SUPPORTING BUSINESS MODEL INNOVATION BASED ON DEEP LEARNING SCENE SEMANTIC SEGMENTATION

Nikolay Neshov
Assistant Professor, Faculty of Telecommunications, Technical University of Sofia, 8 blvd. Kliment Ohridski,
Sofia, 1756, Bulgaria
nneshov@tu-sofia.bg
http://ftk.tu-sofia.bg/

Agata Manolova
Associate Professor, Faculty of Telecommunications, Technical University of Sofia, 8 blvd. Kliment Ohridski,
Sofia, 1756, Bulgaria
amanolova@tu-sofia.bg
http://ftk.tu-sofia.bg/

Krasimir Tonchev
Researcher, Faculty of Telecommunications, Technical University of Sofia, 8 blvd. Kliment Ohridski,
Sofia, 1756, Bulgaria
k_tonchev@tu-sofia.bg

Antoni Ivanov
PostDoc student, Faculty of Telecommunications, Technical University of Sofia, 8 blvd. Kliment Ohridski,
Sofia, 1756, Bulgaria
astivanov@tu-sofia.bg

Abstract

The capacity to create innovative Business Models (BM) has become the foundation for numerous businesses. Business Model Innovation (BMI) grows more significant as digitalization influences our everyday lives and prompts the development of better approaches for working, imparting and collaborating in this computerized universe of Industry 4.0. In this paper we present a conceptual architecture which can be applied in the modern video-conference systems with the help of semantic segmentation. The scene represents an environment, intended for discussion of ideas in business modeling. The semantic segmentation allows each pixel of an image (or video) from the scene to be related or classified to a specific type of object. In this way it is possible to interpret the description of a scene by the machine. Thus, with the help of the proposed architecture, the processes taking place between objects and people in the surrounding environment can be analyzed for the purpose of digitization of BMI by modelling human behavior and cognitive processes into logical expressions that can be digitized and automated. The semantic segmentation is considered as a basic element in this type of interaction. We demonstrate the effectiveness of our algorithm in with real data examples.

Keywords: Deep learning; semantic segmentation; teleconference; business model innovation.

1. Introduction

Organizations have consistently adjusted with the evolving times, yet the deluge of digital innovation technologies, such as mobile, cloud, social, and big data analytics has quickened the pace at which the organizations need to advance and how much they change the manner in which they develop, work, and serve their clients. Business Model Innovation (BMI) and Business Models (BM) have become obligatory for any organization. According to [4], BM and the process of BMI can be described as phenomena with patterns and mechanisms. Building on this, the BM as a conceptually distinct construct may provide theories with new explanatory power and reach.

Artificial intelligence and deep neural network architectures are entering the modern world through the ability to analyze certain types of problems, replacing the need for human intervention and taking action to increase the ability to achieve a goal. In this aspect, deep neural networks can be applied as an innovative approach to support efficiency and effectiveness in business modeling. The modern literature describes various developments at the theoretical and applied level, the main challenge of which is the modeling of human behavior and cognitive processes in an appropriate digital form. This in turn allows computer analysis of complex problems in order to make optimal decisions. In this sense, there is a need to develop effective methods and algorithms to bridge the semantic gap between the way one understands the scene and the way machines interpret it.

The objectives of this paper are precisely in the development of such an algorithm to serve as a tool to support activities in the business environment. With further development, a typical application of our work includes the following scenarios:

- If a user is unable to attend a meeting in the room, he/she can participate remotely via a video conferencing application. Semantic segmentation allows the participant to remove unnecessary visual content around him/her (objects from the background, objects around him/her or other users who do not participate in the conference). On the one hand, this reduces redundant data transmission (which is especially valuable in conditions where the network capacity is very limited), and also hides objects that should not be visible to other users. In addition to participant segmentation, an interactive whiteboard and/or marker can be segmented to be visualized remotely.
- Another example of application is related to the analysis of the work with used tools during a meeting. For example, an assessment can be made of which tools are used and how often they are used [9]. Subsequently, the unnecessary ones can be removed or improvements of the used ones can be offered.

Semantic segmentation methods can be divided into two groups: Classical methods and those based on neural networks with deep learning. Classical methods include treatments such as thresholding, K-means clustering, and contour detection. In modern systems, algorithms based on the second group (neural networks with deep learning) are preferred, which are characterized by higher efficiency. Hence, the proposed algorithm in this work is based on this type of networks.

In these days, there is a lot of research works, based on 3D reconstruction that can be used in teleconferencing systems [1], [2], [5], [9]. A pool of applications exists with attempt to analyze the user behavior [7], [8], [10]. The fundamental elements in these research papers are scene reconstruction and human computer interaction. In our work, the intention is to upgrade the usage of these basic elements, offering additional useful architecture that can be of help in real application scenarios for virtual teleconferencing systems.

The rest of the paper is organized as follows: the next section gives a detailed description of the proposed algorithm. Section 3 presents the experimental results. The final section draws the conclusion and suggests the scope of future work.

1. Algorithm description

On Fig. 1 is shown the conceptual diagram of conference system for supporting business modeling.

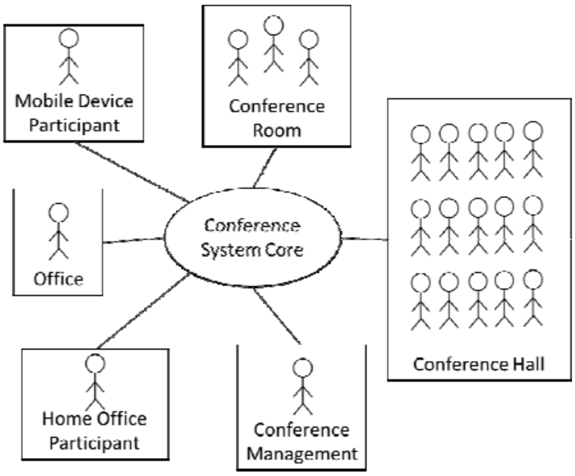

Fig. 1. General conception diagram of the proposed conference system.

The interaction analysis and management is maintained by the "Conference System Core" unit. We assume several real cases in (tele) conference communication, each of which can occur individually or simultaneously. We will describe each case and possible scenarios behind it, so that we can conclude what are the necessarily logical processing steps that the "Conference System Core" unit should take care of. Further, we propose a semantic segmentation based algorithm that is able to handle all the described scenarios. On Fig. 2, are shown the fundamental construction blocks during the analysis. Each of these blocks can be considered relevant or not relevant to the specific case and scenario which we will be discussed further.

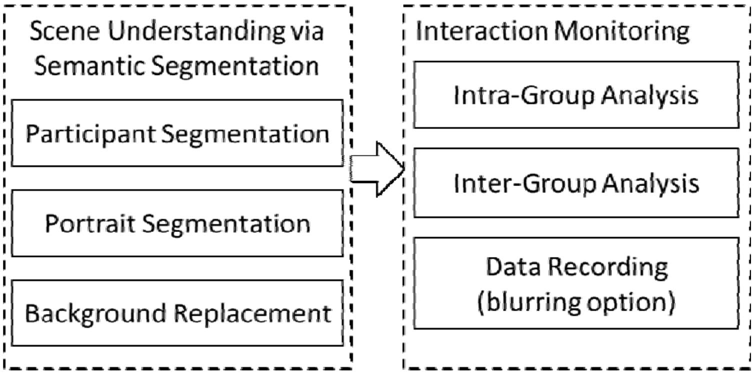

Fig. 2. Basic analyses blocks for supporting different conference systems cases.

1.1. *"Scene Understanding via Semantic Segmentation"*

The left part in Fig. 2 "Scene Understanding via Semantic Segmentation" includes the following major processing modes/steps of the proposed algorithm.

- "Participant Segmentation" - A segmentation masks for all available people in the frame are computed and the background is removed. A special case is considered when only one participant should be segmented for remote conference meeting, so that no other participant in the scene can distract the other participants. The "Conference System Core" module should be set so that either one or more participants can be segmented. In case of one participant, a tracking method should be applied so that the algorithm guaranties that the image of only this participant will be segmented during the meeting.
- "Portrait Segmentation" - A segmentation mask for only the upper part of the body can be useful during the conference. Since in most of the cases we are only interested in the face-to-face communication, the proposed algorithm allows for portrait segmentation mode. This mode has also the ability to reduce the data transferred or recorded during the communication.
- "Background Replacement" - This mode is closely related to "Participant Segmentation". Moreover it provides background substitution, which is beneficial in teleconference systems, where unnecessary parts from the scene are not in need to be transmitted. This allows for hiding some details from the scene that might distract the participants. This mode is helpful especially in home office environment. It can also be used for replacing the background for all participants in the conference with a common appropriate content. This creates an immersive feeling in the participants as if they were in the same place.

1.2. *"Interaction Monitoring"*

The right part in Fig. 2 "Interaction Monitoring" contains the general processing blocks, which implement the participants' interaction analysis. Here we discuss each method individually:

- "Intra-Group Analysis" - This processing block performs interaction analysis locally between participants, which are located in a common meeting room. The goal of this block is finding out where a person is looking at. During the meeting analysis, four possible targets are taken into account. These are:
  - Participant is looking at Working Station-WS (PC, Notebook, Mobile Device or etc.)
  - Participant is looking at other participant
  - Participant is looking at the screen/monitor/TV
  - Participant is distracted (i.e. looking at some other point)

  In order to accomplish such analysis, two contributing factors might be taken into account: head orientation and eyes orientation. Since our goal is to determine the focus of attention, not the exact gaze point, we simplify this processing by choosing to estimate head orientation only. At first all possible targets are found by the "Scene Semantic Segmentation" (See Fig. 2). To achieve this goal, a Deep Learning DL method is applied, so that all Participants, WSs and Whiteboard are detected in the scene. Once all targets locations are determined, we apply Supervised Descent Method SDM [11] for head orientation assessment on each participant. Given the coordinates of all available targets, the focus of attention can be easily estimated from the head orientation. If the participant is not focused on some of the targets, this participant is considered distracted. The statistical information for the attention measured by the aforementioned method is recorded for all participants during the whole meeting.

- "Inter-Group Analysis" – This block includes remote interaction monitoring and analyses how the participants between distinct groups communicate each other. Each participant or group of participants should have a dedicated conference camera (i.e. web camera). The video of each camera is processed so that each face emotion over time is recognized and stored on the server. This can be of help in analysis of how each participant reacts emotionally during the meeting. For emotion recognition, we used the implementation of our previous work described in [5]. The possible emotional states are: neutral, anger, happy, sad and wondering.
- "Data Recording (blurring option)" - The interaction monitoring data computed by the aforementioned blocks (see the right part in Fig. 2) along with the captured video are stored in a dedicated server. For preserving privacy, a blurring option of all detected faces is available, so recording can be performed on a blurred version of the video.

1.3. *Example case scenarios*

From the conceptual diagram presented in Fig. 1, it can be seen that it is very flexible in terms of how the conference system is to be used. The "Conference System Core" can be configured from the "Conference Management" unit for the following example cases:

- "Local conference meeting" – In this case, there is no need for remote data transferring, and all possible interactions between participants are performed at place (room/hall/office). Considering a given presenter in the meeting, the system includes video sensors that capture the whole meeting room, as well as dedicated video sensor for the presenter screen/white board. The "Conference System Core" unit in this case should perform the following operations:
    - Intra-Group Analysis
    - Data recording (blurring option)
- "Remote conference meeting" – Unlike local conference meeting, in this case there is a need of Inter-Group Analysis, so the following operations are allowed:
    - Intra-Group Analysis
    - Inter-Group Analysis
    - Data recording (blurring option)

In Fig. 3 is depicted the main processing steps of the proposed algorithm.

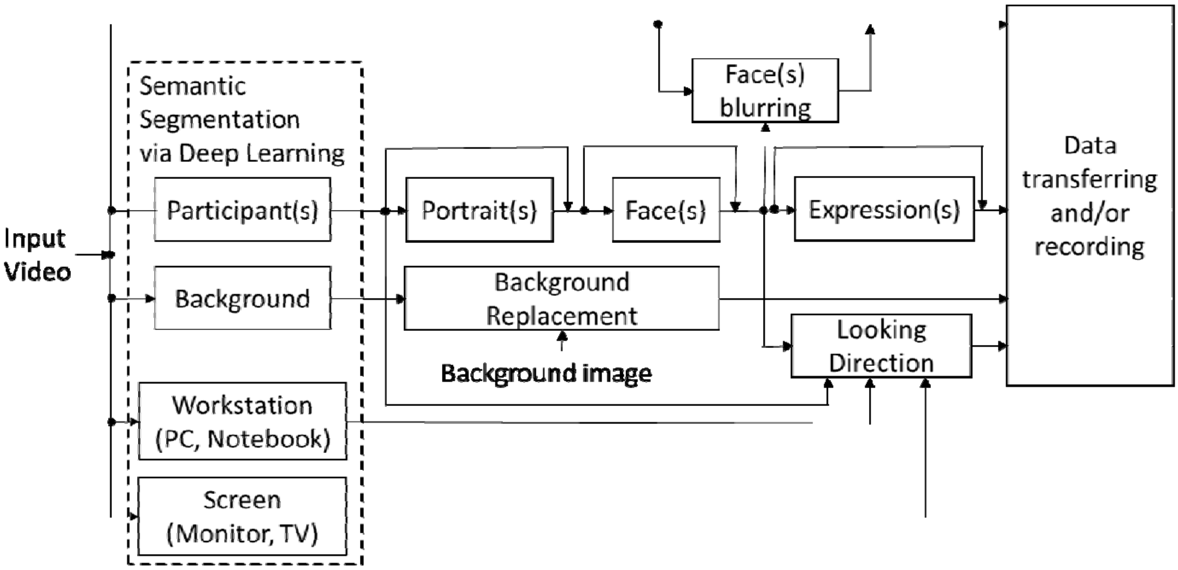

Fig. 3 General steps of the proposed algorithm.

The "Input Video" from the camera is given to the first block (Semantic Segmentation via Deep Learning), where we apply the algorithm based on the DeepLabV3 model with a ResNet-101 backbone [3] in order to distinguish "Participant(s)" from the "Background". If the input camera is positioned so that it can capture the whole scene, it is also possible to detect any WS (i.e. PC or Notebook) and Screen (i.e. Monitor or TV). Their coordinates are used if "Intra-Group Analysis" is desired, finding out the looking direction of the participants. Once the participant(s) contour(s) is(are) found the "Portrait(s)" detection is performed, extracting only the upper found of the body. Afterwards, "Face(s)" detection and "Expression(s)" detection is applied via the algorithm presented in our previous work [5] based on the Supervised Descent Method (SDM) [10]. We also assess the "Looking Direction" of the person(s) using a method based on SDM as mentioned earlier. The "Background Replacement" is also performed (if necessary) using appropriate "Background Image". For providing participant(s) privacy protection functionality a "Face(s) blurring" is also applied. This is done utilizing a Gaussian filter, processing only the facial part from the image. All gathered data is then transmitted through the network and/or recorded on the server (see "Data transferring and/or recording" block in Fig. 3).

2. Experimental part

In this section, we present the experiments conducted for each of the functionalities:

- "Participant Segmentation" – Fig. 4 depicts some example images of participants located by the Deep Learning semantic segmentation algorithm. Only the pixels that belong to the participants are remained, the other content of the image is replaced in white color.

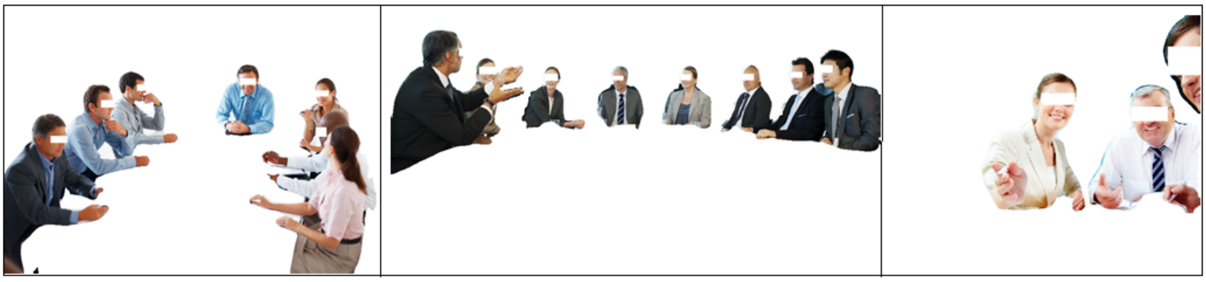

Fig. 4 Some example images of participants segmentation.

- "Portrait Segmentation" – In Fig. 5 are shown some example images, over which portrait segmentation operation is applied.

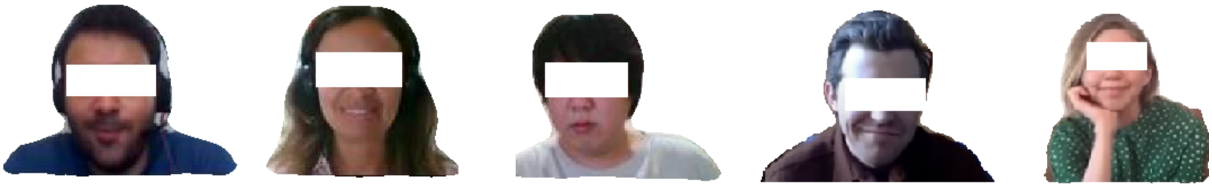

Fig. 5 Some example images of portraits segmentation.

- "Background Replacement" – This Fig. 6, we present the portraits of several participants in a conference call where the background was replaced with a common image.

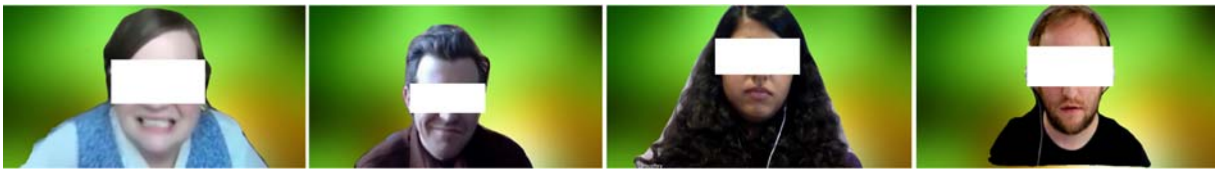

Fig. 6 Some example images of background replacement.

- "Intra-Group Analysis" – The main goal in this block is determining the looking direction. In Fig. 7, we present two picture examples of conference meeting and the corresponding labels of the resulted looking direction.

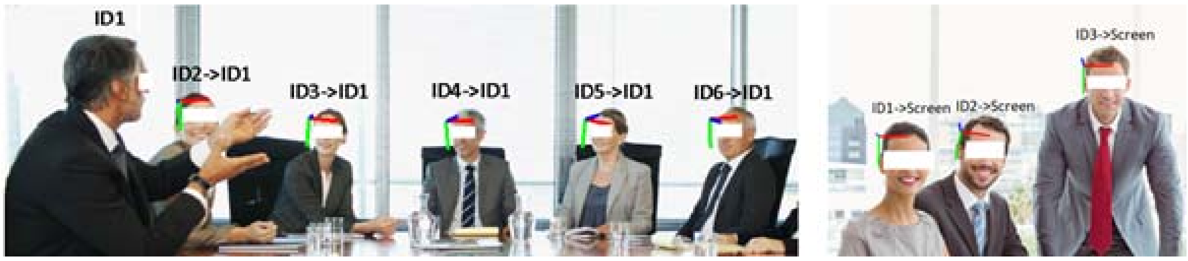

Fig. 7 Intra-group analysis of some participants along with the estimated looking direction.

It should be note that the face for each participant is not fully visible by the camera. In these cases, the looking direction cannot be determined, but if a given participant is identified by the sematic segmentation algorithm, an ID is assigned, and the looking direction of other participants to this participant can still be determined. This is the case in the left part in Fig. 7. We can see that the looking direction of participant with ID1 is not found. But his ID is found by the semantic segmentation. The looking directions of the other participants (ID2 to ID6) are determined and all of them are focused to ID1. For the right side in Fig. 7 all participants look at the screen. It should be noted that the screen is not visible and the algorithm for this particular environment is set so that the camera's plane is aligned to the screen's plane. It is convenient option, if the environment has only one camera that is attached to the top of the presentation screen. On Fig. 8 are depicted the interaction profile statistics of two participants during a conference. The Intra-group analysis determines the looking directions and builds a histogram, where each bin corresponds to the looking direction and its height determines the percentage time, at which this looking direction was preserved by the participant. From Fig. 8, it can be seen that both of the participants were looking at the screen in most of the time. The participant ID2 was more distracted than the participant ID1. It should be noted that it is not possible to determine the looking direction during the whole meeting due to a number of factors (the participant face was not visible from the camera, the participant is partially or fully occluded, the participant is out of the room/meeting, and etc.). Hence, the aforementioned results on Fig. 8 are presented during the 75% of the meeting time for Participant ID1 and for 79% of the meeting time for Participant ID2.

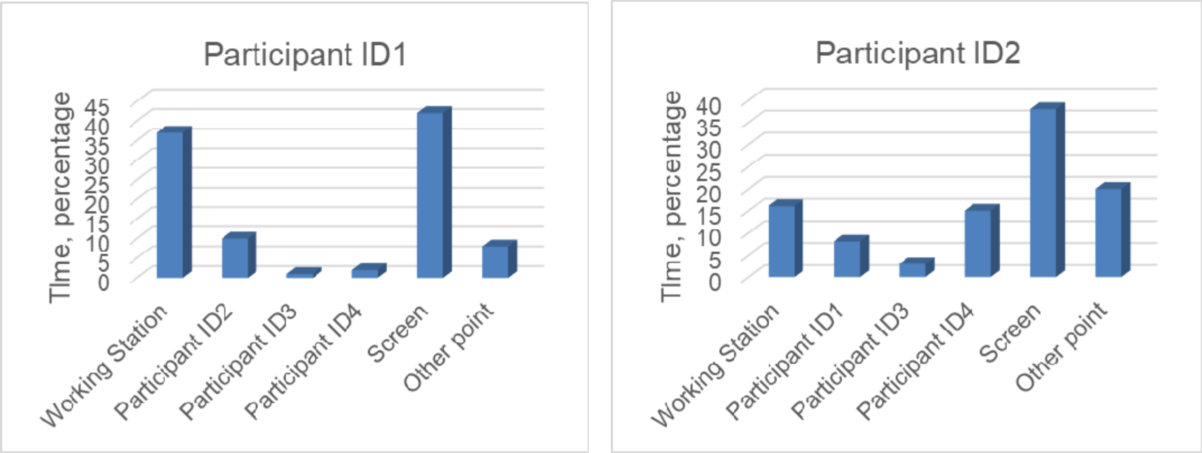

Fig. 8 Intra-group analysis: interaction profile for two participants with respect to the looking direction estimated during the meeting.

- "Inter-Group Analysis" – In Fig. 9 are depicted several participants in remote conference along with the estimated emotions for each of them. In this example, three different emotional states are distinguished (i.e. Neutral, Wondering, and Happy).

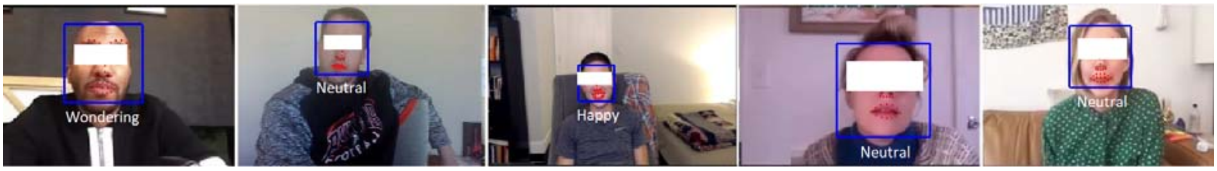

Fig. 9 Inter-group analysis of some participants along with the estimate emotion for each of them.

Fig. 10 depicts the statistical results of the emotional states for two participants during Inter-Group analysis. As mentioned earlier, this type of analysis is also available for the Intra-Group analysis mode.

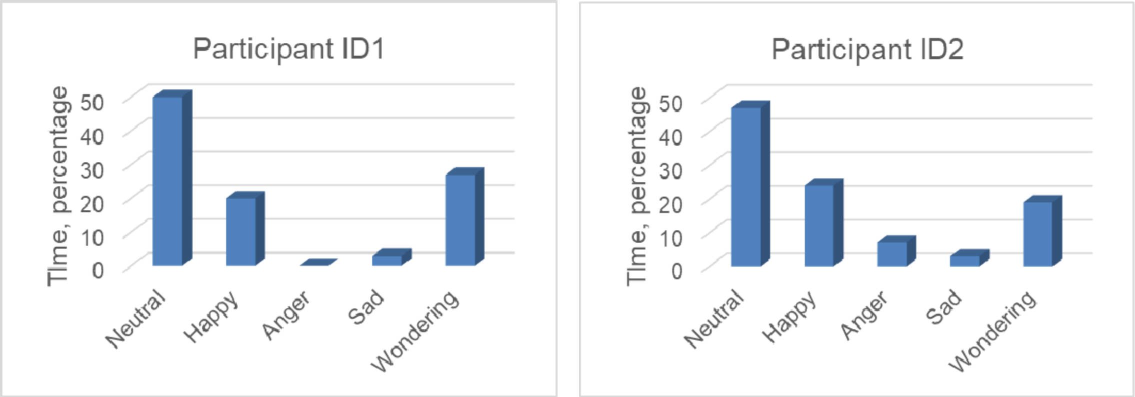

Fig. 10 Intra-group analysis: interaction profile for two participants with respect to the emotional state estimated during the meeting (also available in Inter-group analysis mode).

From Fig. 10, it can be concluded that in most of the time both of the participants' emotional states were estimated as a "Neutral". It should be note that the statistical results are determined only for frames, for which the faces were detected by the algorithm (67% of the meeting time for the Participant ID1 and 79% - for the Participant ID2).

3. Conclusions

In this paper we propose a conceptual architecture and an algorithm, applicable specifically for the purposes of innovation in business modeling. The proposed model includes: participants' extraction, background removing (or replacement), participants' behavior analysis (which we further classify as intra and inter-group analysis) and data recording. Behavior analysis includes: finding out the looking direction (by assessing the orientation of the head) and emotional state assessment (by estimating the facial expression). During the meeting, four possible looking targets are taken into account. These are: participant is looking at the Working Station; participant is looking at other participant; participant is looking at the screen or participant is looking at some other place (i.e. distraction state). We show the statistical information of the estimated profile for the particular participants in a real case scenario.

The proposed conceptual model architecture and an algorithm can be easily adapted by businesses by integrating other sensors (ex. stress wearables) and machine learning techniques to observe, analyse and predict human behaviour and facilitate de the digitalization of the BMI process. This environment should be able to facilitate businesses and organizations by "making" it possible for them to "see", "sense", understand and communicate their Business Models dimensions and components. The business models must "learn" as they are created to enhance the competitive-edge of the businesses.

Acknowledgments

This work was supported by financially by the project UNITe BG05M2OP001-1.001-0004/28.02.2018 (2018-2023).

References

[1] Alexiadis, D. S., Chatzitofis, A., Zioulis, N., Zoidi, O., Louizis, G., Zarpalas, D., & Daras, P. (2016). An integrated platform for live 3D human reconstruction and motion capturing. IEEE Transactions on Circuits and Systems for Video Technology, 27(4), 798-813.

[2] Afzal, H., Aouada, D., Font, D., Mirbach, B., & Ottersten, B. E. (2014, August). RGB-D Multi-view System Calibration for Full 3D Scene Reconstruction. In ICPR (pp. 2459-2464).

[3] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[4] Gassmann, O., Frankenberger, K., & Sauer, R. (2016). Leading business model research: the seven schools of thought. In Exploring the Field of Business Model Innovation (pp. 7-46). Palgrave Macmillan, Cham.

[5] Manolova, A., Neshov, N., Panev, S., & Tonchev, K. (2014, June). Facial expression classification using supervised descent method combined with PCA and SVM. In International Workshop on Biometric Authentication (pp. 165-175). Springer, Cham.

[6] Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., ... & Tankovich, V. (2016, October). Holoportation: Virtual 3d teleportation in real-time. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (pp. 741-754).

[7] Panev, S., & Manolova, A. (2015, September). Improved multi-camera 3D eye tracking for human-computer interface. In 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS) (Vol. 1, pp. 276-281). IEEE.

[8] Soleimani, V., Mirmehdi, M., Damen, D., Hannuna, S., & Camplani, M. (2016, October). 3d data acquisition and registration using two opposing kinects. In 2016 fourth international conference on 3D vision (3DV) (pp. 128-137). IEEE.

[9] Tonchev, K., Lindgren, P., Manolova, A., Neshov, N., & Poulkov, V. (2017, October). Digitizing human behavior in business model innovation. In 2017 Global Wireless Summit (GWS) (pp. 97-101). IEEE.

[10] Tonchev, K., Panev, S., Manolova, A., Neshov, N., Boumbarov, O., & Poulkov, V. (2015). Gaze Tracking, Facial Orientation Determination, Face and Emotion Recognition in 3D Space for Neurorehabilitation Applications. Neuro-Rehabilitation with Brain Interface￼, 40, 51.

[11] Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 532-539).