Étude comparative de réseaux de neurones pour la reconnaissance des émotions avec les images plénoptiques

Sabrine Djedjiga OUCHERIF[1] Mohamad Motasem NAWAF[2] Jean-Marc BOÏ[2] Lionel NICOD[3] Djamal MERAD[2] Séverine DUBUISSON[2]

[1]Institut de Mathématiques de Marseille (I2M), Centre de Mathématiques et Informatique (CMI), Technopôle Château-Gombert, 39, rue Frédéric Joliot-Curie, 13453 Marseille Cedex 13, France

[2]Laboratoire d'Informatique et Systèmes (LIS), Aix Marseille Université – Campus de Saint Jérôme – Bat. Polytech, 52 Av. Escadrille Normandie Niemen, 13397 Marseille Cedex 20, France

[3]Centre d'Études et de Recherche en Gestion d'Aix-Marseille (CERGAM), 424, chemin du Viaduc 13080, Site Pauliane, Aix-en-Provence, France

Résumé – Dans cet article, nous présentons notre contribution dans la reconnaissance des expressions faciales en utilisant des images obtenues à partir de la base de données Light Field Face Dataset. Nous comparons diverses architectures qui sont principalement développées autour d'un réseau de neurones convolutif de type EfficientNetV2-S et combiné avec différents réseaux de neurones récurrents (LSTM, GRU, BiLSTM and BiGRU). Nous exploitons également les différents ensembles d'images de sous-ouverture. Chaque ensemble varie en fonction du nombre d'images et de leur position virtuelle. Les résultats montrent une amélioration significative de la précision dans deux configurations et selon un ensemble d'images de sous-ouverture. La première consiste à utiliser le modèle d'EfficientNetV2-S avec un LSTM en deux branches. La seconde utilise une seule branche, combinée avec un BiLSTM.

Abstract – In this paper, we present our contribution to facial expression recognition by using image data obtained from the Light Field Face Dataset (LFFD). We compared several neural network architectures which are mainly developed around a convolutional neural network of EfficientNetV2-S and combined with different kinds of recurrent neural networks (LSTM, GRU, BiLSTM and BiGRU). Besides, we exploit different sets of sub-aperture images, each vary in terms of number of images and virtual position. The results show a significant accuracy improvement in two used configurations, depending on the sets of sub-aperture images. The first when using the model of EfficientNetV2-S in two branches configuration and composed with an LSTM. The second uses single branch model with a BiLSTM.

1 Introduction

La caméra plénoptique (en anglais, *Light Field (LF) camera*), est un système optique doté d'une matrice de microlentilles placée entre l'objectif et le capteur image [6]. Cette caméra permet d'obtenir des informations sur l'intensité et la direction des rayons de la lumière, ce qui permet d'acquérir à partir d'une seule prise, plusieurs images représentant la même scène mais avec des points de vue différents. Ces images sont appelées images de sous-ouverture. Les caméras LF permettent également de générer une carte de profondeur [10] et une image de superrésolution [9] comme illustrées sur la figure 1.

Durant les dernières décennies, les caméras plénoptiques ont été exploitées dans différents domaines y compris dans la reconnaissance des expressions faciales. Shen *et al.* [8] ont utilisé la caméra Lytro pour réaliser une base de données privée constituée de 46 sujets. Chaque sujet simule les six émotions basiques (colère, dégoût, peur, joie, tristesse, surprise) et l'état neutre. Un histogramme de gradient orienté (HOG) a été employé pour extraire les points caractéristiques sur les cartes de profondeur et les images de superrésolution, et par la suite, un SVM pour classifier les émotions.

Sepas-Moghaddam *et al.* ont proposé d'extraire les points caractéristiques spatiaux en utilisant les CNN et les points caractéristiques angulaires avec les RNN. Pour cela, ils ont réalisé la Light Field Face Dataset[2], une base de données publique pour la reconnaissance des émotions et de visages contenant 50 sujets sur deux sessions et simulant 3 émotions (joie, colère et surprise) ainsi que l'état neutre. Dans [5], les auteurs ont proposé une architecture composée d'un CNN et d'un réseau neuronal à capsules et dans [3, 4], ils ont proposé un réseau composé d'un VGG16 pré-entrainé sur VGG Face Dataset [1] et d'un LSTM bidirectionnel. À travers leurs résultats, les auteurs ont démontré que l'exploitation des images de sous-ouverture améliore la précision de la reconnaissance des émotions.

Dans cet article, nous exploitons les images de sous-ouverture pour effectuer de la reconnaissance des expressions faciales. Pour atteindre ce but, nous utilisons plusieurs architectures composées d'un réseau de neurones convolutif (CNN) de type EfficientNetV2-S [13] pour extraire les points caractéristiques spatiaux et d'un des réseaux de neurones récurrent (RNN), LSTM [11] / GRU [7]/ BiGRU ou BILSTM [12], pour extraire les points caractéristiques angulaires. Chaque réseau prend en entrée un ensemble d'images de sous-ouverture.

L'article est organisé en 4 parties : nous présentons dans la section 2 notre méthode et les différentes architectures de réseaux de neurones ainsi que notre sélection d'ensembles d'images de sous-ouverture. Dans la section 3, nous analysons les résultats obtenus avec nos différentes approches et mettons

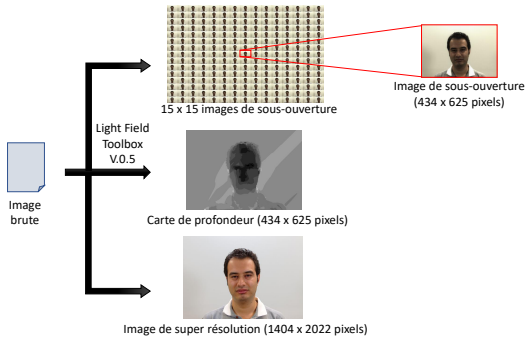

FIGURE 1 : Informations obtenues par la caméra plénoptique pour la base de données LFFD [2].

en évidence les méthodes les plus performantes pour la tâche de la reconnaissance des expressions faciales. Nous concluons dans la section 4 et donnons quelques perspectives.

2 Approche Proposée

2.1 Base de données

La base de données *The IST-EURECOM Light Field Face Database* (LFFD) [2] est utilisée pour étudier la reconnaissance faciale et des émotions. Elle contient des images brutes obtenues avec la caméra Lytro Illum, des rendus d'image $2D$ et des cartes de profondeur. LFFD est composée de deux dossiers : `session1` et `session2`, chaque dossier contient des images de visages prises sur 50 sujets avec différentes expressions faciales et postures. Entre les images de la `session1` et la `session2`, il y a un délais temporel d'un à six mois. Chaque sujet est représenté par 20 images de variations faciales. Nous nous intéressons aux images brutes représentant les émotions (la colère, la joie, la surprise) et l'état neutre (les cartes de profondeur ne sont pas exploitables à cause d'une mauvaise calibration sur la zone du visage).

2.2 Pré-Traitement

Pour extraire les images de sous-ouverture à partir des images brutes, nous utilisons Light Field Toolbox V. 0.5. Pour chaque image brute, nous obtenons une mosaïque de 15×15 images (figure 1). Toutes les images de sous-ouverture ont une résolution de 434×625 pixels. Elles sont rognées sur la zone du visage et redimensionnées en une résolution de $60 \times 60 \times 3$ pixels. Ce choix vient d'un compromis entre la mémoire du GPU (24 GB) et la capacité d'EfficientNetV2-S sans perte de performance.

2.3 Ensembles d'Images de Sous-Ouverture

L'un de nos objectifs est de comparer la précision de la reconnaissance des expressions faciales en fonction de la sélection d'images de sous-ouverture. L'utilisation de l'ensemble d'images de sous-ouverture représentant un sujet avec une émotion est coûteuse en terme de calcul et non nécessaire car entre deux images voisines, il y a une faible variation, et donc peu d'informations angulaires significatives.

Parmi les 225 images de sous-ouverture de la mosaïque, nos ensembles sont selectionnés comme suit :
- *Image unique :* image localisée au centre de la mosaïque d'images (figure 2.(a)).
- *5 images verticales et horizontales :* 5 images de la ligne et de la colonne centrale avec un pas de 3 images (figure 2.(b)).
- *5 images des diagonales :* 5 images de la diagonale ascendante et de la diagonale descendante avec un pas de 3 (figure 2.(c)).
- *15 images verticales et horizontales :* toutes les images de la ligne et de la colonne centrale (figure 2.(d)).
- *15 images des diagonales :* toutes les images de la diagonale ascendante et de la diagonale descendante (figure 2.(e)).

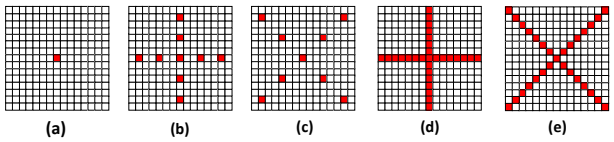

FIGURE 2 : Ensembles d'images de sous-ouverture. (a) Image unique, (b) 5 images verticales et horizontales, (c) 5 images des diagonales, (d) 15 images verticales et horizontales et (e) 15 images des diagonales.

2.4 Architecture Profonde

Dans cet article, nous comparons 5 architectures neuronales pour la reconnaissance des expressions faciales. Ces architectures sont composées principalement d'EfficientNetV2-S pré-entraîné avec ImageNet. Certaines d'entre elles sont composées également d'un RNN (LSTM,GRU, BiLSTM ou BiGRU) qui permettent d'obtenir des informations clés d'un point caractéristique sur toutes les images de sous-ouverture. Le LSTM avec son architecture plus complexe que celle du GRU, capture des dépendances à plus long terme.

La figure 3 illustre les différentes architectures et qui sont caractérisées comme :

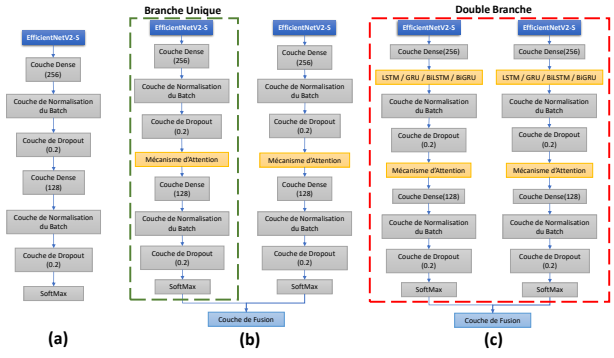

FIGURE 3 : Architectures proposées. (a) Modèle d'EfficientNetV2-S simple, (b) Modèle d'EfficientNetV2-S en branche unique et double, (c) Modèle d'EfficientNetV2-S avec RNN en branche unique et double.

- *EfficientNetV2-S simple :* l'architecture est composée d'un EfficientNetV2-S suivi de deux couches denses, chacune suivie d'une normalisation de batch et d'un dropout pour prévenir du surapprentissage. La classification s'effectue à travers un softmax.
- *EfficientNetV2-S en double branche :* deux modèles d'EfficientNetV2-S sont utilisés, l'un prendra en entrée les

images de la ligne horizontale/diagonale ascendante et l'autre prendra en entrée les images de la ligne verticale/diagonale descendante. Chaque CNN est suivi d'une couche dense, de normalisation de batch et d'un dropout. Nous ajoutons un mécanisme d'attention pour améliorer les résultats, suivi de couches denses, de normalisation de batch et de dropout. Chaque branche se termine avec un softmax pour ensuite se regrouper dans une couche de fusion afin d'obtenir la moyenne entre les deux résultats des deux branches.

- *EfficientNetV2-S en double branche avec RNN :* un RNN est ajouté après l'EfficientNetV2-S et avant la première couche dense. Dans notre étude le RNN est un LSTM, GRU, BiLSTM ou un BiGRU (figure 3.(c)). Le bidirectionnel exploite l'information en amont et en aval des séquences d'entrée et le mécanisme d'attention va permettre de capturer et mettre en évidence les dépendances contextuelles afin d'assister le RNN.

- *EfficientNetV2-S en branche unique :* semblable au modèle EfficientNetV2-S en double branche. Cependant, nous nous intéressons à une seule branche qui prendra comme entrée l'ensemble des images de sous-ouverture de la sélection.

- *EfficientNetV2-S avec RNN en branche unique :* semblable au modèle EfficientNetV2-S avec RNN en double branche. Cependant, l'architecture se compose d'une seule branche qui prend en entrée toutes les images de sous-ouverture de la sélection.

3 Résultats Expérimentaux

3.1 Protocole Expérimental

Nous définissons un protocole afin de comparer les performances de nos modèles avec les differents ensembles d'images de sous-ouverture. Nous prenons `session1` pour l'entraînement et `session2` pour les tests (voir section 2.1). Le processus est répété en interchangeant les sessions et la moyennes des résultats est prise.

3.2 Hyper-Paramètres

La résolution des images de sous-ouverture est de $60 \times 60 \times 3$ pixels. La taille du batch est fixée à 45 et nous considérons 100 epochs. Un arrêt anticipé est ajouté afin d'éviter le surapprentissage en mettant le paramètre d'attente à 10 et nous sauvegardons ainsi les meilleurs poids.

3.3 Analyse des Performances

Le tableau 1 regroupe les résultats obtenus avec les différentes architectures proposées pour les ensembles d'images de sous-ouverture décrites dans la section précédente.

Tout d'abord, nous observons que notre modèle d'EfficientNetV2-S, qui prend en entrée une image de sous-ouverture, donne une meilleure précision (81.8%) en comparaison aux modèles VGG16-EmotiW (75.5%), VGG19-PAM (78.8%) ou AlexNet-PAM (78.4%) de Sepas-Moghaddam *et al.* [4]. Ces architectures sont employées pour la classification des émotion à partir d'une image 2D. Cela démontre que l'EfficientNetV2-S est plus performant que les autres CNN de l'état de l'art.

Par ailleurs, il est important de noter que l'architecture qui a donné un meilleur score de précision est l'architecture EfficientNetV2-S avec un LTSM en double branche (82.9%). Dans ce qui suit, nous donnons une analyse et une comparaison des différentes caractéristiques de nos approches :

- *Branche unique par rapport à double branche d'EfficientNetV2-S avec un RNN :* même si les meilleurs scores sont obtenus avec l'architecture à double branche, l'utilisation d'images de sous-ouverture verticales/diagonales ascendantes et d'images de sous-ouverture horizontales/diagonales descendantes comme deux entrées distinctes n'est pas nécessaire.

- *LSTM par rapport au GRU :* l'utilisation d'EfficientNetV2-S avec un LSTM en double branche pour la reconnaissance donne une plus grande précision qu'en utilisant un GRU. Cependant, l'architecture à une seule branche de réseaux neuronaux composée de CNN et d'un GRU fonctionne mieux qu'avec un LSTM.

- *RNN par rapport au RNN Bidirectionel :* la comparaison entre RNN et RNN bidirectionnel est complexe. Pour 30 images en entrée, un double EfficientNetV2-S avec LSTM a donné de meilleurs résultats que toutes les autres méthodes. Cependant, une seule branche d'EfficientNetV2-S avec un BiLSTM a également atteint une précision de 82,5% $\pm$5,55% .

- *Images Verticales/horizontales par rapport aux images de la diagonale ascendante et descendante :* les résultats obtenus avec les images diagonales sont généralement meilleurs que ceux obtenus avec les images verticales/horizontales. Cela est dû à la grande disparité entre les images obtenues avec la caméra plénoptique.

- *10 images par rapport à 30 images :* l'utilisation de 10 images en entrée est meilleure que l'utilisation d'une seule image compte tenu de l'écart-type. Par exemple, le modèle à double branche d'EfficientNetV2-S avec BiLSTM a une précision de 80,5% et 81,5% avec un écart-type respectif de 5,72% et 6,4%. Cela signifie que ce modèle présente moins de disparité, donc il est plus stable que l'architecture unique. Cependant, l'utilisation de 30 images donne des résultats meilleurs. En effet, le modèle à double branche d'EfficientNetV2-S avec LSTM (82,88% $\pm$ 6,47%) donne le meilleur score.

Certaines images représentant les émotions de colère, neutre ou de surprise n'ont pas de variations significatives d'expressions faciales. C'est pourquoi leur reconnaissance est moins bonne que celle de l'émotion joie.

4 Conclusion et Perspectives

Dans cet article, nous avons introduit plusieurs architectures d'apprentissage profond utilisant des images LF et nous avons comparé leurs performances pour la reconnaissance des expressions faciales. Le modèle EfficientNetV2-S simple avec une image 2D unique obtient une meilleure reconnaissance que VGG16, VGG19 et AlexNet avec une précision de 81,8% $\pm$ 8,9 %. L'utilisation de la sélection d'images diagonales permet d'obtenir une meilleure précision par rapport aux images verticales et horizontales. L'EfficientNetV2-S peut effectuer la tâche avec une faible résolution d'images. En utilisant ce modèle en deux branches avec un LSTM, on obtient le meilleur résultat avec une précision de 82,9% $\pm$ 6,5%. Nous avons également constaté que l'utilisation de deux branches au lieu

TABLE 1 : Performances des différentes architectures utilisant des images de sous-ouverture du jeu de données LFFD pour la reconnaissance des expressions faciales

Architectures	Ensembles d'Images de Sous-Ouverture	Méthodes Proposées	Colère(%)	Joie (%)	Neutre (%)	Surprise (%)	Moyenne (%)	Écart-Type (%)
Architecture Simple	**Image unique**	EfficientNetV2-S	76.5	93.5	83.5	73.5	**81.8**	8.9
Architectures à Double Branche	**5 images verticales et horizontales**	EfficientNetV2-S	69.0	89.5	76.5	72.5	76.9	9.0
		EfficientNetV2-S + LSTM	72.5	92.5	81.0	**77.0**	**80.8**	8.6
		EfficientNetV2-S + BiLSTM	73.5	89.0	**81.5**	68.0	78.0	9.2
		EfficientNetV2-S + GRU	**75.0**	85.5	76.5	74.5	77.9	5.2
		EfficientNetV2-S + BiGRU	72.5	**93.0**	75.0	64.0	76.1	12.2
	5 images des diagonales	EfficientNetV2-S	70.5	85.0	78.0	80.0	78.4	6.0
		EfficientNetV2-S + LSTM	72.0	**90.0**	75.0	76.5	78.4	8.0
		EfficientNetV2-S + BiLSTM	74.5	87.5	**82.5**	77.5	**80.5**	5.7
		EfficientNetV2-S + GRU	**78.0**	87.0	65.0	79.0	77.3	9.1
		EfficientNetV2-S + BiGRU	75.0	89.5	72.5	**80.5**	79.4	7.5
	15 images verticales et horizontales	EfficientNetV2-S	70.5	86.0	77.0	80.0	78.4	6.5
		EfficientNetV2-S + LSTM	**83.5**	88.0	71.0	72.0	78.6	8.4
		EfficientNetV2-S + BiLSTM	82.5	91.5	75.5	76.5	**81.5**	7.4
		EfficientNetV2-S + GRU	76.5	88.0	**81.0**	78.0	77.3	9.1
		EfficientNetV2-S + BiGRU	75.0	**94.5**	66.5	**81.5**	79.4	11.8
	15 images des diagonales	EfficientNetV2-S	77.5	90.5	77.0	77.0	80.5	6.7
		EfficientNetV2-S + LSTM	**80.0**	**92.5**	78.5	**80.5**	**82.9**	6.4
		EfficientNetV2-S + BiLSTM	78.5	88.0	81.0	73.0	80.1	6.2
		EfficientNetV2-S + GRU	75.5	88.0	**82.5**	80.0	81.5	5.2
		EfficientNetV2-S + BiGRU	75.5	89.0	75.5	80.0	80.0	6.4
Architectures à Branche Unique	**5 images verticales et horizontales**	EfficientNetV2-S	74.5	85.5	**80.5**	72.5	78.3	5.9
		EfficientNetV2-S + LSTM	78.0	91.0	74.5	78.0	80.4	7.3
		EfficientNetV2-S + BiLSTM	76.0	**90.5**	79.0	**82.0**	**81.9**	6.3
		EfficientNetV2-S + GRU	**84.0**	88.5	73.0	78.0	80.9	6.8
		EfficientNetV2-S + BiGRU	72.0	88.0	74.5	76.5	77.75	7.1
	5 images des diagonales	EfficientNetV2-S	73.5	89.0	72.0	83.0	79.4	8.1
		EfficientNetV2-S + LSTM	79.0	89.5	67.0	**84.0**	79.9	9.6
		EfficientNetV2-S + BiLSTM	**82.0**	**95.0**	73.0	71.5	80.4	10.8
		EfficientNetV2-S + GRU	70.5	90.0	**85.0**	78.0	**80.9**	8.5
		EfficientNetV2-S + BiGRU	70.5	87.5	80.0	76.0	78.5	7.2
	15 images verticales et horizontales	EfficientNetV2-S	77.0	90.5	76.5	76.5	80.1	6.9
		EfficientNetV2-S + LSTM	74.5	90.5	**82.5**	69.5	79.3	9.2
		EfficientNetV2-S + BiLSTM	78.5	90.5	82.0	79.0	**82.5**	5.6
		EfficientNetV2-S + GRU	77.0	**96.5**	72.5	77.5	80.9	8.5
		EfficientNetV2-S + BiGRU	**80.0**	91.5	76.5	**78.0**	81.5	6.8
	15 images des diagonales	EfficientNetV2-S	**81.0**	86.0	74.5	76.5	79.5	5.1
		EfficientNetV2-S + LSTM	70.5	**90.5**	80.0	79.5	80.1	8.2
		EfficientNetV2-S + BiLSTM	69.0	90.0	80.0	**80.5**	79.9	8.6
		EfficientNetV2-S + GRU	**81.0**	90.0	**84.5**	71.5	**81.8**	7.8
		EfficientNetV2-S + BiGRU	79.5	89.0	75.5	80.0	81.0	5.7

d'une peut être avantageuse. Le regroupement de toutes les images en entrée donne de bons résultats, similaires à ceux du modèle EfficientNetV2-S avec BiLSTM qui atteint une précision de ($82{,}5\% \pm 5{,}6\%$). L'image centrale sélectionnée à partir d'une mosaïque 15x15 d'images de sous-ouverture peut être assimilée à l'image 2D obtenue avec un appareil photo numérique standard. Dans ce contexte, nous pouvons affirmer que l'utilisation du système plénoptique améliore les performances de reconnaissance des expressions faciales.

Références

[1] Q. CAO *et al.* : Vggface2 : A dataset for recognising faces across pose and age. *In 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74, 2018.

[2] A. Sepas-Moghaddam *et al.* : The ist-eurecom light field face database. *In 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2017.

[3] A. Sepas-Moghaddam *et al.* : A deep framework for facial emotion recognition using light field images. *In 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

[4] A. Sepas-Moghaddam *et al.* : Facial emotion recognition using light field images with deep attention-based bidirectional lstm. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3367–3371. IEEE, 2020.

[5] A. Sepas-Moghaddam *et al.* : Capsfield : Light field-based face and expression recognition in the wild using capsule routing. *IEEE Transactions on Image Processing*, 30:2627–2642, 2021.

[6] G. Wu *et al.* : Light field image processing : An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017.

[7] K. Cho *et al.* : Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*, 2014.

[8] T. W. Shen *et al.* : Facial expression recognition using depth map estimation of light field camera. *In IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pages 1–4. IEEE, 2016.

[9] Y. Wang *et al.* : Lfnet : A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Transactions on Image Processing*, 27(9):4274–4286, 2018.

[10] H. G. JEON, J. PARK, G. CHOE et J. PARK : Accurate depth map estimation from a lenslet light field camera. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1547–1555, 2015.

[11] F. KARIM *et al.* : Lstm fully convolutional networks for time series classification. *IEEE access*, 6:1662–1669, 2017.

[12] M. SCHUSTER et K. K. PALIWAL : Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[13] M. TAN et Q. LE : Efficientnet : Rethinking model scaling for convolutional neural networks. *In International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.