The VISOR project: Virtual coachIng Services for OldeR adults

Evaggelos Spyrou*†, Eirini Mathe*‡, Ioannis Vernikos*†, Panagiotis Dedousis†, Anna Vlachou† and Phivos Mylonas‡
*Institute of Informatics and Telecommunications, National Center for Scientific Research - "Demokritos," Athens, Greece
Email: espyrou@iit.demokritos.gr
†Department of Computer Science and Telecommunications, University of Thessaly, Lamia, Greece
Email: ivernikos@uth.gr, panagiwtisdedousis@gmail.com, annavlaxou1993@hotmail.com
‡Department of Informatics, Ionian University, Corfu, Greece
Email: {c17math,fmylonas}@ionio.gr

***Abstract*—In this paper we present an approach for virtual coaching services. In brief, coaching consists a method for development, involving a coach and a learner and focuses on a specific goal. The coach supports, shapes, reinforces some desired behavior, by training and guidance, while the learner sets a goal of development, aims to achieve it by interacting with the coach. Moreover, health/life coaching is a means of achieving some health-related/wellness goals in order to enhance the well-being of individuals. We provide an end-to-end approach for virtual coaching. Our approach is IoT-ready and is based on the analysis of multimodal, multisensory data. Upon analysis of sensor measurements, decision making takes place and the users' behavior (activities, emotions, events etc.) is recognized. All these data are then used for virtual coaching, through the users' smartphone or smartwatch.**

***Index Terms*—virtual coaching, human behavior recognition**

I. Introduction

Generally speaking, coaching consists a method for development, involving two parts: a) a coach, whose role is to support, shape and/or reinforce some kind of desired behavior, by providing training and guidance; and b) a learner (coachee), who sets a specific goal of development and aims to achieve it by interacting with the coach. The main competence of a coach is his/her experience and expertise, which guides the learning processing. Coaching is different from learning, in the sense that the former focuses on a specific, specialized goal, while being continuously supported and guided by the coach. On the other hand, learning may set more general goals, may not require a tutor and may be achieved even without conscious awareness. Several areas of living, directly associated with some kind of coaching are business, sports, and personal life.

In the context of healthy living, Palmer [11] describes "health-coaching" as a means of achieving some health-related goals in order to enhance the well-being of individuals. Therein, health coaching is directly related to health education and promotion, while in practice it combines best practices from the fields of both health and coaching psychology. Several interventions that are encountered include nutritious diet, increased physical activity, weight management, medication adherence, etc. [10]. The overall goals are focused on designing interventions for the prevention and management of chronic diseases. Complementary to health coaching is "life-coaching." Contrary to the former, the latter focuses on healthy individuals, aiming to maintain or even promote their wellness. More specifically and as stated by Williams et al. [13], life coaching focuses on the person's whole life and on wellness rather than pathology.

In this work we present the vision of the VISOR project, which in brief aims to provide virtual coaching services for the elderly. By using a set of available information that will be collected either by observing the users or their environment, it will provide an affordable solution, mainly based on off-the-shelf products, while following an Internet-of-Things (IoT) [8] architecture. Cloud-based services will be responsible for the collection and the analysis of the multimodal and heterogeneous sensory information, aiming to extract physiological environmental measurements and also for the analysis of audiovisual information, in order to recognize patterns of behavior, emotions and events. Then, a decision making system will provide coaching through an avatar within the users' phone or smartwatch.

The rest of this paper is organized as follows: section II provides a brief overview of virtual coaching. Then, section III describes the objectives and the goals of VISOR. Section IV presents novel methodologies that have already been implemented for activity and emotion recognition, while conclusions are drawn in section V, where plans for future work are also presented.

II. Related Work

Virtual coaching approaches have been used in several areas. Notable examples include coaching of teachers during a course. For example, a university professor may coach a teacher via a bluetooth headset, by suggesting words to be used or by alerting for student behaviors [1]. More challenging use cases such as supporting teachers for students with disabilities have also been proposed [2]. Also virtual coaching has been applied for support of nurses [4], or for improving wheelchair transfer techniques of wheelchair users [3]. The majority of virtual coaching applications have been applied in both health and wellness domains. Notable examples include coaching for

quitting smoking [5], management of nutrition [6], walking [7] etc.

III. The VISOR Project

The VISOR (Virtual coachIng Services for OldeR adults) project aims to provide virtual coaching services for the elderly people. Its aim is to guide and lead them to a better lifestyle. It is based on an IoT-ready architecture, which shall be discussed shortly. Therefore, all sensors and processing algorithms will be implemented as exposed web services running on the cloud. A typical IoT categorization is consisted of three distinctive, yet interdependent, types of services: a) *sensing* services, acquiring measurements and capturing properties of the physical world, providing the aforementioned raw or slightly processed measurements; b) *processing* services whose input consists of the output of sensing services, while their output consists of the inferred results; and c) *actuating* services that enable actions based on the results of processing services.

In the context of VISOR, *sensing* services will be based on low-cost sensors which will be divided into two categories: a) sensors that will be installed within the users environment (i.e., her/his home), such as video cameras, microphone arrays, environmental sensors etc.; and b) worn by the user, such as the heart rate sensor, the accelerometer and the gyroscope of a smart watch or a smart bracelet. The raw measurements collected will be further processed and given as input to *processing* services, which will be categorized into: a) audio feature extraction; b) video feature extraction; c) physiological processing; d) environmental processing; e) emotion recognition; f) behaviour recognition module; and g) event recognition services. The output of processing will be sent to a decision support system that will identify patterns of behavior, health issues and emotions with the aid of machine learning algorithms. By using a smart app, *actuation* will take place as personal alerts will be provided in mobile phones and smartwatches of users. In Fig. 1, we illustrate the architecture of the proposed system and show the flow of information from the sensorial components to the services.

For example, the decision making module will collect input from the physiological and environmental sensors and also from the detected events, behaviour and emotions. It will also consider contextual information (such as temperature, weather), the users location (indoors, outdoors) and the users history (e.g., whether she/he has not exercised for days) and will make appropriate recommendations (e.g., if the user is in the house, the weather is good and she/he has not walked for several days, a walk to the park may be suggested). It will also make decisions based on users emotions (e.g., if the users affective state suggests that he is sad, the coaching system may suggest him to participate in a social event or visit his/her family). In case physiological parameters indicate a health issue (e.g., higher than normal heart rate) or emotional state indicates situations such as depression, her/his carers will be informed. In case a dangerous event such as fall is detected, the system will inform her/his carers, or another person indicated by the carers, such as a neighbor.

Note that emphasis will be given to ensure data protection and privacy preservation. Evaluation will be both quantitative and qualitative. The former will take place into two phases; each service shall be evaluated using publicly available datasets and upon integration a real-like experiment will take place into an artificial environment that simulates a real home. The latter will focus on usability and user satisfaction, targeting real users.

IV. Novel Methods for Behavior Recognition and Preliminary Results

In this section we will present a set of novel algorithms that have been developed in the context of VISOR. More specifically, we will introduce a) a novel algorithm for emotion recognition from speech; b) a novel end-to-end deep learning neural network architecture for human activity recognition from 3D skeletal data; and c) an approach for human activity recognition using data acquired from inertial measurement unit (IMU) of mobile phones.

A. Emotion Recognition from Speech

The recognition of emotion in the context of real assisted living or coaching environments has only been recently proposed as proof-of-concept and in a few experiment setups [15]. Within the VISOR project, our main goal is to experiment with a wide range of verbal and non-verbal content of various types and duration from audio-visual information and with early and late fusion so as to combine prosody and other acoustic features, focusing on deep methods for multimodal fusion [16]. We will also use computer vision inspired methodologies, e.g., [17], which is based on the bag-of-visual words method, applied on audio segment spectrograms. Random segments of length t_s are extracted for any given audio sample and then by applying the short-time Fourier transform, pseudocolored images of spectrograms are created. Then, the SURF features [18] are extracted to create a visual vocabulary and train an SVM classifier.

B. Human Activity Recognition from 3D Skeletal Data

Human action recognition is a promising research field in the broader areas of computer vision and pattern recognition. Within the context of VISOR, one of our main goals is to dig out novel visual representations of 3D skeletal information, in order to create images, which will be then used to feed convolutional neural networks (CNNs) [9], which will be used both for feature extraction and recognition. Among the other types of visual sensors, we will use RGB and depth cameras, such as the Microsoft Kinect. The latter is equipped with a powerful SDK, able to extract the 3D positions of human skeletal joints, in real-time. In previous work [12] we proposed a novel representation by creating pseudocolored images. These images contained all the necessary information, regarding the 3D inter-joint distances during the performance of an action. More specifically, we used the 25 available joints and their coordinates x,y and z and created images by assigning the aforementioned coordinates to R,G,B image

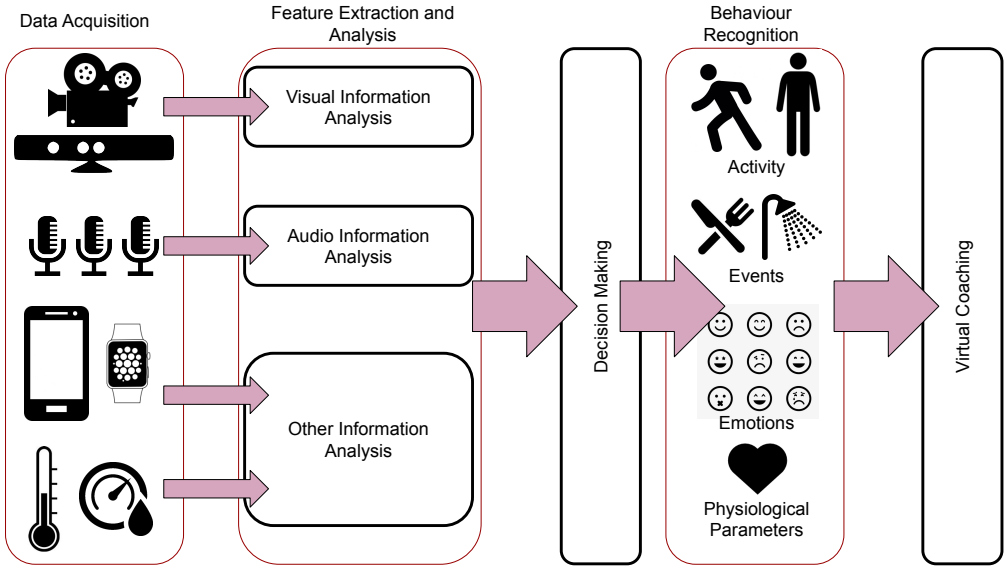

Fig. 1. An overview of the proposed approach.

channels, respectively. Note that it is common sense that different persons may perform the same action at different duration. To eliminate this problem we enforced a linear interpolation step at each video sequence, i.e., set the duration of all videos equal to $N = 150$ frames.

C. Human Activity Recognition from wearable sensors

For recognizing actions based on wearable devices, based on IMU data (e.g., accelerometers, gyroscopes, etc.), we have implemented a novel deep approach. More specifically, motivated by the work of Jiang and Yin [19], we propose a novel deep CNN-based approach that does not require a feature extraction step. Instead of training a single CNN into a multi-class problem, we opted for training one CNN for each class, solving a binary one-vs-all problem. More specifically, each CNN was trained to recognize one class vs all the other available. We first created signal images, based on the raw acceleration measurements. These measures included triaxial linear acceleration and angular velocity signals that had been pre-processed with a median and a low-pass filter for noise reduction. As discussed in subsection IV-B, we also imposed a linear interpolation step. Then, we applied FFT so as to create an image representation which we refer to as "activity image," that captures the spectral properties of these measurements. We have experimented with both greyscale and pseudocolored activity images. Note that although measurements used in this work have been collected by the embedded IMU of a mobile phone, our approach may easily be adapted to any IMU measurements, e.g., of wearables.

D. Datasets

For emotion recognition from speech, we have used 3 publicly available datasets, namely EMOVO [20] (Italian), SAVEE [21] (English) and EMO-DB [22]. All these datasets have been created by actors, performing utterances of neutral content. Selected emotions were anger, boredom, disgust, fear, happiness, sadness and neutral. For human action recognition using skeletal data, we used the PKU-MMD [23] dataset. It consists a large scale dataset captured via the Kinect v2 sensor, that focuses on human action understanding. It provides 5 categories of resources, i.e., RGB images, RGB videos, depth maps, infrared sequences and skeletal data. In our experiment, we used Phase #1 which contains approximately 20K action instances from 51 action categories, spanning into 5.4M video frames. Finally, for human activity recognition from wearable data, we used the dataset of [24], which consists of 6 actions (walking, walking_upstairs, walking_downstairs, sitting, standing, laying_down), performed by 30 subjects.

E. Experimental Results

Emotion recognition from speech achieved average an F_1 score of 57% (ranging between 50% and 65%) by using visual vocabularies of size $1100 \leq N \leq 1500$ from the same language samples and 45% (ranging between 43% and 48%) by using visual vocabularies of size $100 \leq N \leq 400$ from multilingual samples.Concerning the human action recognition task, early experiments in order to evaluate our approach have been performed under three approaches. First, we achieved an accuracy up to 89% in single view experiments, then in cross-view experiments we achieved an accuracy up to 82% and finally, in cross-subject experiments, accuracy was up to 85%. Our approach for human action recognition from wearables achieved average accuracy 85.8% (ranging between 81.8% and 91.2%) when using pseudocolored images and 89.3% (ranging between 81.6% and 96.4%) when using greyscale images.

V. Discussion and Future Work

In this paper we presented an affordable end-to-end approach for virtual coaching for elderly. Our system, namely

VISOR aims to exploit information gathered from sensors placed within the users' environment and also from sensors worn by them, while follows an IoT-ready architecture. Sensor measurements will be processed by services running in a cloud infrastructure. A decision support system will provide personalized coaching to users through their smartphone or smartwatch aiming to preserve their wellness or tackle health issues or diseases.

We feel that VISOR consists an innovative system for virtual coaching. Its notable aspects are: a) it integrates novel multimodal behavior recognition algorithms; b) it uses cheap/off-the-shelf sensory devices, thus consisting an affordable solution; c) it is unobtrusive, since user will only wear a smartwatch and carry a smartphone which are typical daily-life objects – cameras, microphones and environmental sensors shall be minimally exposed; d) it will integrate open IoT protocols enabling scalability and seamless interoperability between heterogeneous services, including novel approaches to (semi-)automated service composition approaches using fully semantic descriptions of services; e) it will be easily upgradeable since services will run on the cloud, making it also easily expandable.

Future work will firstly focus on the integration and deployment of the presented methods into a real-like environment and evaluation by users. We aim to convert algorithms into exposed, re-usable web services and use the cloud infrastructure that has been developed by the SYNAISTHISI project [25]. Moreover, we plan to continue the investigation on novel representations and architectures of deep networks, which will be applied into the aforementioned problems, to fit the goals of the VISOR project. In the process, other types of sensors, such as environmental may be added and other types of information may me inferred, such as physiological parameters and/or events. All available information will be used by the decision support system to a coaching scenario. We plan to evaluate this scenario in a real-life assistive living environment.

ACKNOWLEDGMENT

We acknowledge support of this work by the project VISOR "Virtual coachIng Services for OldeR adults" which is implemented under the "1st HFRI Call for Scholarships for PhD Candidates," funded by the General Secretariat for Research and Technology and the Hellenic Foundation for Research and Innovation (HFRI).

REFERENCES

[1] Rock, M. L., Gregg, M., Gable, R. A., & Zigmond, N. P. (2009). Virtual coaching for novice teachers. Phi Delta Kappan, 91(2), 36-41.

[2] Israel, M., Carnahan, C. R., Snyder, K. K., & Williamson, P. (2013). Supporting new teachers of students with significant disabilities through virtual coaching: A proposed model. Remedial and Special Education, 34(4), 195-204.

[3] Hwang, S., Tsai, C. Y., & Koontz, A. M. (2017). Feasibility study of using a Microsoft Kinect for virtual coaching of wheelchair transfer techniques. Biomedical Engineering/Biomedizinische Technik, 62(3), 307-313.

[4] Maritz, J. E., & Roets, L. (2013). A virtual appreciative coaching and mentoring programme to support novice nurse researchers in Africa. African Journal for Physical Health Education, Recreation and Dance, 19(Supplement 2), 80-92.

[5] Grolleman, J., van Dijk, B., Nijholt, A., & van Emst, A. (2006, May). Break the habit! designing an e-therapy intervention using a virtual coach in aid of smoking cessation. In International Conference on Persuasive Technology (pp. 133-141). Springer.

[6] Jarvinen, P., Jarvinen, T. H., Lahteenmaki, L., & Sodergard, C. (2008, January). HyperFit: hybrid media in personal nutrition and exercise management. In 2008 Second Intl Conf. on Pervasive Comp. Technologies for Healthcare (pp. 222-226). IEEE.

[7] Albaina, I. M., Visser, T., van der Mast, C. A., & Vastenburg, M. H. (2009, April). Flowie: A persuasive virtual coach to motivate elderly individuals to walk. In 2009 3rd Intl Conf. on Pervasive Comp. Technologies for Healthcare (pp. 1-7). IEEE

[8] Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. Computer networks, 54(15), 2787-2805.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[10] Olsen, J. M., & Nesbitt, B. J. (2010). Health coaching to improve healthy lifestyle behaviors: an integrative review. American Journal of Health Promotion, 25(1), e1-e12.

[11] Palmer, S. (2004). Health coaching: A developing field within health education. Health Education Journal, 63(2), 189-191.

[12] Vernikos, I., Mathe, E., Papadakis, A., Spyrou, E., & Mylonas, Ph. (2019) An Image Representation of Skeletal Data for Action Recognition using Convolutional Neural Networks. In Proc. of Int'l Conf. on PErvasive Technologies Related to Assistive Environments (PETRA).

[13] Williams, P., & Davis, D. C. (2007). Therapist as life coach: An introduction for counselors and other helping professionals (revised and expanded). WW Norton & Company.

[14] Wolever, R. Q., Simmons, L. A., Sforzo, G. A., Dill, D., Kaye, M., Bechard, E. M., ... & Yang, N. (2013). A systematic review of the literature on health and wellness coaching: defining a key behavioral intervention in healthcare. Global advances in health and medicine, 2(4), 38-57.

[15] Konstantinidis, E. I., Billis, A., Savvidis, T., Xefteris, S., & Bamidis, P. D. (2017). Emotion Recognition in the Wild: Results and Limitations from Active and Healthy Ageing cases in a Living Lab. In eHealth 360 (pp. 425-428). Springer.

[16] Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., ... & Ferrari, R. C. (2015). Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces, 1-13

[17] Spyrou, E., Nikopoulou, R., Vernikos, I., & Mylonas, P. (2019). Emotion Recognition from Speech Using the Bag-of-Visual Words on Audio Segment Spectrograms. Technologies, 7(1), 20.

[18] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool,Speeded-up robust features (SURF), Computer Vision and Image Understanding, 110(3),pp.346359, 2008.

[19] Jiang, W., & Yin, Z. (2015, October). Human activity recognition using wearable sensors by deep convolutional neural networks. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 1307-1310). Acm.

[20] Costantini, G.; Iaderola, I.; Paoloni, A.; Todisco, M. Emovo corpus: An italian emotional speech database. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 2631 May 2014; pp. 35013504.

[21] Haq, S.; Jackson, P.J.; Edge, J. Speaker-dependent audio-visual emotion recognition. In Proceedings of the 2009 International Conference on Auditory-Visual Speech Processing, Norwich, UK, 1013 September 2009; pp. 5358.

[22] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 48 September 2005.

[23] Liu, C., Hu, Y., Li, Y., Song, S., & Liu, J. (2017). PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475.

[24] Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013, April). A public domain dataset for human activity recognition using smartphones. In Esann.

[25] Pierris, G., Kothris, D., Spyrou, E., & Spyropoulos, C. (2015, October). SYNAISTHISI: An enabling platform for the current internet of things ecosystem. In Proceedings of the 19th Panhellenic Conference on Informatics (pp. 438-444). ACM.