一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NC2023-13,IBISML2023-13(2023-06)

基本感情の関係性を活用した表情認識のための Hierarchical Classification Transformer

三好　遼† 秋月　秀一† 飛谷　謙介†† 長田　典子††† 橋本　学†

† 中京大学大学院工学研究科　〒466–8666 愛知県名古屋市昭和区八事本町 101–2
††† 関西学院大学　〒699–1330 兵庫県三田市学園上ケ原 1 番
†† 長崎県立大学　〒851–851 長崎県西彼杵郡長与町まなび野 1–1–1
E-mail: †{miyoshi, mana}@isl.sist.chukyo-u.ac.jp, †s-akizuki@sist.chukyo-u.ac.jp, ††tobitani@sun.ac.jp, †††nagata@kwansei.co.jp

あらまし　表情認識を利用した感情認識技術は自然さの観点からさまざまな場面での活用が期待されている．しかし，表情の個人差が原因で依然と困難なタスクである．従来の表情認識手法においては，表情のみを表現する特徴を抽出することに着目されて取り組まれている．また，表情は独立したカテゴリ変数として扱われている．心理学分野において，基本感情は類似性や対極性という性質を持つと示唆されている．例えば，「怒り」と「嫌悪」や類似感情である．また，基本感情は，「ネガティブ」，「ポジティブ」，「その他」の感情に分けられる．本研究では，これらの基本感情の関係性に基づき，それを階層的に分類する transformer ベースの表情認識手法を提案する．大規模な自然環境下におけるデータセットである DFEW によって提案手法を評価したところ，従来手法よりも高い認識率を示した．
キーワード　表情認識，Transformer, 基本感情, 感情モデル

Hierarchical Classification Transformer for Facial Expression Recognition Utilizing Basic Emotion Relationships

Ryo MIYOSHITOKYO†, Shuichi AKIZUKI†, Kensuke TOBITANI††, Noriko NAGATA†††, and Manabu HASHIMOTO†

† Graduate School of Engineering, Chukyo University　101–2 Yagoto-honmachi, Showa-ku, Aichi, 466–8666 Japan
††† School of Science and Technology, Kwansei Gakuin University　1 Gakuenuegahara, Sanda-shi, Hyogo, 699–1330 Japan
†† Faculty of Information Systems, University of Nagasaki　1–1–1 Nagayo-chou, Nishisonogi-gun, Nagasaki, 851–1330 Japan
E-mail: †{miyoshi, mana}@isl.sist.chukyo-u.ac.jp, †s-akizuki@sist.chukyo-u.ac.jp, ††tobitani@sun.ac.jp, †††nagata@kwansei.co.jp

Abstract　Emotion recognition technology based on facial expression recognition(FER) is expected to be used in various situations from the viewpoint of naturalness. However, it is a difficult task due to individual differences in facial expressions. Conventional FER methods have focused on extracting features that express only facial expressions. In addition, facial expressions are treated as independent categorical variables. In the field of psychology, it has been suggested that basic emotions have properties of similarity and opposition. For example, "anger" and "disgust" and similar emotions. In addition, basic emotions can be divided into "negative," "positive," and "other" emotions. We propose a transformer-based FER method that hierarchically classifies these basic emotions based on their relationships. We evaluated the proposed method on the DFEW dataset, which is a large-scale natural environment dataset, and found that the accuracy of the proposed method is higher than that of conventional methods.
Key words　Facial Expression Recognition，Transformer,Basic Emotion, Emotion model

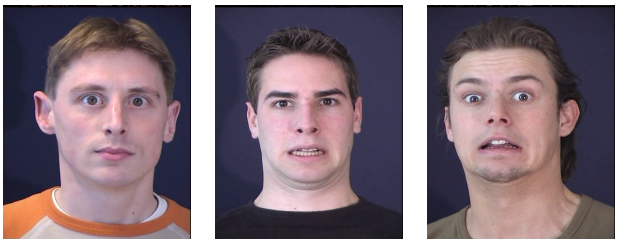
1. はじめに

表情は，感情や意図を伝達するための重要な非言語的情報の 1 つである．表情認識による感情推定は，自然さという観点からヒューマンコンピュータインタラクション（HCI）[1] やモビリティ [2] といった幅広い分野での活用が期待されている．

しかし，表情認識タスクは依然として困難なタスクである．その原因として，表情の個人差があげられる．そして，表情の個人差は次の 2 つの要因に分解できると考えられる．一つ目は，個人の顔の作りの違いである．同じ表情を表現していたとしても，個人間で顔の作りが違うことから表情の違いが生まれてしまう．二つ目は，表情の表現方法の違いである．同じ感情を表現する表情をしていたとしても，個人間で表情表現が異なることにより表情の違いが生まれてしまう．図 1 に表情表現の違いの例を示す．図 1(a)，図 1(b)，ともに同じ感情を表現している表情である．図 1(a) が実験室環境下における表情，図 1(b) が自然環境下における表情である．実験室環境下においても個人間で表情が大きく異なり，その違いは自然環境下になるとより顕著である．

画像認識分野における表情認識は，データ処理の利便性や関連する学習およびテストデータの入手可能性から，時間情報を考慮せずに静止画像に基づく表情認識について取り組まれてきた．近年では，表情は顔の時空間変化によって表現されることから，時間情報をより直接的に活用することが有効と考えられ，動画ベースの表情認識が活発なっている [3]．多くの研究において，無表情と表情が喚起されている顔の変化，すなわち，表情のみを表現する特徴の獲得することによってロバスト性を向上させる手法が提案されている [4]〜[6]．文献 [4] では，表情変化を捉えるために有効な顔の時空間的領域に着目するための graph neural network を提案している．文献 [5] では，表情変化を捉えるために，現在の時間ステップからより前の時間における情報を参照可能な Convolutionlan LSTM を提案している．文献 [6] では，差分画像を入力とすることによって，個人の顔の作りに依存しない手法が提案されている．さらに，無表情と表情が喚起されている顔画像の特徴量との相互情報量を最小化することによって表情のみを表現する特徴量を抽出している．これらの手法のアプローチは，個人差における個人の顔の違いを取り除くアプローチとみなすことができる．また，画像および動画認識において transformer [7] を応用した手法が高い精度を示している [8], [9]．そして，動画ベース表情認識においても transformer を応用した手法が提案されており，高い精度を示している [10], [11]．従来手法では，表情は独立したカテゴリ変数として扱っている．

表情は感情が喚起されることによって発現される非言語的行動であるため，感情と密接な関わりを持っている．心理学分野において，人の基本感情は円環状に配列されているというモデルが提唱されている [12], [13]．図 2 にラッセルの感情円環モデルを示す [12]．これらのモデルにおいて，感情は類似性（e.g. 怒りと嫌悪）や対極性（e.g. 喜びと悲しみ）といったような関係性を持っている．ここで，基本感情は類似性をもとにカテゴライズできる．「喜び」はポジティブ感情，「恐れ」，「怒り」，「嫌悪」，「悲しみ」はネガティブ感情，「驚き」はどちらにも属さないその他の感情と分類できる．



(a) 実験室環境下における個人間の表情の違い

(b) 自然環境下における個人間の表情の違い

図 1　個人間の表情の違いの例

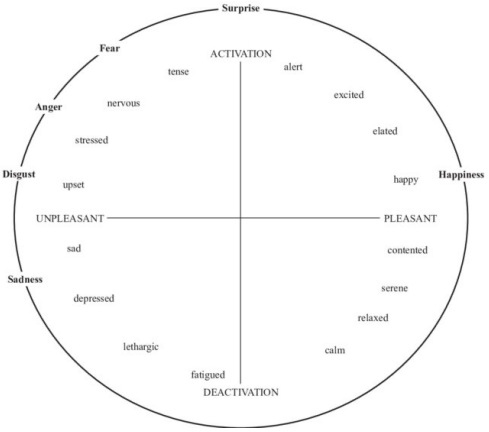
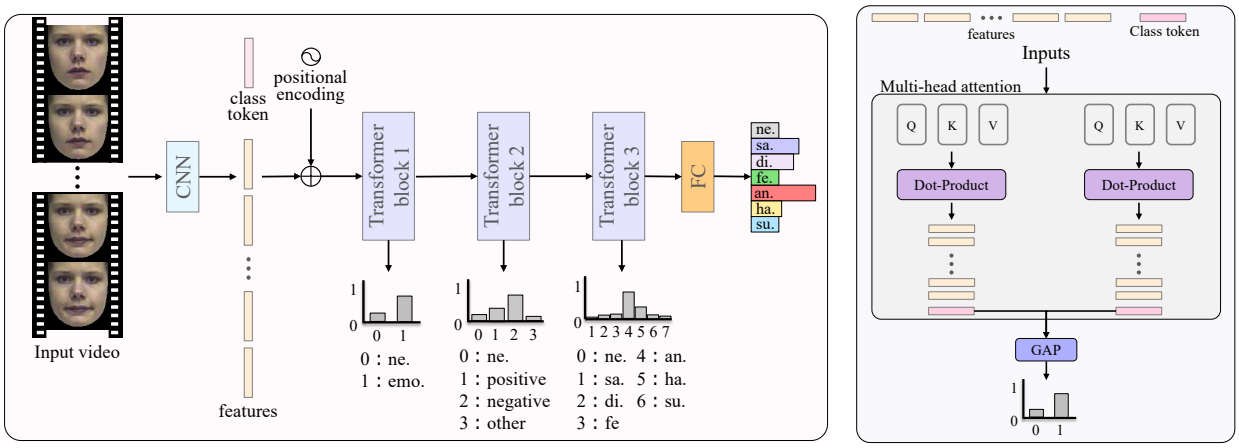


図 2　ラッセルの感情円環モデル [12]

本研究では動画ベース表情認識における transformer の台頭や心理学分野における感情に関する知見を背景に，感情の類似性といった関係性を活用した transformer ベースの表情認識手法を提案する．従来手法において，表情は独立したカテゴリ変数として扱われていたが，基本感情の円環構造に基づく感情の類似性をデータに対する制約，すなわち，帰納バイアスとして活用する．提案手法では表情のクラス分類の学習に加えて，transformer の中間レイヤーにおいて，感情の関連性に基づくカテゴリの分類を学習するサブタスクを導入する．さらに，そのサブタスクを導入することによって，モデルの解釈性の向上を目指す．

提案手法は，大規模な自然環境下における表情認識データセットによって評価された．実験の結果，提案手法は，baseline 手法や従来手法より高い認識率を示した．さらに，提案手法における中間レイヤーでの誤分類について分析したところ，最終的に誤認識をしている場合は，「ポジション」，「ネガティブ」，



(a) 提案する Hierarchical Classification Transformer

(b) Hierarchical Classification Transformer block の概要

図 3　提案手法の概要

「その他」といったカテゴリの分類において失敗している傾向があった．

2. 提案する Hierarchical Classification Transformer

本研究では，感情の類似性という関係性を帰納バイアスとして活用する．提案する Hierarchical Classification Transformer（HCT）の概要を図 3 に示す．

提案手法は，$X \in \mathbb{R}^{T\times 3\times H\times W}$ である動画を入力とする．ここで，T はフレーム数，H は各フレームの画像の高さ，W は各フレームの横幅を示す．

提案手法の流れは，まず，動画を CNN に入力し，各フレームごとに token 特徴を抽出する．i フレーム目からえられる token 特徴量は $F_i \in R^{512}$ である．そのため，CNN によって得られる token 特徴群は $F \in R^{T\times 512}$ である．次に抽出した各 token 特徴に class token を追加する．そして，positional encoding によって位置情報を付加し，それらの特徴ベクトル群を transformer block に入力する．各 Transformer block において，multi-head attention による特徴抽出とともに感情の類似性に基づいた感情クラスの分類をおこなう．そして，最後の transfomer block から得られる class token 特徴を fully connected layer（FC）に入力することによって表情を分類する．

Transformer bolock における処理を図 3(b) に示す．各 transformer block における multi-head attention は，分類する感情クラス数分の Dot-Product によって構成される．図 3(b) は，transformer block1 を示す．transformer block1 では，入力された動画が感情（emo.）かそうでないか（ne.）を分類する．この transformer block では 2 クラス分類を解くため，2 つの Dot-Product が用意される．そして，各 head から得られる class token 特徴を grobal average pooling（GAP）に入力することによって，感情の類似性に基づいた高次の感情クラスの尤度を算出する．各 head から得られた特徴ベクトル群は，concat された後に多層パーセプトロンによって融合され，次の transformer block に渡される．本手法で用いられる Dot-Product は，文献 [14] で用いられているものと同様である．

HCT は，表情クラスの分類の損失と各 transformer block における高次の感情クラスの分類の損失によって学習される．式 1 に HCT の損失関数を示す．

$$L = L_c(x_i) + \sum_{i=1}^{N} L_b(x_i) \tag{1}$$

ここで，x_i は i 番目の入力サンプル，$L(\cdot)$ はクロスエントロピー損失，N は transformer block の数を示す．また，$L_c(\cdot)$ は表情分類における損失，$L_b(\cdot)$ は各 transformer block における高次の感情クラスの分類における損失を示す．$L_b(\cdot)$ によって transformer block の multi-head attention の各 head が対応するクラスの特徴を表現するように学習される．

今回の実験では，基本 6 表情に無表情を加えた 7 表情のデータセットを用いた．そのため，感情/無感情の 2 クラス分類，ポジティブ/ネガテイブ/その他/無感情の 4 クラス分類，7 表情を分類する 3 つの transformer block を用いた．

3. 実　　験

3.1 実験条件

本実験では，DFEW [15] という公開データセットを用いた．DFEW は，自然環境下における大規模なデータセットである．DFEW のビデオクリップは，世界中の 1,500 以上の映画から収集されており，激しい照度変動，頭部姿勢の変化，顔のオクルージョンなどのさまざまな難易度の高いシーンが含まれている．さらに，DFEW の各ビデオは，専門家の指導の下においてアノテーターによってラベル付けされ，7 つの表情（喜び，悲しみ，中立，怒り，驚き，嫌悪，恐怖）のうちの 1 つが割り当てられている．DFEW には，12,059 個のビデオクリップが含まれ，すべてのサンプルは重複することなく 5 つの同じサイズの部分（fd1〜fd5）に分割されている．評価プロトコルとして，5-fold cross validation が採用されている．すべての実

表 1　従来手法，baseline 手法および提案手法の認識率

Method	Sampling	WAR	UAR
C3D [18]	DS	53.54%	42.74%
R3D18 [19]	DS	53.22%	42.79%
P3D [20]	DS	54.47%	43.79%
I3D-RGB [21]	DS	54.27%	43.40%
VGG11-LSTM [15]	DS	53.70%	42.39%
ResNet18-LSTM [15]	DS	53.08%	42.86%
EC-STFL [15]	DS	54.72%	43.60%
Former-DFER [10]	DS	65.70%	53.69%
DCPNet [22]	GWS	66.32%	**57.11%**
EST [11]	SS	65.85%	54.30%
Baseline(Transformer) [7]	DS	63.78%	52.15%
HCT(Ours)	DS	**66.93%**	53.15%

表 2　fd1 における各 transformer block における感情クラスの認識率

Transformer block	Baseline	HCT(Ours)
block1	32.42%	77.19%
block2	21.70%	71.94%
block3	6.84%	67.41%

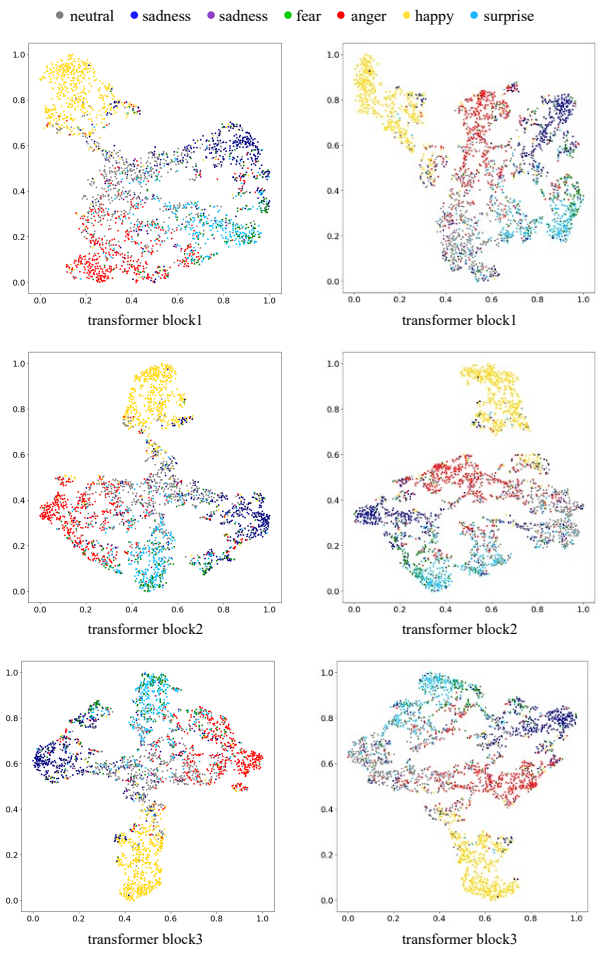
験において，評価には Unweighted Average Recall（UAR）と Weighted Average Recall（WAR）が使用される．

提案手法において用いる CNN は，VGGFace2 [16] によって事前学習された Inception ResNetV1 [17] を用いた．入力する動画のフレーム数は 16 フレーム，画像サイズは 160×160 ピクセルである．学習のバッチサイズは 64，epoch 数は 200，学習率は 0.0005 とした．Baseline 手法は，シンプルな transfomer [7] である．提案手法と baseline 手法の違いは，各 transfomer block において分類を学習するか否かである．提案手法および baseline 手法の multi-head attention の head 数は同じである．

3.2 実験結果

従来手法，baseline 手法および提案手法の認識率を表 1 に示す．ここで，表中の DS は down sampling，GWS は group wise sampling，SS は snippet sampling を示す．Baseline 手法の WAR および UAR はそれぞれ 63.78%，52.15% であった．一方，提案手法の WAR および UAR はそれぞれ 66.93%，53.15% であり，各指標においてそれぞれ 3.15%,1.00% の認識率の向上を確認した．従来手法の最も精度が高い手法と比較すると，WAR においては 0.61% の向上が確認されたが，UAR においては 3.96% 低かった．この原因として，入力する動画のサンプリング方法が影響していると考えられる．提案手法は単純なダウンサンプリングを適用しているのに対して，DCPNet [22] では文献中で提案している特殊なサンプリングを適用している．そのため，提案手法においてもそのサンプリング方法を適用することによって更なる精度の向上が期待される．

表 2 に fd1 における各 transformer block における高次の感情クラスの認識率を示す．Baseline 手法は各 transformer block においてほとんど分類する能力を獲得できていない．それに対して，提案手法は各 block において 70% 程度の分類性能を示している．このことから，提案手法によって translformer block の Multi-head attention の各 head は対応するクラスの特徴を獲得できていると考えられる．



(a) Baseline 手法の特徴空間　　(b) 提案手法の特徴空間

図 4　各 transformer block から得られた特徴量の可視化

提案手法は，各 transformer block において分類問題を解いている．そこで，誤分類した際に，どの block において失敗したのかを確認した．その結果，誤分類する際の多くは block2 において分類に失敗する傾向があることが確認された．また，一度分類に失敗してしまうとそれ以降の block でも分類に失敗してしまうことが多いことが確認された．

さらに提案手法と baseline 手法の違いを分析するために，各 transformer block において得られる class token を t-SNE [23] によって可視化した．図 4 にその結果を示す．表 2 における実験では，提案手法の各 transformer block がクラス固有の特徴を獲得していることが確認されたが，この実験においては大きな差は見られなかった．この原因として，表 2 における実験で使用した特徴量と本実験における特徴量が異なることが考えられる．Transformer block では，multi-head attention によって得られた特徴量を feed forward network（FFN）によって融合してる．そのため，各実験に用いられた特徴が異なることか

らこのような結果になったと考えられる．また，この結果からFFN が最終的な特徴の抽出に大きな役割を担っていることが考えられる．

4. おわりに

本研究では，表情の個人差を吸収することを目的として，人の基本感情の類似性に着目し，その情報を学習における帰納バイアスとして活用した Hierarchical Classification Transformer（HCT）を提案した．提案する HCT における各 transformer block において感情の類似性によってカテゴライズした高次の感情クラスを分類するサブタスクを追加することによって精度向上を図った．大規模な公開データセットである DFEW において提案手法を評価したところ，baseline 手法および従来手法より高い認識率であることを確認した．また，提案手法は中間の transfomer block においても分類タスクを解いているため，モデル内のどのタイミングで誤認識しているかを分析可能であり，モデルの解釈性が向上した．

今後の方針として，中間 block における誤分類を改善することによって更なる精度の向上に取り組む．また，特徴量同士の関係性に対しても明示的に感情の関係性を反映させることによるロバスト化に取り組む予定である．

謝　辞

本研究は，JSPS 科研費 JP22H03681，および中京大学戦略的研究事業の助成を受けたものです．

文　献

[1] M. S. Bartlett, G. Littlewort, I. Fasel and J. R. Movellan: “Real time face detection and facial expression recognition: Development and applications to human computer interaction.”, 2003 Conference on computer vision and pattern recognition workshop, Vol. 5IEEE, pp. 53–53 (2003).

[2] J. Yang, T. Qian, F. Zhang and S. U. Khan: “Real-time facial expression recognition based on edge computing”, IEEE Access, **9**, pp. 76178–76190 (2021).

[3] S. Li and W. Deng: “Deep facial expression recognition: A survey”, IEEE Transactions on Affective Computing, pp. 1–1 (2020).

[4] D. Liu, H. Zhang and P. Zhou: “Video-based facial expression recognition using graph convolutional networks”, 2020 25th International Conference on Pattern Recognition (ICPR)IEEE, pp. 607–614 (2021).

[5] R. Miyoshi, N. Nagata and M. Hashimoto: “Enhanced convolutional lstm with spatial and temporal skip connections and temporal gates for facial expression recognition from video”, Neural Computing and Applications, **33**, pp. 7381–7392 (2021).

[6] X. Liu, L. Jin, X. Han and J. You: “Mutual information regularized identity-aware facial expression recognition in compressed video”, Pattern Recognition, **119**, p. 108105 (2021).

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin: “Attention is all you need”, Advances in neural information processing systems, **30**, (2017).

[8] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al.: “A survey on vision transformer”, IEEE transactions on pattern analysis and machine intelligence, **45**, 1, pp. 87–110 (2022).

[9] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund and A. Clapés: “Video transformers: A survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[10] Z. Zhao and Q. Liu: “Former-dfer: Dynamic facial expression recognition transformer”, Proceedings of the 29th ACM International Conference on Multimedia, pp. 1553–1561 (2021).

[11] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen and Y. Zhan: “Expression snippet transformer for robust video-based facial expression recognition”, Pattern Recognition, **138**, p. 109368 (2023).

[12] J. A. Russell: “Evidence of convergent validity on the dimensions of affect.”, J. personality and social psychology, **36**, 10, p. 1152 (1978).

[13] R. Plutchik: “The emotions”, University Press of America (1991).

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al.: “An image is worth 16x16 words: Transformers for image recognition at scale”, arXiv preprint arXiv:2010.11929 (2020).

[15] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu and J. Liu: “Dfew: A large-scale database for recognizing dynamic facial expressions in the wild”, Proceedings of the 28th ACM international conference on multimedia, pp. 2881–2889 (2020).

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman: “Vggface2: A dataset for recognising faces across pose and age”, 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)IEEE, pp. 67–74 (2018).

[17] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi: “Inception-v4, inception-resnet and the impact of residual connections on learning”, Proceedings of the AAAI conference on artificial intelligence, Vol. 31 (2017).

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri: “Learning spatiotemporal features with 3d convolutional networks”, Proceedings of the IEEE international conference on computer vision, pp. 4489–4497 (2015).

[19] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri: “A closer look at spatiotemporal convolutions for action recognition”, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018).

[20] Z. Qiu, T. Yao and T. Mei: “Learning spatio-temporal representation with pseudo-3d residual networks”, proceedings of the IEEE International Conference on Computer Vision, pp. 5533–5541 (2017).

[21] J. Carreira and A. Zisserman: “Quo vadis, action recognition? a new model and the kinetics dataset”, proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017).

[22] Y. Wang, Y. Sun, W. Song, S. Gao, Y. Huang, Z. Chen, W. Ge and W. Zhang: “Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos”, Proceedings of the 30th ACM International Conference on Multimedia, pp. 101–110 (2022).

[23] L. Van der Maaten and G. Hinton: “Visualizing data using t-sne.”, Journal of machine learning research, **9**, 11 (2008).