J-023

対話中の頭部運動機能を認識する特徴膨張収縮ニューラルネットワーク
Feature Inflation-Deflation Neural Networks for Recognizing Head-Movement Functions in Conversations

武田 一輝 *　　大塚 和弘 †
Kazuki Takeda　　Kazuhiro Otsuka

1 はじめに

対話中の人物行動や心情の認識はコミュニケーションの質的な評価や人間関係を分析するうえで重要な課題である。この課題に対して，身振り手振りや頭部の動きなどの非言語情報が豊富な手掛かりを持つことから注目を集めている。特に頭部運動は対話中で重要な機能を持つことが知られ，その運動は状況によって様々な機能を担う．また頭部運動は一度に一つの機能が表出されるだけでなく，複数の機能が多重化され同時に現れる場合がある．

このような点に着目し，Otsuka & Tsumori は対話中の頭部運動が持つ多様な機能を定義し，複数の外部評定者により重複を許容したアノテーションを行い頭部運動機能コーパス（Functional Head-Movement Corpus, 以後 FHM コーパスと呼ぶ）を作成した．彼らはこのコーパスを利用して頭部運動と発話状況の時系列データから各機能の有無の 2 値分類を行う問題を定式化し，この問題のための畳み込みニューラルネットワーク（Convolutional Neural Networks, CNN）を提案した．実験の結果，F 値で約 0.3∼0.9 と機能間において認識性能のばらつきが大きく，性能向上の余地が大きく残されている．

本稿では，頭部運動機能の認識という課題に対して，CNN や他の深層ニューラルネットワーク（以降 CNN を含めて DNN と呼ぶ）の認識性能向上を目指し，DNN に付加する新しい機構として「特徴膨張収縮機構」（feature Inflation-Deflation module，I/DeF 機構）を提案する．この機構は，DNN に入力される時系列データに対して，微視的スケール方向および入力時系列の時間窓の前後方向に膨張と収縮を繰り返し行うことで，稠密かつ子細な時間構造の特徴学習の促進を狙った機構である．

本稿では，この I/DeF 機構を CNN，VGG[1]，Residual Network（ResNet）[2] の入力部に付加した新しい DNN を提案する．頭部運動機能カテゴリ 10 種に対して，提案モデルと従来法である CNN モデルの認識性能を比較した結果，F 値で最大 4.5 ポイント，平均 2.3 ポイントの性能向上を確認し，I/DeF 機構の有効性が示唆された．

* 横浜国立大学 大学院 理工学府 Graduate School of Engineering Science, Yokohama National University
† 横浜国立大学 大学院 工学研究院 Faculty of Engineering, Yokohama National University

2 関連研究

CNN は当初，画像認識を対象として提案された DNN であるが，近年では認識対象を広げ，時系列データの認識へも適用されている．特に，日常生活で生じる人物行動を対象とした時系列認識においても，標準的なモデルとして用いられ [3]，高い認識性能を持つことが知られている [4].

従来，この CNN を改良する種々の方法が提案されてきた．例えば VGG や ResNet のように畳み込み層を増やす深層化があげられる．この深層化は畳み込み層を重ねることで，特徴表現の能力が向上し，データに内在する複雑なパターンの認識が可能となる．Hu らが提案した CNN への付加機構である Squeeze and Excitation 機構（SE 機構）[5] もまた有効な手段として知られている．これは CNN 内の畳み込み層の出力に対して Global Average Pooling(GAP) による平均値の計算を行い，その値と入力されたデータを乗算することで，重要な特徴に対して注意を配分する一種の Attention 機構として働く．それにより，CNN の性能が大きく向上することが確認されている [5].

本稿で提案する I/DeF 機構は以上とは異なる方法論として位置づけられる．この機構では入力時系列データに対して，膨張過程によるアップスケーリングと収縮過程による特徴圧縮の 2 つを繰り返し行う．この機構は，従来，主に画像認識や画像生成で用いられていた畳み込みオートエンコーダ [6][7] から着想を得ている．画像認識における畳み込みオートエンコーダでは複数回の畳み込みにより，情報を抽象化する．その後，複数回の転置畳み込みにより，抽象化された情報を元の画像次元に復元している．I/DeF 機構ではこの畳み込みオートエンコーダの後半に行われる画像復元の過程を，DNN の入力の前段に配置し，元の入力時系列のサンプル間隔よりも密なデータを獲得することで，入力時系列の詳細な時間構造について特徴学習を促進させることを狙っている．

また I/DeF 機構は VGG や ResNet のような深層化された DNN や SE 機構と合わせて統合することが可能であり，各方法論との相乗効果が期待される．

3 提案手法

3.1 全体構成

本稿で提案する I/DeF 機構は DNN に入力される時系列データに対して，時間軸上の膨張および収縮を行う機構である．この機構の膨張過程では，ある窓幅を掛けて入力された時系列データに対して，転置畳み込みを用いて，窓幅の前後

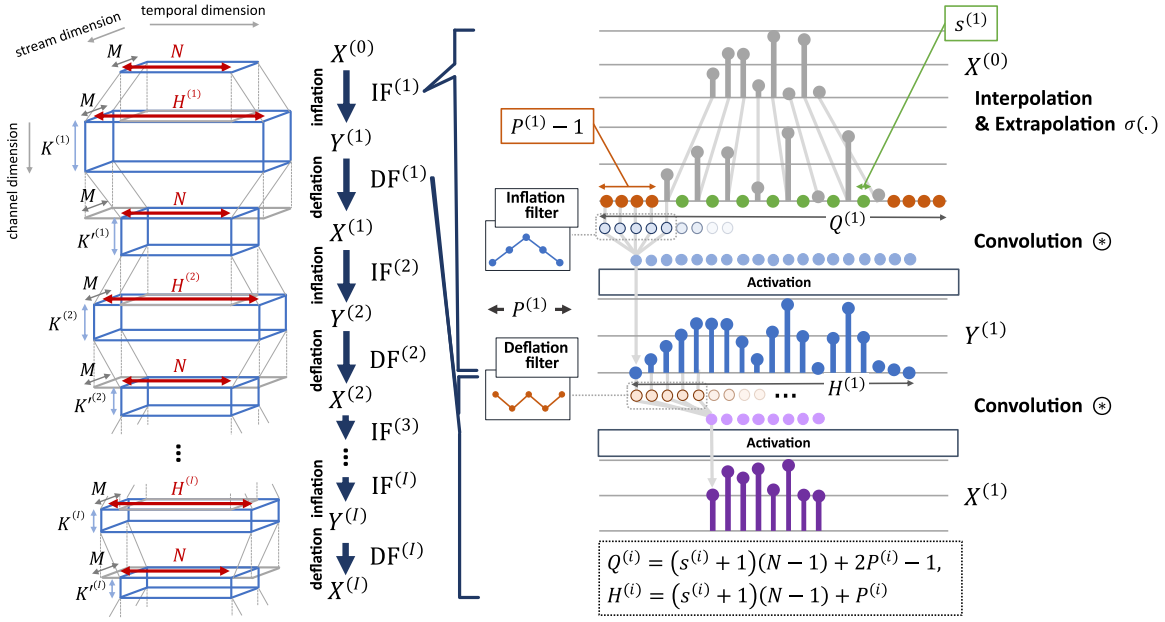

図 1　特徴膨張収縮機構．左側は膨張・収縮によってどのようにデータが拡大・圧縮されるかを表しており，右側は $\mathbf{IF}^{(1)}$, $\mathbf{DF}^{(1)}$ によるデータの変化を表す．

方向へのデータの拡張（外挿）および微視的方向へのアップスケーリング（内挿）を行う．収縮過程では，膨張した時系列データに対して畳み込み処理を行うことで，元の窓幅のサンプル数を持つデータへと次元削減を行う．I/DeF 機構ではこれらの膨張過程，収縮過程を繰り返し行う構造を持つ．

I/DeF 機構の概念図を図 1 に示す．I/DeF 機構では複数の時系列データを入力対象とする．ここで入力時系列の個数を M，窓幅が N である入力を $\boldsymbol{X}^{(0)}$ と記す．この入力に対して膨張収縮を 1 度行ったときの出力を $\boldsymbol{X}^{(1)}$，また複数回繰り返した時の I/DeF 機構の出力を $\boldsymbol{X}^{(I)}$（I は繰り返し回数）と記すと，膨張収縮過程は

$$\boldsymbol{X}^{(i)} = \mathrm{IDF}^{(i)}(\boldsymbol{X}^{(i-1)}),\ i = 1, \ldots, I \tag{1}$$

と再帰的に表せられる．ここで $\mathrm{IDF}^{(i)}$ は i 回目の膨張収縮過程を表し，この中の膨張過程を $\mathrm{IF}^{(i)}$，収縮過程を $\mathrm{DF}^{(i)}$ と記すと，1 回の膨張収縮処理は

$$\begin{aligned} \mathrm{IDF}^{(i)}(\boldsymbol{X}^{(i-1)}) &= \mathrm{DF}^{(i)}(\boldsymbol{Y}^{(i)}), \\ \boldsymbol{Y}^{(i)} &= \mathrm{IF}^{(i)}(\boldsymbol{X}^{(i-1)}), \end{aligned} \tag{2}$$

のように表すことができる．ただし，式 (2) における $\boldsymbol{Y}^{(i)}$ は i 回目での膨張過程による出力を表す．

式 (1),(2) では，まず膨張処理 $\mathrm{IF}^{(1)}$ により時間幅 N の時系列入力 $\boldsymbol{X}^{(0)}$ が，まず時間幅 $H^{(1)}$ のデータ $\boldsymbol{Y}^{(1)}$ へと膨張される．その後，収縮処理 $\mathrm{DF}^{(1)}$ によって，$\boldsymbol{Y}^{(1)}$ から元の時間幅 N のデータ $\boldsymbol{X}^{(1)}$ へと収縮する．またこの膨張過程での畳み込みフィルタ数を $K^{(i)}$ 個，収縮過程でのフィルタ数を $K'^{(i)}$ 個と表す．以後，膨張，収縮過程での畳み込みフィルタを膨張フィルタ，収縮フィルタとそれぞれ呼ぶ．このような膨張フィルタの個数 $K^{(i)}$，収縮フィルタの個数 $K'^{(i)}$ に応じてもデータが拡大収縮される．

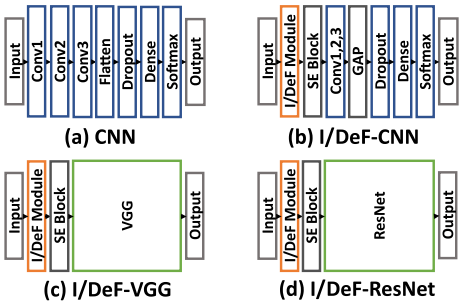

図 2　CNN および I/DeF 組み込みモデルの構造

3.2　特徴膨張収縮

3.2.1　特徴膨張過程

i 回目の入力を $\boldsymbol{X}^{(i-1)}$ と表す．また i 回目の膨張過程での出力を $\boldsymbol{Y}^{(i)}$ と表す．この入力 $\boldsymbol{X}^{(i-1)}$ に対して出力 $\boldsymbol{Y}^{(i)}$ は，転置畳み込みによって，

$$\boldsymbol{Y}^{(i)} = g\left(\boldsymbol{W}^{(i)} \circledast \sigma\left(\boldsymbol{X}^{(i-1)}\right) + \boldsymbol{B}^{(i)}\right) \tag{3}$$

のように得られる．ただし，g は活性化関数，$\boldsymbol{W}^{(i)}$ は重み係数，$\circledast$ は畳み込み処理，$\sigma(.)$ は入力に対して過去未来方向に $P^{(i)} - 1$ 個の 0 を外挿し，サンプル間に $s^{(i)}$ 個の 0 を内挿する処理，$\boldsymbol{B}^{(i)}$ はバイアスを表す．ここで $P^{(i)}$ は畳み込みフィルタの時間方向の長さである．また $s^{(i)}$ を内挿幅と呼ぶ．なおこの式 (3) 中の $\sigma(.)$ により入力データのサンプル数は $Q^{(i)}$ に変化し，式 (3) によって膨張処理の出力サンプル数は $H^{(i)}$ となる．ここでサンプル数 $Q^{(i)}$ および $H^{(i)}$ は図 1 下部に示される式で定義される．この膨張過程により，時間窓の前後方向に拡張され、また、サンプル間が内挿されアップスケーリングが施された時系列データを得る。

表 1　内挿幅を可変とする I/DeF 機構のパラメータセット Sδ

	Sδ	
パラメータ	$i=1$	$i=2$
膨張フィルタ数 $K^{(i)}$	50	20
収縮フィルタ数 $K'^{(i)}$	50	20
フィルタ長 $P^{(i)}$	10	10
内挿幅 $s^{(i)}$	δ	δ
活性化関数 g	ReLU	

3.2.2　特徴収縮過程

収縮過程は直前の膨張過程の出力 $\boldsymbol{Y}^{(i)}$ を入力とし，畳み込みフィルタにより出力 $\boldsymbol{X}^{(i)}$ を得る処理であり，

$$\boldsymbol{X}^{(i)} = g\left(\boldsymbol{W'}^{(i)} \circledast \boldsymbol{Y}^{(i)} + \boldsymbol{B'}^{(i)}\right) \tag{4}$$

のように書ける．ただし g は活性化関数，$\boldsymbol{W'}^{(i)}$ は重み係数，$\circledast$ は畳み込み処理，$\boldsymbol{B'}^{(i)}$ はバイアスを表す．ここで式 (4) の畳み込みの際の収縮フィルタの時間方向の長さを膨張過程におけるフィルタの時間方向の長さ $P^{(i)}$，ストライドを $s^{(i)}+1$ とすることで，畳み込みを行った際に，入力時と同じサンプル数を持つデータへと圧縮される．この収縮過程により，直前の膨張過程で拡張されたデータに対し，元のサンプル数へと次元削減を行うと同時に，拡張された時系列データに関する特徴表現を獲得する．

3.3　深層ニューラルネットワークとの統合

I/DeF 機構は DNN の入力の前段に付加される．本稿では，CNN，VGG，ResNet に I/DeF 機構をそれぞれ統合したモデルとして I/DeF-CNN，I/DeF-VGG，I/DeF-ResNet を提案する．

I/DeF-CNN は図 2(a) に示す文献 [8] の CNN に，I/DeF 機構を統合したモデルである．このモデルの構造を図 2(b) に示す．なお Conv1，Conv2 では畳み込み処理の後に，最大プーリングを行う．従来モデルと異なり，I/DeF-CNN では膨張収縮によりデータが多重化されるため，過学習を防ぐ役割として Conv3 の出力に対して GAP を付加した．また I/DeF-CNN では I/DeF 機構と CNN との間に SE block [5] を挿入し，I/DeF 機構の各チャンネルのデータに重みをつけ，より重要な特徴に着目するように促す．

図 2(c),(d) に I/DeF-VGG と I/DeF-ResNet の構造をそれぞれ示す．なお VGG は文献 [1] の VGG-16 を用いる．また ResNet として，文献 [2] の ResNet-110 を用いる．ここで VGG，ResNet は画像認識に対して構築されたモデルであるため，元の構造から畳み込みフィルタのサイズを変更することで時系列データ認識への適用を図る．I/DeF 機構の統合は I/DeF-CNN と同様とし，入力部に I/DeF 機構および SE 機構が付加される．

4　実験

本稿では，頭部運動機能認識のため FHM コーパス [8] を用いた．このコーパスでは，女性 4 人による対面対話 4 セッションを対象として，会話映像のフレーム単位に 32 種類の頭部運動機能ラベルが非排他的に付与されている．今回対象とする頭部運動機能カテゴリは，従来研究と同様にリズム取り (s1)，強調 (s2)，反応確認 (s5)，思考（発話時）(s8)，相槌 (r1)，応答 (r2)，思考（受け手）(r5)，理解 (r6)，肯定 (r11)，正の感情表出 (c1) の出現頻度上位 10 種類とし，これらを学習時および評価時の正解ラベルとした．

実験にあたり，各機能の有無を出力する 2 値分類モデルを構築した．モデルへの入力には，FHM コーパスに含まれる頭部姿勢情報と，発話の状況の時系列データを用いた．頭部姿勢情報としては，対話者の頭部に装着されたセンサにより得られる方位角，仰角，ロール角の頭部姿勢 3 成分について，それぞれフレーム間差分により計算される角速度を用いた．発話の状況としては，元々 2 値である発話の有無の情報に，移動平均フィルタを用いて平滑化を行い，実数値化された時系列データを用いた．頭部姿勢角速度と平滑化された発話状況の時系列データを入力として扱う際，認識対象となるフレームを中央に，その前後のフレームでのデータも含むような時間窓を設定し，抽出した．

提案した I/DeF-CNN，I/DeF-VGG，I/DeF-ResNet に対して，比較を行うモデルとして，3.3 に示した CNN，VGG，ResNet を実装した．また，I/DeF 機構のパラメータセットは表 1 に従い，表 1 中の Sδ を用いた I/DeF-CNN を I/DeF-CNN(Sδ) と記す．

モデルの認識性能を評価するため，各機能における各フレームでの 2 値分類の結果と，正解ラベルにより算出された F 値を用いた．データセット内のデータ数が少数であることから，従来研究に従い交差検証法を行った．なお個人による正解ラベルの偏りを考慮し，micro-F 値を用いた．

5　結果と考察

表 2 に I/DeF-CNN, I/DeF-VGG, I/DeF-ResNet, CNN, VGG，および ResNet の評価結果を示す．

まず表 2A について，従来法である CNN と比較すると，I/DeF-CNN(Sδ)($\delta=0,1$) では全ての機能にて F 値の向上が確認できた．向上幅は最大で I/DeF-CNN(S0) では r5 の 4.5 ポイント，I/DeF-CNN(S1) では r11 の 3.7 ポイントであった．一方，I/DeF-CNN(Sδ)($\delta=2,3$) では CNN より F 値が低下した機能が存在し，平均で見た向上幅も小さくなった．以上，CNN に対する性能向上から，頭部運動機能認識に対して I/DeF 機構は有効に機能したことが示唆された．また，内挿幅に関してはある程度有効な範囲があることがわかった．

次に表 2B について，VGG では CNN と比べて，全ての機能にて F 値の低下が生じた．これは，複雑な頭部運動の時系列パターンを VGG が学習できなかったためだと考えられる．一方，VGG と比べて，I/DeF-VGG では 8 割の機能にて F 値が向上した．特に c1 においては，14.1 ポイント向上した．依然として I/DeF-VGG の性能は従来の CNN よりも劣っているが，I/DeF 機構による性能向上は頭部運動の特徴学習が促進された結果であると考えられる．

表 2C について，ResNet においては CNN に比べて 7 割の機能で F 値が向上している．この性能向上は，ResNet の特

表 2　頭部運動機能カテゴリに対する認識性能（F 値）．表中の太文字は比較モデルの中で最高の F 値を示す

番号	カテゴリ名	従来法	A)I/DeFs-CNN（内挿幅の違い）				B)VGG		C)ResNet		
		CNN	I/DeFs-CNN(S0)	I/DeFs-CNN(S1)	I/DeFs-CNN(S2)	I/DeFs-CNN(S3)	VGG	I/DeFs-VGG(S1)	ResNet	I/DeFs-ResNet(S0)	I/DeFs-ResNet(S1)
s1	リズム取り	0.750	**0.769**	0.763	0.752	0.757	0.716	0.746	0.760	0.759	0.766
s2	強調	0.555	0.573	0.575	0.568	0.574	0.530	0.539	0.570	0.580	**0.590**
s5	反応確認	0.576	**0.590**	**0.590**	0.589	0.584	0.525	0.562	0.574	0.582	0.583
s8	思考（発話時）	0.374	0.382	0.378	0.368	0.365	0.226	0.227	**0.385**	0.339	0.357
r1	相槌	0.874	**0.885**	0.884	0.873	0.877	0.789	0.871	0.873	0.884	0.883
r2	応答	0.330	**0.347**	0.339	0.332	0.318	0.224	0.212	0.326	0.329	0.318
r5	思考（受け手）	0.400	**0.445**	0.433	0.424	0.416	0.272	0.383	0.426	0.402	0.418
r6	理解	0.236	0.265	0.269	0.248	0.237	0.192	0.188	0.251	**0.278**	0.174
r11	肯定	0.254	0.284	**0.291**	0.267	0.290	0.158	0.166	0.266	0.230	0.216
c1	正の感情表出	0.411	0.431	0.431	0.412	0.397	0.171	0.312	**0.441**	0.418	0.366
	Average	0.476	**0.497**	0.495	0.483	0.482	0.380	0.421	0.487	0.480	0.467

長が頭部運動機能の認識にも有効であったことの証左といえる．また表 2C には I/DeF-ResNet(Sδ)($\delta = 0, 1$) の結果も併せて示す．これらと ResNet を比べると，F 値が向上した機能は I/DeFs-ResNet(S0) では半数, I/DeFs-ResNet(S1) は 4 割と一部の機能にとどまった．このことは ResNet では既に CNN に比べて顕著な性能向上を果たしていたため，I/DeF 機構による性能向上は限定的であったことが示唆される．

以上のモデルを全て比較した際に，I/DeF-CNN は 10 機能中 6 機能で F 値が最高となった．中でも，F 値の向上幅が最高であったのは，r5 における I/DeF-CNN(S0) の 4.5 ポイントであった．

6　議論

本稿では頭部運動機能認識の性能向上を目的として，時系列データの特徴学習を促進させるための DNN の付加機構として I/DeF 機構を提案した．今後さらなる認識性能向上には，I/DeF 機構の改良やデータセットの増加などが考えられる．このうち，I/DeF 機構の改良に関しては，機構の動作や構造を解析し，膨張や収縮の過程によりどのようにデータが拡張、抽象化されているかを明らかにすることが機構改良の手がかりにつながると考えられる．

また本研究は，対話中の非言語行動のモダリティとして頭部運動に焦点を絞ったが，非言語行動には視線や表情など多数存在する．そのため，これら複数モダリティの行動を時系列情報として入力とすることで，提案モデルをより多様な非言語行動の機能の解析に利用することができると考える．

7　結び

対話中の頭部運動機能認識のため新たな深層ニューラルネットワークの機構として特徴膨張収縮機構 (I/DeF 機構) を提案した．この機構は，転置畳み込みによる膨張と畳み込みによる収縮によって，入力時系列データに含まれる時間構造の特徴学習を促進することを狙った機構である．I/DeF 機構を統合した I/DeF-CNN について，頻出する頭部運動機能 10 種を対象として，各機能の 2 値分類を行った結果，従来法である CNN の性能を全機能において上回った．また I/DeF 機構を VGG，ResNet にも統合し比較した結果，I/DeF 機構を統合した CNN が 8 割の機能カテゴリにおいて最も優れた性能を示した．以上，本稿にて提案した I/DeF 機構の有効性が確認された．今後，様々な時系列データ認識において，活用が期待される．

謝辞

本研究をご支援いただいた日本電信電話株式会社コミュニケーション科学基礎研究所に感謝いたします．本研究は一部，栢森情報科学振興財団，立石科学技術振興財団，及び JSPS 科研費 JP21K12011 の助成を受けて遂行されました．

参考文献

[1] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *2015 Int. Conf. Learning Representations (ICLR)*, Vol. abs/1409.1556, pp. 1–14, 2015.

[2] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[3] Salwa O. Slim, Ayman Atia, Marwa M.A. Elfattah, and Mostafa-Sami M.Mostafa. Survey on human activity recognition based on acceleration data. *Int. J. Advanced Computer Science and Applications*, Vol. 10, No. 3, pp. 84–98, 2019.

[4] Carlos Avilés-Cruz, Andrés Ferreyra-Ramírez, Arturo Zúñiga-López, and Juan Villegas-Cortéz. Coarse-fine convolutional deep-learning strategy for human activity recognition. *Sensors*, Vol. 19, No. 7, p. 1556, 2019.

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, June 2018.

[6] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proc. 2015 IEEE Int. Conf. Computer Vision (ICCV)*, pp. 1520–1528, 2015.

[7] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. Proc. Int. Conf. Learning Representations (ICLR), 2016.

[8] K. Otsuka and M. Tsumori. Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, Vol. 8, pp. 217169–217195, 2020.