CEUR-WS.org/Vol-3116/Paper_2.pdf

A System for Real-Time Emotion Recognition in Smart Cities

Davide Andreoletti[1], Felipe Cardoso[1], Andrea Arzillo[1], Luca Luceri[1], Achille Peternier[1] and Silvia Giordano[1]

[1]*Information Systems and Networking Institute, Department of Innovative Technologies, University of Applied Sciences of Southern Switzerland, Viganello, Switzerland*

Abstract

The paradigm of Smart Cities is based on the idea of enhancing citizens' life by means of digital technologies. The widespread generation of data is seen as the main enabling factor of this paradigm, as it makes possible to create data-driven services to be offered to the citizens. While objective measurable data are generally considered (e.g., measurement of air pollution), we argue that subjective data may also play a relevant role in the context of Smart Cities. Indeed, services could be better tailored to citizens, provided that their subjective experience can be effectively assessed. In this paper, we consider citizens' emotions as the subjective data that can be exploited to improve services, and we propose a system for inferring this data from the movements of the citizens themselves. We note that movements can be acquired in various ways, such as by using cameras or non-invasive motion capturing technologies (e.g., the Kinect) that can be easily deployed in a Smart City. Specifically, we propose a system that, from the analysis of movements acquired using the Kinect 2.0, can effectively i) disambiguate between portions of movements characterized by non-negative or negative emotions and ii) identify in real time the instants of transitions between such states. Results, obtained after extensive simulations, make us confident that the proposed system can find application in a real Smart City context (e.g., to automatically assess if a person in a public place is too much nervous). Finally, we also observe that citizens' movements are not as privacy sensitive as other types of data that are generally considered in the emotion recognition task (e.g., facial expressions). In the paper, we briefly discuss these privacy issues, arguing that, because of general privacy concerns, a system for emotion recognition from movements is the most suitable for the Smart City context.

Keywords

Smart Cities, Emotion Recognition, Body Motion

1. Introduction

The aim of a Smart City is to improve citizens' well-being by enhancing services through the use of digital technologies. Examples of these services are electricity supply and public transportation. The cornerstones of this transformation are mainly three: the first is the widespread diffusion of data acquisition tools, such as Internet of Things (IoT) devices and smartphones, which allow sensing various aspects of a city's life (e.g., IoT can be used to collect information about traffic and public transportation services can be optimized accordingly). The

Proceedings of FTAL 2021, October 28–29, 2021, Lugano, Switzerland

davide.andreoletti@supsi.ch (D. Andreoletti); felipe.cardoso@supsi.ch (F. Cardoso); andrea.arzillo@supsi.ch (A. Arzillo); luca.luceri@supsi.ch (L. Luceri); achille.peternier@supsi.ch (A. Peternier); silvia.giordano@supsi.ch (S. Giordano)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

second is the improvement of artificial intelligence, which allows inferring valuable insights from the collected data. The third is the consolidation of next generation telecom network (i.e., the 5G), which enables low-latency data communication and processing. The combination of these factors enables real-time data-driven service optimization.

In general, the optimization of services in a Smart City is not performed based on subjective data. For instance, traffic data is used to optimize the route of a public transportation, but the feelings of drivers are not even taken into consideration. While in most of the cases objective data are more relevant than subjective ones, in several scenarios subjective data might help crafting user-tailored services. In this paper, we focus on a specific type of subjective data, i.e., citizens' emotions, and we propose a system for the real-time identification of emotions of a person from her body's movements. We observe that the emotions of a person are an entry point for understanding her perception of the situations she lives. Indeed, emotions carry relevant information about the unverbalized concerns of a person, and are considered a very relevant implicit feedback in a number of scenarios (e.g., the possibility to adapt a service to the current emotion of its user is a desirable feature of new-generation human-computer interfaces). Let us note that the emotions of citizens can be valuable in the context of Smart Cities as well. For instance, emotions can be analyzed in a smart shop to evaluate the appreciation of a potential customer towards some products or, in relation to public safety, to early assess if a person in a public place is dangerously getting angry. Several existing works have considered the task of emotion recognition in a Smart City scenario (see Section 2).

We observe that most of the data considered in existing emotion recognition solutions carry very sensitive information (e.g., facial expression). Hence, their application might not be desirable in a Smart City scenario, where privacy is a general concern. A less privacy-intrusive approach consists in inferring some person's emotions from the movements of her body. Besides being less privacy intrusive than other techniques, however, this approach is also less effective than others. Indeed, the emotions of a person are much more evident from her facial expression than from her movements. However, given the aforementioned privacy issues, in this work we chose to sacrifice effectiveness in the name of privacy, and we propose a system for the real-time inference of emotion from body's movements.

We observe that, among existing methods, those based on the analysis of body's movements (as we also do in this paper), are by far the least widespread. To explain why this data is less considered than others, various reasons can be speculated. In our view, the most important reason is that inferring the emotion of a person from her body's movements is extremely challenging. To reduce the complexity of the task, in this work we only infer if an emotion is negative or not. We argue that this choice is reasonable in the context of the Smart Cities, where having a rough estimate of one's emotional state (i.e., whether it is a negative one) may suffice.

Please note that, with respect to existing methods, the observed data (i.e., positions and orientations of joints) do not carry a single emotion, but a sequence of non-negative and negative emotional states. The main contribution of this paper is an inference system that detects the instants of transitions between these two emotional states. The proposed system is based on a combination of i) signal pre-processing techniques (employed to find a suitable representation of the observed data), ii) Gradient Boosting classifier (employed to classify portions of the observed data into non-negative and negative classes) and iii) a real-time postprocessing heuristic that identifies relevant transitions between different emotional states. The performance of the system is measured considering the effectiveness of two sub-tasks: i) classification of portion of the signals into the right emotional class and ii) segmentation of the signal into homogeneous portions, i.e., portions of the observed data that carry the same emotional content. Results, obtained by means of extensive simulations, show that the proposed system can effectively perform both the sub-tasks.

The rest of the paper is structured as follows. In Section 2 we briefly review existing works in the field emotion recognition, with a particular interest for those employed in a Smart City scenario. Section 3 formally describes the problem statement that we aim to solve, while Section 4 presents the system that we propose to solve it. We show the effectiveness of our system in Section 5, and we provide conclusive remarks in Section 6.

2. Related Work

Emotion Recognition is a well-investigated research area. An extensive survey of existing methods to perform the emotion recognition task can be found at Ref. [1]. Existing methods to perform this task mainly differ with respect to i) the type of data from which emotions are inferred and ii) the considered inference system. As for point (i), examples of data from which existing solutions infer emotions are facial expressions [2], electroencephalogram [3], speech [4] and body's movement [5]. Multi-modal solutions that combine different types of data have also been explored, e.g., [6].

As for point (2), a combination of signal processing and machine learning techniques is generally employed. Signal processing techniques are used to pre-process the signal, eventually to extract hand-crafted features designed by experts to capture relevant properties of the signal (e.g., the authors of [7] extract the Mel-frequency cepstral coefficients from speech and infer emotions from them). The pre-processed signal is then fed to a machine learning architecture, such as Recurrent Neural Networks and Convolutional Neural Networks in [8] and Support Vector Machine [9]. In this work, we employ a Gradient Boosting Classifier that, to our knowledge, was never previously considered to perform emotion recognition.

Emotion recognition has been considered in relation to Smart Cities only recently. In particular, several works have elaborated on the concept emotion-driven service as a prominent characteristic of futuristic Smart Cities (e.g., [10, 11]). Most of existing works consider emotion recognition as an enabling technology of improved healthcare systems, in which the physical health status of patients is complemented with the psychological one [12]. The authors of Ref. [13] propose a system that applies sentiment analysis techniques to optimize the organization of some events (e.g., a cultural event that takes place in the city) based on the feedback that citizens provide on their online social networks. A gamification approach enhanced with emotion recognition is proposed in the context of Smart Cities in [14]. Finally, levering on the seminal work for emotion recognition from the analysis of wireless signals reflected by the human body [15], the authors of [10, 11] propose the concept of IoT barriers, i.e., an array of IoT devices deployed in a Smart City to automatically sense the emotions of citizens.

3. Problem Statement

The aim of this work is to infer the instants of transitions between two emotional states of a person from the analysis of her movements. We refer to the set of emotional states that a person can perceive as $\mathcal{E} = \{0, 1\}$, where 0 and 1 encode non-negative and negative emotional states, respectively (e.g., both joy and neutral state are encoded as 0, while frustration is encoded as 1). The movements of a person are represented as the trajectories and orientations of a set of salient joints, such as head, left and right hands, left and right knees, ... More specifically, the state of each joint (say the i-th) at a given time instant t is univocally identified by a 3-components position tuple, i.e., $(x_i(t), y_i(t), z_i(t))$, and by a 4-components orientation tuple (or *quaternion*), i.e., $\left(w_i^{(0)}(t), w_i^{(1)}(t), w_i^{(2)}(t), w_i^{(3)}(t)\right)$. Therefore, the state of the $i - th$ joint at time t is represented by a vector with 7 components, derived from the juxtaposition of the position and orientation tuples.

We develop an inference system, referred to as $\mathcal{S}$, that, from the analysis of the movements of a person, performs a real-time estimation of the instants of transitions between her emotional states. More formally, the proposed system is designed to perform the following estimation:

$$\mathcal{S}\left(x_i(t), y_i(t), z_i(t), w_i^{(0)}(t), w_i^{(1)}(t), w_i^{(2)}(t), w_i^{(3)}(t)\right) \\ = \left\{t^{(\star)} \,|\hat{\mathcal{E}}\left(t^{(\star)} - dt\right) \neq \hat{\mathcal{E}}\left(t^{(\star)} + dt\right)\right\} \quad (1)$$

where $\hat{\mathcal{E}}\left(t^{(\star)}\right)$ is the estimated emotional state at time $t^{(\star)}$ and dt is the infinitesimal delta of time. By real-time estimate, we refer to the fact that each candidate transition instant $t^{(\star)}$ has to be detected with a maximum delay of γ seconds. In other words, the detection of instant $t^{(\star)}$ should not be performed later than $t^{(\star)} + \gamma$. In our experiments, we consider $\gamma = 6s$.

4. Methodology

4.1. Data Acquisition

The data acquisition is performed by a set of actors and an assistant. The assistant communicates to the actor the emotional state to be acted (i.e., non-negative or negative) and employs a software specifically tailored to label movements according to this state. Note that each actor performs her activity one at a time.

The dataset has been collected with the help of 6 actors, whose movements have been acquired with the Kinect 2.0 [16], which provides information about positions and orientations of head and hands at a frequency of 30 Hz. The generic sample of the resulting dataset consists of the sequence of current positions and orientations of the head and hands (and is encoded as a vector of 21 = 7X3 components, i.e., 3 positions and 4 orientations for the three considered joints) and its associated binary state (i.e., 0 for the neutral and 1 for the negative state). In total, the dataset is made of 159 samples with an average duration of 120 seconds. During this period, on average, 5 negative emotions are acted, each with an average duration of 3 seconds. The remaining time

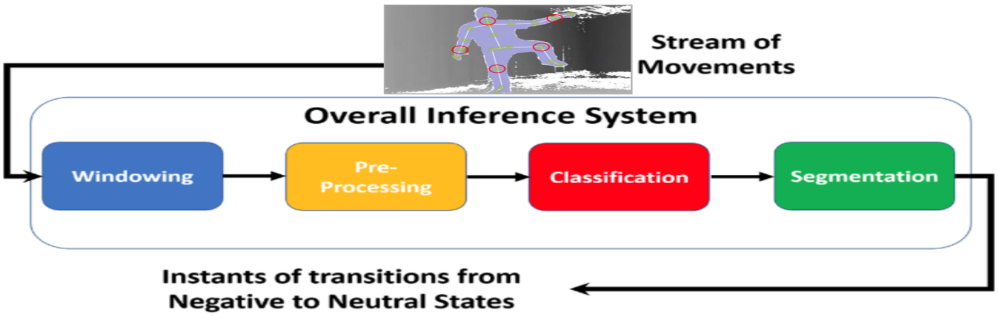

Figure 1: Representation of the proposed inference system

actors are in neutral state. It is important to remark that, during the neutral states, actors have been asked to be as much natural as possible, and to avoid rigid postures that would have made the disambiguation between neutral and negative emotions trivial.

4.2. Inference System

The objective of the proposed inference system is to identify the instants of transitions between a negative and a neutral state or, in other words, to detect when negative states finish. To make this possible, we exploit signal processing and machine learning strategies to develop a data processing pipeline, which consists in the following four main phases, which are performed in real-time: i) windowing, ii) preprocessing, iii) classification and iv) segmentation. These phases are described in the following, and depicted in Fig. 1.

4.2.1. Windowing

In the windowing phase, the signal (i.e., movements of positions and orientations) is sliced into overlapping windows of fixed length. The type, length and overlapping percentages of the windows are hyperparameters of our system that, after some heuristic optimization, have been set to: rectangular window, length of 2 seconds and overlapping of 50%, respectively.

4.2.2. Pre-Processing

The windowed signals represent the evolution over some time (and, specifically, of a window's length time, i.e., 2 seconds) of positions and orientations of head and hands. The signal in the temporal domain already carries significant information for the considered inference task, but other relevant information is more evident from the signal in the frequency domain. Hence, the module of the Fourier Transform is also computed. Other features have also been extracted (e.g., the distance between hands over time), but from numerical assessments they did not prove able to increment the effectiveness of the system. Hence, they have been discarded. The signal resulting from the pre-processing phase is therefore a window both in the time and frequency domains. This signal is then fed to the machine learning module described in the following subsection.

4.2.3. Classification

The machine learning module is based on a Gradient Boosting Classifier composed of 100 estimators. This machine algorithm has been selected mainly because of its capability to avoid over-fitting issues, which are likely to occur given the limited size of the considered training set. The machine learning module takes as input the signal described above (i.e., a window of the stream of movement), and returns a binary value encoding the inferred emotional state (i.e., 0 for the neutral and 1 for the negative states, respectively). Hence, the classification module outputs a stream of $0s$ and $1s$, which correspond to the inferred class for each window of the overall stream.

4.2.4. Segmentation

The segmentation module aims to identify only the relevant instants of transitions between negative and neutral states given the sequences of classes predicted by the classification module. Let us firstly note that, in case the previous classification was perfect, the segmentation task would be trivial. Indeed, in this case it would have been enough to identify transitions from class 1 (i.e., negative state) to class 0 (i.e., neutral state). However, the classification process is prone to errors, which in turn affects also the segmentation. To tackle this issue, we propose a heuristic approach capable of identifying relevant transitions only and to discard the spurious ones (e.g., occasional transitions between states).

5. Results

The effectiveness of the proposed inference system is evaluated considering its ability to detect only the relevant transitions between negative and neutral states. Specifically, the set of instants of detected transitions (i.e., the output of the segmentation module) is compared with the ground truth set of instants of actual transitions. An inferred transition instant (say $t^\star$) is considered to be correct if there is a corresponding t in the ground truth set such that $t - \eta < t^\star < t + \eta$, where η is a value of tolerance set to 3 seconds. The considered metrics to evaluate the performance of the proposed system are the Precision, Recall and F-score. The precision is the percentage of predicted instants that are actually transition points. The recall is the percentage of the actual transition instants that are correctly predicted. The F-score is a metric that balances precision and recall. While the performance of the overall system is measured as described, it is also important to show the result of the classification process, which is a fundamental building block for the overall inference system. The performance of the classification is also evaluated considering Precision, Recall and F-score. In this case, the meaning of these metrics is as follows: the precision is the percentage of predicted classes that are correctly estimated; the recall is the percentage of ground truth classes that are correctly predicted; the F-score balances precision and recall. Performance of both the overall inference system (i.e., segmentation part) and of the intermediate classification process are summarized in Table 1.

From these results, it is possible to observe that the inference system i) effectively classifies the windows in which the stream of movements is sliced, and ii) effectively identifies relevant transitions between negative and neutral states. In fact, the F-scores of both the processes are

	Precision	Recall	F-Score
Classification	0.81	0.80	0.81
Segmentation	0.78	0.80	0.79

Table 1
Performance of the proposed system, measured considering Precision, Recall and F-score, in the sub-tasks of Classification and Segmentation

fairly high, which makes the overall system suitable for being used in real contexts. Moreover, we note a high balance between performance of the classification and of the segmentation. Given that the segmentation highly relies on the previous classification process, it is possible to conclude that the proposed heuristic is able to effectively exploit the output of the classification, without introducing other significant errors. Finally, it is also important to note that precision and recall are also fairly balanced, both in the classification and in the segmentation process. This implies that the overall system does not have a biased behavior (e.g., either to over-segmented or to down-segment).

6. Conclusion

In this paper, we have proposed an inference system that estimates the instants of transitions between non-negative and negative affective states of a person from her body movements. The system is a two-layers architecture, where the first layer classifies portions of the observation (i.e., body's movements) into one of the aforementioned classes, while the second layer identifies the transitions between two successive effective states (process that we refer to as segmentation). Given that the classification process introduces errors that affect the overall system, the segmentation process is aimed to select only the relevant transitions between the detected emotional states. In our work, segmentation is performed by means of an heuristic approach. The system has been evaluated on data that we collected using the Kinect 2.0 technology, and its performance is evaluated in both the first phase (i.e., the classification into non-negative or negative emotions) and in the second one (i.e., segmentation). Results show that Precision, Recall and F-score are fairly high and balanced (i.e., around 0.8), which suggests that the proposed system is already applicable in a real Smart City scenario. As a future work, we plan to optimize the overall system by using a higher volume of training data, as well as by implementing other data processing strategies.

Acknowledgement

This research was funded by the EU EIT Manufacturing initiative through the V-Machina project.

References

[1] A. Saxena, A. Khanna, D. Gupta, Emotion recognition and detection methods: A comprehensive survey, Journal of Artificial Intelligence and Systems 2 (2020) 53–79.

[2] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, S. D. Kollias, Emotion recognition through facial expression analysis based on a neurofuzzy network, Neural Networks 18 (2005) 423–435.

[3] H. Chao, L. Dong, Y. Liu, B. Lu, Emotion recognition from multiband eeg signals using capsnet, Sensors 19 (2019) 2212.

[4] S. G. Koolagudi, K. S. Rao, Emotion recognition from speech: a review, International journal of speech technology 15 (2012) 99–117.

[5] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, G. Anbarjafari, Survey on emotional body gesture recognition, IEEE transactions on affective computing (2018).

[6] Y. Huang, J. Yang, P. Liao, J. Pan, Fusion of facial expressions and eeg for multimodal emotion recognition, Computational intelligence and neuroscience 2017 (2017).

[7] S. Lalitha, D. Geyasruti, R. Narayanan, M. Shravani, Emotion detection using mfcc and cepstrum features, Procedia Computer Science 70 (2015) 29–35.

[8] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomedical Signal Processing and Control 47 (2019) 312–323.

[9] P. Shen, Z. Changjun, X. Chen, Automatic speech emotion recognition using support vector machine, in: Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, volume 2, IEEE, 2011, pp. 621–625.

[10] H. Kim, J. Ben-Othman, S. Cho, L. Mokdad, A framework for iot-enabled virtual emotion detection in advanced smart cities, IEEE Network 33 (2019) 142–148.

[11] H. Kim, J. Ben-Othman, Toward integrated virtual emotion system with ai applicability for secure cps-enabled smart cities: Ai-based research challenges and security issues, IEEE Network 34 (2020) 30–36.

[12] Y. Jiang, W. Xiao, R. Wang, A. Barnawi, Smart urban living: Enabling emotion-guided interaction with next generation sensing fabric, IEEE Access 8 (2019) 28395–28402.

[13] A. Elabora, M. Alkhatib, S. S. Mathew, M. El Barachi, Evaluating citizens' sentiments in smart cities: A deep learning approach, in: 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), IEEE, 2020, pp. 1–5.

[14] M. Rodrigues, R. Machado, R. Costa, S. Gonçalves, Smart cities: Using gamification and emotion detection to improve citizens well fair and commitment, in: Science and Information Conference, Springer, 2020, pp. 426–442.

[15] M. Zhao, F. Adib, D. Katabi, Emotion recognition using wireless signals, in: Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, 2016, pp. 95–108.

[16] A. Napoli, S. Glass, C. Ward, C. Tucker, I. Obeid, Performance analysis of a generalized motion capture system using microsoft kinect 2.0, Biomedical Signal Processing and Control 38 (2017) 265–280.