Motion Consistency Map for Facial Expression Spotting

Ouala Benjemaa, Amel Aissaoui, Benjamin Allaert, Ioan Marius Bilasco

Motion Consistency Map for Facial Expression Spotting

Ouala Benjemaa*, Amel Aissaoui*, Benjamin Allaert*, Ioan Marius Bilasco†
*IMT Nord Europe, Institut Mines-Télécom, Univ. Lille, Centre for Digital Systems, F-59000 Lille, France
{ouala.benjemaa, amel.aissaoui, benjamin.allaert}@imt-nord-europe.fr
†Univ. Lille, CNRS, Centrale Lille, MR 9189 – CRIStAL – Centre de Recherche en Informatique, Signal et Automatique de Lille. F-59000 Lille, France
{marius.bilasco}@univ-lille.fr

***Abstract*—Facial expression spotting is an effective metric for categorizing human behavior changes. It refers to the precise localization of the temporal intervals in a sequence where a visual event occurs in a face. In this paper, we propose an innovative framework, which relies on the consistency in terms of orientation and intensity of the local facial motions. First, we build local facial motion consistency maps to differentiate expression-related facial motion from facial noise. Then, these maps are fed into a recurrent neural network to precisely delineate the temporal progression of facial expression activation. Extensive evaluations were undertaken on SNAP-2DFE dataset demonstrating the effectiveness of the proposed framework in temporally segmenting expression activation in presence of low or high head pose variations**

***Index Terms*—facial expression, spotting, motion, deep learning**

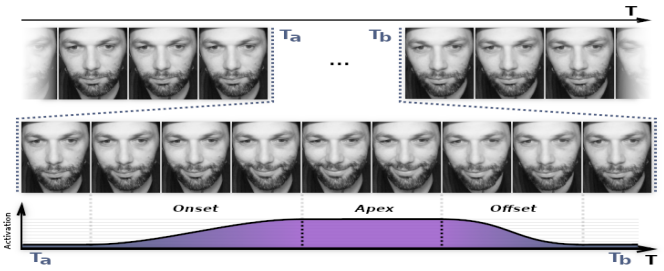

Fig. 1. Example of facial expression spotting. First line - video sequence; Second line - activation sequence of the facial expression within the video from its onset (activation of the expression) to the apex (moment when the expression is most intense) and back to the offset (end of the expression and return to neutral)).

I. Introduction

Automatic visual event spotting is a very popular and a current trendy research topic [1]. It is applicable to many areas, such as behavioral changes detection or to anticipate risks (fall detection, suspicious behavior), where accuracy is important in order to provide a quick response. Among the many possible application domains, our focus is on facial expression analysis. Facial expression (FE) is a non-verbal communication clue that makes possible the analysis of personal emotional states [2]. Automatic FE analysis usually includes two tasks: spotting and recognition. FE recognition associates a class to a given FE. Spotting consists in finding the temporal intervals in a video sequence which contain the FE (from the onset to the offset) (see Figure 1). Originally, FE recognition was only conducted on curated and well segmented video sequences. However, real-world application systems require dealing with long non-segmented videos. The spotting of FE in video sequences is a prerequisite for advanced human behavior analysis [3].

Spotting facial expressions can be challenging due to the multitude of factors involved: 1) the facial movement intensity can vary depending on the specific expression and the individual ranging from micro and macro movements [4] ; 2) the occurrence of subtle changes due to other factors (i.e., head movement, light variation) ; 3) the scarcity of labelled data due to the difficulty for human observers to perceive the onset phase of a FE, especially when variations in facial pose are encountered - the small amount of available training data is a serious limitation for evaluating current FE spotting systems.

In this paper, we propose an innovative framework which accurately spots the FE, while taking into account the challenges mentioned above (see Figure 2). The particularity of our approach lies in the ability to train a model to accommodate facial movement noise induced by other sources (e.g. variations in head pose) that are not related to facial muscle activation. Assuming that facial movement is governed by strong physical constraints, the movement flows are expected to be locally consistent in terms of intensity and direction. We rely on these constraints to train a neural network to faithfully capture genuine facial motions. The initial network, focused on disentangling genuine facial motion from noise, generates a consistent motion map, which is subsequently integrated into a recurrent neural network to precisely track the temporal activation of facial expressions. Moreover, this temporal model allows for the exclusion of disruptive movements such as eye blinks or head movements, that exhibit coherent movement patterns but are irrelevant for expression spotting analysis. Our paper delves into the analysis of spotting within more dynamic environments, where head movements are also considered.

We demonstrate how our method, centered on encoding facial expressions through facial local motion and recurrent neural networks, addresses the challenges posed by the variability of facial expression intensity, the presence of motion noise induced by head pose variation, and the lack of training

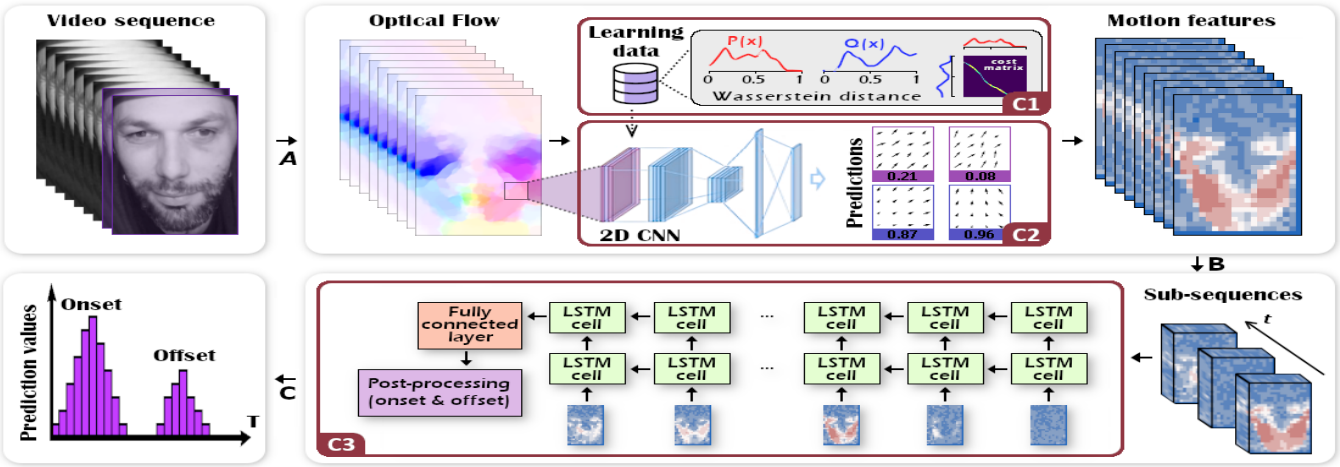

Fig. 2. Overview of the FE spotting analysis process. A) Facial motion is calculated using a dense optical flow approach. The optical flow is then encoded using a 2DCNN architecture, driven by a loss function based on the Wasserstein distance, to keep only the motion induced by the FE. B) The extracted facial motion maps are fed into a LSTM to detect the FE phase. C) The output of the system corresponds to a probability series reflecting the FE spotting.

data. For this, we use the SNaP-2DFe database [5] due to its comprehensive annotations covering all six facial expressions and providing detailed information about expression activation (i.e., onset, apex, offset). This investigation encompasses sequences captured under two conditions: sequences with minimal pose variation (Cam 1) and those with pronounced pose variations (Cam 2). It's worth mentioning that the sequences from Cam 1 and Cam 2 are synchronized, offering a valuable opportunity to examine the challenges posed by head movements.

The paper is structured as follows. Section II provides an overview of the recent approaches for facial movement encoding and FE spotting. Section III presents our approach for extracting consistent facial motion maps. Section IV presents how the recurrent neural network perform FE spotting in a video sequence exploiting the consistent motion maps. Section V illustrates the outcomes achieved on SNAP-2DFE in presence or absence of head pose variations. Finally, we discuss the results and future work in Section VI.

II. Related Work

A. Facial expression Features Extraction

The majority of frameworks for facial expression (FE) spotting analyze feature differences between images in a time window. These features are either handcrafted or learned. Handcrafted approaches utilize texture and motion temporal descriptors like Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [6], [7] and optical flow [8] to encode FE. Optical flows have garnered significant interest due to their ability to encode micro-movements [9]–[11]. Micro-movements are very important in FE spotting as they occur at the onset and the offset phases. Recent works focus on filtering optical flows to isolate FE-induced movements [9], [12], demonstrating that analyzing both movement intensity and direction adequately encode micro-movements. While handcrafted approaches offer good performance, their generalization diminishes in the presence of complex data. Learning-based approaches, in contrast, employ joint feature learning and classification pipelines. These solutions benefit from the invariance and robustness of the extracted features during the learning phase with regard to the available learning data. They exploit directly raw images [13] or descriptors such as LBP [14] or optical flow [15]. Temporal extensions [14], [16]–[18] have been proposed in order to harness the temporal dimension of FE. However, their efficacy remains limited because they require substantial training data, which is lacking in FE spotting.

B. Facial Expression Peak Event Detection

Spotting can be addressed as a classification task, where each image is classified to neutral, onset, apex or offset state [3]. Two main classes exist: Thresholding-based approaches and Learning-based approaches. Thresholding-based methods analyze feature differences along the sequence using appearance-based [20] or motion-based features [8], [21]. Image differences are computed, and a sliding window detects peaks using a threshold [20], [22]. However, the success of these approaches heavily relies on threshold parameter selection and may not improve when more data is available [23]. Learning-based approaches, including traditional machine learning (ML) methods like Support Vector Machine (SVM), Random Forest (RF), Decision Trees (DTs) [12], [24]–[26], and modern deep learning techniques such as Convolutional Neural Networks (CNNs) [27], [28] and Recurrent Neural Networks (RNNs) [23], have been proposed to overcome the limitations of threshold-based methods. For example, Verburg et al. [23] proposed an RNN-based FE spotting method to encode temporal changes in facial regions. However, research in this area remains limited due to the lack of training data.

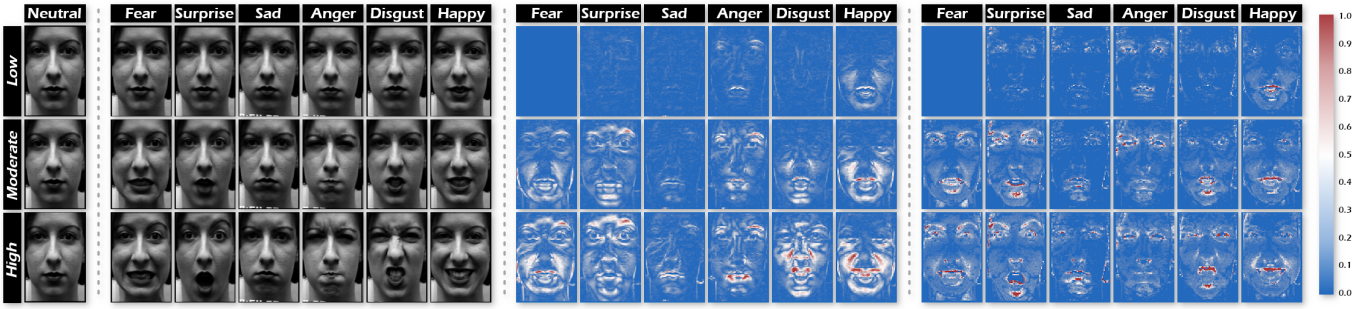

Fig. 3. The model is trained using a subset of CK+ [19] to encode consistent movements (rated between 0-1) associated with facial expression activation (i.e., facial motion between neutral and onset phase - at different intensities (left)). Motion coherence maps generated using Wasserstein (middle) and ChiSquare (right) highlight the significance of considering local motion propagation when labeling facial motion consistency in presence of a facial expression

C. Handling Limited Training Data for Expression Spotting

Many methods proposed for FE spotting struggle due to insufficient training data [3]. Researchers explore augmentation techniques, often involving translations, rotations, and scaling directly on raw images [29]. However, these methods are limited in FE spotting, especially during the onset phase where facial motion and noise intensity are similar. To address this, temporal augmentation methods adapted to facial expression analysis have been introduced. For example, Li et al. [30] employ temporal interpolation and motion magnification. However, these approaches may lead to significant deformations impacting performance in presence of motion noise. In response, some researchers compare various methods for selecting representative frame pairs [31] or combine multiple optical flow methods to compute facial movement [32], enhancing data variability and aiding in model generalization. Despite these advancements, most optical flow augmentation approaches rely on smoothing techniques, leading to information loss as the difference between motion related to expression activation and noise becomes minimal.

III. Motion Consistency Map

In order to preserve the facial motion induced by the FE, we rely on the properties of the facial motion highlighted in [9]: motion within a localized facial area cannot undergo significant variation due to existing bio-mechanical laws. This indicates that there should be no abrupt motion changes neither in direction nor intensity, between consecutive images. Based on these observations, we provide labeled data and we train a model which is intended to dissociate consistent and inconsistent local facial motion.

In order to establish a robust metric for training a model to encode facial movement ranging from micro to macro-movements, it is crucial to closely examine how facial expressions influence the dynamics of facial movements. We selected a set of facial expressions to analyse facial movements at various levels of intensity (see Figure 3). In order to reduce the biases in describing a coherent facial movement, the selected data is acquired under excellent conditions (no pose variation, no illumination change).

In each sequence, the optical flow between the first image (neutral state) and different images of the sequence between the onset phase and the apex were generated to observe the facial movement patterns during the activation phase of the expression.

In contrast with other works, our approach consists in splitting the face into a set of small motion patches of dimension $k * k$. This alleviates difficulties, such as the size of the training set and the inter-individual variations as many identity-disentangled patches can be collected. Utilized human motion properties facilitate convergence and model re-use with a limited number of participants. Recent work has demonstrated that the use of motion modality for facial expression analysis, can reduce the complexity of the learning process [31]. Indeed, the same expression or action, the appearance or geometry varies greatly between different individuals. However, there is a strong similarity between the data when analyzing the motion induced by the facial muscles or the joints of the human body, which are governed by strong biomechanical constraints. Consider two successive face images i and i' of dimensions $w * h$, $OF(i, i')$ represents the dense optical flow computed between these two images. The images i and i', as well as $OF(i, i')$ are identically split into a set of small patches of dimension $k * k$. Each patch is characterized by an optical flow matrix of dimension $k * k * channel$, where $channel$ equals two (i.e., direction and magnitude of the motion in each pixel). Each small patch $OF(i, i')(x, y)$ is then processed in order to measure the motion consistency. In order to facilitate the generalization of the model to handle new datasets, the magnitude of motion is normalized by a common clipping technique. The clipping threshold can be chosen efficiently by plotting the distribution of the magnitude and then choosing a value that prevents large information loss (e.g., the third quartile Q_3).

The motion consistency map is calculated as follows: first the face is detected, cropped, then the optical flow is calculated, on high resolution images, and resized by average mean pooling. For each patch $OF(i, i')(x, y)$, the annotation of a patch reflects the consistency of motion distribution.

The patches are fed into a 2DCNN which encode the optical flow features (direction, intensity and distribution) in order

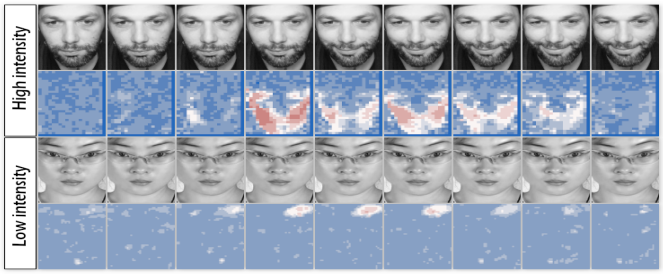

Fig. 4. Qualitative results of the local consistent facial motion extracted by the proposed 2DCNN model learned on the generated training data. High motion intensities - CK+ [19]. Low motion intensities - CASME^2 [6].

to learn motion consistency. Inspired by [9], the Wasserstein distance [33] is used as a loss function within the 2D CNN. This methodology appears to be better suited than others, as it captures the nature of motion propagation over time, as shown in Figure 3. Consider a sequence of three successive images i_1, i_2 and i_3, where two patches $OF(i_1, i_2)(x, y)$ and $OF(i_2, i_3)(x, y)$ are extracted. The label (0 or 1) associated to a patch depends on the Wasserstein distance between the two patches to guarantee local motion coherence $\mathcal{L}_{local}$, and on the distance between the patch and its direct neighbors in 4 connected at $\pm k$ pixels offset on x and y $\mathcal{L}_{prop}$ to guarantee motion coherence in terms of propagation. Global coherence $\mathcal{L}_{global}$ can be formulated as follows: $\mathcal{L}_{global} = \lambda_1 \mathcal{L}_{local} + \lambda_2 \mathcal{L}_{prop}$ where λ_1 and λ_2 are scaling factors. The qualitative results of training a basic 2DCNN are shown in Figure 4. Learning based on local motion propagation seems well suited to dissociating low- and high-intensity coherent motions.

IV. Facial Expression Spotting

A. Adapting FE spotting through motion analysis

The spotting method proposed in this work only considers the onset and offset phases as we deal with motion patterns that do not occur during the apex phase, where often we observe no motion at all. As shown in Figure 5, depending on the person and the expression, the activation intervals (onset, apex, offset) vary strongly from one sequence to another, both in terms of duration and intensity (represented by the dotted line).

To design the training data, the SNaP-2DFe database [5] is used. Each video V in the dataset starts and ends with a neutral phase, which results in a phase with no movement. Cumulated with the apex phases, where the behavior is similar, we decide to delineate each video according to the onset and the offset of the activation. So, for each V we only consider the sequence S that starts at the image situated at the onset minus a small temporal delta Δ_T in order to catch the premises of the activation sequence and ends at offset plus a small temporal delta Δ_T in order to capture the end of the activation sequence. This provides a balance between motionless sequences and onset and offset sequences, which we intend to learn. Using a sliding window, each sequence S is sub-sampled into a set of sub-sequences S_i of l images neighboring the onset and offset points. A time step d is fixed for the window shift in order to guarantee an overlap of images between two successive sub-sequences. Each sub-sequence S_i is annotated depending to the overlap with the onset/offset phase. The labels is equal to 1 if if the ratio between the overlap and the sub-sequence size is greater than a given threshold k and to 0 either-wise.

B. Peak Event Detection

In order to spot the FE activation phases in a given video, we proceed in two steps: the first step consists of motion encoding by the 2DCNN proposed in Section III to obtain a series of facial motion consistency feature maps $M \in \Re^{H*W*C}$ over the considered sequence. The second step consists in sequences classification. Starting from the encoded feature maps M, a subsampling is performed in order to generate for each sub-sequence S_i(section IV-A) the local consistency maps M_i which are fed to a LSTM model. The model architecture is composed of three stacked LSTM layers with intermediate batch normalization layers and two MLP layers. A GELU non-linear activation function is adopted throughout all these layers. The last layer outputs a confidence score between 0 and 1. A threshold T is then applied on the confidence score to predict if the sub-sequence belongs or not to the onset/offset phase.

Frame-based decision - Once the LSTM model predicts a score for each sub-sequence, we need to aggregate all these results by tackling the issue of overlapped predictions at frame level. In our scenario, the model processes the same frame multiple times, depending on the stride of the sliding window. As the frame position change further from the onset phase towards the apex phase, the score associated with each frame within the sliding window increases. Once the frame exits the sliding window, a global score is computed by averaging all the obtained scores. Additionally, to calculate the score per frame, we sum all the confidence scores of the LSTM model and divide by the number of times that frame has been seen by the model. Subsequently, a threshold is applied to this average score to determine whether the frame belongs to the onset - offset phase.

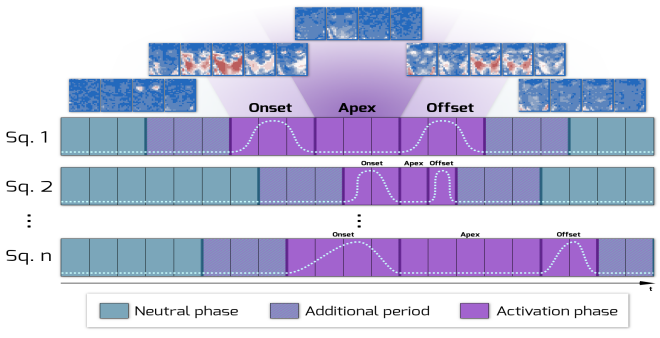

Fig. 5. Details of the expression activation phases. The duration of various phases (onset, apex, offset) varies strongly from one sequence to another, both in terms of duration and intensity. The facial motion maps at the top of the figure represent the outputs returned by our 2DCNN model.

V. Evaluation

A. Datasets

We use the SNaP-2DFe database [5] due to its comprehensive annotations providing detailed information about expression activation (i.e., onset, apex, offset) and head pose changes. The dataset contains 1260 sequences of 15 subjects. Each sequence correspond to one of the six basic facial expressions (i.e., happiness, anger, disgust, fear, sadness, surprise) enacted by untrained persons with head pose and facial expression intensity variations, closely resembling real-world scenarios. For each subject, six head pose variations combined with seven expressions were recorded by two cameras, which results in a total of 630 recordings captured with a helmet camera (i.e., without head movement) and 630 recordings captured with a regular camera placed in front of the user (i.e., with head movements). This ensures that the dataset accurately reflects the intricacies of facial expressions and biomechanical constraints, making it exceptionally suitable for addressing the challenges posed by head pose variation for facial motion spotting.

B. Performance metrics

To evaluate the model, we use the F1 Score. For each subsequence S_i of l frames, the model returns a probability between 0 and 1. A sub-sequence S_i is considered as belonging to the onset or the offset phase, if the confidence score is greater than a predefined threshold T. Hence, for a given subsequence S_i), True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) are defined in terms of whether the sub-sequence S_i belongs to the onset or offset phase. For each evaluation, the F1 score is calculated, along with recall, precision and accuracy.

C. Implementation Details

Following the study of optical flows for FE recognition [32], the facial motion is calculated using Farnebäck method [34]. We use a clipping value of 10.0 for magnitudes. This value was calculated according to the average distribution of the facial movement and motion intensities on SNaP-2DFe dataset, and we normalize both magnitudes ($[0..1]$) and directions ($[-\pi..\pi]$).

The 2DCNN model undergoes training on 9,000 patches, ensuring an equal distribution between the two classes (i.e., consistent or inconsistent) using a loss function derived from the Wasserstein distance. The training spans 10 epochs with a batch size of 64 and encompasses 8,769 parameters.

To fed our LSTM model, we build our training dataset by considering sequences with a temporal delta Δ_T equal to 10 and for each sequence setting up the sub-sequences length l to 10 and the temporal stride d to 1. For the annotation, k is set to 0.5. This results in generating 34,020 sub-sequences distributed equally between two classes (i.e., presence or absence of FE activation). The training consists of 20 epochs with a batch size of 128.

D. Quantitative results

To assess the proficiency of our method in spotting the activation phase, we conduct a study following the Leave-One-Subject-Out (LOSO) protocol. This approach helps mitigate learning biases across various subjects within the SNaP-2DFe database. Each model is subject to the following data distribution of sub-sequences : train - 29,484, validation - 2,268 and test - 2,268. The data is balanced, guaranteeing an identical ratio between sequences belonging to the onset and offset phases, and the other phases. The data is also stratified to ensure an identical ratio of sequences specific to an expression or head movement.

This investigation encompasses sequences captured under two conditions: limited pose variation (Cam 1) and pronounced pose variations (Cam 2). It's worth mentioning that the sequences from Cam 1 and Cam 2 are synchronized, offering a valuable opportunity to examine the challenges posed by head movements. Table I shows the results obtained on all configurations for each model trained on the different subjects in the database. The results are obtained with the following parameters: $T = 0.2$.

In regards to performance, the data from Cam1 are more straightforward for the model to interpret. This is due to the presence of small head pose variation. Analyzing Cam2 data proves to be more challenging, resulting in an average accuracy decrease of $13.5pp$. This difficulty arises from large head pose variations. Application of the LOSO protocol enables us to identify significant differences between subjects. This is mainly due to the variability of activation patterns, and more specifically to the speed of activation and the duration of the onset and offset phases, which depend on the reaction time of the subject. This is particularly true in the presence of pose variation, where concentrating on making the right head movement, while making a facial expression

TABLE I
PERFORMANCE METRICS OBTAINED WITH THE LOSO VALIDATION PROTOCOL IN PRESENCE OR ABSENCE OF HEAD MOTION.

Subject	Cam1 - without head motion				Cam2 - with head motion			
	Accuracy	F1	Recall	Precision	Accuracy	F1	Recall	Precision
SN001	81.0%	66.0%	69.0%	64.0%	63.0%	51.0%	74.0%	39.0%
SN002	84.0%	75.0%	84.0%	68.0%	72.0%	13.0%	07.0%	62.0%
SN003	64.0%	59.0%	97.0%	42.0%	36.0%	45.0%	97.0%	29.0%
SN004	80.0%	75.0%	83.0%	67.0%	73.0%	58.0%	54.0%	63.0%
SN005	78.0%	48.0%	36.0%	74.0%	38.0%	47.0%	97.0%	31.0%
SN006	86.0%	61.0%	55.0%	70.0%	72.0%	50.0%	71.0%	38.0%
SN007	80.0%	59.0%	56.0%	61.0%	63.0%	52.0%	82.0%	38.0%
SN008	82.0%	63.0%	63.0%	63.0%	74.0%	11.0%	06.0%	34.0%
SN009	74.0%	30.0%	23.0%	41.0%	74.0%	42.0%	42.0%	43.0%
SN010	59.0%	56.0%	89.0%	41.0%	50.0%	51.0%	88.0%	36.0%
SN011	76.0%	56.0%	45.0%	74.0%	38.0%	50.0%	92.0%	34.0%
SN012	58.0%	58.0%	94.0%	41.0%	63.0%	57.0%	81.0%	44.0%
SN013	76.0%	74.0%	89.0%	63.0%	68.0%	64.0%	77.0%	54.0%
SN014	77.0%	68.0%	59.0%	82.0%	53.0%	63.0%	94.0%	48.0%
SN015	45.0%	58.0%	99.0%	41.0%	65.0%	23.0%	14.0%	73.0%
Mean	73.5%	60.4%	69.5%	59.5%	60.0%	45.2%	65.2%	44.4%

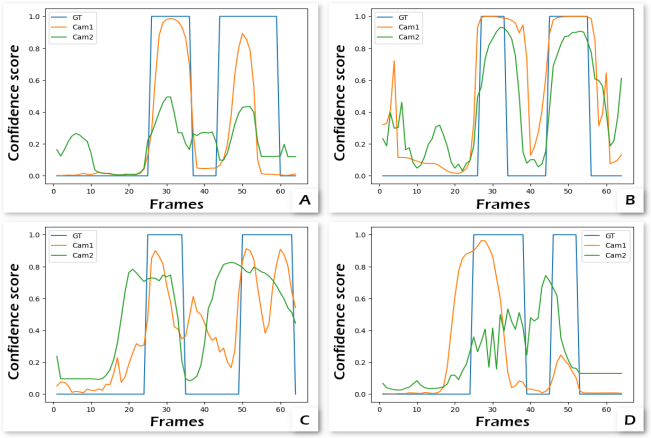

Fig. 6. Prediction of the LSTM model according to the two camera on different sequences of SNaP-2DFe following the corresponding activation pattern: neutral - onset - apex - offset - neutral. A) Nothing-Happy-SN001; B) Nothing-Happy-SN010; C) Yaw-Anger-SN012; D) Diag-Disgust-SN007.

at the right moment, causes significant variation in activation patterns. On average, the performance of Cam2 is worse than that of Cam1, particularly in terms of model precision. The confidence scores returned by the LSTM model are less reliable, making it difficult to delimit the onset and offset phases using thresholding, resulting in a wider range of false positives (see Figure 6).

To deepen the studies, we propose a more detailed analysis of the expressions (Table II) and the different head pose variations (Table III). In these tables, values correspond to average scores calculated on all subjects on both Cam1 and Cam2 in Table I. The results are obtained with $T = 0.2$.

In terms of expression (Table II), 'sadness' and 'anger' are more difficult for the model to interpret, whether or not there is head movement. This is because the same regions around the eyes and mouth are activated on these expressions, making analysis more complex. This difficulty is reinforced in the presence of pose variation, where the small movements induced by these expressions are mixed up with the movement of the head pose. Figures 6-A and 6-B illustrate predictions obtained by the models on Cam1 and Cam2, on the same expression 'happy', without head pose variation and on two different subjects. Under these conditions, decrease in performance can be explained by noise coming from eye blinks, which are perceived as coherent movements. This problem is probably linked to the lack of variability of the activation phases in the training data. On these same configuration, the model trained on Cam2 encounters an additional difficulty in the confidence score.

With regard to head pose variations (Table III), the faster the movement in the 2D plane (Tx and Roll), which blurs the image, or the greater the movement in the 3D plane, which tends to occlude the face (Yaw), the more difficult it becomes for the model to delimit the activation of the expression. Figures 6-C and 6-D illustrate predictions obtained by the models on Cam1 and Cam2, on two different head movements.

TABLE II
PERFORMANCE METRICS PER EXPESSION.

Expression	Cam1 - without head motion				Cam2 - with head motion			
	Accuracy	F1	Recall	Precision	Accuracy	F1	Recall	Precision
Happiness	72.0%	64.0%	72.0%	57.0%	59.0%	52.0%	64.0%	43.0%
Sadness	70.0%	56.0%	67.0%	47.0%	61.0%	47.0%	65.0%	37.0%
Anger	74.0%	57.0%	63.0%	53.0%	60.0%	47.0%	64.0%	37.0%
Disgust	75.0%	64.0%	75.0%	56.0%	59.0%	49.0%	66.0%	39.0%
Fear	75.0%	63.0%	70.0%	57.0%	60.0%	50.0%	65.0%	41.0%
Surprise	74.0%	65.0%	77.0%	56.0%	60.0%	53.0%	72.0%	41.0%
Mean	73.5%	61.4%	70.9%	54.2%	60.0%	49.6%	65.9%	39.8%

TABLE III
PERFORMANCE METRICS PER HEAD POSE.

Pose	Cam1 - without head motion				Cam2 - with head motion			
	Accuracy	F1	Recall	Precision	Accuracy	F1	Recall	Precision
Nothing	74.0%	65.0%	76.0%	57.0%	65.0%	53.0%	62.0%	47.0%
Tx	74.0%	59.0%	67.0%	53.0%	61.0%	48.0%	64.0%	38.0%
Yaw	73.0%	58.0%	66.0%	51.0%	58.0%	47.0%	65.0%	36.0%
Pitch	73.0%	65.0%	77.0%	57.0%	61.0%	52.0%	65.0%	43.0%
Roll	73.0%	57.0%	63.0%	51.0%	55.0%	46.0%	69.0%	35.0%
Diag	73.0%	63.0%	75.0%	55.0%	60.0%	52.0%	70.0%	41.0%
Mean	73.5%	61.3%	70.7%	54.1%	60.0%	49.7%	65.9%	40.1%

Under these conditions, we notice that the two models have more difficulty predicting the activation phases of expressions. The Cam1 model, trained on data without pose variation, manages to delimit the phases, but greater uncertainty appears, sometimes reducing the confidence to a very low score. As for the Cam2 model, it tends to be more sensitive to movements which results in a less clear dissociation of the activation phases. In these two cases the application of the threshold is limited.

VI. Conclusion

In this paper, we propose an approach for FE spotting based on facial movement consistency maps. The exploitation of motion features allows models to finely encode the micro and macro movements characterizing a FE. The use of the SNaP-2DFe database provides an ideal framework for working on the analysis of spotting of FE in the presence of pose variations through its system for synchronous capture of expressions with and without pose variations of the face, making this study innovative in the context of spotting FE. The experiments raised interesting avenues to explore. In particular, the exploitation of facial movement consistency maps tends to identify phases of expression activation in the presence of significant head pose variation. Future work on post-processing and learning methods can be explored to improve performance in the presence of strong variations in head pose and reduce processing time.. In particular, the use of an end-to-end method, taking into account both spatial encoding of motion and temporal encoding related to expressions based on self-supervised spatio-temporal graphs and transformers.

REFERENCES

[1] Y. Guo, S. Guo, Z. Jin, S. Kaul, D. Gotz, and N. Cao, "A survey on visual analysis of event sequence data," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[2] A. Tcherkassof and D. Dupré, "The emotion–facial expression link: evidence from human and automatic expression recognition," *Psychological research*, vol. 85, no. 8, pp. 2954–2969, 2021.

[3] H. Pan, L. Xie, Z. Wang, B. Liu, M. Yang, and J. Tao, "Review of micro-expression spotting and recognition in video sequences," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 1–17, 2021.

[4] Z. Zhang, T. Chen, Y. Liu, C. Wang, K. Zhao, C. H. Liu, and X. Fu, "Decoding the temporal representation of facial expression in face-selective regions," *NeuroImage*, vol. 283, p. 120442, 2023.

[5] B. Allaert, J. Mennesson, I. M. Bilasco, and C. Djeraba, "Impact of the face registration techniques on facial expressions recognition," *Signal Processing: Image Communication*, vol. 61, pp. 44–53, 2018.

[6] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS one*, vol. 9, no. 1, p. e86041, 2014.

[7] A. K. Davison, M. H. Yap, N. Costen, K. Tan, C. Lansley, and D. Leightley, "Micro-facial movements: An investigation on spatio-temporal descriptors," in *European conference on computer vision*. Springer, 2014, pp. 111–123.

[8] D. Patel, G. Zhao, and M. Pietikäinen, "Spatiotemporal integration of optical flow vectors for micro-expression detection," in *International conference on advanced concepts for intelligent vision systems*. Springer, 2015, pp. 369–380.

[9] B. Allaert, I. M. Bilasco, and C. Djeraba, "Micro and macro facial expression recognition using advanced local motion patterns," *IEEE Transactions on Affective Computing*, 2019.

[10] Y. He, S.-J. Wang, J. Li, and M. H. Yap, "Spotting macro-and micro-expression intervals in long video sequences," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 742–748.

[11] Y. Han, B. Li, Y.-K. Lai, and Y.-J. Liu, "Cfd: A collaborative feature difference method for spontaneous micro-expression spotting," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1942–1946.

[12] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 254–267, 2017.

[13] C. H. Yap, M. H. Yap, A. K. Davison, and R. Cunningham, "Efficient lightweight 3d-cnn using frame skipping and contrast enhancement for facial macro-and micro-expression spotting," *arXiv preprint arXiv:2105.06340*, 2021.

[14] T.-K. Tran, Q.-N. Vo, X. Hong, and G. Zhao, "Dense prediction for micro-expression spotting based on deep sequence model," *Electronic Imaging*, vol. 2019, no. 8, pp. 401–1, 2019.

[15] A. J. R. Kumar and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1511–1520.

[16] S. Nag, A. K. Bhunia, A. Konwer, and P. P. Roy, "Facial micro-expression spotting and recognition using time contrasted feature with visual memory," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2022–2026.

[17] J. Hong, C. Lee, and H. Jung, "Late fusion-based video transformer for facial micro-expression recognition," *Applied Sciences*, vol. 12, no. 3, p. 1169, 2022.

[18] L. Zhang, X. Hong, O. Arandjelović, and G. Zhao, "Short and long range relation based spatio-temporal transformer for micro-expression recognition," *IEEE Transactions on Affective Computing*, 2022.

[19] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[20] A. Moilanen, G. Zhao, and M. Pietikäinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *2014 22nd international conference on pattern recognition*. IEEE, 2014, pp. 1722–1727.

[21] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, 2017.

[22] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 51–56.

[23] M. Verburg and V. Menkovski, "Micro-expression detection in long videos using optical flow and recurrent neural networks," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–6.

[24] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *2011 international conference on computer vision*. IEEE, 2011, pp. 1449–1456.

[25] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE, 2013, pp. 1–6.

[26] R. Danescu, D. Borza, and R. Itu, "Detecting micro-expressions in real time using high-speed video sequences," in *Intelligent Video Surveillance*. IntechOpen, 2018.

[27] E. Sariyanidi, H. Gunes, and A. 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, 2014.

[28] Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, "Smeconvnet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos," *IEEE Access*, vol. 6, pp. 71 143–71 151, 2018.

[29] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 382–386.

[30] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Reading hidden emotions: spontaneous micro-expression spotting and recognition," in *CVPR*, 2015, pp. 217–230.

[31] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Dynamic facial expression recognition under partial occlusion with optical flow reconstruction," *IEEE Transactions on Image Processing*, vol. 31, pp. 446–457, 2021.

[32] B. Allaert, I. R. Ward, M. Bilasco, C. Djeraba, and M. Bennamoun, "A comparative study on optical flow for facial expression analysis," *Neurocomputing*, 2022.

[33] L. Rüschendorf, "The wasserstein distance and approximation theorems," *Probability Theory and Related Fields*, vol. 70, no. 1, pp. 117–129, 1985.

[34] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 2003, pp. 363–370.