Temporal Triplet Mining for Personality Recognition

Dario Dotti, Esam Ghaleb and Stylianos Asteriadis
Department of Data Science and Knowledge Engineering, Maastricht University, Maastricht, The Netherlands
{dario.dotti, esam.ghaleb, stelios.asteriadis}@maastrichtuniversity.nl

***Abstract*— One of the primary goals of personality computing is to enhance the automatic understanding of human behavior, making use of various sensing technologies. Recent studies have started to correlate personality patterns described by psychologists with data findings, however, given the subtle delineations of human behaviors, results are specific to predefined contexts. In this paper, we propose a framework for automatic personality recognition that is able to embed different behavioral dynamics evoked by diverse real world scenarios. Specifically, motion features are designed to encode local motion dynamics from the human body, and interpersonal distance (proxemics) features are designed to encode global dynamics in the scene. By using a Convolutional Neural Network (CNN) architecture which utilizes a triplet loss deep metric learning, we learn temporal, as well as discriminative spatio-temporal streams of embeddings to represent patterns of personality behaviors. We experimentally show that the proposed Temporal Triplet Mining strategy leverages the similarity between temporally related samples and, therefore, helps to encode higher semantic movements or sub-movements which are easier to map onto personality labels. Our experiments show that the generated embeddings improve the state-of-the-art results of personality recognition on two public datasets, recorded in different scenarios.**

I. INTRODUCTION

Extensive studies in the field of psychology showed that attitude, mood, and personality are directly connected to human behavioral patterns [22]. Since these human characteristics are often subtle, the affective computing field still faces several challenges. With the recent advances in computational resources and data availability, attention is now shifting towards personality analysis (i.e. mapping data findings to personality labels), more intensively than in the previous years. Recent works show that personality and, in general, affective computing, can contribute significantly to several applications in areas like surveillance [13], Human-Computer Interaction [14], and healthcare [11]. Furthermore, the mapping of behavioral patterns to personality labels allows systems to be more interactive and adaptive [2], avoiding unnecessary effort in manual interaction ensuring smoother and more personalized interactions.

Recent personality computing applications achieved reliable results in analyzing faces [18], body postures [15], and multimodal information [23]. Despite this growing attention, most of the models focus mainly on specific contexts where behaviors can be mapped ad-hoc to personality labels. For example, facial expressions can be linked to personality attributes for applications like job screening [24], however, the user position has to be constantly in front of the camera in a quiet environment. Recently, authors in [16] proposed a CNN model for personality recognition using body and contextual information in different scenarios. Motion and context samples are extracted in time windows, and mapped to personality labels, yet, the temporal correlation of human motion dynamics is not exploited. Building on these findings, in this paper, we propose a novel framework that further expands the use of body information, context learning and their interaction in time, using Deep Metric Learning (DML) [12]. As stated by [12], DML on human motion data improves the measurement of motion similarities. Hence, by adding the temporal analysis to our DML framework, we help the system to discover higher semantic movements that enhance the discovery of discriminative personality patterns, and therefore, improve the personality recognition task.

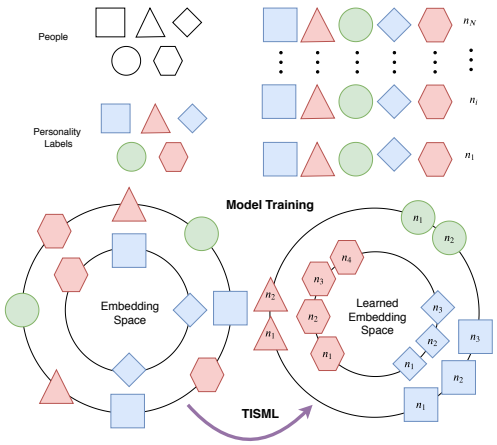


Fig. 1: High level description of the proposed model. The general goal of our approach is to create an embedding space optimized for the personality recognition task. During training time, the model is encouraged to match similar short-term spatio-temporal descriptors using the proposed TISML framework to create discriminative behavioral sequences with varied temporal relation (bottom row, right).

In Fig. 1, we show a high level description of the proposed model. The analyzed data is composed of people performing activities in certain scenarios. Every person is different in the way they act and move (indicated by the empty shapes; top left), however, there exist common behavioral patterns that can be categorized into discrete personality classes (indicated by colored shapes; top left). Human motion, as well as proxemics features, are extracted in a time-window approach (top right of the figure).

The framework is trained through the Temporal Identification Similarity Metric Learning (TISML) component, which contains two signals: The first one is an identification signal based on personality labels, while the second one is a similarity signal based on Deep Metric Learning (DML). The general goal of the DML approach is to bring

samples with similar labels (positive examples) closer together, while pushing apart samples with different labels (negative examples). Additionally, in the training stage, our intuition is to select temporally related positive examples to encourage the model to generate embeddings with temporal relation while maintaining a high discriminative power for personality recognition. The bottom of Fig. 1 illustrates the training process, where, before training, samples are randomly distributed in the embedding space (bottom left). Our proposed approach employing TISML helps the model to generate temporally, as well as semantically related embeddings (bottom right).

We found that learning the temporal similarity allows the model to assemble longer sequences that contain higher semantic value than the input features. Moreover, as we do not add any constraint on the temporal relations, the model automatically learns sequences of varied temporal lengths (bottom right of the figure). As the same human movements can be performed at varying speeds and durations, our results are relevant in improving the similarity measurements of human motion to select more discriminative personality patterns. Finally, our contributions are as follows:

- We build a novel deep framework that learns temporal and discriminative motion patterns in real-world scenarios. We experimentally show that our generated embeddings perform better than state-of-the-art short-term motion samples.
- Using TISML, we encode the relation of temporally adjacent spatio-temporal samples, hence, without introducing any temporal constraint or alignment during training time, motion dynamics carrying similar semantic values are matched via deep metric learning.
- Extensive experiments are conducted to investigate the relation between local motion features, global context features and their interactions in time using two real-world datasets.

II. Related Work

Body information and personality. Recent works showed that body expressions are a powerful personality indicator [20], [16]. Moreover, human body communication can be better integrated with (visual) sensing technologies as it is more robust to camera positioning, noise and occlusions [15]. Authors in [6] investigate nonverbal behavioral cues using optical flow and Neural Networks for personality and leadership recognition. Authors in [15] extracted body information using skeleton tracking methods, and learned body motion dynamics using an Autoencoder-LSTM framework. In the recent paper proposed by [16], body motion information as well as context features are fused in a CNN framework for personality recognition in different scenarios. Features are extracted every n frames and mapped independently to personality labels.

Context information and personality. How people use and share their interpersonal space has been shown to be a discriminative cue for personality understanding [34]. Interpersonal distance has been studied extensively in social scenarios [34], additionally, in a recent study proposed by [16], authors linked the use of personal space to personality patterns also in a nonsocial environment (i.e. when people are not surrounded by others). Given these findings, we further study the temporal interaction between body motion and context features for improved personality recognition performance.

Deep Metric Learning. DML has become popular with the advances and success of deep learning [8]. It projects embeddings produced by mapping functions ($f(x)$) such as a CNN, onto a manifold space where similar samples are closer while the dissimilar ones are placed apart from each other. In this space, a Euclidean distance between two samples ($d(f(x_i), f(x_j) = \|f(x_i) - f(x_j)\|_2^2$) can be employed directly as a distance metric, for classification or retrieval tasks. Deep learning can be used explicitly in learning mapping functions for metric learning through a set of non-linear transformations, with a metric loss to exploit the similarity and dissimilarity between data samples [21].

The common learning procedure of DML frameworks includes the selection of positive and negative examples using the class labels [27]. Recently, DML has also been used in self-supervised learning where the similarity is indicated using data attributes [32], [4]. Authors in [4] apply DML to construct a motion dictionary capturing the high-level similarity between the sequence of motions that vary in terms of joint-angles, timing, and ordering. The proposed framework is trained using a triplet-loss strategy, where positive examples are motion words that appear temporarily close in the training data. Authors in [12] proposed a metric objective to measure the similarity of two motion sequences to address the limitation of standard triplet-based DML when dealing with human motion data. They aim to enforce the separation of embeddings with respect to the means of associated distribution moments (time-windows). Similarly, we propose a triplet-loss learning framework to improve the similarity measurements of motion features for personality recognition.

III. The Proposed Framework

We propose a framework to encode local motion dynamics from the human body in combination with global interpersonal distances (proxemics) to encode personality-dependent behavioral patterns. Our work employs DML to map spatio-temporal descriptors to an optimized latent space, where, behaviors with discriminative power are learned and grouped together whereas non-informative sequences are positioned far apart. As human behaviors are very dynamic and change according to the situation, it is very difficult to find semantic similarities between them [12]. Therefore, a novel Temporal Triplet Mining (TTM) strategy tailored for behavioral data is proposed. We argue that taking advantage of the triplet mining scheme, short-term spatio-temporal descriptors are implicitly matched creating longer sequences with higher semantic value. As a result, our embedding space encodes

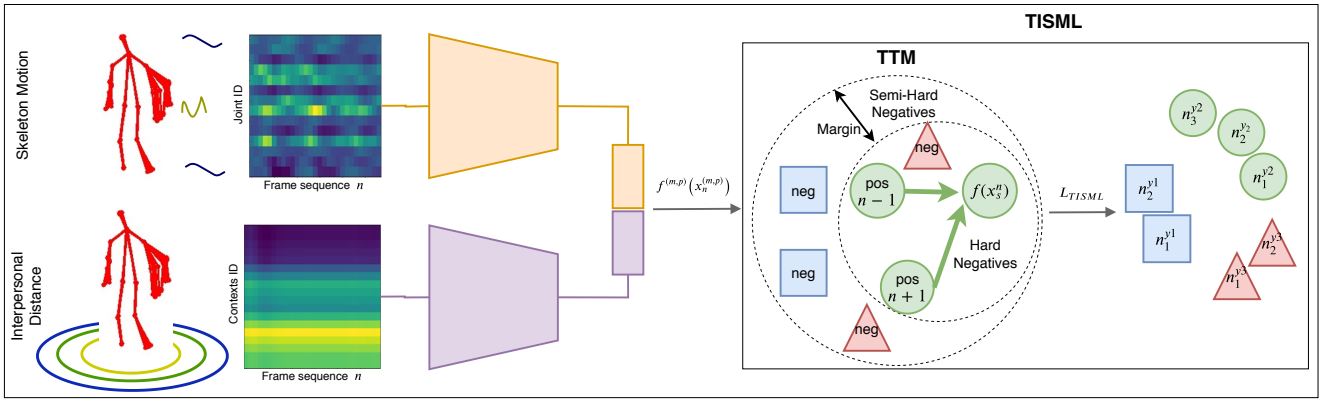


Fig. 2: The proposed Architecture. Two descriptors representing the skeleton temporal motion as well as the spatial interaction are extracted every frame sequence n. The descriptor images show the evolution over time (x-axis) of the reference information (y-axis). The reference information is joints motion evolution for the person descriptor and proxemics to the surrounding contexts for the context descriptor. Deep CNN models are then used to obtain a compact representation of each spatio-temporal information. The outputs of the CNN models are concatenated and fed into the proposed learning framework TISML. An effective Temporal Triplet Mining (TTM) is employed to select temporally related positive samples encouraging the model to learn meaningful behavioral sequences that bear a higher discriminative power. Finally, a double objective loss function L_{TISML} is adopted for personality recognition and personality retrieval.

behavioral patterns of varying sizes optimized to retrieve personality-conditioned behaviors.

Fig. 2 shows our framework architecture, where skeleton motion, as well as proxemics descriptors, are extracted for every frame sequence of size n. As the two descriptors describe the motion and the spatial dynamics of a sequence, two separate CNN architectures are leveraged to obtain compact representations of the input features. The obtained representations are concatenated and fed to the Temporal Identification Similarity Metric Learning (TISML) loss component. TISML aims to project the concatenated motion (m) and proxemics (p) embeddings produced by the two CNNs (which serve as mapping functions of the raw features) $f^{(m,p)}(x^{(m,p)}) : \mathbb{R}^{d^{(m,p)}}$ onto a shared feature space $\mathbb{R}^d$ in which similar features are positioned closely and dissimilar ones are put far apart from each other based on data similarity and personality class. To do so, within TISML, a simple but effective Temporal Triplet Mining (TTM) is proposed to facilitate the overall learning effort.

A. Motion Features

In this work, skeleton information is extracted from every frame using the OpenPose library proposed by [9]. As this method does not provide a skeleton re-identification function, following [16], a frame by frame tracking information is added. Then, local temporal information is extracted from every skeleton joint in terms of joint motion and rotation. As explained in [19], similarities between short-term motions are easier to learn in respect to long-term sequences as they embed less noise. For every detected joint j at a frame f, we compute its spatial as well as rotation evolution in a narrow frame sequence of length n. By combining all the skeleton joints in a sequence $n_{1,2..f}$, we obtain a matrix of size $J \times n$, where J is the total number of the detected joints. The generated matrices describe the motion in 3D Cartesian coordinates. Lately, numerous works showed that, by converting from Cartesian to Cylindrical coordinates [33], a more invariant motion descriptor could be obtained. Therefore, following [16], 3D Cartesian motion values are transformed into 3D Cylindrical values.

Finally, to leverage the learning power of CNN models, we utilize motion image clips, in which we treat Cylindrical coordinate values as pixel values in an image [5], [19]. Hence, the values are converted between 0 and 255 using a linear transformation and the matrices are reshaped to be suitable for a CNN architecture. Fig. 2, top stream, shows the motion descriptor construction, where given a skeleton sequence of size n, frame by frame motion values (x-axis) of all the detected joints (y-axis) are organized in a motion image. In this example, the highest motion values (yellow color) correspond to the skeleton arms.

B. Interpersonal Distances (Proxemics)

In this work, we aim to build a general system that could follow the users in different situations (i.e. at home or at a social event), hence, in addition to local skeleton motion image, we build interpersonal distance (proxemics) images that can be applied to both social as well as nonsocial scenarios.

Social Proxemics. For the social scenario, for every detected subject in the scene, we compute the Euclidean distance between the current subject s and the rest of the subjects S present in the scene. Specifically, we use as a reference point the joint j^i, which corresponds to the body torso (empirically chosen as the most robust to noise), and we compute the Euclidean distance between the coordinates of j^i_s and j^i_S. By combining the interpersonal distances between subjects within the frame sequence n, we obtain a matrix of size $S \times n$, where S is the total amount of subjects in the dataset. Please note that, to overcome the problem of finding different amounts of subjects in the scene at different times, we construct the images considering the total amount of subjects in the dataset $\mathbb{D}$. In the situation when not all the subjects are present in the scene, the maximum distance

values are assigned.

Nonsocial Proxemics. For the nonsocial scenario, we use the intuition proposed in [16], in which proxemics is intended as "the way people use their personal space in relation to objects". There exist several studies showing how people engage with objects as they engage with other humans (called Anthropomorphism). Therefore, an interesting idea is to extract how people move and interact with the surroundings (i.e. proxemics towards objects instead of people). For example, we hypothesize that an Overcontrolled personality that has a high level of the Conscientiousness trait is more meticulous in the searching of objects than a Resilient personality.

Following [16], fist we locate semantic regions of the scene in an unsupervised way. Given a discrete set of r regions, we compute the Euclidean distances between the position of the subject s and the semantic regions r for every frame sequence n, obtaining a final matrix $r \times n$. Finally, the final values are transformed into Cylindrical coordinates, converted between 0 and 255 using a linear transformation and the matrices are reshaped to be suitable for a CNN architecture. Fig. 2, bottom stream, shows the motion descriptor construction process, where, the interpersonal distance between the given subject s and the context entities (r in the nonsocial scenario and S in the social scenario) are depicted on the y-axis and temporal information is depicted on the x-axis.

IV. Temporal Identification Similarity Metric Learning (TISML)

A. Definitions

In this study, a given sequence of motion and proxemics features of a subject s ($x_s^{n(m,p)}$) is associated to a discrete label (y_s) to estimate the corresponding subject personality label. Each frame sequence is represented by motion and proxemics embeddings $f^{(m,p)}(x_s^{n(m,p)})$, where $f^{(m,p)}$ denotes the motion and proxemics mapping function. For simplicity, we refer to $f^{(m,p)}(x_s^{n(m,p)})$ as $f(x_s^n)$, which includes both motion and proxemics embeddings.

B. Formulation

TISML optimizes $f(x_s^n)$ to generate embeddings correlated with a personality class. In our work, the personality learning task is guided through two signals, the first signal is a similarity measure based on a DML loss which positions semantically related embeddings closer to each other (decreasing the intra-class variations) and positions the semantically unrelated embeddings far apart (increasing the inter-class variations) [4]. We apply DML based on the triplet loss strategy.

Triplet loss uses triplet sets: $\{f(x_s^n), f(x_{s+}^{n+}), f(x_{s-}^{n-})\}$, where $f(x_s^n)$ is an anchor (baseline), $f(x_{s+}^{n+})$ is a positive (similar) sample to $f(x_s^n)$, and $f(x_{s-}^{n-})$ is a negative sample (i.e. different label) to $f(x_s^n)$. As shown in (1), the optimization procedure aims to minimize the distance between the anchor (baseline) input to a positive sample while maximizing the distance from the anchor to the negative sample within a margin [27].

$$L_{Sim}(f(x_s^n), f(x_{s+}^{n+}), f(x_{s-}^{n-}) = \|f(x_s^n) - f(x_{s+}^{n+})\|_2^2 - \|f(x_s^n) - f(x_{s-}^{n-})\|_2^2 + \text{margin} \quad (1)$$

The second signal in our work is an identification signal, which classifies a given embedding into one of the given personality type labels (e.g. *Y=3*). The identification signal is achieved by an n-way softmax-layer to predict the probability distribution over the n-personality labels [31]. In our work, the network is trained to minimize the cross-entropy that servers as the identification loss and is defined as:

$$L_{Ident}(f(x_s^n), y, \theta) = -\sum_{i=1}^{Y} -p_i \log \hat{p}i \quad (2)$$

where $f(x_s^n)$ refers to the mapping functions that produced the motion and proxemics embeddings, y is the target class, and θ denotes the parameter of the softmax layer. p_i is the target probability distribution, where $p_i = 0$ for all i except $p_i = 1$ for the target class i. $\hat{p}i$ is the predicted probability distribution. Finally, the optimization of the network is achieved through the joint loss and formulated as follows:

$$\underset{f^{(m,p)}}{\text{argmin}}\ L_{TISML} = L_{Sim} + L_{Ident} \quad (3)$$

The goal of this formulation is to optimize the proposed behavioral descriptors mapping the function $f^{(m,p)}$ during the training process to generate temporal personality-related embeddings. This strategy utilizes, simultaneously, the similarity measure between short-term descriptors through L_{Sim}, and the supervised class information through the identification signal L_{Ident}. Note that the two losses are equally weighted.

C. Temporal Triplet Mining (TTM)

A prominent problem when using the triplet mining strategy is that the possible number of triplet sets could be extremely large, and training the DML can be challenging and prohibitively expensive. As a result, one of the main challenges in the triplet loss based DML is the slow-convergence during the training process. Without a careful and smart strategy to select the triplets, DML could only learn to map correctly easy sequences with little discriminative power. Therefore, in this work, we adopt a semi-hard triplet-sets mining strategy to guide the training process during the selection of the triplet sets. Moreover, as temporally adjacent short-term descriptors are likely to belong to the same semantic behavior, we propose a Temporal Triplet Mining (TTM) strategy.

The training process is displayed at Fig. 2 (TTM). For a given anchor $f(x_s^n)$ at a frame sequence n, we restrict the selection of its positive samples to the temporal vicinity (i.e. within a temporal window t). For example, if we set $t = 3$ centered to the anchor temporal position n, the positive samples will be selected at $n-1$ and $n+1$. Regarding the negative samples, they are randomly chosen from other personality classes and could be from any time-window. Clearly,

the choice of t is critical to obtain the best optimization performance and its impact is discussed in the experiments section (Sec. VI-C).

The optimization process is based on the online DML where the selection of triplet sets is based on mini-batches in each iteration during the training phase [28], [30]. Specifically, at every batch, we compute the loss on all the triplets that satisfy the constraint expressed in eq. 4. As also shown in Fig. 2 (TTM), the loss is computed on the hard-negative as well as semi-hard negative samples. A crucial step is to not take into account the easy negatives (i.e. $d(f(x_s^n), f(x_{s-}^{n-})) > d(f(x_s^n), f(x_{s+}^{n+})) + margin$) which would give a small loss, and therefore, yielding little information to the learning procedure.

$$d(f(x_s^n), f(x_{s-}^{n-})) < d(f(x_s^n), f(x_{s+}^{n+})) + margin \qquad (4)$$

Since the proposed TTM minimizes the distance between samples in the temporal vicinity, adjacent short-term semantically related descriptors are aggregated forming an informative series of sequences with different lengths. One advantage of this approach is that unlike approaches like Dynamic-Time-Warping (DTW), we do not need any explicit time synchronization or alignment to find similarities between sequences of different lengths [12].

V. Implementation Details

We use the Keras [10] and Tensorflow frameworks [1] for all computations in this work. As the datasets used are recorded using different frame rates, we experimentally set the frame sequence duration to $n = 180$ and $n = 90$ frames for the nonsocial dataset and the salsa dataset respectively (note that due to different frame rate both of the sequences contain 6 seconds of data). For our image descriptors, we resize the final images to a 32×32 image to be a suitable input for the CNN architecture. Since our descriptor is not a real image, this dimensionality has the advantage of not being computationally expensive while still preserving the discriminative information. Given the motion as well as proxemics images as input, the VGG19 architecture [29], pre-trained on ImageNet [26], is adopted for feature extraction and learning. Although CNN models demonstrated to learn discriminative and generic features applicable in novel domains [19], early convolutional layers learn more low-level generic features, while higher convolutional layers learn more task-specific features.

Moreover, as our features describe temporal information, they are not optimized for standard pooling strategy. Hence, we extract a compact representation from the Conv3 layer and, following [19], we apply the Temporal Mean Pooling (TMP) to exploit the temporal information in our spatio-temporal image descriptors. The output descriptors are concatenated, and fed to two Fully-Connected Layers with Batch normalization. Finally, our embeddings dimension of size 128 are used as input to our TISML.

Training. The proposed framework is trained on a Nvidia TITAN V GPU for 80 epochs. The batch size is obtained from two parameters: the number of randomly selected anchors a and the size of the temporal-window t. In order to obtain batches containing a balanced amount of data for each personality labels, we set $a = 15$ (i.e. 5 anchors for each label), while t is set to be $t = 5$ (more details on this parameter selection in section VI-D). We use Adam optimizer with a $7e-6$ learning rate and the margin in the triplet loss is set to 0.1.

VI. Experiments

To evaluate the proposed study, we present personality recognition experiments on two public datasets recorded in different scenarios.

A. Datasets and Labels

The Salsa dataset [3] contains multimodal data from two social events (30 minutes each) in an university scenario. The two parts contain the same participants recorded during a poster session as well as a cocktail party. Their personality scores were collected using the Big-Five personality questionnaire [17].

Personality in a nonsocial context dataset [15] provides video data of 45 participants in an unconstrained indoor scenario. Every subject performed six tasks resembling Activities of Daily Living (ADL) and filled the short version of the Big-Five personality questionnaire [25].

Personality labels: We use as labels the three personality categories provided by [16]. In this model, the Big Five personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness) are projected onto three semantically higher categories called personality types [7] (Resilient, Overcontrolled and Undercontrolled). This model has the advantage of representing the commonly used one-dimensional independent traits (e.g. high/low Extraversion independent from high/low Neuroticism) as multidimensional dependent factors. For example, the Resilient personality type is represented by high Extraversion and Openness traits, and low Neuroticism. This multidimensional representation of personality behaviors was shown to be more similar to human judgments of behavioral characteristics [18].

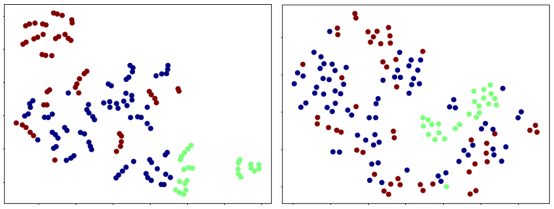
B. Evaluation protocol

As our TISML framework contains two objective functions (i.e. L_{Sim} and L_{Ident}), following the experimental setup described in [31], we use the prediction of the softmax layer when comparing to the state-of-the-art results on the personality labels. On the other hand, the discriminative power of the generated embeddings is evaluated separately, as to the authors' knowledge this is the first work that employs a DML strategy for personality recognition. In all of our experiments (e.g. Table I and Table II), we follow a leave-subjects-out based evaluation, in which a set of 6 subjects for the nonsocial dataset, and a set of 3 subjects for the salsa dataset, are left out from the training procedure and used only for testing. All results are reported in terms of f1 score. We compare our performance against various state-of-the-art results for both datasets, note that we report only the

TABLE I: F1 score on the personality recognition task using different triplet mining strategies.

Triplet mining strategy	Salsa	Nonsocial
Random triplet mining RTM	72.0%	71.6%
Triplet mining through TTM	75.6%	74.9%

results from [16] that used the same experimental protocol as ours (i.e. leave-subjects-out, and f1 accuracy).



(a) (b)

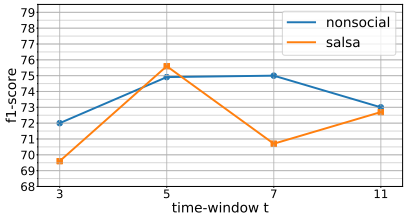
Fig. 3: (a) Temporal Tripet Mining (TTM) Embeddings. (b) Random Triplet Mining (RTM) Embeddings. TTM helps in creating a more explicit separation between the personality classes (red, blue, green). Moreover, short-terms spatio-temporal descriptors are temporally aligned via TISML creating higher semantical sequences that are easier to map to personality labels.

C. TMM versus Random Triplets Mining (RTM)

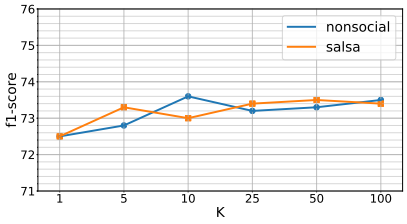
In this section, we investigate the effect of selecting positive samples that are temporally related to the anchor (as explained in section IV-C). To do so, we compare the proposed TTM approach (section IV-C) to a standard random positive selection (RTM). In RTM, we select random samples from other subjects with the same personality. Formally, given an anchor $f(x_s^n)$ with a personality label y_s, positive samples $f(x_{s+}^{n+})$ are chosen given the following constraints: $y_{s+} = y_s$ and $s+ \neq s$.

The results are reported in Table I. Results show that selecting random samples from different subjects with the same personality ensures the model to learn similarities invariant to the identity of a subject. However, as the samples embed only short sequences (90 or 180 frames depending on the dataset) the discrimination between the personality classes is harder. On the other hand, selecting triplets with temporal constraint forces the model to learn similarities over samples further away in time, and therefore, learning more comprehensive behaviors which results in a stronger personality recognition performance.

Furthermore, in Fig. 3, we provide a visual example of the embeddings generated through different triplet mining strategies. In particular, Fig. 3a depicts the short-term spatio-temporal descriptors of 5 batches from the nonsocial dataset [15] encoded via the proposed TTM, while Fig. 3b shows the short-term spatio-temporal descriptors of the same 5 batches encoded using RTM. It is easy to notice that the separation between the $Y = 3$ personality classes (depicted using red, blue, and green color) is more explicit in Fig. 3a, confirming the results of Table I. Moreover, TTM embeddings are organized in sequences of varying lengths, as can be noticed in the subclusters formed in Fig. 3a, while the embeddings generated by RTM do not present any visible structure. As the short-term sequences are temporally aligned during the TISML learning, more discriminative behavioral patterns are determined to enhance the overall understanding of personality displays.



(a) Impact of temporal time-window (t).



(b) KNN results on the embeddings of TISML.

Fig. 4: Parameters investigation.

D. Impact of Time-window Selection

As explained in section IV-C, the temporal range of the time-window t is crucial to capture the affective behaviors of the analyzed subjects. Positive samples that are temporally too far away from the anchor risk to carry little similarity, and therefore, deceive the final goal of aggregating semantically related descriptors. On the other hand, positive samples that are too temporally close to the anchor risk to be "too similar", and therefore, yield an insignificant contribution to the learning objective.

In Fig. 4a, we show the impact of the time-window selection. The nonsocial dataset [15] contains data of subject performing problem-solving activities in an indoor environment. As the subjects are moving to complete the given tasks, the behavioral data contains several active and fast interactions, hence, selecting positive samples in a large time-window range, (e.g. $t = 11$), is not beneficial (blue line). As a matter of fact, fast interactions have a short duration, and therefore, highly informative samples have to be selected from a shorter time-window (e.g. $t = 5$). On the other hand, the Salsa dataset [3] contains interaction from a poster session and a cocktail party in a university environment. As subjects are engaged in social interactions, movements are slower and the impact of longer time-windows is less visible. Given the results, $t = 5$ is selected for the rest of the evaluation in both datasets.

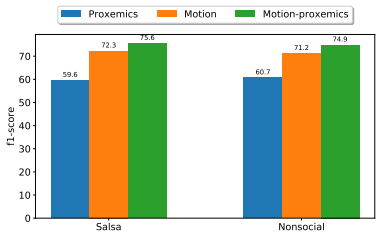


Fig. 5: Ablation study to evaluate the contribution of the input features. The framework is trained solely on motion, proxemics, and on both motion and proxemics features.

E. Ablation study

An ablation study was conducted to verify the contribution of the chosen input descriptors (Fig. 5). In this experiment, the framework is trained using a single input descriptor per time (i.e. skeleton motion or proxemics features), or using the combination of the two cues as displayed in Fig. 2. The results show that TISML achieves higher accuracy using the two descriptors combined, confirming our initial hypothesis and, thus, it will be used in the rest of our experiments.

F. Embedding Evaluation

To evaluate the discriminative value of the generated embeddings ($f(x_x^n)$), we use the conventional K-Nearest Neighbor (*KNN*) classifier with Euclidean distance. Fig. 4b shows the results of several K for the personality recognition task on the analyzed datasets. Good performance is obtained when K is set to higher values, in a range between $[25, 100]$. For example, at $k = 50$, we obtain 73.5% and 72.8% f1-score, for both, Salsa and non-social datasets, respectively.

G. Baseline Comparison

We compare our performance against various state-of-the-art results for both datasets. For a fair comparison, Table II is organized according to the input features. The first part of the table indicates the performance of methods that uses solely skeleton motion features. In particular, $LSTM_{cl}$, proposed by [15] uses an Autoencoder-LSTM framework to learn skeleton motion dynamics. $Clips + MTLN$, proposed by [19], uses similar skeleton motion descriptors as input to a CNN framework called MTLN. EL-LMKL [6] was proposed for leadership recognition as well as personality trait recognition using optical-flow motion information. As El-LMKL uses optical flow-based features, it cannot be applied to the nonsocial dataset.

In the second part of Table II, we report the performance obtained using motion-proxemics features. In particular, Person-Context CNN [16] maps short-term motion-context descriptors to personality labels using a multi-stream CNN framework.

Additionally, to evaluate the effect of each term in our objective function L_{TISML} (equation 3), we also train the model with individual loss functions, L_{Sim} and L_{Ident}, separately. The results of this evaluation are indicated as "Similarity Signal" when trained using the first term L_{Sim}, while we refer to the results as "Identification Signal" when the model is trained using the second term L_{Ident}.

TABLE II: F1 score on the personality recognition task using different features of TISML compared to baselines and other approaches.

Feature Type	Method	Salsa	Nonsocial
Motion	$LSTM_{cl}$ [15]	59.6%	55.3%
	EL-LMKL [6]	61.2%	-
	Clips+MTLN [19]	68.5%	70.7%
	TISML (ours)	**72.3%**	**71.2%**
Motion-proxemics	Person-Context CNN [16]	73.0%	72.6%
	TISML with only Identification Signal (ours)	68.2	67.7
	TISML with only Similarity Signal (ours)	73.2	73.0
	TISML (ours)	**75.6%**	**74.9%**

Results show that the proposed TISML framework achieves higher results in all the tested feature settings. Specifically, we achieve higher results compared to the state-of-the-art models that use only motion by 3.8% for the salsa dataset and by 0.5% on the nonsocial dataset. When using skeleton motion and proxemics, we improve the personality recognition state-of-the-art results by 2.6% on the salsa dataset, and by 2.3% on the nonsocial dataset.

Furthermore, the TISML trained using a double objective loss reaches higher performance results than when trained using individual signals, proving that using a double term is beneficial to create more informative embeddings leading to better recognition performance on both datasets.

VII. Conclusion

In this paper, we propose a model that analyzes human behavioral patterns from different real-world scenarios to tackle the personality recognition task. Using a CNN framework, we extract compact representations of local skeleton motion features, as well as interpersonal distance features. The learning task is accomplished using the novel TISML component. In TISML, a Temporal Triplet Mining (TTM) strategy is employed to leverage the similarity between temporally adjacent short-term descriptors as they are likely to belong to the same semantic behavior and, thus, have higher chances to lead to robust modeling of personality labels. Finally, a double term objective function is used for personality recognition and personality retrieval tasks. Experiments show that our framework generates embeddings that are aligned in temporal sequences and therefore, creates more meaningful behavioral patterns that improve state-of-the-art results.

Acknowledgment

This work has been partially funded by the European Union's Horizon2020 project: PeRsOnalized Integrated CARE Solution for Elderly facing several short or long term conditions and enabling a better quality of LIFE (PROCare-Life), under Grant Agreement N.875221.

References

[1] M. Abadi, A. Agarwal, P. Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] H. Achten. Buildings with an attitude. In *Stouffs, R. andSariyildiz, S.(eds.), ComputationandPerformance–Proceedings of the 31st eCAADe Conference*, volume 1, pages 477–485, 2013.

[3] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1707–1720, 2016.

[4] A. Aristidou, D. Cohen-Or, J. K. Hodgins, Y. Chrysanthou, and A. Shamir. Deep motifs and motion signatures. In *SIGGRAPH Asia 2018 Technical Papers*, page 187. ACM, 2018.

[5] A. Aristidou, D. Cohen-Or, J. K. Hodgins, and A. Shamir. Self-similarity analysis for motion capture cleaning. In *Computer Graphics Forum*, volume 37, pages 297–309. Wiley Online Library, 2018.

[6] C. Beyan, M. Shahid, and V. Murino. Investigation of small group social interactions using deep visual activity-based nonverbal features. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 311–319. ACM, 2018.

[7] J. Block and J. H. Block. The role of ego-control and ego-resiliency in the organization of behavior. In *Development of cognition, affect, and social relations*, pages 49–112. Psychology Press, 2014.

[8] F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1861–1870, 2019.

[9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[10] F. Chollet et al. Keras, 2015.

[11] E. Cocoradă, C. I. Maican, A.-M. Cazan, and M. A. Maican. Assessing the smartphone addiction risk and its associations with personality traits among adolescents. *Children and Youth Services Review*, 93:345–354, 2018.

[12] H. Coskun, D. Joseph Tan, S. Conjeti, N. Navab, and F. Tombari. Human motion analysis with deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–683, 2018.

[13] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 290–297. IEEE, 2011.

[14] E. S. de Lima, B. Feijó, and A. L. Furtado. Player behavior and personality modeling for interactive storytelling in games. *Entertainment Computing*, 28:32–48, 2018.

[15] D. Dotti, M. Popa, and S. Asteriadis. Behavior and personality analysis in a nonsocial context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2354–2362, 2018.

[16] D. Dotti, M. Popa, and S. Asteriadis. Being the center of attention: A person-context cnnframework for personality recognition. *Arxiv*, 2019.

[17] O. P. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[18] J. C. S. J. Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. Van Gerven, R. Van Lier, et al. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, 2019.

[19] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017.

[20] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2012.

[21] J. Lee, S. Abu-El-Haija, B. Varadarajan, and A. P. Natsev. Collaborative deep metric learning for video understanding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 481–490, New York, NY, USA, 2018. ACM.

[22] K. Loewenthal and C. A. Lewis. *An introduction to psychological tests and scales*. Psychology press, 2018.

[23] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM, 2008.

[24] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision*, pages 400–418. Springer, 2016.

[25] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[28] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[31] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[32] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[33] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, 104(2-3):249–257, 2006.

[34] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pages 37–42. ACM, 2010.