J-027

オプティカルフローとニューラルネットを用いた顔の動き情報検出

Detection of facial motion information using optical flow and neural network

井上 真紀子† 小田 英介† 伊藤 昭† 寺田 和憲†
Makiko Inoue Eisuke Oda Akira Ito Kazunori Terada

1. はじめに

人はコミュニケーションにおいて，顔から多くの情報を得て心の状態や感情を推測しており，機械においても顔情報を利用することで人の心の状態を推測することが可能になると考えられる．これに関しては，従来から表情認識の研究が広く行われているが，表情認識では静止画像としての情報を用いるものが多く，顔の動き情報を用いるものは少ない．

そこで我々は，対話における顔の動きとそのときの「心の動き」の関連に着目し，顔の動き情報の検出と，その意味理解の研究を行っている．本研究では，まず顔の動きを捉えるため対話時の顔動画像に対してオプティカルフロー推定を行い，フローが一定の大きさ以上の連続フレームを抽出した．次に複数の人にその意味を評価してもらい，それを「教師データ」とすることで，顔の動き情報の検出を行った．

2. 顔の動き情報

2.1 対話映像

本研究で用いた動画像は，実験者と映像提供者との対話の様子をビデオカメラで撮影したものである．自然な対話の映像を得るため，映像提供者は実験者の友人から選出した．また実験目的を偽装することで，録画されていることを意識しない状態での対話映像を取得した．動画は画像サイズ 640×480，フレームレート 30fps であり，図 1 のように映像提供者の顔付近のみが撮影されている．

2.2 オプティカルフロー推定

対話映像から顔領域を検出し，オプティカルフロー推定を行う．本研究では勾配法を用い，第 2 の拘束条件には処理が高速な Lucas-Kanade 法を採用する．推定結果を図 1 に示す．図中の線分の長さが顔の移動量を表す．

2.3 特徴量

動き情報検出に用いる特徴量を各画素のオプティカルフローより求める．算出する特徴量を以下に示す．ただし，画像中の点 (x,y) におけるオプティカルフロー $\boldsymbol{f}_{(x,y)}$

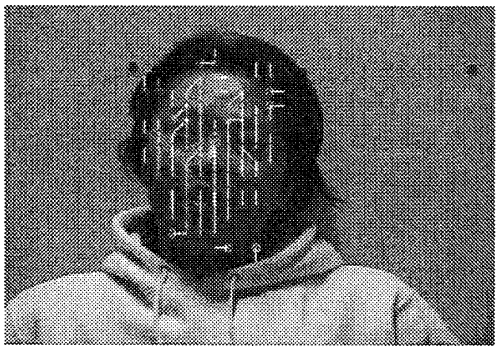


図 1: オプティカルフロー推定

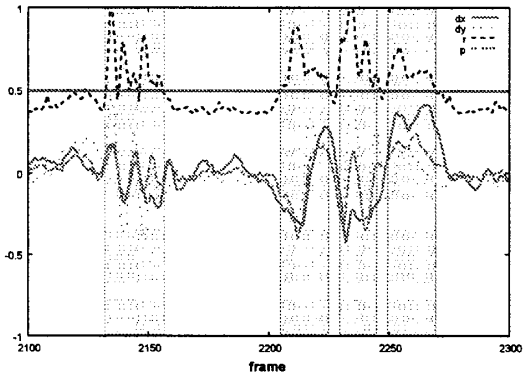


図 2: 特徴量のフレーム変化

の水平，垂直成分をそれぞれ $dx_{(x,y)}$，$dy_{(x,y)}$，画像の全画素数を N とする．

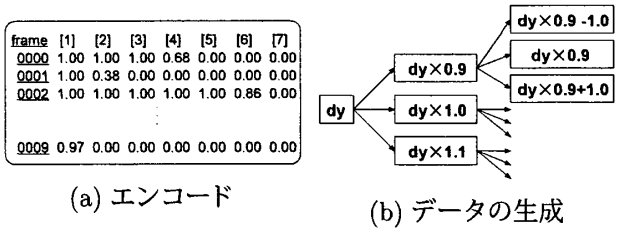
- $\overline{dx} =< dx >= \frac{1}{N}\sum_{x,y} dx_{(x,y)}$
- $\overline{dy} =< dy >= \frac{1}{N}\sum_{x,y} dy_{(x,y)}$
- $\overline{r} =< \|\boldsymbol{f}\| >= \frac{1}{N}\sum_{x,y}\sqrt{dx^2_{(x,y)} + dy^2_{(x,y)}}$
- $\overline{p} =< \boldsymbol{f}\cdot\boldsymbol{e} > \frac{1}{N}\sum_{x,y}\frac{dx_{(x,y)}\cdot x - dy_{(x,y)}\cdot y}{\sqrt{x^2+y^2}}$

3. ニューラルネット

3.1 入力データ

対話映像の中で，顔全体に一定以上の動き（特徴量 $\overline{r}$ が閾値 $\overline{r}_{th}$ 以上）がある連続フレームを 1 つのブロック

†岐阜大学大学院工学研究科



(a) エンコード　　(b) データの生成

図 3: 入力データ

として抽出し，この領域での特徴量の変化を用いて顔の動きを検出する (図 2).

今回は頷きのみを検出するため，入力値にはブロック内の $\overline{dy}$ をエンコードした値を用いる．具体的には各フレームでの $\overline{dy}$ を 7 ビットで表現し，それをブロック長だけ集めたものを入力値とする．ただしブロックのサイズが様々であるため，10 フレーム以上のブロックについてはブロック内の先頭 10 フレームのみを用いた．また 10 フレーム未満のブロックでは，不足フレーム分の値をすべて 0 とした (図 3(a)).

学習データの不足を補うため，図 3(b) のように入力値を 1.0 ± 0.1 倍し，さらに ± 1.0 したデータを各ブロック毎に生成することで学習データを 9 倍に増やす．これは評価に影響しない微妙な差異を学習させることを意味する．

3.2 学習方法

ニューラルネットにはフィードフォワード型三層パーセプトロンを用い，学習には誤差逆伝播法を採用する．ノード数は，入力層で 70(7×10)，隠れ層で 100，出力層で 2(頷き，その他) とする．

4. 実験

4.1 頷き検出

2. で述べたビデオ映像から頷きを検出した．ひとり約 5 分間 (12600frame)，3 人分の映像から $\overline{r} > \overline{r}_{th}$ のブロックを抽出し，これを人に評価してもらったものを教師データとしてニューラルネットを学習させた．抽出したブロック数はそれぞれ 230 個，140 個，185 個，そのうち 38〜39%が頷きである．各人のデータで 5-fold cross validation を用いて精度を検証した結果を表 1 の self に示す．これを見ると，どの映像提供者のデータにおいても人の評価と 85%以上の一致が見られ，本手法の有効性が示された．

次に，映像提供者間で cross validation を行った結果を表 1 の other に示す．例えば提供者 1 の結果は，提供者 2 及び提供者 3 のデータで学習させたニューラルネットで提供者 1 のデータをテストした結果である．これを

表 1: ニューラルネットによる検出結果

	提供者 1	提供者 2	提供者 3
self	86.5%	87.5%	89.4%
other	83.7%	87.5%	72.9%

見ると概ね人の評価と一致しており，個人によらず頷きを検出できていることが分かる．提供者 3 が他の提供者に比べて低くなっているが，これはブロックを抽出する際の $\overline{r}$ の閾値が異なるからであると考えられる．

4.2 全フレーム検出

これまでは，$\overline{r} > \overline{r}_{th}$ で切り出された領域だけを対象に検出を行ってきたが，ここでは全映像からの頷きの検出を試みる．

具体的には，前節で生成した提供者 1 の学習データを，先頭 10 フレームだけではなく，先頭から 1 フレームずつずらして作れる長さ 10 フレームの領域全てを学習データとして，ニューラルネットを学習させ，検出器とした．次に，全映像に対して，先頭から長さ 10 フレームのウィンドウを 1 フレームずつずらして作り，上記の検出器で頷きの検出を行った．その結果，全データ 12,589 個のうち，学習データでの誤検出は 1 個 (0.000079%)，未検出は 24 個 (0.0019%)，未知データで頷きと検出した領域は 504 個 (4.0%) であった．未知データで頷きと検出した領域のうち，5 個以上連続して頷きとして検出した領域は 37 個であった．実際の映像を見ると，これら全てが誤検出ではなく，小さな頷きと認識される領域がかなり含まれることが分かった．現在，この領域のデータについての評価によりラベル付けを行っている．

5. まとめ

本研究では，オプティカルフローから求めた特徴量でニューラルネットを学習させ，顔の動きのひとつである頷きを検出した．顔の動きが一定以上の大きさのフレームを抽出することにより，ニューラルネットで時系列情報を扱うことが可能になった．また，本手法が個人に寄らず有効であることを示した．さらに，本手法を用いて全フレーム検出を試み，リアルタイムでの検出が可能であることを示唆した．

今後は頷き以外の動き情報「首を傾げる」「首を横に振る」「笑う」などの検出を行う．これは，検出に $\overline{dy}$ 以外の特徴量を用いることで実現可能であると考えている．さらに，これらをリアルタイムで検出することを目指す．長期的には，心の動きのモデルを考え対応づけを行いたい．