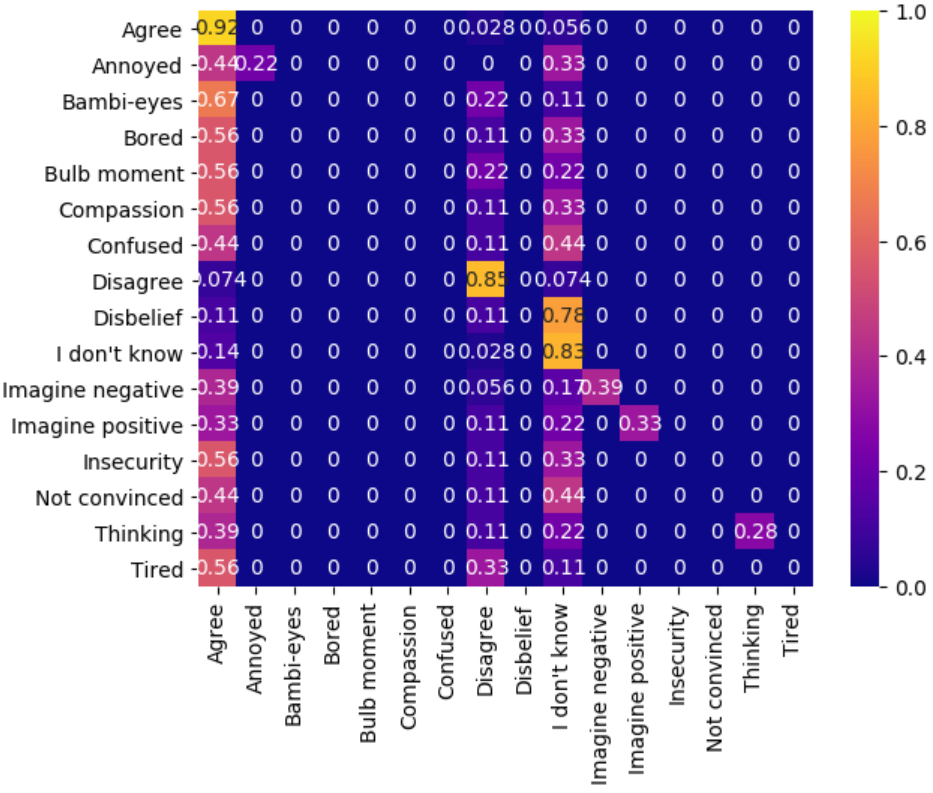


図 4.1: 推定結果の混合行列 (提案手法)

CK+データセットでの認識精度を以下の表に示す．ここでは深層学習を用いている関連研究の認識成功率を比較対象として挙げている．

表 4.2: CK+データセットの認識正解率

手法	認識正解率 (%)
先行研究 3DCNN[40]	93.6
2-Stream (提案手法)	75.3
Facial Image Stream	70.8
Landmark Image Stream	49.6

4.3.2 考察

MPI と CK+の両方のデータセットにおいて，認識精度が最も高かった手法は表情と頭部の動きの両方を学習させた提案手法であり，次に表情の動きを学習したネットワーク，頭部の動きを学習したネットワークであった．

MPI データセットを用いた提案手法の推定結果の混合行列では”Agree”, ”DisAgree”, ”I don’t know”のラベルに多くの動画が振り分けられている．これは，これらのラベルの頭部の動きが他のラベルに比べて大きく，視覚的特徴が他のラベルと大きく異なるため，ラベル当たりの動画数が他のラベルに比べて多いためであると考える．

今回の実験で LSTM の層の数を減らし各層のユニット数を増やした構成の場合，層の数を増やし各層のユニット数を減らした構成より，早く過学習が起こることがわかった．これは 1 クラス当たりのサンプル数が少なくネットワークのユニット数が多いため，クラスごとに共通した表情変化の特徴ではなく各サンプルごとの視覚的特徴を学習したためである．そのため，ユニット数を減らす，前処理により表情変化以外の視覚的特徴を画像中から取り除くことにより過学習を防ぐことができる．

また，今回の実験では，頭部の動きを学習するために顔特徴点画像を用いたが，頭部の動き情報をより詳細化するためにフレーム間の物体の移動を考慮したオプティカルフロー画像を用いることでより高い認識精度が期待できる．

4.4　比較実験

多クラス少サンプルであり，分類器に表現力が必要なデータセットに対して，データを多く必要としないハンドメイドの手法と表現力の高い深層学習手法を組み合わせどのような手法が有効かを検証する．ハンドメイドの手法として HOG 特徴を用い，深層学習の手法として CNN と学習済み CNN (VGG)，LSTM を用いて実験を行う．

4.4.1　各手法の流れ

各実験の手法を以下に示す．この実験での LSTM のネットワーク構成は中間層が 6 層，出力層が 1 層で構成されており，それぞれ中間層のユニット数は 50，出力層のユニット数は 16 である．損失関数として softmax cross entropy を用いた．

HOG + LSTM

入力動画から顔領域抽出を行い，顔領域画像から OpenCV を用いて 8100 次元の HOG 特徴ベクトルを抽出し，PCA により 1024 次元の特徴ベクトルを生成する．生成した特徴ベクトルを LSTM に入力し認識を行う．処理の流れを図 4.2 に示す．HOG 特徴抽出処理では，ImageSize, BrockSize, CellSize, BrockStride をそれぞれ $(224, 224)$, $(16, 16)$, $(8, 8)$, $(8, 8)$ としている．

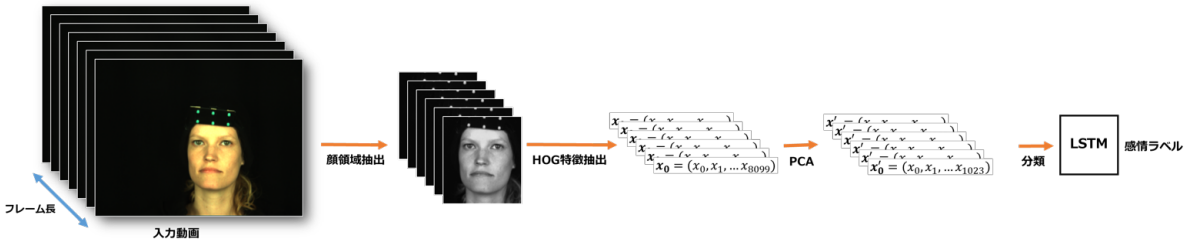


図 4.2: HOG+LSTM の処理の流れ

CNN + LSTM

入力画像から顔領域抽出を行い，顔領域画像を CNN に入力する．全結合層で得られる 1024 次元のベクトルを LSTM に入力し認識を行う．処理の流れを図 4.3 に示す. この手法では，LSTM だけでなく CNN も学習を行いパラメータの最適化を行っている．

図 4.3: CNN+LSTM の処理の流れ

VGG + LSTM

少サンプル多クラス問題を解決するため約 2622 人の約 200 万枚からなる顔画像によって学習された VGG16[49] を特徴抽出に用いる．このネットワークは 13 層の畳み込み層，3 層の全結合層から構成されている．最終層を除いたすべての活性化関数は ReLu 関数であり，最終層は SoftMax 関数となっている．この実験では 15 層目の全結合層の出力を特徴ベクトルとしている．

入力画像から顔領域抽出を行い，顔領域画像を VGG に入力する．全結合層で得られる 1024 次元のベクトルを LSTM に入力し認識を行う．処理の流れを図 4.4 に示す. この手法では，VGG のパラメータ最適化は行わず LSTM のみ学習を行っている．

図 4.4: VGG+LSTM の処理の流れ

4.4.2 実験結果

得られた認識正解率を以下の表に示す．

表 4.3: 比較実験の実験結果

手法	認識正解率 (%)
HOG + LSTM	44.4
CNN + LSTM	11.8
VGG + LSTM	35.3

4.4.3 考察

比較実験を行った結果，最も認識精度が高かった手法はハンドメイド手法の HOG 特徴と深層学習の LSTM を組み合わせた手法であった．続いて，学習済みである VGG と LSTM を組み合わせた手法，CNN と LSTM を組み合わせた手法の順に認識精度が高かった．

HOG + LSTM 手法が VGG + LSTM 手法よりも認識精度が高いのは，HOG 特徴の場合，見た目の情報がそのまま特徴ベクトルに含まれており表情の動き特徴など必要な情報が保持されているのに対し，顔認識で用いられていた VGG は個人の顔特徴を抽出し，表情認識に必要な情報を全て抽出できていないためであると考える．また，CNN+LSTM の認識結果が他の 2 手法より大幅に低い原因として，全ての手法において学習ループ数が等しい条件で実験を行っているのにも関わらず最適化を行うパラメータ数が多くパラメータ最適化のための学習ループ数が不足している点，すべてのパラメータを最適化させるのに学習データの数が充分でない点が挙げられる．

第 5 章

予備実験

5.1　表情動画データセット

本章では，現在ある表情動画データセットの特性と問題点，その解決手法の提案と評価実験について述べる．LSTM や 3DCNN の躍進により高精度の動画解析が可能となっている．表情認識の分野でもそれらの手法が用いられており，動画の表情データセットの需要が増している．本実験で用いているデータセットを含む多くの表情動画データセットは，無表情から表情が表れる過程を撮影している．しかし，アノテーションとして与えられている感情ラベルは 1 動画あたり 1 つであるため，連続した無表情フレームを無視したアノテーションが与えられていることになる．異なる感情ラベルが与えられているにもかかわらず同様の特徴が得られる無表情フレームがすべての動画に含まれているのは，学習を妨げる要因の一つになると考える．また，無表情から表情が表れるまでの長さに個人毎，動画毎に差があるため，表情が表れるまでの長さが認識結果に影響を及ぼすことも考えられる．これらの問題を解決するために，2 つの手法が考えられる．

- 無表情フレームが連続している区間を動画から取り除く．
- 既に与えられている感情ラベルに加えて，無表情ラベルをフレーム単位で追加する．

前者の場合，無表情フレームと表情有りフレームの教師無し分類が可能なため，アノテーションをつける必要がない．後者の手法では一定数のデータセットに対して無表情か表情が表れているかを判断しアノテーションをつけなければならない．表情が表れているか否かを主観で判断しなければならないため，今回は前者の手法を提案する．無表情フレームを取り除いたデータセットを学習と評価の両方に用いて評価実験を行う．

5.2　提案手法

表情動画から連続する無表情フレーム区間を取り除く手法を提案する．この提案手法は，表情動画データセットが表情動画が必ず無表情フレームから始まり，表情が現れた後に無表情フレームは表れない場合に使用できる．

動画のすべてのフレームに対してアライメント処理を行う．次に 0 フレーム目とそれ以外のフレームとの差分画像を生成する．生成した差分画像から HOG 特徴を抽出し，k-means クラスタリングによりすべての特徴ベクトルを 2 クラスに分類する．0 フレーム目と 1 フレーム目の差分画像から得られた特徴ベクトルが含まれるクラスを無表情クラスとし，含まれていないクラスを表情クラスとする．0 フレームから表情クラスに分類されたフレームまでを取り除く．

手法の流れを図 5.1 に示す．

図 5.1: 無表情フレーム除去処理の流れ

5.3　評価実験

無表情フレーム区間を取り除くことによって，認識結果にどのような影響があるのかを調査する．本実験では，無表情から表情が表れるまでの過程が録画されている CK+データセットを用いて評価実験を行う．用いたネットワークモデルと評価方法は提案手法と同じである．認識結果を以下の表に示す．

表 5.1: 無表情フレームの有無を比較した認識結果

データセット	認識正解率
無表情フレーム有り (元データ)	75.3
無表情フレーム無し (提案手法適用)	71.0

5.4 考察

提案手法により無表情フレームを取り除いたデータセットを用いた認識精度より，元データセットを用いた認識精度のほうが高かった．その原因として無表情フレームだけでなく表情フレームも誤って取り除いている点が挙げられる．アライメント処理の際，”驚き”ラベルには大きく口を開ける動作が含まれているが，顔の輪郭形状が大きく変わることから正しくアライメントされない場合がある．図 5.2 と図 5.3 で示すようにアライメントを失敗したフレームと無表情フレームの絶対値差分画像は，他の絶対値差分画像と比べて大きく異なるため意図した特徴を得ることができず，無表情と表情有りの 2 クラスにクラスタリングすることが困難になる．それが表情フレームが取り除かれている原因であると考える．これを解決する手法として，アライメント処理を行わず元画像から HOG 特徴を抽出しクラスタリングを行う手法が挙げられる．

図 5.2: アライメントに成功した場合の絶対値差分画像

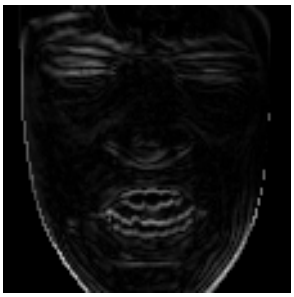


図 5.3: アライメントに失敗した場合の絶対値差分画像

第 6 章

結論

6.1　本研究のまとめ

本論文では，従来の表情認識に関する研究が対象としていない会話中の表情に有効な 2-stream の手法を提案した．提案手法は従来の基本的な表情にはなかった頭部の動きを学習可能であり，表情と頭部の動きを学習することで表情の動きのみを学習する手法と比較し，認識精度向上を実現した．

また，非基本的な表情データセットにある少サンプル多クラスかつ多種多様な表情変化に対してどのような手法が有効かを比較実験した結果，ハンドクラフト手法である HOG 特徴と深層学習である LSTM の組み合わせがもっとも有効であるとわかった．

6.2　今後の課題

今回の実験では，51 種類から 16 種類の感情ラベルに再編した感情ラベルを認識対象としたが，より日常生活でみられる表情に限定し，学習させる必要がある．

また，対照実験では HOG 特徴量と隠れマルコフモデルを組み合わせる等のハンドクラフトの手法のみを用いた手法での実験を行っていないので，今回の実験結果と比較しどれほど有効かを確かめる必要がある．

今回の研究では動画単位の表情推定を行っている．しかし，動画単位での表情推定は応用範囲が狭いため，リアルタイムの認識ができるフレーム単位の表情推定が望ましい．動画内の各フレームに表情なしのラベルと感情ラベルを付与し，LSTM で学習することによって，フレーム単位で表情なしか，その会話中のどの表情かを推定することが可能になる．そのため，表情の有無分類の高精度化が今後の課題となる．

付録 A

研究で用いたデータの参照場所

研究に用いたすべてのプログラムとデータはヒューマンインタフェース研究室サーバー内の

```
/net/xserve0/users/fukiage/Master
```

のディレクトリ下に存在する．本ディレクトリの構成は以下のとおりである．

```
Master/            #研究用ディレクトリ
|
|--codes/           #作成したコード保存用ディレクトリ
|
|--Data/            #使用したデータ保存用ディレクトリ
```

詳しくは本ディレクトリ下の readme.md に示す．

付録 B

発表資料

修士論文発表では, 以下の資料を用いた.

2019/3/8

Mie Univ.

会話動画中の表情認識に関する研究

三重大学工学研究科情報工学科
ヒューマンインタフェース研究室
417M520　吹上　雅樹

1

Mie Univ.

表情認識

表情認識は盛んに行われている研究分野

表情認識による人の多彩な感情推定は、
人とロボットの高度なコミュニケーションの実現

ヒューマン・ロボット・インタラクションの高度化

応用例

- **コミュニケーションロボット**
- **介護用ロボット**

2

Mie Univ.

表情認識の関連研究

- **ハンドクラフト特徴量**を用いた表情認識手法[1]
- **深層学習**を用いた表情認識手法[2]

基本表情(喜び,悲しみ,恐怖,怒り,軽蔑,驚き)を対象とした研究が多い

➡**基本表情以外の表情**も認識できるシステムの実現と精度向上は重要な課題

本研究では、**会話中の表情**を認識対象とする。

[1]Shan et al, Facial expression recognition based on Local Binary Pattern, Image Vision Computer2009, pp 803 - 816
[2]Hasani et al, Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks, CVPR2017, Vol. abs/1705.07871,

3

Mie Univ.

基本表情と会話中の表情

基本表情	**会話中**の表情
6 ラベル **短時間**の表情変化 **表情**の動きのみ	**51** ラベル **長時間**の表情変化 **表情**と**頭部**の動き

4

Mie Univ.

課題と目的

会話中の表情認識における課題

- 従来の表情認識手法が適用できない

➡首を振る,うなずく等の**頭部の動き**がある

目的

- **頭部の動き**がある会話中の表情を認識可能なシステムの実現と精度向上

研究成果

国際会議 : IW-FCV 2019(査読有り)

5

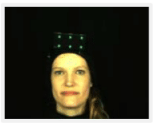
Mie Univ.

表情動画データセット

The large MPI Facial Expression Database[3]
会話中の表情を再現したデータセット

- 感情ラベル数 : 51
- 被験者数 : 10人(男性 : 5人 女性 : 5人)
- 動画数 : 510

無表情➡表情➡無表情 の表情変化を記録



[3]"The large MPI Facial Expression Database",BTU Cottbus-Senftenberg, https://www.b-tu.de/en/graphic-systems/databases/the-small-mpi-facial-expression-database

6

2019/3/8

Mie Univ.

認識対象とする感情

- 見た目では識別不可なラベルを統合
  →Imagine negative と Remember negative
- 認識対象とする16表情

Expression			
Agree	Aha light bulb moment	Annoyed	Bored
Compassion	Confused	I don't know	Disagree
Disbelief	Imagine negative	Imagine positive	Insecurity
Not convinced	Thinking	Tired	Bambi-eyes

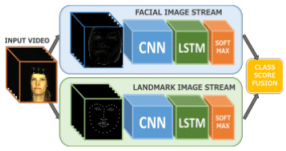
7

Mie Univ.

提案手法 – 概要

会話中の表情に有効な手法を提案

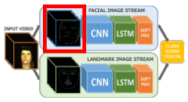
- **表情の動き**と**頭部の動き**を学習する
  ネットワークを組み合わせたアーキテクチャ
- **動画像**を入力し,**1つの感情ラベル**を出力する



8

Mie Univ.

提案手法 – 前処理



Facial Image stream

表情の動きを学習するネットワーク
入力 : **差分画像**

差分画像生成手法

OPENFACEによりアライメント処理を行い、フレーム毎の目鼻の位置を揃える

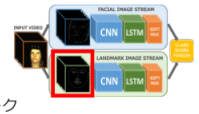
↓

入力フレームと0フレームの差分の絶対値をとる

9

Mie Univ.

提案手法 – 前処理



Landmark Image stream

頭部の動きを学習するネットワーク
入力 : **顔特徴点画像**

顔特徴点 - 顔や目などの輪郭の代表点

顔特徴点画像生成手法

顔画像ライブラリDlibを用いて顔特徴点画像を生成

↓

重み関数を用いて顔特徴点を拡大

$\omega(L,P) = 1 - 0.1 \cdot d_{M(L,P)}$

L : 顔特徴点座標
P : 注目画素
d_M : マンハッタン距離

10

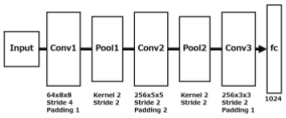
Mie Univ.

提案手法 – CNN

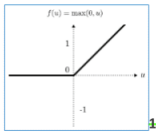
CNN

畳み込み層と全結合層により構成されるNN
学習を行うことで、識別に有効な特徴を抽出

CNNの構成



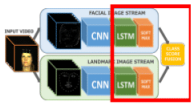
活性化関数 : ReLu関数



11

Mie Univ.

提案手法 –LSTM



LSTM

再帰構造を持つNN
短期、長期の系列データを識別可能

LSTMの構成

層 : 2　ユニット数 : 150
損失関数 : softmax cross entropy関数

Class Score Fusion

各ネットワークの事後確率を平均し、ラベルを決定する。

12

2019/3/8

Mie Univ.

提案手法 – 評価実験

- **データセット**
  The large MPI facial expression database
- **実験条件**
  Leave One Person Out 交差検証
  - 1人分を評価用, 1人分を検証用, 8人分を学習用,
  学習ループ数 : 2000
- **評価方法**

$$\text{認識成功率} = \frac{\text{推定成功動画数}}{\text{全動画数}} \times 100$$

13

Mie Univ.

提案手法 – 実験結果

表情と頭部の動き両方を学習したネットワーク、
表情の動きのみを学習したネットワーク、
頭部の動きのみを学習したネットワークを比較

MPI データセット

手法	認識成功率[%]
提案手法(2-stream)	**40.4**
Facial Stream	25.0
Landmark Stream	14.2

14

Mie Univ.

まとめ

頭部の動きがある会話中の16表情に対応した
2-streamアーキテクチャを提案した。

評価実験の結果, **表情だけを学習したネットワークで25.0%**だった認識精度が、**表情と頭部の両方を学習したネットワークでは40.4%**に向上した。

- **今後の課題**
  サンプル数が少ないため、ハンドクラフト特徴量と組み合わせる。

15

謝辞

本研究を進めるにあたり, 適切な御助言や様々な専門知識や技術をご指導, ご指摘頂いた若林哲史教授, 学会発表時における発表資料や投稿論文の添削, 授業を通して論文の書き方をご教授くださった大山航准教授, 研究のアドバイスをして下さった白井宙伸助教, そしてお忙しい中ディスカッションに参加してくださり, 積極的にアドバイスをして下さった三宅康二名誉教授に深く感謝します. また, 諸連絡の刑事や備品の貸し出し等研究をしやすい環境を作って下さった, ヒューマンインターフェース研究室事務員, 田中みゆきさん, 中塚沙智子さんに深く感謝します. そして, 本研究や研究生活への多くのアドバイスを与えて下さった研究室の先輩, 後輩の皆様, お互いの研究について議論し合い, 技術の共有を行い, ともに切磋琢磨した同期の皆様に深く感謝いたします．皆様のおかげでこの学生生活が私にとって非常に楽しく, 有意義なものとなりました．最後になりましたが, ここまで私の学生生活を支えてくれた家族に今一度の感謝の意を表して, 本論文の結びといたします.

参考文献

[1] A. Mehrabian. *Silent Messages: Implicit Communication of Emotions and Attitudes*. Wadsworth Publishing Company, 1981.

[2] Google Home. https://store.google.com/jp/product/google_home. (accessed Dec. 25, 2018).

[3] Apple Siri. https://www.apple.com/jp/siri/. (accessed Dec. 25, 2018).

[4] 大川弥生. 介護分野におけるコミュニケーションロボットの活用に関する大規模実証試験報告書. 2017.

[5] C Vinola and K Vimaladevi. A survey on human emotion recognition approaches, databases and applications. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, Vol. 14, No. 2, pp. 24–44, 2015.

[6] Ashok Samal and Prasana A Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition*, Vol. 25, No. 1, pp. 65–77, 1992.

[7] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, Vol. 36, No. 1, pp. 259–275, 2003.

[8] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, Vol. 1, No. 1, pp. 56–75, 1976.

[9] Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, Vol. 12, No. 5, p. 1161, 2012.

[10] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, Vol. 44, No. 3, pp. 572–587, 2011.

[11] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 71–82. Springer, 2007.

[12] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, Vol. 4, pp. 3437–3443. IEEE, 2005.

[13] Jukka Kortelainen, Suvi Tiinanen, Xiaohua Huang, Xiaobai Li, Seppo Laukka, Matti Pietikäinen, and Tapio Seppänen. Multimodal emotion recognition by combining physiological signals and facial expressions: a preliminary study. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 5238–5241. IEEE, 2012.

[14] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 37, No. 6, pp. 1113–1133, 2015.

[15] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, No. 6, pp. 681–685, 2001.

[16] Bihan Jiang, Michel F Valstar, Brais Martinez, and Maja Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Trans. Cybernetics*, Vol. 44, No. 2, pp. 161–174, 2014.

[17] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, Vol. 27, No. 12, pp. 1797–1803, 2009.

[18] Yan Tong, Wenhui Liao, and Qiang Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 29, No. 10, 2007.

[19] Yunfeng Zhu, Fernando De la Torre, Jeffrey F Cohn, and Yu-Jin Zhang. Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–8. IEEE, 2009.

[20] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, Vol. 27, No. 6, pp. 803–816, 2009.

[21] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1, pp. 886–893. IEEE, 2005.

[22] Zhen Wang and Zilu Ying. Facial expression recognition based on local phase quantization and sparse representation. In *Natural Computation (ICNC), 2012 Eighth Inter-*

national Conference on, pp. 222–225. IEEE, 2012.

[23] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pp. 428–441. Springer, 2006.

[24] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, Vol. 61, No. 1, pp. 38–59, 1995.

[25] Mohammad Reza Mohammadi, Emad Fatemizadeh, and Mohammad H Mahoor. Pca-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, Vol. 25, No. 5, pp. 1082–1092, 2014.

[26] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, Vol. 91, No. 1-2, pp. 160–187, 2003.

[27] Mohammed Yeasin, Baptiste Bullot, and Rajeev Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, Vol. 8, No. 3, pp. 500–508, 2006.

[28] Y Zhu, Liyanage C De Silva, and Chi Chung Ko. Using moment invariants and hmm in facial expression recognition. *Pattern Recognition Letters*, Vol. 23, No. 1-3, pp. 83–91, 2002.

[29] Yi Sun, Xiaochen Chen, Matthew Rosato, and Lijun Yin. Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 40, No. 3, pp. 461–474, 2010.

[30] Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. Authentic facial expression analysis. *Image and Vision Computing*, Vol. 25, No. 12, pp. 1856–1863, 2007.

[31] Behzad Hasani, Mohammad M Arzani, Mahmood Fathy, and Kaamran Raahemifar. Facial expression recognition with discriminatory graphical models. In *Signal Processing and Intelligent Systems (ICSPIS), International Conference of*, pp. 1–7. IEEE, 2016.

[32] Behzad Hasani and Mohammad H Mahoor. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pp. 790–795. IEEE, 2017.

[33] Suyog Jain, Changbo Hu, and Jake K Aggarwal. Facial expression recognition with

temporal modeling of shapes. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1642–1649. IEEE, 2011.

[34] Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, Vol. 104, No. 2-3, pp. 210–220, 2006.

[35] Sy Bor Wang, Ariadna Quattoni, Louis-Philippe Morency, David Demirdjian, and Trevor Darrell. Hidden conditional random fields for gesture recognition. In *null*, pp. 1521–1527. IEEE, 2006.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[38] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pp. 1–10. IEEE, 2016.

[39] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. Facial expression recognition from world wild web. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 58–65, 2016.

[40] Behzad Hassani and Mohammad H. Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. *CoRR*, Vol. abs/1705.07871, , 2017.

[41] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

[42] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[43] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 445–450. ACM, 2016.

[44] W Cunningham Douglas. The large mpi facial expression database. https://www.b-tu.de/en/graphic-systems/databases/the-large-mpi-facial-expression-database.

[45] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pp. 568–576, 2014.

[46] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 94–101. IEEE, 2010.

[47] OpenCV. https://opencv.org/. (accessed Feb. 13, 2019).

[48] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, Vol. 10, No. Jul, pp. 1755–1758, 2009.

[49] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, Vol. 1, p. 6, 2015.