IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 26, 2021, accepted December 24, 2021, date of publication December 29, 2021, date of current version January 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3139078

Mining Inconsistent Emotion Recognition Results With the Multidimensional Model

AGNIESZKA LANDOWSKA, TERESA ZAWADZKA, AND MICHAŁ ZAWADZKI
Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 80-233 Gdańsk, Poland

Corresponding author: Teresa Zawadzka (tegra@eti.pg.edu.pl)

This work was supported in part by the Polish-Norwegian Financial Mechanism Small Grant Scheme under Contract Pol-Nor/209260/108/2015; and in part by the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Bioethics Committee at the District Medical Chamber in Gdańsk, "Biomeasurements of Emotional States Symptoms in Human–Computer Interaction."

ABSTRACT The paper deals with the challenge of inconsistency in multichannel emotion recognition. The focus of the paper is to explore factors that might influence the inconsistency. The paper reports an experiment that used multi-camera facial expression analysis with multiple recognition systems. The data were analyzed using a multidimensional approach and data mining techniques. The study allowed us to explore camera location, occlusions and algorithm factors in the late fusion of emotion recognition results. We proposed to use a multidimensional data model for mining the various interdependencies between the factors of inconsistency. The study allowed the exploration of challenges in multichannel emotion recognition. It was achieved by comparing the consistency of obtained emotions and identification of rules determining conditions when the obtained emotions are consistent. However, the main novelty of the paper is the method of mining the inconsistencies. The study might be interesting both for researchers dealing with integration in emotion recognition, as well as for practitioners who use automatic emotion analysis software and expect to get valid results.

INDEX TERMS Data mining, emotion recognition, facial expression analysis, inconsistency, late fusion, multidimensional model.

I. INTRODUCTION

Affective computing could potentially revolutionize the way we interact with technologies and even with other humans using media. Practical and research applications are taking place in teaching and learning, therapy, job interviews and marketing [26]. Nowadays, emotion recognition holds the promise of improving business as well as individual well-being. However, the challenges that remain in the discipline might jeopardize the prospects. In this paper, we deal particularly with uncertainty and inconsistency in automatic emotion recognition, which was reported among significant challenges before [21]. Emotion recognition is based on symptoms of emotions, that are observable or measurable. The symptoms might be captured simultaneously by several devices in a variety of channels, and then analysed by a number of algorithms. Multimodal and multichannel emotion recognition was reported to improve accuracy [5]. At the same time, multichannel observation introduces some confounding variables to the recognition processes, including inconsistencies among observed modalities and recognition algorithmic results. There are basically two generic approaches to multichannel integration, i.e. early and late fusion [11]. The first one integrates the data on symptoms forming a wide feature vector, before classification. The latter one processes (classifies) each channel and modality separately, integrating recognition results. The combination of the approaches called hybrid fusion might be considered as well. Whatever the integration paradigm, the inconsistencies might occur at the level of features or recognition results. Regarding the variety of application areas mentioned above, the inconsistencies shall be tracked and reported to mitigate the risk of misinterpretation of recognition results. In this paper, we are mining inconsistencies among multiple observation channels and multiple recognition solutions. An experiment with four parallel cameras and three competitive facial expression recognition systems was explored in detail using the multidimensional model, statistical and data mining analysis.

The associate editor coordinating the review of this manuscript and approving it for publication was Kah Phooi (Jasmine) Seng.

The research goal of this study is to identify inconsistency factors and to confirm the applicability of the multidimensional data model and mining techniques for that purpose. Some confounding factors have been identified and quantified using statistical and data mining methods upon the multidimensional model. The novelty of the approach is the integration of the data obtained within the emotion recognition experiment into a multidimensional model, and the application of statistical and data mining methods on this model (further called multidimensional analysis) to get more insight into confounding factors in the recognition task.

The hypothesis *It is possible to identify inconsistency factors using multidimensional analysis of facial expression data* is being proved within the paper.

The research methodology of this study consists of the following steps:

1) Identification of potential inconsistency factors and ways to explore them.
2) Designing and executing the experiment to obtain estimated emotional states (emotional states recognized by the specified algorithm from the specified channel) using the multichannel approach to explore inconsistency factors.
3) Establishing the process of calculating the potential dependencies between estimated emotional states and potential inconsistency factors.
4) Analysis of obtained results.

The hypothesis of the research is treated as proven if the utilization of the multidimensional model defined in the third step allows the identification of the inconsistency factors and interconnections between them.

The paper is organized as follows: Section II provides a review of related research identifying possible inconsistency factors in emotion recognition. It is followed by Sections III and IVwhere research design and execution are presented. In the Section III, both experiment and data processing design are described. Analogically, Section IV contains experiment execution and implementation aspects of the established process. Analysis of obtained results is described in Section V. Finally, the discussion and conclusion sections follow.

II. RELATED WORK

The previous research works that contribute to this study comprise studies on automatic emotion recognition including multimodal/multichannel recognition and factors that influence accuracy and consistency of classification results.

There are numerous emotion recognition algorithms that differ on input channels and modalities, output labels (or affect representation model), and classification method. The most frequently used emotion recognition methods include facial expression analysis [2], audio (voice) signal analysis in terms of prosody [12], physiological signal interpretation [1] and behavioural patterns analysis [16]. Among those, video input is the most commonly used channel for emotion recognition as it is a universal and not disturbing method of human monitoring. There are many algorithms, both developed within research studies or as commercial products, that differ significantly on the number of features and methods of data extraction, feature selection and classification process [2], [6], [10].

There are three major model types of emotional state representation: discrete, dimensional and componential [31]. Discrete models distinguish a set of basic emotions and describe each affective state as belonging to a certain emotion from the predefined set. One of the best known and extensively adapted discrete representation models is Ekman's six basic emotions model, which includes happiness, anger, disgust, surprise, sadness, and fear [9]. Dimensional models represent an emotional state as a point in a multi-dimensional space and the most adapted model uses continuous valence and arousal dimensions, which is also frequently extended with an additional dimension of dominance [32]. Componential models use several factors that constitute or influence the resulting emotional state and the mostly adopted OCC model defines a hierarchy of emotion types representing all possible states which might be experienced [34]. The analysis of the emotion recognition solutions reveals that there is no one commonly accepted standard model for emotion representation. Continuous adaptation of Ekman's six basic emotions and the valence-arousal models are those widely used in emotion recognition solutions [31]. Recently, more papers have been discussing the nature and modelling of emotions in affective computing [30], [39], but most of the research is still focused on dichotonomous models of emotions [35].

Classifiers are usually built on one of the known artificial intelligence tools and algorithms, including decision trees, neural networks, Bayesian networks, linear discriminate analysis, linear logistic regression, Support Vector Machine, Hidden Markov Models [27], and lately also convolutional networks [25]. Deep learning becomes a new trend in emotion recognition, especially applied to facial expression recognition [13], [28], [29]. Depending on the classification method, input channels and selected features, the accuracy of affect recognition differs significantly, sometimes achieving more than 90 percent, but mostly in a cross-validation scheme on a single dataset. The emotion classifiers differ in terms of accuracy of the obtained results, moreover, most of them are prone to diverse uncertainty factors, especially while not applied on a well-prepared dataset, but in so-called ''in-the-wild'' conditions [21]. ''In-the-wild'' studies try to build natural and not posed emotional expression datasets and explore confounding factors that influence emotion recognition accuracy [17], [18]. For example, emotion recognition from facial expressions is susceptible to illumination conditions, angle towards the camera, size of the face in the recording (details visibility), and occlusions of the face parts [22], [33].

Multimodal and/or multichannel observation are seen as solutions to availability and accuracy challenges in emotion recognition [11], [14]. Channel in this context is the type of signal recorded for analysis, f.e. video, while modality is the type of information processed to find emotion symptoms, f.e. facial expressions. Thus, multichannel observation is,

for example, using multiple cameras, while an example of multimodal observation would be using facial expressions combined with heart rate variability analysis. The main goal of the multimodal observation studies is to improve the recognition accuracy. A review of such studies [5] reported that for 86% of systems (26 out of 30) multimodal analysis improved accuracy with an average of 9.83% (median: 6.6%) gain compared to the best unimodal result. In this study, we have intentionally chosen to explore multiple channels instead of multiple modalities, as the purpose of the study was to propose a method to explore confounding factors rather than accuracy improvement.

There are early (feature level) fusion, late (decision level) fusion, and hybrid fusion approaches used to integrate multiple observations [11], [14], whereby each of these introduces some additional challenges. Poria *et al.* [27] reports the challenge of time synchronisation between the observation channels. Another study reports how to handle missing data [36]. In late and hybrid fusion the challenge is that the individual classifiers might report non-consistent (or even contradictory) results. Poria *et al.* [27] reports three types of late fusion mechanisms: (1) rule-based methods that combine information using statistical rules such as linear combination, majority voting, or others, (2) classification-based methods, for example, using decision trees [37], and (3) estimation-based methods, for example, using Kalman filtering [23], [27]. The inconsistency of the results in most of the studies was treated as a problem to solve and all papers proposed a solution of some kind. None of the papers explored the confounding factors in order to find the reasons or the rules behind the inconsistency.

III. RESEARCH DESIGN

The research methods applied in this study comprise a quasi-experiment on emotion recognition using a multichannel approach and a multidimensional analysis of the inconsistency factors. The design of both methods is provided in Section III-A and III-B respectively.

A. EXPERIMENT DESIGN

This section describes the design of a quasi-experiment that was held at the Gdansk University of Technology, at an Emotion Monitor Stand [19]. It was a quasi-experiment, as it was conducted with human participants, and not all confounding variables could be potentially controlled within such conditions, however, we will refer to it as an experiment in the paper.

In the experiment, we chose to address a single modality (facial expressions), but captured simultaneously with four observation channels (video). The reason behind such a decision was a result of the previous experiments, in which the availability of a single channel observation turned out to be time-dependent, participant-dependent and even task-dependent with single camera use [20].

Discrepancies between modalities (eg. emotional expressions derived from facial expressions vs. body postures), although important and worth exploring, were not the subject of this study. We focused on analysing a single modality (facial expressions), captured with four observation channels (cameras). Another decision was to capture natural - not posed - expressions, during normal tasks in human-computer interaction.

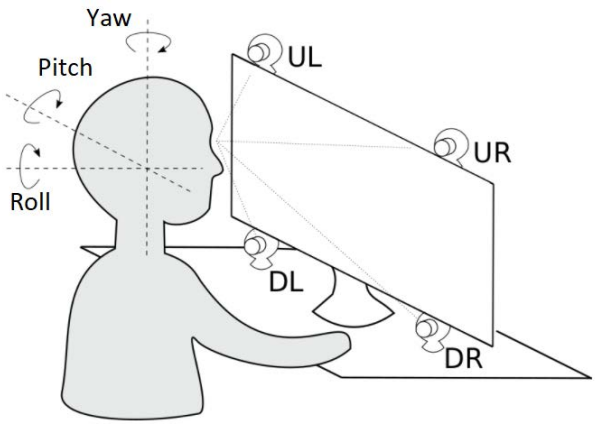

FIGURE 1. Experimental setup for multichannel facial expression recognition with location of the cameras (UL - upper left, UR - upper right, DL - down left, DR - down right).

Accuracy of emotion recognition from facial expressions depends on multiple contextual factors and among those is the angle of face plane towards a camera. Off-the-shelf software is claimed by the producers to work best while a user's face is exactly in front of the camera, although they also claim, that software still works while the angle is up to 40 degrees. In practical computer-based settings, when a user is expected to have a screen in front, it is hard to locate a camera directly in the middle of it. Screen sideways locations are used instead and the angle is not optimal (although within application range) for facial expression analysis purposes. In the experiment on educational activities, reported here, there were four cameras used, that were located on screen edges - above and below the monitor, left and right side, symmetrically. The cameras were fully synchronized. The setting is depicted in Figure 1.

The experiment was held at the Emotion Monitor Stand, which is a usability and human-computer interaction monitoring stand equipped with devices and software solutions enabling capturing additional observation channels for emotion recognition [19]. The stand is located in a closed room and experimenters/observers are separated from the participant by a furniture wall. The professional photographic lamp was used for optimal face illumination.

In the experiment, educational activities were chosen to be performed by the participants, as we wanted to invite students as participants and that was a natural activity for them. The students followed a linear scenario, starting with initial metric questions, logging into a learning platform (task 1), finding a course (task 2), filling in an emotional state questionnaire, listening to a lecture (task 3), filling in an emotional state questionnaire, adding a forum entry (task 4), solving a quiz on the topic covered by the lecture part (task 5), filling in an

emotional state questionnaire and user experience questionnaire, evaluating subjective camera disturbance.

The emotional state questionnaire was based on the Self-Assessment Manikin (SAM) scheme [3]. It consisted of three questions, each one covering one of the PAD model dimensions of emotional states, namely P-pleasantness (valence), A- arousal, and D-dominance [24]. The answers were provided using a single question: ''Please rate your emotional state'' and three 7-point Likert scales, from 1 to 7 (with 4 considered as a neutral condition). SAM questionnaire is reported to be a proper tool for capturing momentary emotional states, and its additional feature is a schematic visualisation of the emotional states assigned to labels in the scale, which helps the participants in state evaluation.

Subjective camera disturbance was rated on a 5-point scale (1-not disturbing at all, 5-very disturbing). Conducting the experiment and capturing the activities as well as questionnaires was supported with a Morae Recording Tool. The four cameras were synchronized using an iSpy software solution. The video recordings were then independently analyzed with three emotion recognition solutions: Noldus Face Reader, QuantumLab Express Engine and Luxand-based. The first two of them are off-the-shelf solutions for analysing facial expressions, while Luxand is a library to use in your own application (based on an open licence). They allow obtaining emotional state estimates based on facial expressions using the following models. Noldus Face Reader provides six emotions: happiness, anger, sadness, disgust, fear, surprise (called Ekman's basic emotions) and neutral state, followed by a mapping of those into the two-dimensional model of emotions. This is a PA model with pleasantness and arousal dimensions, where pleasantness is called valence. QuantumLab Express Engine software provides five out of six basic emotions (without fear) and a neutral state, while Luxand library provides six basic emotions and a neutral state. All systems recognize emotional states in the continuous model [7], [8] and the ranges of values that can be taken by the variables representing emotional states are depicted in Table 1 within the the category of the dependent variable. The facial expression solutions provide additional information, apart from the estimate of an emotional state:

- Noldus Face Reader: sex, age, presence of occlusions (beard, moustache and glasses), head orientation in three dimensions, quality of the video, facial actions, location of face parts;
- QuantumLab Express Engine: 3D location of face (three variables), 3D rotation of head (three variables);
- Luxand: the single value of the angle.

For each recording of each camera location, the following emotional state estimates are provided: six Ekman's emotions, neutral state as well as valence and arousal dimensions values. Please note that in a specific moment of time multiple, Ekman's emotions can be detected. The number of time series obtained via experiment for one participant for the same period of time - reflecting the time of completing the tasks - are depicted in Figure 2.

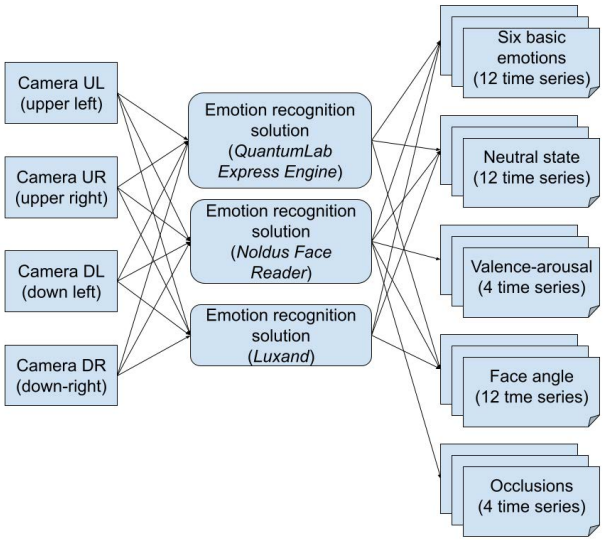

FIGURE 2. Multichannel data obtained from the experiment.[2]
1. For fear only 8 time series are obtained as this emotional state is not recognized by QuantumLab Express Engine.

Summing up the experiment design, the definition of the variables that were addressed by the experiment and are explored further are depicted in Table 1.

TABLE 1. Variables definition.

type of variables	description of variables	name of variable	possible values
independent variable	location of the camera	location	UL (upper left), UR (upper right), DL (down left), DR (down right)
dependent variables	estimated emotional state in Ekman model	happiness	<0,1>
		anger	<0,1>
		sadness	<0,1>
		disgust	<0,1>
		fear	<0,1>
		surprise	<0,1>
	estimated neutral emotional state	neutral	<0,1>
	estimated emotional state in two-dimensional model	valence	<-1,1>
		arousal	<0,1>
confounding variables [3]	automatic facial expression analysis solution	system	FR (Noldus Face Reader), XP (QuantumLab Express Engine), LU (Luxand)
	sex	sex	Male, Female
	presence of occlusion	beard	None/Some/Heavy
		moustache	None/Some/Heavy
		glasses	Yes/No
	head orientation	yaw	<-90,90>
		pitch	<-90,90>
		roll	<-90,90>

B. DATA PROCESSING AND ANALYSIS

Having multiple dependent and confounding variables, we dealt with some challenges in data analysis. Having multiple cameras and then multiple facial expression

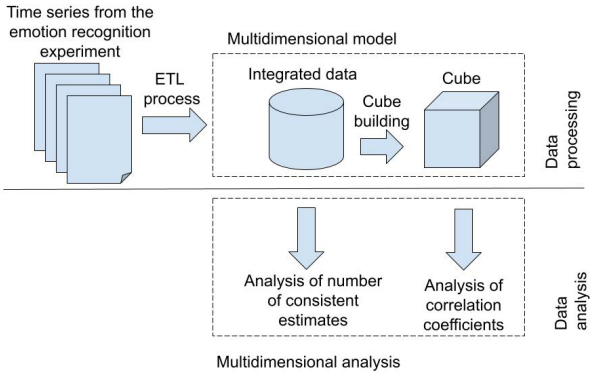

FIGURE 3. Design view.

recognition systems, we found significant inconsistencies between the obtained results. Therefore, we proposed to use the data mining approach. In the paper, we treat each output from the facial expression analysis software as an estimated emotional state (and perhaps not the true one), whereby we refer to it as estimated emotion or simply estimate. By inconsistency, we mean the situation when the estimates obtained for the same person and point in time, differ. Having defined the camera location as the independent variable, we found that it explains only part of the inconsistencies. Traditional statistical approaches like ANOVA analysis only partially allowed the finding of the causes of inconsistencies in particular cases, therefore we decided to introduce the analysis using a multidimensional model. The process of calculating the potential dependencies between estimated emotional states and potential inconsistency factors is depicted in Figure 3. Within this process, the analysis of estimated emotional states is done only with respect to the inconsistency factors, which are not recognized or inferred. During the emotion recognition experiment, the set of data containing estimated emotional states is generated. These data are depicted in Figure 3 as *Time series from the emotion recognition experiment*.

Due to the fact that these data are generated from various recognition systems, the data and their formats vary between datasets. The main differences include various estimated emotions, various models of emotions (Ekman or two-dimensional) or various sampling times. According to the Kimball architecture [15] for building data warehouses - understood as an implementation of the multidimensional model - the process of *ETL* (Extract, Transform, Load) must be executed to integrate data. However, data integration is a wider need, necessary to be done for almost every analytical task extracting data from various sources. Depicted in Figure 3 *Integrated data* are the result of this process and are stored in a relational star schema (the first layer of the dimensional model) being a relational representation of the multidimensional model. Having the data integrated a *cube building* process is done. A *Cube* is a set of facts understood as a single event described with numeric and descriptive values (the second layer of the multidimensional model). The numeric values are called measures and descriptive values are called dimensions. Let's notice that the discretized values of measures are also called dimensions. The measures can be aggregated with aggregation functions (e.g. sum or max) and further analysed versus various dimensions. It is connected with the fact that cubes are strongly focused on cube slicing - choosing only the subset of facts and calculating aggregations for this subset. Depending on the needs, sometimes it is easier to analyse data stored in the star schema and sometimes stored in the cube. In the presented paper, the two analyses were performed:

1) ***Analysis of a number of consistent estimates*** with respect to various inconsistency factors, for dominant emotions (emotion with the highest estimated value) and four quadrants of the valence-arousal (VA) space, namely HVHA, HVLA, LVHA and LVLA (H, L, A and V stand for high, low, arousal and valence respectively). In the Ekman model, the two estimates are treated as consistent if the dominant emotion is the same. In a two-dimensional model, the two estimates are treated as consistent if they belong to the same VA quadrant. This analysis is done for the relational data - *Integrated data* using SQL language.
2) ***Analysis of correlation coefficients*** with respect to various inconsistency factors. This analysis is done for the data stored in the cube using MDX language and decision trees. It is done for the whole set of estimates as well as for the subset of them. The subsets are created by categorizing values of emotion estimates.

1) ETL PROCESS DESIGN

When designing the ETL process the following tasks were identified:

1) Choosing the needed data and format unification - according to the experiment assumptions retrieving automatic facial expression analysis solution, location of the camera, type of the occlusion, participant, emotional states.
2) Unifying sampling timestamps - all time series for emotion estimates should be sampled with the same time rate.
3) Calculating estimated emotion states for unified timestamps.
4) Discretization of estimates - calculating VA quadrants and categorizing basic emotions estimates for those less than 0.2 and greater than or equal 0.2. The 0.2 threshold was chosen as that which indicates that some elicitation of the estimated emotion is detected.
5) Determining a dominant emotion.

As a result of the ETL process, the two files are to be prepared. The first file describing the participants contains information about their sex, moustache, beard and glasses - the ground truth about them, not the one inferred by the systems. The second file contains information about all samples i.e. the participant the sample concerns, sample timestamp, values of estimates, system and location variables, ranges of

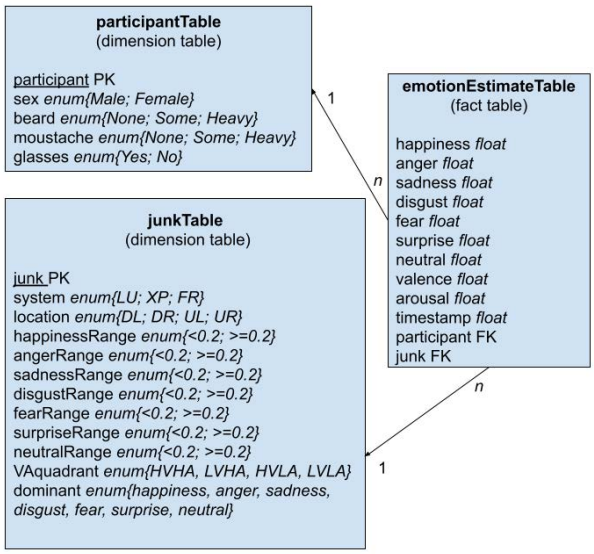

FIGURE 4. Integrated data in the relational model (star schema).

basic emotion estimates and neutral state (<0.2 and >=0.2), VA quadrant and the name of the dominant emotion.

2) MULTIDIMENSIONAL MODEL DESIGN

The multidimensional model, as it was mentioned previously, consists of two layers. The first step of building a multidimensional model is to define a fact. From the application perspective, the atomic fact is the fact describing the event with the lowest grain [15]. When analysing estimated emotional states - after the ETL process - the atomic fact is the single estimation obtained from the specified camera location, produced by the specified system, concerning the specified participant and estimated at the specified time. The applied fact definition (the lowest grain) allows for the widest spectrum of possible analysis. This is because there are no measures defined for the higher grain, which cannot be calculated on the lowest grain.

The relational model - a star schema, consists of two types of tables: the fact table (*emotionEstimateTable*) and the two-dimension tables (*participantTable* and *junkTable*), which are depicted in Figure 4. The fact table stores measures - estimate values, sample timestamp and foreign keys to the dimension tables. For the purpose of efficiency, all data cannot be kept in the same fact table - only numeric values are stored there. The exception is timestamp, as the levels of time hierarchies are not needed in the further analysis [4]. The two dimension tables storing descriptive values, are defined. The one - *participantTable* contains ground truth about the sex and occlusions for participants and the other one - *junkTable* contains the information about system and location variables, ranges of emotion estimates, VA quadrant and the name of the dominant emotion.

The measures in multidimensional models are understood in two ways. Firstly, these are numeric values associated with the fact. And in that way, they are understood in the relational model. Secondly, these are pairs consisting of these numeric values and aggregation functions associated with them. In that way, they are understood in the cube. Moreover, it is possible to create more than one pair from one numeric value associating two different aggregation functions e.g. sum or max. In the presented model, for each estimate, two pairs are created (the word *Estimate* in the equations is used to denote the specific estimated emotional state):

$$EstimateSum = sum(estimates) \tag{1}$$

and

$$EstimateCount = count\ non\ empty(estimates) \tag{2}$$

These measures are defined to count the average value of each estimate according to the equation

$$EstimateAvg = estimateSum/estimateCount \tag{3}$$

The example is:

$$\begin{aligned} HappinessSum &= sum(happiness) \\ HappinessCount &= count\ non\ empty(happiness) \\ HappinessAvg &= HappinessSum/HappinessCount \end{aligned} \tag{4}$$

However, it must be noticed that the *EstimateSum* measures are non-additive. This means that the values of these measures can not be aggregated via any dimensions, because it has no analytical sense. In addition to numerical values, each fact is also described by dimensions. For the need of this experiment, the dimensions correspond to the dimension tables.

3) MULTIDIMENSIONAL ANALYSIS DESIGN

The aim of the multidimensional analysis is to identify inconsistency factors. This analysis is done twofold.

a: ANALYSIS OF THE NUMBER OF CONSISTENT ESTIMATES

The analysis of the number of consistent estimates is done separately for the two-dimensional model and Ekman model with a neutral state. In the Ekman model, the two estimates are treated as consistent, if for the same participant, for the same moment of time the dominant emotion is the same. In the two-dimensional model, the two estimates are treated as consistent, if for the same participant, for the same moment of time the VA quadrant is the same. The calculation is always done as a percentage value representing the percentage number of consistent estimates vs. a number of all estimates. The set of estimates taken to the analysis can vary. For the two models the following calculations are done:

1) What is the number of consistent estimates for various camera locations for each system?
2) What is the number of consistent estimates for various camera locations for various systems for the participants categorized versus beard, glasses, moustache and sex?
3) What is the number of consistent estimates in the Ekman model for various systems for each camera?

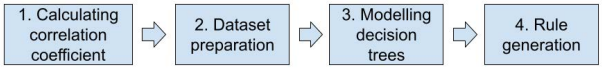

FIGURE 5. Workflow for analysis of correlation coefficients.

4) What is the number of consistent estimates in the Ekman model for various systems for various camera locations for the participants categorized versus beard, glasses, moustache and sex?

For the first two questions for a two-dimensional model, only the Noldus Face Reader system is analysed.

b: ANALYSIS OF CORRELATION COEFFICIENTS WITH RESPECT TO VARIOUS INCONSISTENCY FACTORS

The aim of this analysis is to find the response to the following questions:

1) Which inconsistency factors, and in what way, influence the consistency of the recognized emotions from various camera locations?
2) Which inconsistency factors, and in what way, influence the consistency of the recognized emotions from various systems?

The process of these analyses is depicted in Figure 5. The influence of the inconsistency factors on the estimated emotional states is analysed by *calculating correlation coefficients* (step one in Figure 5) for:

- Question 1: all combinations of pairs of camera location elements, for each emotion estimate in both models vs. system, participant, beard, sex, moustache, glasses, and additionally, emotion range for the Ekman model and neutral state and VA quadrant for the two-dimensional model. For happy emotion as an example, the *happyRange* dimension is taken into account. For valence and arousal in a two-dimensional model *VAquadrant* is taken.
- Question 2: all combinations of pairs of system elements, for each emotion in the Ekman model vs. location, participant, beard, sex, moustache, glasses and analogically as for Question 1 emotion range.

These calculated correlation coefficients must be further processed to prepare datasets: two for the first question and one for the second question (step two in Figure 5). For the second question, only one dataset is prepared, as valence and arousal are recognized by Face Reader only. In each data set, additionally to calculated correlation coefficients, dimensions describing them and emotion estimates for which they are calculated, the categorized value of correlation coefficient must be determined. This categorization is done twice:

$$\begin{aligned}(-1) &\rightarrow \textit{perfect negative},\\ (-1, -0.7>) &\rightarrow \textit{strong negative},\\ (-0.7, -0.5>) &\rightarrow \textit{moderate negative},\\ (-0.5, -0.3>) &\rightarrow \textit{weak negative},\\ (-0.3, 0, 3) &\rightarrow \textit{no},\\ <0.3, 0.5) &\rightarrow \textit{weak positive},\\ <0.5, 0.7) &\rightarrow \textit{moderate positive},\\ <0.7, 1) &\rightarrow \textit{strong positive and}\\ (1) &\rightarrow \textit{perfect positive}. \end{aligned} \tag{5}$$

and

$$\begin{aligned}<-1, -0.3> &\rightarrow \textit{yes negative},\\ (-0.3, 0.3) &\rightarrow \textit{no},\\ <0.3, 1> &\rightarrow \textit{yes positive}. \end{aligned} \tag{6}$$

In the third step for both data sets, for the two-class columns (two categorizations of correlation coefficients defined by equations (5) and (6)) *decision trees are modelled*. From such prepared models in step four, the *rules are generated* and evaluated.

IV. STUDY EXECUTION

This section presents the study execution and is organized analogically as the previous one. Experiment execution is described in Section IV-A and implementation aspects of the data processing and analysis are presented in Section IV-B.

A. EXPERIMENT EXECUTION

Ten people took part in the study and the intensity of recognized emotional states was estimated over time.

In the experiment, ten participants took part (four women and six men). The appearance characteristics of participants are depicted in Table 2, focusing on the features corresponding to identified occlusions. The age of the participants varied from 23-28 and the average age was equal to 24. The group of students only encompassed those who were a year or two before the end of their Master Studies of Computer Science or Biomedical Engineering.

TABLE 2. Participants' characteristics.

participant	sex	beard	moustache	glasses
P01	Female	None	None	No
P02	Male	Heavy	Some	No
P03	Male	Some	Some	Yes
P04	Male	None	None	Yes
P05	Female	None	None	No
P06	Female	None	None	No
P07	Female	None	None	No
P08	Male	Some	Some	No
P09	Male	Heavy	Some	Yes
P10	Male	None	None	Yes

The tasks were performed on the Moodle e-learning platform, thus during the initial questions, the overall experience of the participants with this platform was estimated. Participants had used this platform for about 4/5 years, most of them several times a month mainly to check course scores, view teaching materials or check pass rules. Most of them had finished 6-10 courses on the Moodle platform and during the experiment were enrolled to between 1-5 courses. Moreover, only three of them had experience with other e-learning platforms.

The emotional state questionnaires were filled by the participants three times, after the second, third and fifth task.

According to the assumptions, the answers were given in the 7-point Likert scale. The answers were presented in Table 3. For most participants, the pleasure dimension value did not change during the experiment. There were two participants (P05 and P06) whose pleasure value changed over time. For the P05 participant, it decreases after the third task and again increases after the fifth one. Also, participant's dominance level increases at the end of the experiment. For the P06 participant valence decreases along with the execution of subsequent tasks at the same time her arousal increases. The change in arousal is also noticed for P02, P09 and P10 participants. For the P02 participant it is the highest after the third task, but the dominance decreases during the whole experiment. For P09 and P10 participants, dominance is the highest at the end of the experiment. The P09 participant has no dominance after the third task.

In Table 3 the camera disturbance given by the participants is also depicted. Half of the participants reported a disturbance, but only for three of them, it was greater than 2.5 on the 1-5 scale where 1 means no disturbance.

TABLE 3. Questionnaires results.

participant	1st QA			2nd QA			3rd QA			camera disturbance
	P	A	D	P	A	D	P	A	D	
P01	5	4	4	5	5	5	5	4	4	1
P02	6	2	7	5	5	6	5	3	5	3
P03	6	4	5	6	5	5	6	5	5	2
P04	6	4	4	6	4	4	6	5	5	1
P05	5	5	2	2	4	2	6	4	5	4
P06	5	2	3	3	5	3	2	6	3	3
P07	4	3	4	3	3	3	4	3	4	1
P08	7	2	4	7	2	4	7	2	4	1
P09	4	3	4	4	1	2	5	5	5	1
P10	4	2	4	3	2	3	3	4	3	2

Within the experiment, participant's faces were recorded with the use of four identical cameras - Logitech Creative, located according to the experiment assumption in the upper left, upper right, down left or down right corner of the screen, as presented in Figure 1. Figure 6 presents the exemplary frame for the P06 participant in the 12th minute, 15th second and 1st millisecond of the experiment.

The movies were recorded as.mp4 files with AVC1 codec and automatically split by iSpy into several files lasting for about 15 minutes (excluding the last file of the recording). Duration times of recordings for each participant for each camera are presented in Table 4. The maximum difference between the duration is 614 milliseconds for the P07 participant.

For each split file, video recordings were analyzed with Noldus Face Reader, QuantumLab Express Engine and Luxand. The video recordings were initially analysed by *ffprobe* tool, to identify a number of frames for each movie and list the beginning time for each one.

The *ffprobe* tool shows the number of frames and lists time frames. What is strange is the number of listed time frames greater by 1 from the returned number of frames.

FIGURE 6. Exemplary time moment recorded by four cameras located in UL - upper left, UR - upper right, DL - down left, DR - down right.

TABLE 4. Analysis of video duration in milliseconds.

	DL	DR	UL	UR	maximum difference between durations
P01	2361895	2362413	2361856	2361869	***557***
P02	2285988	2285870	2285894	2285942	***118***
P03	2510861	2510878	2510821	2510848	***57***
P04	2504420	2504361	2504381	2504397	***59***
P05	2232157	2231994	2232077	2231971	***186***
P06	2125821	2125320	2125842	2125370	***522***
P07	943266	943305	943880	943289	***614***
P08	2959959	2959939	2959723	2959950	***236***
P09	2705855	2705775	2705884	2705859	***109***
P10	2255755	2255626	2255643	2255625	***130***

Express Engine and Luxand are recognizing emotional states in the Ekman model with a neutral state for each frame. The number of frames recognized by Express Engine is compatible with a number of frames given by *ffprobe* tool, whereas the number of frames recognized by Luxand is compatible with the number of listed frames. Thus, the number of recognized frames differs between Express Engine and Luxand by 1 (in Luxand there is always one additional frame). Noldus Face Reader is also recognizing emotional state in a two-dimensional model. Other than for Express Engine and Luxand, the recognized emotional states generated from Face Reader were marked with timestamps - not frame number. The generated emotional state time series had significant differences in recording duration and time series length. The average difference was 486 milliseconds while the maximum was 1426 millisecond. Consultation with Noldus provides the information that the AVC1 codec is not supported by Face Reader. Thus, the recordings files were converted by *ffprobe* tool to its default codec, which is presented in Listing 1.

```
ffmpeg -i <input_file> -codec:a copy -preset medium <output_file>
```

Listing 1. Video conversion.

After this conversion, the average difference between the length of generated time series and recordings was 39 milliseconds and the maximum was equal to 94 milliseconds.

The analysis is based on the comparison of the estimated emotional state derived for the specified camera and system. Thus, for a specified moment in time, we have 12 estimates of emotional state - four cameras and three systems - and missing data is the percentage of recording time without missing values in the set of estimates. It is worth noticing that tools define missing value as the lack of data for all estimates. There is no situation that only some estimates are provided. However, we must remember that Express Engine does not recognize the emotion of fear - this value is missed for all time series generated by it. Express Engine and Face Reader report missing values as NaN in contrast to Luxand, which reports them with all emotional states equal to 0. In Table 5, the percentage of recording time without missing values at the same time for each system is presented, separately for each participant and camera location. In Table 6, the percentage of recording time without missing values at the same time for each camera location is presented separately for each participant and system. To calculate these values it is assumed that each provided estimate lasts from the detection timestamp to the next timestamp in the time series.

TABLE 5. The percentage of recording time without missing values at the same time for all systems.

Participant	DL [%]	DR [%]	UL [%]	UR[%]
P01	5.89	4.46	72.17	83.19
P02	80.08	83.56	60.56	82.99
P03	48.52	63.83	99.90	99.81
P04	86.97	97.53	97.82	97.98
P05	37.89	46.22	91.68	94.59
P06	28.16	27.69	58.41	64.55
P07	94.07	96.70	98.60	98.88
P08	60.84	47.16	53.07	50.76
P09	66.01	65.12	30.05	32.69
P10	45.44	42.22	94.43	79.06
average for all participants	55.39	57.45	75.67	78.45
average for all female participants	41.50	43.77	80.22	85.30
average for all male participants	64.64	66.57	72.64	73.88
average for all participants with glasses	39.99	42.79	56.10	52.89
average for all participants without glasses	51.16	50.97	72.42	79.16
average for all participants with heavy beard	73.05	74.4	45.31	57.84
average for all participants with some beard	54.68	55.50	76.49	75.29
average for all participants without beard	49.74	52.47	85.52	86.38
average for all participants with some moustache	63.86	64.92	60.90	66.56
average for all participants without moustache	49.74	52.47	85.52	86.38

The basic analysis of missing values presented in Tables 5 and 6 shows that the major influence on the

TABLE 6. The percentage of recording time without missing values at the same time for all camera locations.

Participant	FR [%]	LU [%]	XP [%]
P01	45.64	8.15	4.64
P02	96.93	73.63	62.25
P03	97.79	42.43	72.93
P04	98.75	87.32	94.45
P05	86.76	67.98	33.22
P06	56.00	23.99	41.72
P07	99.97	95.42	94.00
P08	37.27	55.53	36.08
P09	57.19	40.76	31.39
P10	90.93	23.43	99.94
average for all participants	76.72	51.86	57.06
average for all female participants	72.09	48.89	43.40
average for all male participants	79.81	53.85	66.17
average for all participants with glasses	61.48	26.66	51.07
average for all participants without glasses	70.43	54.12	45.32
average for all participants with heavy beard	77.06	57.20	46.82
average for all participant with some beard	67.53	48.98	54.51
average for all participants without beard	79.68	51.05	61.33
average for all participants with some moustache	72.30	53.09	50.66
average for all participants without moustache	79.68	51.05	61.33

percentage of time for which all analysed emotional estimates are available has the system used to recognize emotional states. The percentage of recording time without missing values at the same time for each camera location is the highest for Face Reader where Express Engine and Luxand have a similar percentage of them. Moreover, in general, the percentage of recording time without missing values at the same time for each system is higher for upper cameras than those located at the bottom of the screen. It seems that glasses also have some influence on the percentage time without missing values at the same time, both for each system and each camera location. It seems there is no correlation of this percentage with sex or other occlusions.

B. EXECUTION OF DATA PROCESSING AND ANALYSIS

In this Section the data processing implementation aspects are described. Although they follow design assumptions presented in Section III-B, we found it valuable to add implementation details of the multidimensional model-based analysis. The data processing consists of: execution of the data extraction (ETL) process, construction of the multidimensional model, and implementation of the inconsistency analysis - they are described in subsections that follow.

1) EXECUTION OF ETL PROCESS

The first step of the ETL Process was to unify the file formats i.e. standardize the column set and column names, as well as matching frames with timestamps. Due to the

fact that the video files were split into 2 to 4 parts, the files with recognized emotional states were split as well. Therefore, the next step was to combine them for each participant, with camera location and system separately (120 files identically formatted were obtained). As stated before, each of Luxand emotional state time series - for each part of the recording - contained one additional estimate for excess frame vs. Express Engine. The Express Engine was taken as a reference for the timeline as the number of frames and the length of the emotional state time series match exactly the recording duration as reported by *ffprobe*, which is treated as an industry standard. The command line used to gather the statistics is presented in Listing 2.

```
ffprobe -v error -select_streams v
-show_entries stream=nb_frames,duration:frame=best_effort_timestamp_time
```

Listing 2. Frames statistics.

The additional estimates for Luxand were removed. A similar approach was used for Face Reader to comply with the referenced Express Engine timeline. However, this time exceeding estimates were removed (or the last estimate was prolonged), to match the reference time. All time series obtained in this way were then enhanced with timestamps corresponding to starting time for frames obtained from *ffprobe* - except Face Reader which already contained the timestamps. To avoid accumulating potential time misalignment during merging various parts of the recording, the timestamps in the next part were increased by the duration of previous parts of recordings (and not calculated from the emotional states time series). It means that the first timestamp in which the emotional state was recognized is equal to the length of the previous recordings (or 0 for the first recording) and not equal to the length of the generated time series from previous recordings.

Knowing that the maximum difference between the length of the time series generated by Face Reader and recording duration is 94ms and any of exceeded frames for Luxand was not longer than this time we can estimate the maximal potential misalignment to 94 ms.

To compensate the variable framerate used in the source recordings (which resulted in different frame lengths) in generated time series and to avoid data loss due to re-sampling - the aim was to unify the estimate length for all 120 files - the Nyquist theorem was applied. Due to the fact the shortest estimate length was 4 ms, the constant new estimate length during resampling was set to 2 ms. The estimate values between two consequent original timestamps were equal to the estimated value in an earlier timestamp. It meets the assumption that the emotional estimate is valid for the whole frame time.

A short analysis of missing values presented in IV-A shows that the percentage of recording time without missing values at the same time for each system/camera location is generally not high enough to take into consideration only estimates with no missing value from any camera or system at the same time. Therefore, the replacement operation for missing values was applied. Because micro-expressions are difficult to interpolate, the nearest valid value - in terms of timeline - was used.

In order to analyse micro-expressions, it is required to have the estimate representing what is happening in the specified period of time (for micro-expressions the mean duration of this period is equal to 313,91 ms with a standard deviation of 85,81 ms [38]). It is not possible to split the estimated emotion time series into the windows, which one starts after the previous one. In such a situation we would not be able to detect some micro-emotions, as they could start within one window and end within the other one. To solve this problem the rolling window approach is applied. The size of the rolling window corresponds to the duration of the analysed period. In our case, this is equal to 314ms (value rounded to the whole millisecond). The resulting value is equal to an average of values inside the window - half of the window size before and after the timestamp. For boundaries, the window size is adjusted to cover existing data only.

For designed analysis, the length of twelve generated emotional state time series for each participant should be equal, thus obtained time series are truncated to the shortest one.

Furthermore, the discretization process takes place. For each basic and neutral emotional states, an additional column is created that contains discretized values according to the experiment design assumptions: less than 0.2 or greater than or equal to 0.2. Analogically, the VA quadrant is stored in the next column. Additionally, the dominant emotion is calculated and stored.

Finally, the resulting 120 files were merged together into a single file, obviously retaining the originating information (camera location, system, and participant). The data from that file and the data about the participants (depicted in Table 2) were the inputs used to populate the first layer of the multidimensional model i.e. *Integrated data*.

2) IMPLEMENTATION OF MULTIDIMENSIONAL MODEL

Implementation of the multidimensional model was done in MS SQL Server 2017. The relational layer was implemented in a relational database and the cube in Analysis Services. The chosen storage mode was MOLAP (Multidimensional Online Analytical Processing). This storage mode allows storing all data in an OLAP server in a column-oriented database meaning that the query response time is the shortest one. Measures *estimateSum* defined in Equation (1) and *estimateCount* defined in Equation (2) were implemented as a standard measures. The *estimateAvg* measures defined in Equation (3) were defined as calculations expressed in MDX (Multidimensional Expressions) language. It is worth noticing that the dimensions, dimension attributes and measures are named with a capital letter in the Cube to distinguish multidimensional notions from relational ones.

3) IMPLEMENTATION OF MULTIDIMENSIONAL ANALYSIS

For the first analysis i.e. **analysis of a number of consistent estimates**, all of the answers for the four defined analytical

questions are obtained similarly, according to the following steps:

1) Only needed samples are filtered.
2) Number of consistent estimates is determined.
3) Percentage of consistent estimates is calculated.

In the first step, depending on the type of the analysis (for various camera locations or for various systems), the temporary tables are created, containing only filtered data (for the particular camera or system). This step is done using the simple SQL query which creates tables with an analogical structure such as the *emotionsEstimateTable*. In Listing 3 the exemplary query is depicted. In that example the *FREmotionsEstimateTable* is populated with tuples obtained from the Noldus Face Reader.

```
INSERT INTO FREmotionsEstimateTable
select * from emotionsEstimateTable where System='FR';
```

Listing 3. Inserting filtered data.

In the second step, the number of consistent estimates between the two cameras (*camera1* and *camera2*) or systems (*system1* and *system2*) is calculated. To do that, again, the temporary tables are created - Listing 4 presents an exemplary table creation clause for *camera1* and *camera2* for Noldus Face Reader.

```
create table FRNoConsistentEstimatesTable(
        participant varchar(3),
        camera1 varchar(2),
        camera2 varchar(2),
        numberOfTuples integer
        )
```

Listing 4. Table creation storing number of consistent estimates.

In Listing 5, the exemplary query, which calculates the number of consistent estimates for Noldus Face Reader for the two-dimensional model, is presented. The insert clause calculates the number of consistent estimates for all combinations of camera locations for a given participant. The consistency is determined here with respect to the two-dimensional model so the consistency of the *VAquadrant* is checked. The delete clause is then used to remove duplication of camera locations (e.g. DL:DR is the same as DR:DL).

In the third final step, the percentage of the consistent emotional estimates is calculated. It can be either calculated for all filtered estimates obtained from the table achieved in the first point or calculated separately with respect to the values of the specified confounding variable. In Listing 6 the second option is used to determine the percentage of consistent estimates for Noldus Face Reader in a two-dimensional model with respect to *the beard* variable. In that clause, the fraction of consistent emotional estimates in all emotional estimates is calculated and presented as a percentage. The *counterTable* used in the statement consists of the whole number of estimates for each participant.

The second analysis i.e. **analysis of correlation coefficients with respect to various inconsistency factors** is done according to the assumptions depicted in Figure 5. Firstly within the first step, the *correlation coefficients are calculated*. The MDX queries are issued to the Cube. The exemplary query for Question 1, where the correlation coefficient

```
insert into FRNoConsistentEstimatesTable
select participant, camera1, camera2, count(*) as numberOfTuples from
    (select participant, timestapmp1, camera1, camera2 from
        (select e1.timestamp As timestamp1, e1.location As camera1,
                e1.participant AS participant, e1.VAquadrant As range,
                e2.timestamp As timestamp2, e2.location As camera2  from
                    FREmotionsEstimateTable e1, FREmotionsEstimateTable e2,

                    JunkTable j1, JunkTable j2
                    where e1.participant = e2.participant and
                          e1.junk = j1.junk and
                          e2.junk = j2.junk and
                          j1.VAquadrant = j2.VAquadrant and
                          j1.location <> j2.location and
                          e1.timestamp = e2.timestamp) t
         where timestamp1 = timestamp2) t2
         group by participant, camera1, camera2

delete from FRNoConsistentEstimatesTable
where CONCAT(camera1, camera2) in (
    select
        case when et.camera1 > et.camera2 then
            CONCAT(et.camera2, et.camera1) else
            CONCAT(et.Camera1, et.camera2)
        end
    from FRNoConsistentEstimatesTable et
    where et.camera1 in
        (select it.camera2 from FRNoConsistentEstimatesTable it
         where et.camera2 = it.camera1 and et.camera1 = it.camera2)
    group by
        case when et.camera1 > et.camera2 then
            CONCAT(et.camera2, et.camera1) else
            CONCAT(et.camera1, et.camera2)
        end
    having count(*) > 1
);
```

Listing 5. Statements for calculating number of consistent estimates for Noldus Face Reader in two-dimensional model.

```
select bc.beard, bp.camera1, bp.camera2, x*100/bc.numberOfTuples from
    (select beard, sum(counterTable.numberOfTuples) As numberOfTuples from
        counterTable, participantTable
        where counterTable.participant = participant.participant
        group by beard
    ) bc
    JOIN
    (select beard, camera1, camera2, sum(numberOfTuples) as x from
        FRNoConsistentEstimatesTable, participant
        where FRNoConsistentEstimatesTable.participant = participant.participant
        group by beard, camera1, camera2
    ) bp
    on bc.beard = bp.beard
```

Listing 6. Calculation of the percentage of consistent estimates for Noldus Face Reader.

is calculated between the emotion estimates time series for various pairs of camera locations, is depicted in Listing 7.

```
with
member ActualMeasure AS [Measures].[HappinessAvg]
member dl AS ([Location].[Location].&[DL], ActualMeasure)
member dr AS ([Location].[Location].&[DR], ActualMeasure)
member ul AS ([Location].[Location].&[UL], ActualMeasure)
member ur AS ([Location].[Location].&[UR], ActualMeasure)
member DL:DR as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, dl, dr),
    Format_String="Standard"
member DL:UL as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, dl, ul),
    Format_String="Standard"
member DL:UR as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, dl, ur),
    Format_String="Standard"
member DR:UL as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, dr, ul),
    Format_String="Standard"
member DR:UR as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, dr, ur),
    Format_String="Standard"
member UL:UR as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, ul, ur),
    Format_String="Standard"

select
{   DL:DR, DL:UL ,DL:UR, DR:UL, DR:UR, UL:UR    } on 0,
{   ([Participant].[Participant].[All],
    [Junk].[System].&[FR],
    [Junk].[HappinessRange].&["<0.2"])    } on 1
from [Cube]
```

Listing 7. Query for correlation coefficient calculation between various locations.

In the listing the *ActualMeasure* defines the name of emotion estimate, for which the correlation coefficient is to be calculated (as previously given in Equation 3). In exemplary query, it is the average value of happiness - *happinesAvg*. The next four calculated variables (members in MDX query) (*dl*, *dr*, *ul*, *ur*) allow the calculation of values of

ActualMeasure but only for appropriate camera location. The next six variables are responsible for calculating correlation coefficients between appropriate camera locations. The central point of the query is the *select* clause, which defines that the correlation coefficients are displayed on the first axe, whereby the second axe defines how the emotional state time series are created. In the presented Listing they are calculated for all participants - average happiness for all of them - but only estimates obtained from Noldus Face Reader, and less than 0.2 are taken into consideration.

The Listing 8 is analogical to the Listing 7 and correspond to Question 2, where the correlation coefficient is calculated between the emotion estimates time series for various pairs of systems. In this listing, the *ActualMeasure* denotes average anger - *AngerAvg*. The calculated variables are defined for systems, not camera locations as in the Listing 7. The emotion estimate time series are calculated only for estimates obtained for the P08 participant from the down left camera. All estimates, whatever value they have, are taken into consideration.

```
with
member ActualMeasure AS [Measures].[AngerAvg]
member fr AS ([Junk].[System].&[FR], ActualMeasure)
member lu AS ([Junk].[System].&[LU], ActualMeasure)
member xp AS ([Junk].[System].&[XP], ActualMeasure)
member FR:LU as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, fr, lu),
    Format_String="Standard"
member FR:XP as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, fr, xp),
    Format_String="Standard"
member LU:XP as Correlation (
{   [EmotionEstimate].[Timestamp].Members } as times, lu, xp),
    Format_String="Standard"

select
{   FR:LU, FR:XP, LU:XP } on 0,
{   ([Participant].[Participant].&[P08],
    [Junk].[Location].&[DL],
    [Junk].[AngerRange].&[All])} on 1
from [Cube]
```

Listing 8. Query for correlation coefficient calculation between various systems.

It is possible to expand the queries with additional axes defining multiple ways of constructing the emotion estimates time series. These queries were issued to the Cube for six basic emotion estimates and neutral state - for both various camera locations and systems as well as for valence and arousal. The obtained answers were then transformed within the *Dataset preparation* step into four files:

1) **Six basic emotions and neutral state for various camera locations:** name of estimated emotional state - *emotion*, pair of camera locations - *cameras*, *system*, *sex*, *beard*, *moustache*, *glasses*, *range*, *correlationCoefficient*.
2) **valence and arousal for various camera locations:** *emotion*, *cameras*, *sex*, *beard*, *moustache*, *glasses*, *VA quadrant*, *correlationFactor*.
3) **Six basic emotions and neutral state for various systems:** *emotion*, pair of systems - *systems*, camera location - *camera*, *sex*, *beard*, *moustache*, *glasses*, *range*, *correlationCoefficient*.

The files were then expanded with two additional columns containing the categorized value of correlation coefficient for categorizations given with equations (5) and (6) respectively.

Using the KNIME tool for each of such obtained files, within the third step *Modelling decision tree*, the two decision tree models were built (C4.5) - one for each categorization. The *correlationCoefficient* column was removed. The pruning method was set to MDL. The next fourth step was to generate rules using the *Decision Tree to Ruleset* task.

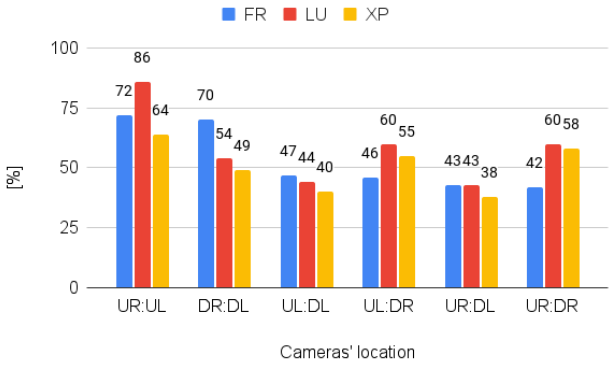

FIGURE 7. Percentage of consistent estimates for various cameras' location for the specified systems in Ekman model.

V. RESULTS

This Section presents the results of the analysis performed - the analysis of a number of consistent estimates in Section V-A and analysis of correlation coefficients in Section V-B. Within all presented results the notation is analogical to the one presented in Table 1 (DL, DR, UL, UR - for camera locations, and FR, LU, XP - for recognition systems). Additionally, when the correlation between two cameras or systems is presented, the notation with a colon is used (e.g. UL:UR).

A. ANALYSIS OF THE NUMBER OF CONSISTENT ESTIMATES

In the following points, the answers to the four defined analytical questions are presented. If possible - for the two models; if not - only for the Ekman model.

1) WHAT IS THE NUMBER OF CONSISTENT ESTIMATES FOR VARIOUS CAMERA LOCATIONS FOR EACH SYSTEM?

a: THE EKMAN MODEL WITH NEUTRAL STATE

In the Ekman model, whatever system recognizes emotional states, the highest percentage of consistent estimates is for upper cameras. Also, for all systems, this percentage is lower than 50% for left cameras and upper-right camera vs. the down-left camera. For the right cameras, the highest consistency is for Luxand and slightly less for Express Engine. All percentages between various camera locations and for all three systems are presented in Figure 7.

b: TWO-DIMENSIONAL MODEL

For the two-dimensional model, only the Noldus Face Reader provides the recognized estimates of valence and arousal. As presented in Figure 8, the highest percentage of consistent estimates is for both the upper and down cameras. However, for any combination of camera locations, this percentage is not less than 68%.

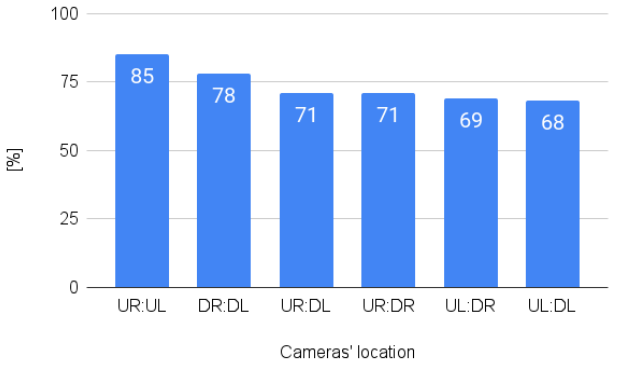

FIGURE 8. Percentage of consistent estimates for various camera locations for Noldus Face Reader in two-dimensional model.

2) WHAT IS THE NUMBER OF CONSISTENT ESTIMATES FOR VARIOUS CAMERA LOCATIONS FOR THE PARTICIPANTS CATEGORIZED VERSUS BEARD, GLASSES, MOUSTACHE, AND SEX?

The presentation of results for this question is done in radar charts, as this type of chart illustrates well how sex and occlusions influence the number of consistent estimates.

a: THE EKMAN MODEL WITH NEUTRAL STATE

The influence of sex on consistency for the Ekamn model is presented in Figure 9. In the Ekman model, the influence of the participant's sex on the consistency depends on the systems: for Noldus Face Reader, a slight influence can be noticed only when consistency between down cameras are analysed, whereas for Luxand no influence is observed. For Express Engine, this influence is seen for every pair of camera locations, although upper cameras and higher consistency is observed for female participants.

When analysing the influence of beard and moustache occlusions on consistency, an analogy can be observed on radar charts presented in Figures 10 and 11 for Face Reader and Luxand. For Noldus Face Reader, the heavy beard and some moustache have a negative influence on consistency when the cameras are diagonally located. In other cases, no influence is observed. For Luxand, some beard and some moustache influence consistency, although, this is excluding the upper cameras. What is interesting is that the influence of a heavy beard is much smaller than some beard. For Express Engine, the highest negative influence on consistency has some beard for down cameras, left cameras and upper-right camera versus down left camera. The significant influence on consistency also has some moustache, but for down cameras only.

Glasses are the last analysed occlusion. The results of the analysis are depicted in Figure 12. For both Luxand and Face Reader systems - for all pairs of camera locations - consistency is higher for participants with glasses than without them. For the Luxand system, the decrease of consistency is similar for all pairs of camera locations. For Face Reader, the decrease is higher for diagonally located cameras. In Express Engine, glasses have no influence on consistency when

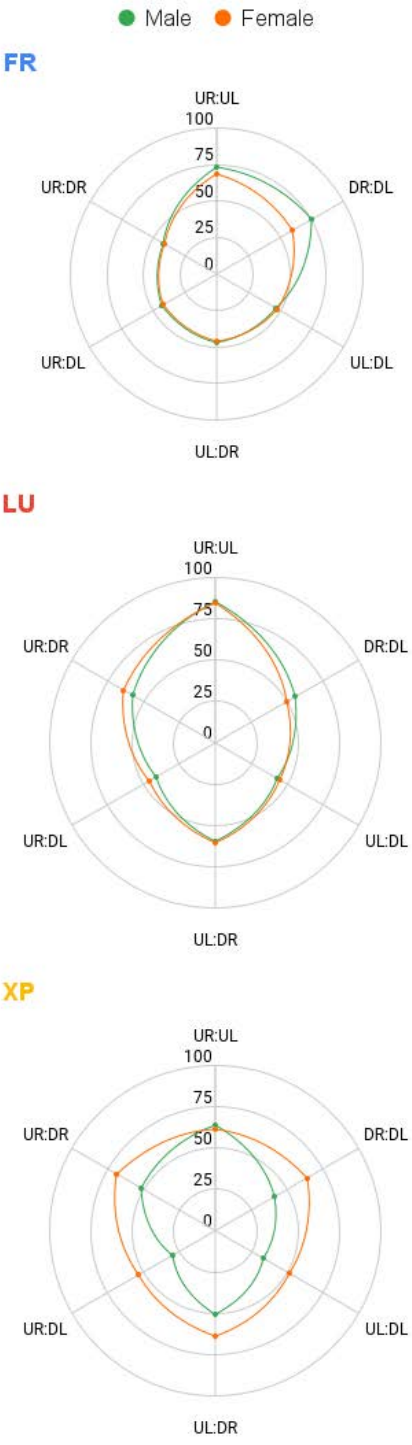

FIGURE 9. Percentage of consistent estimates for various camera locations for the specified systems in the Ekman model versus participant sex.

consistency between DR:DL, UR:DR or UL:DR camera location is analysed. Higher consistency for participants with glasses is observed only for upper cameras. In the last two cases (UR:DL and UL:DL), the higher consistency is for participants without glasses.

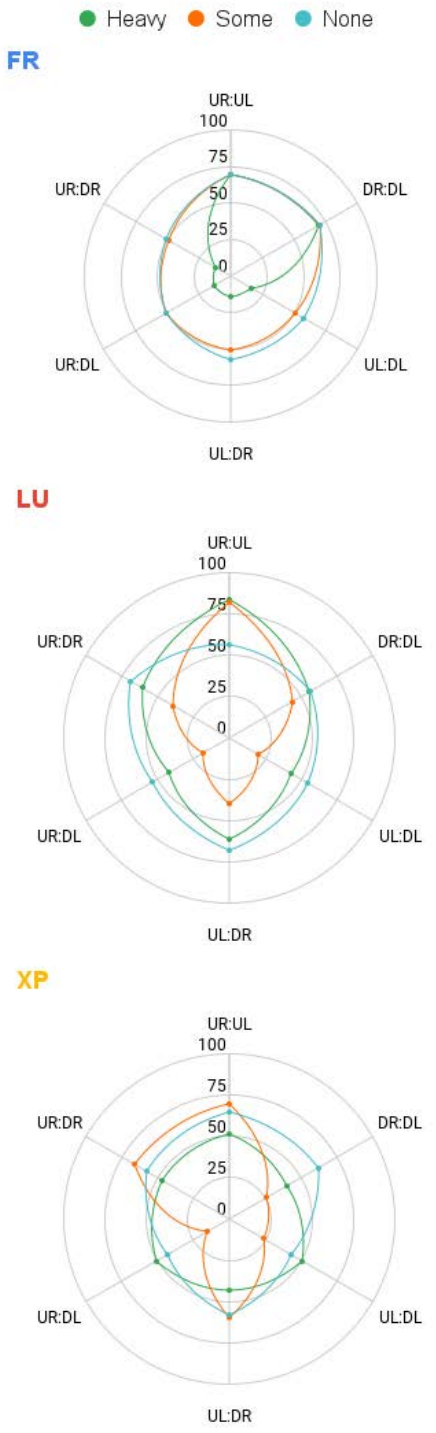

FIGURE 10. Percentage of consistent estimates for various camera locations for the specified systems in the Ekman model versus beard occlusion.

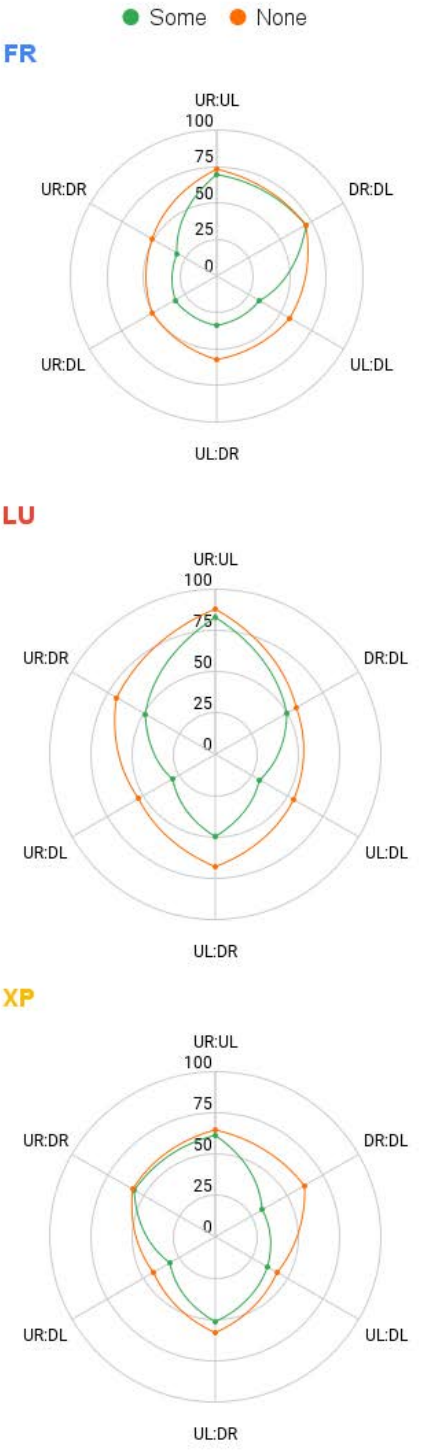

FIGURE 11. Percentage of consistent estimates for various camera locations for the specified systems in the Ekman model versus moustache occlusion.

b: THE TWO-DIMENSIONAL MODEL

The influence of sex on the percentage of consistent estimates is presented in Figure 13. This diagram shows that sex has no influence on consistency when upper or down cameras are analysed. For other camera locations, the consistency is higher for female participants.

When analysing the influence of the beard occlusion on the consistency, it can be noticed that it has a slight influence when comparing upper cameras, as is presented in Figure 14. The highest negative influence on consistency has some beard. Moustache has almost no influence on consistency,

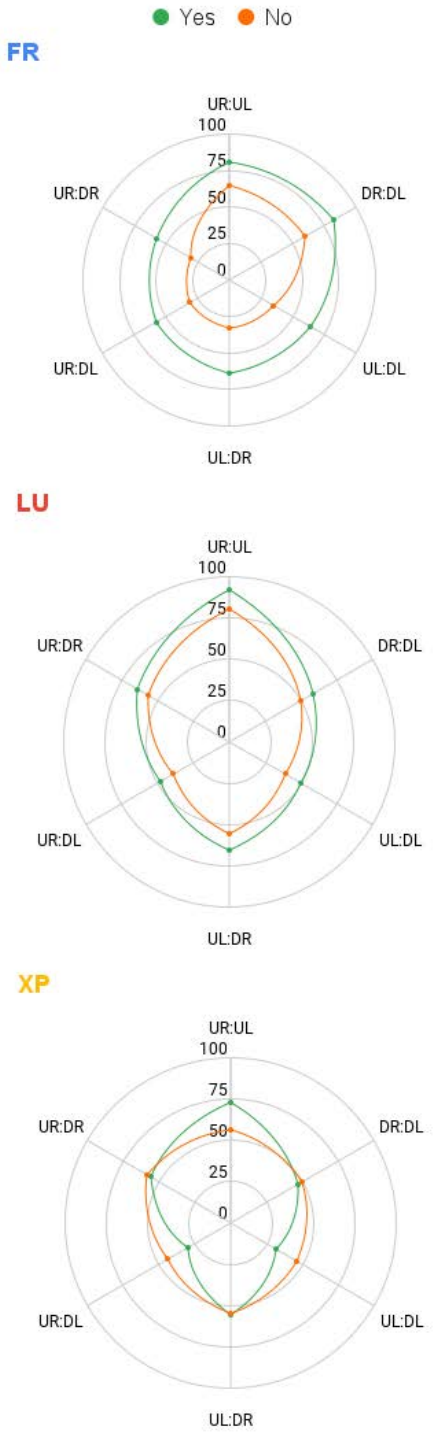

FIGURE 12. Percentage of consistent estimates for various camera locations for the specified systems in the Ekman model versus glasses occlusion.

as it is presented in Figure 15. The influence of the glasses occlusion depicted in Figure 16 is analogical to the one observed for sex (almost no influence for down and upper cameras). The existence of glasses influences negatively on consistency.

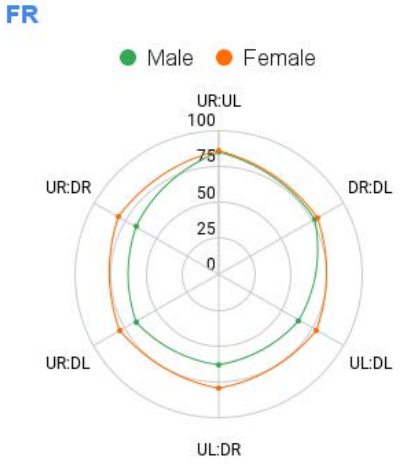

FIGURE 13. Percentage of consistent estimates for various camera locations versus sex for Noldus Face Reader in a two-dimensional model.

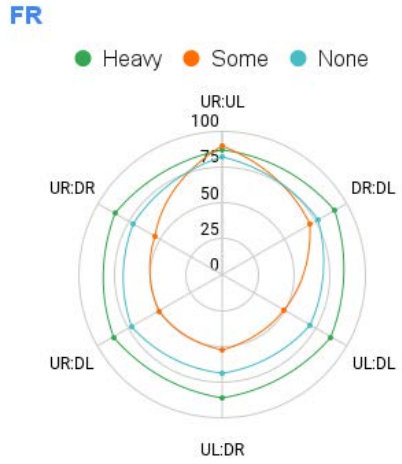

FIGURE 14. Percentage of consistent estimates for various camera locations versus beard for Noldus Face Reader in a two-dimensional model.

3) WHAT IS THE NUMBER OF CONSISTENT ESTIMATES IN THE EKMAN MODEL FOR VARIOUS SYSTEMS FOR EACH CAMERA?

The percentage of consistent estimates for various systems for the specified camera location in the Ekman model is depicted in Figure 17. The highest consistency is observed for upper cameras and the smallest for the down-left camera. The consistency between various pairs of systems with regard to camera location is quite similar - the highest difference is equal to 11%.

4) WHAT IS THE NUMBER OF CONSISTENT ESTIMATES IN THE EKMAN MODEL FOR VARIOUS SYSTEMS IN VARIOUS CAMERA LOCATIONS FOR THE PARTICIPANTS CATEGORIZED VERSUS BEARD, GLASSES, MOUSTACHE AND SEX?

The results of the analysis are depicted in Figures 18, 19, 20 and 21, analogically as for various camera locations firstly for sex, then for occlusions: beard, moustache and glasses.

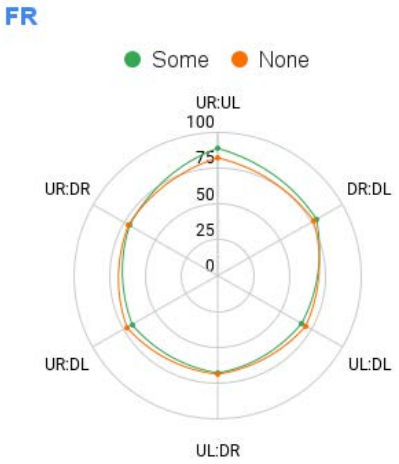

FIGURE 15. Percentage of consistent estimates for various camera locations versus moustache for Noldus Face Reader in a two-dimensional model.

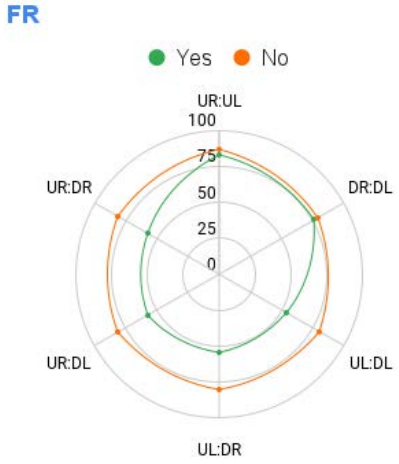

FIGURE 16. Percentage of consistent estimates for various camera locations versus glasses for Noldus Face Reader in a two-dimensional model.

The sex factor has almost no influence on consistency for each pair of systems when upper cameras are analysed, and for both Luxand and Face Reader systems for down cameras. The negative influence for down cameras for other pairs of systems is observed for male participants.

When analysing the consistency with respect to beard occlusion it is noticed that some and no beard have a slight influence on consistency for upper cameras. Heavy beard has a negative influence on consistency for upper left and upper right camera locations. Otherwise, for down camera locations, the highest negative influence on consistency has some beard. The only exception is for Express Engine and Face Reader systems for the down-right camera.

For moustache occlusion for all pairs of systems and all camera locations, the negative influence on consistency has some moustache. The smallest influence is observed for both Express Engine and Face Reader for upper cameras.

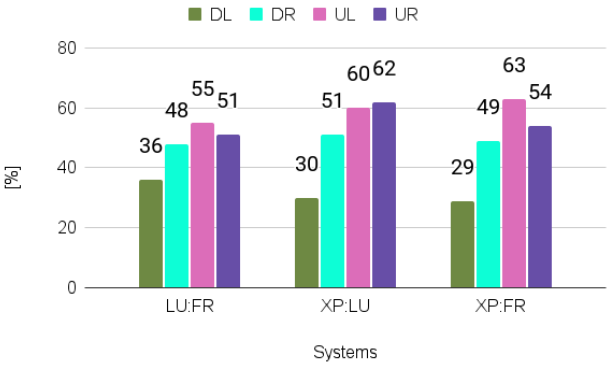

FIGURE 17. Percentage of consistent estimates for various systems for the specified camera location in the Ekman model.

For the last analysed occlusion, an absence of glasses has a negative influence on consistency for all pairs of analysed systems for upper cameras. For down cameras, it depends on pairs of analysed systems. For the down-left camera, a lack of glasses positively influences consistency when Express Engine and the other system is analysed, and negatively for the LU:FR pair of systems. For the down-left camera, the absence of glasses negatively influences consistency when FR versus any other system is analysed, and positively for the XP:LU pair of systems.

B. ANALYSIS OF CORRELATION COEFFICIENTS WITH RESPECT TO VARIOUS INCONSISTENCY FACTORS

The results of the analysis are presented for the two defined questions. For the first question, the analysis is done for both models and for the second question only for the Ekman model, as valence and arousal were only recognized by Noldus Face Reader.

1) WHICH INCONSISTENCY FACTORS INFLUENCE THE CONSISTENCY OF THE RECOGNIZED EMOTIONS FROM VARIOUS CAMERA LOCATION, AND IN WHICH WAY?

a: THE EKMAN MODEL WITH THE NEUTRAL STATE

For the Ekman model with a neutral state, the results of classification are done for two correlation coefficient categorizations. The rules detected for the first categorization from Equation (5) show mainly when there is no correlation (for DL:DR with a 0.62 accuracy ratio and DL:UL, DL:UR, DR:UL, DR:UR with an accuracy ratio of about 0.84). The rules detecting positive correlation - negative correlation was not detected - with an accuracy ratio greater than 0.5 are depicted in Table 7: one for strong correlation and two for both moderate and weak correlation. All of them are for upper cameras. The strongest correlation is for the participants with glasses when a happy emotion is detected, whereby accuracy ratio is equal to 0.69. The moderate correlation is for neutral and anger, but for the second emotion only for the Luxand system. Also, the fifth rule is interesting showing that the weak correlation is detected for disgust, but only when the

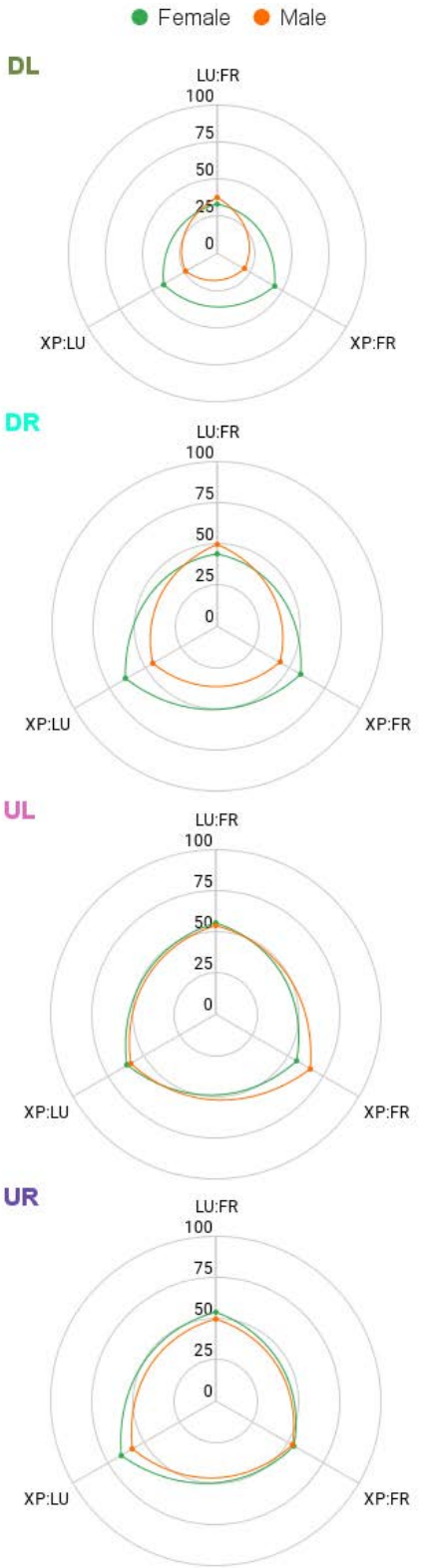

FIGURE 18. Percentage of consistent estimates for various systems for the specified camera location in the Ekman model versus participant sex.

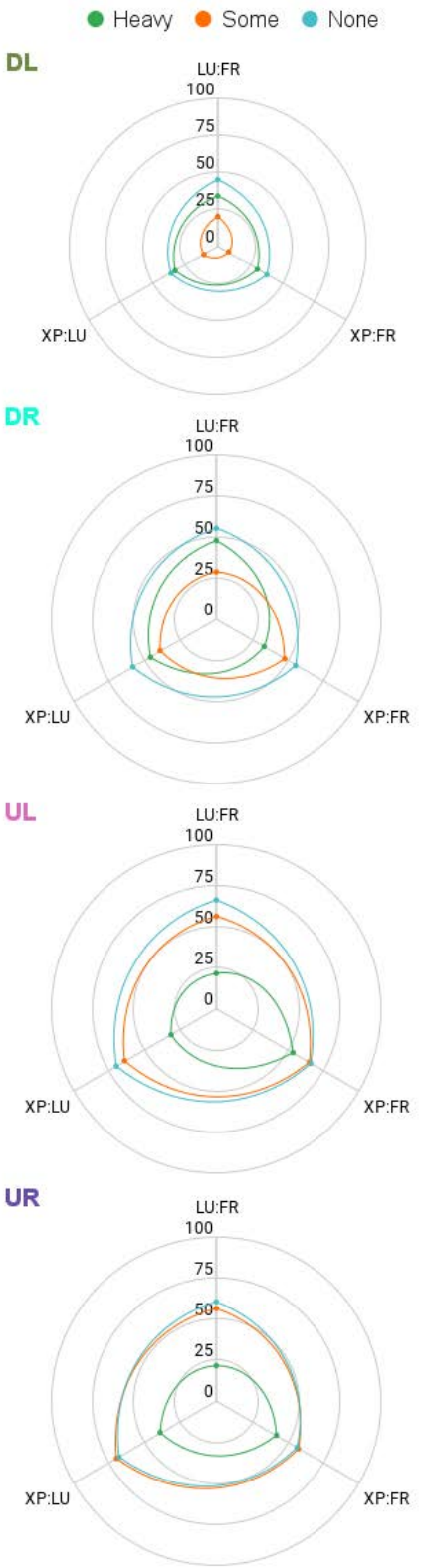

FIGURE 19. Percentage of consistent estimates for various systems for the specified camera location in the Ekman model versus beard occlusion.

value of this emotional estimate is higher or equal to 0.2 and for participants without glasses.

The application of the simplified categorization of correlation coefficient defined by Equation (6) led to the detection of a greater number of rules showing when some correlation is detected and accuracy ratio is greater than 0.5. These rules are presented in Table 8.

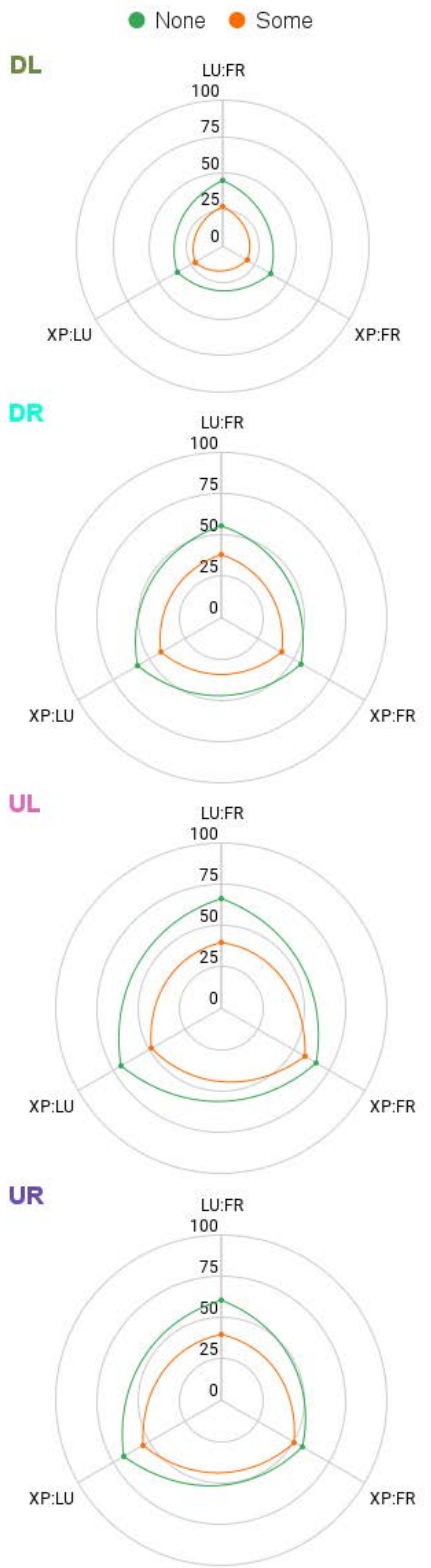

FIGURE 20. Percentage of consistent estimates for various systems for the specified camera location in the Ekman model versus moustache occlusion.

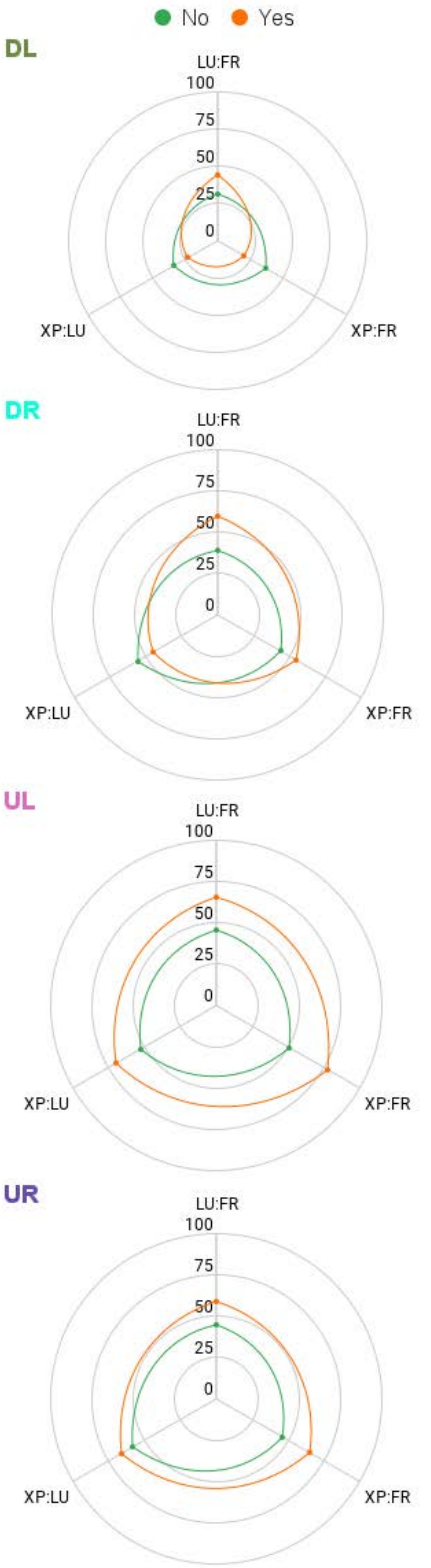

FIGURE 21. Percentage of consistent estimates for various systems for the specified camera location in the Ekman model versus glasses occlusion.

Some correlation is detected only between cameras located in down corners (DL:DR) and cameras located in upper corners (UL:UR) of the screen. The lack of correlation between other camera locations is proved with the detected rules with an accuracy ratio analogical as for the first categorization of correlation coefficient (about 0.84).

TABLE 7. Rules presenting a correlation between camera locations for categorization defined by Equation (5) in the Ekman model.

No.	Rule	Accuracy ratio
1	glasses="yes" AND range="all" AND emotion="happiness" AND cameras="UL:UR" → "strong positive"	0.69
2	range = "all" AND system = "all" AND emotion = "neutral" AND cameras = "UL:UR" → "moderate positive"	0.7
3	range = "all" AND system = "LU" AND emotion = "anger" AND cameras = "UL:UR" → "moderate positive"	0.6
4	range = ">=0.2" AND system = "all" AND emotion = "neutral" AND cameras = "UL:UR" → "weak positive"	0.6
5	glasses = "no" AND range = ">=0.2" AND emotion = "disgust" AND cameras = "UL:UR" → "weak positive"	0.59

Firstly, the rules detecting correlation existence between down cameras are analysed. The correlation is detected for sadness (5th, 12th, 15th rules in Table 7), happiness (18th rule in the same table), surprise (19th rule) and neutral state (8th rule). For sadness, a correlation is detected for all systems, however, the accuracy ratio is the highest for Noldus Face Reader (0.8) and Luxand (0.7). For happiness and surprise, the accuracy ratio is equal to 0.55. When the neutral state is recognized the accuracy ratio is higher and equal to 0.7. Additionally, for Noldus Face Reader the correlation is detected when the recognized emotional state is higher or equal to 0.2 with an accuracy ratio equal to 0.6.

Analysing rules detecting correlation for upper cameras it can be noticed that the correlation exists additionally to sadness (3rd, 16th and 20th rules in table 7), happiness (9th rule) and neutral state (4th, 7th and 10th rules) also for anger (1st, 2nd, 6th, 13th rules) and disgust (11th and 14th rules). In contrast to down cameras, no correlation for surprise is detected. The accuracy ratio (without notifying the system) for sadness is 0.6, for happiness 0.75, for neutral state 0.77, for anger 1 and for disgust 0.65. Additionally, for sadness and neutral state, the accuracy ratio is higher for the Luxand system. Even if any other rules are detected the accuracy ratio is not better. The occlusions appear only twice (2nd and 11th rules), whereby not improving the accuracy ratio.

b: THE TWO-DIMENSIONAL MODEL

Analogically as for the Ekman model with a neutral state, the results of classification are done for the two correlation coefficients categorizations. The rules detecting some correlation with accuracy ratio greater than 0.5 are depicted in Tables 9 and 10. The correlation is detected for down and upper cameras only. For upper and down cameras the moderate correlation is detected (range = ''all'' AND cameras = ''DL:DR'' => ''moderate positive'', range = ''all'' AND cameras = ''UL:UR'' => ''moderate positive''). However, these rules are not depicted in Table 9 as the accuracy ratio is equal 0.5 and 0.45 respectively. For other camera

TABLE 8. Rules presenting the existence of a correlation between camera locations for categorization defined by Equation (6) in Ekman model.

No	Rule	Accuracy ratio
1	range = "all" AND system = "all" AND emotion = "anger" AND cameras = "UL:UR" → "yes positive"	1.00
2	moustache = "some" AND range = ">=0.2" AND system = "all" AND emotion = "anger" AND cameras = "UL:UR" → "yes positive"	1,00
3	system = "LU" AND emotion = "sadness" AND cameras = "UL:UR" → "yes positive"	1.00
4	system = "LU" AND emotion = "neutral" AND cameras = "UL:UR" → "yes positive"	0.93
5	system = "FR" AND emotion = "sadness" AND range = "all" AND cameras = "DL:DR" → "yes positive"	0.80
6	system = "FR" AND emotion = "anger" AND cameras = "UL:UR" → "yes positive"	0.80
7	system = "all" AND emotion = "neutral" AND cameras = "UL:UR" → "yes positive"	0.77
8	emotion = "neutral" AND range = "all" AND cameras = "DL:DR" → "yes positive"	0.75
9	emotion = "happiness" AND cameras = "UL:UR" → "yes positive"	0.75
10	system = "FR" AND emotion = "neutral" AND cameras = "UL:UR" → "yes positive"	0.73
11	glasses = "no" AND range = ">=0.2" AND emotion = "disgust" AND cameras = "UL:UR" → "yes positive"	0.73
12	system = "LU" AND emotion = "sadness" AND range = "all" AND cameras = "DL:DR" → "yes positive"	0.70
13	system = "LU" AND emotion = "anger" AND cameras = "UL:UR" → "yes positive"	0.66
14	range = "all" AND emotion = "disgust" AND cameras = "UL:UR" → "yes positive"	0.65
15	system = "all" AND emotion = "sadness" AND range = "all" AND cameras = "DL:DR" → "yes positive"	0.60
16	system = "all" AND emotion = "sadness" AND cameras = "UL:UR" → "yes positive"	0.60
17	system = "FR" AND range = ">=0.2" AND cameras = "DL:DR" → "yes positive"	0.60
18	emotion = "happiness" AND range = "all" AND cameras = "DL:DR" → "yes positive"	0.55
19	emotion = "surprise" AND range = "all" AND cameras = "DL:DR" → "yes positive"	0.55
20	system = "FR" AND emotion = "sadness" AND cameras = "UL:UR" → "yes positive"	0.50

locations, no correlation is detected (with accuracy ratio from 0.63 to 0.67). The only two rules, which detect some correlation with an accuracy ratio greater than 0.5 concern the LVLA quadrant and depend on the beard. When participants have no beard, the moderate positive correlation is detected and conversely, when participants have a heavy beard, the correlation is a weak positive.

For the second categorization, the only rules showing a correlation for down cameras and upper cameras are detected (without the influence of sex or occlusions). The higher accuracy ratio is for upper cameras, as depicted in Table 10. As in the case of the first categorization of correlation coefficient defined by Equation (5), a lack of correlation for other camera locations is detected with analogical values of accuracy ratios.

TABLE 9. Rules presenting correlation between camera locations for categorization defined by Equation (5) for the two-dimensional model.

No.	Rule	Accuracy ratio
1	beard = "heavy" AND range = "LVLA" AND cameras = "DL:DR" => "weak positive"	0.75
2	beard = "none" AND range = "LVLA" AND cameras = "DL:DR" => "moderate positive"	0.58

TABLE 10. Rules presenting the existence of a correlation between camera locations for (6) categorization in a two-dimensional model.

No	Rule	Accuracy ratio
1	cameras = "UL:UR" => "yes positive"	0.8
2	cameras = "DL:DR" => "yes positive"	0.64

2) WHAT INCONSISTENCY FACTORS INFLUENCE THE CONSISTENCY OF THE RECOGNIZED EMOTIONS FROM VARIOUS SYSTEMS, AND IN WHAT WAY?

This analysis is done for the Ekman model with a neutral state only. When the first categorization defined by Equation (5) is applied, no correlation is detected for all of the analysed emotions. For all emotions other then happiness, the accuracy ratio is greater than 0.8. For happiness, the accuracy ratio is equal to 0.51. Thus, for the second categorization defined by the Equation (6), a single rule detecting correlation is identified (depicted in Table 11). The positive correlation is detected for happiness with an accuracy ratio equal to 0.64.

TABLE 11. Rules presenting existence of correlation between systems locations for categorization defined by Equation (6) in the Ekman model with the neutral state.

No	Rule	Accuracy ratio
1	range = "all" AND emotion = "happiness" => "yes positive"	0.64

3) VALIDATION RESULTS

The generated rules were evaluated by validation of decision tree models built to generate them. The validation was done using the cross-validation method with ten folds and stratified sampling.

The achieved accuracy for the model predicting the correlation coefficient between time series of recognized Ekman's emotional states (obtained from various cameras) is equal to 73.79% for the categorization based on Equation 5 and 77.62% for the categorization based on Equation 6.

The achieved accuracy for the model predicting the correlation coefficient between a time series of recognized Ekman's emotional states (obtained from various systems) is equal to 84.46% for both categorizations, as the sets of generated rules in both cases are identical.

When the 2D model is considered, the accuracy is significantly less than for the Ekman model. It is equal to 52.07% for the categorization based on Equation 5 and 66.92% for the categorization based on Equation 6.

It should be emphasized that the aim of the research is not to find a classifier that would allow to build a classification model with high accuracy and other statistical measures, e.g. F-score, but the identification of rules with a high accuracy ratio. It is worth noting that the general assessment of the classifier (added here for completeness only at a basic level) does not affect the correctness of the detected rules. This is due to the fact that the accuracy or F-score values do not refer to the correctness of a single detected rule, but evaluate the entire classifier. Therefore, a reliable assessment of a particular detected rule, and thus its usefulness, is defined by the accuracy ratio. Hence, no further analysis of the classifier parameters is carried out.

VI. SUMMARY OF RESULTS AND DISCUSSION

The core challenges that we have encountered in analysing the multichannel data from the experiment might be summarized as follows:

- the data (time series) obtained from cameras and further from facial expression analysis systems require synchronisation;
- the time series for a specific person and moment in time might have missing values;
- although typical face occlusions (like beard or glasses) should not differ along with the time series for the same person, the software recognized them with fluctuating value (occlusion-based inconsistency);
- the estimated emotional states for the same person and point-in-time differed for the four cameras (camera-based inconsistency);
- the estimated emotional states for the same camera, person and point-in-time differed depending on the facial expression system used (system-based inconsistency);
- it is hard to resolve the inconsistencies as the ''ground truth'' is unknown - we have used the self-report. However, it is important to emphasize that the report gives a value for a point in time, usually between the tasks, and it might be not valid to interpolate it for the preceding task;
- interconnection between the obtained inconsistency and multiple dependent and confounding variables is unknown.

Some of the challenges mentioned above are characteristic for the affective computing domain - for example finding the ''ground truth'' in emotion recognition is a well-known problem [21]. The inconsistencies in multichannel and multimodal recognition are also known [5]. In those terms, observations from this study are compatible with previous findings. What surprised us, was the differing level of inconsistencies depending on multiple factors. Some of the other challenges listed above might be considered present for all multiple time series analysis, including missing values or time synchronisation. However, the purpose of the study was to explore the inconsistency of the emotion estimates and the main findings

could be divided into those that refer to emotion recognition itself and those that refer to the method used. Regarding multichannel emotion recognition, we have found that:

- The availability of a specific channel (ex. video from a single camera) is user- and time-dependent. Multiplication of the devices is advisable, however, one must then deal with the inconsistencies.
- The upper cameras (above the monitor) give more consistent results (with each other) than any other combination of cameras. That might indicate that the location above the monitor is a better one from the perspective of emotion recognition. However, this conclusion is based only on a consistency level, as the ''ground truth'' was not available in this study.
- The processing of the video channel by different software solutions differs - some systems are more prone to being disturbed by occlusions and camera angle, for example, Face Reader is much less consistent with participants wearing glasses.
- The consistency does not change significantly with sex for any system or camera combination.
- A heavy or some beard is the most disturbing occlusion, especially from the perspective of consistency between the upper and down camera pairings.
- A moustache causes higher inconsistencies, however, this is true only for estimates in Ekman's model and not for the two-dimensional model. This result is quite surprising, when it is clear that the Face Reader calculates dimensional model values from Ekman's basic model.
- Depending on occlusions, diverse recognition systems are more or less prone to generate inconsistencies, which might also indicate that it is advisable to use more than one system in a study. However, in such a case, one must deal with inconsistencies.

Summing up, it might seem like a good idea to multiply observation channels and recognition systems in order to increase time-based availability and reliability of emotion recognition results. However, such a solution introduces possible inconsistencies that must be resolved.

The second part of our results refers to the approach we have chosen to analyse the study outputs. In this paper, we proposed an approach based on a multidimensional model - derived from OLAP (Online Analytical Processing), and data mining. Some of our findings on using such an approach follow:

- The chosen approach allowed to analyze multiple factors that influence consistency in the multichannel emotion recognition. There are other methods, based on a statistical approach, that would allow us to analyze those factors separately as well.
- An advancement of the chosen approach is that it allows us to do multidimensional analysis and to explore the concurrence of the factors.
- Two different analysis types were performed based on the multidimensional model - the first was based on consistent estimate counts, while the other combined correlation coefficients with decision trees. The findings of the two analysis types are consistent.
- The analysis based on consistent estimate count allowed the analysis of multiple variables at the same time - independent variable and multiple confounding variables as defined in Table 1.
- The second analysis allowed the building of rules in the decision tree that show the inter-dependencies between the values of specific variables. Moreover, we were also able to analyze the consistencies for different basic emotions separately.
- The rules that were defined within the second analysis are assigned a value of accuracy that hold the information on the percentage of cases the rule is valid for. There are some rules that have an accuracy of 1, which indicates they ''work'' for this specific dataset for all of the cases. This rule accuracy measure is an interesting approach to quantify the certainty of the research-based findings.

Summing up, the proposed approach allowed the exploration of multiple factors that cause an inconsistency in multichannel emotion recognition. According to our knowledge, such an approach was not previously used in mining inconsistencies in multimodal affect recognition and we consider it the main novelty of the paper.

The validity threats of our study are related to the the construction of the quasi-experiment, to the categorisation of the correlation coefficient, and to the analysis of consistency itself.

Regarding the experiment, in the study we were asking for a self-report of emotional states. However, a SAM questionnaire was applied between the tasks and not within them. The reason for such a decision was that we did not want the tasks to be interrupted. As a result, however, the questionnaire results reflect the overall mood or disposition of the day rather than momentary emotional state - it would be especially risky to assume that the state reported after the task was actually induced by the task. Therefore, we focused on the analysis of consistency rather than accuracy of emotion recognition. The differences in SAM responses did not differ much over the pace of the experiment for some of the participants. Another validity threat is respectively low diversity of the participant group.

Regarding the correlation coefficient, we have proposed two methods of its categorisation (given with equations (5) and (6)). More or different approaches to that categorisation might be applied, which might result in a different outcome. The categorisation has a significant influence on the rules of the decision tree.

Regarding consistency, we have made an assumption of consistency criteria. In the analysis of the number of consistent estimates, the two estimates are regarded as consistent when the same dominant emotion is recognized. It is possible to define the consistency of estimates in various ways. For example, let us imagine that one algorithm recognized the happiness and neutral state with certainty 0.4 and

0.39 respectively and the other algorithm recognized the happiness and neutral state with certainty 0.39 and 0.4. In the assumed definition of consistency, the two estimates are treated as not consistent. To widen the analysis and the set of possible conclusions various consistency definitions could be applied.

Even though there are some validity threats that apply to the study presented, we still find it valuable in terms of analysing concurrence of factors influencing multichannel emotion recognition. We also find it valuable in terms of the evaluation of our approach to data analysis based on a multidimensional model.

VII. CONCLUSION

The paper confirms the applicability of the multidimensional data model and mining techniques for the identification of inconsistency factors in multichannel emotion recognition. Data integrated into the multidimensional model were used as an input for the two analyses, mining inconsistency in emotion recognition results. The two analyses were executed according to the well-defined steps and described in detail. Both these analyses were done for the specified purpose, for the specified variables. However, the steps executed within the analysis can be generalized and further used to define generic methods allowing the identification of various inconsistency factors in multichannel emotion recognition, independently on used variables. Such methods designed and developed could provide an easy and semi-automatic way of identification of inconsistency factors.

The general idea of the proposed approach to inconsistency detection is to (1) formally describe the time series with recognized emotions e.g. using ontologies and then (2) use these formal descriptions to capture new knowledge about consistency. Our future works will focus on developing such ontologies for emotion representation. The inconsistency detection methods vary according to the type of results obtained. Among others, results can be numerical metrics or classification rules. However, in the proposed approach, the formal description is used to define method inputs, e.g. inconsistency factors and consistency variables. In the future, we want to develop more generic steps for the proposed inconsistency detection methods, which would allow the application of methods for various datasets obtained within one or more experiments.

The use of the multidimensional models for data exploration might be considered as the main novelty of the proposed paper. Our future works in using the model for such an exploration might go beyond emotion recognition and might be applied to any multichannel or multimodal recognition systems that deal with inconsistency due to multiple factors.

Mining inconsistencies and confounding factors in emotion recognition has significant practical implications. The facial expression solutions are used in a contemporary setting, for example, during job interviews. Depending on the camera location or one's wearing of glasses, the recognized emotion might be different and possibly influence hiring process results. Therefore not only the confounding factors should be explored, but they also should be reported along with the recognition results. Reporting requires quantification of the influence of the confounding factor on the recognition results. The proposed approach to the exploration of inconsistencies might be a step towards the kind of solutions that report the uncertainty of the recognized state along with the recognized emotion.

REFERENCES

[1] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion classification based on biophysical signals and machine learning techniques," *Symmetry*, vol. 12, no. 1, p. 21, Dec. 2019. [Online]. Available: https://www.mdpi.com/2073-8994/12/1/21

[2] K.-I. Benta and M.-F. Vaida, "Towards real-life facial expression recognition systems," *Adv. Electr. Comput. Eng.*, vol. 15, no. 2, pp. 93–102, 2015, doi: 10.4316/AECE.2015.02012.

[3] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005791694900639, doi: 10.1016/0005-7916(94)90063-9.

[4] L. Corr and J. Stagnitto, *Agile Data Warehouse Design: Collaborative Dimensional Modeling, From Whiteboard to Star Schema*. Devon, PA, USA: DecisionOne Press, 2011.

[5] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–36, Apr. 2015, doi: 10.1145/2682899.

[6] A. Dinculescu, C. Vizitiu, A. Nistorescu, M. Marin, and A. Vizitiu, "Novel approach to face expression analysis in determining emotional valence and intensity with benefit for human space flight studies," in *Proc. E-Health Bioeng. Conf. (EHB)*, Nov. 2015, pp. 1–4, doi: 10.1109/EHB.2015.7391378.

[7] P. Ekman, "Are there basic emotions?" *Psychol. Rev.*, vol. 99, no. 3, pp. 550–553, 1992, doi: 10.1037/0033-295X.99.3.550.

[8] P. Ekman, "Expression and the nature of emotion," in *Approaches to Emotion*, P. Ekman and K. R. Scherer, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1984, pp. 319–344.

[9] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.

[10] *Facereader Solution Description*. Accessed: Mar. 30, 2016. [Online]. Available: http://www.noldus.com http://www.noldus.com

[11] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early fusion vs. late fusion," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 4, Oct. 2005, pp. 3437–3443, doi: 10.1109/ICSMC.2005.1571679.

[12] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2367–2371, doi: 10.1109/ICASSP.2017.7952580.

[13] G. Hongxiang, S. An, J. Li, and C. Liu, "Deep balanced learning for long-tailed facial expressions recognition," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May/Jun. 2021, pp. 11147–11153.

[14] I. Hupont, S. Ballano, S. Baldassarri, and E. Cerezo, "Scalable multimodal fusion for continuous affect sensing," in *Proc. IEEE Workshop Affect. Comput. Intell. (WACI)*, Apr. 2011, pp. 1–8, doi: 10.1109/WACI.2011.5953150.

[15] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN, USA: Wiley, 2013. [Online]. Available: https://www.safaribooksonline.com/library/view/the-data-warehouse/9781118530801

[16] A. Kołakowska, W. Szwoch, and M. Szwoch, "A review of emotion recognition methods based on data acquired via smartphone sensors," *Sensors*, vol. 20, no. 21, p. 6367, Nov. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/21/6367, doi: 10.3390/s20216367.

[17] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 907–929, Jun. 2019, doi: 10.1007/s11263-019-01158-4.

[18] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, Sep. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885617300379, doi: 10.1016/j.imavis.2017.02.001.

[19] A. Landowska, "Emotion monitor—Concept, construction and lessons learned," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, 2015, pp. 75–80.

[20] A. Landowska, "Towards emotion acquisition in IT usability evaluation context," in *Proc. Mulitimedia, Interact., Design Innnovation (MIDI)*, 2015, pp. 1–9.

[21] A. Landowska, "Uncertainty in emotion recognition," *J. Inf., Commun. Ethics Soc.*, vol. 17, no. 3, pp. 273–291, Aug. 2019, doi: 10.1108/JICES-03-2019-0034.

[22] A. Landowska and J. Miler, "Limitations of emotion recognition in software user experience evaluation context," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, 2016, pp. 1631–1640.

[23] M. Glodek, S. Reuter, M. Schels, K. Dietmayer, and F. Schwenker, "Kalman filter based classifier fusion for affective state recognition," in *Multiple Classifier Systems* (Lecture Notes in Computer Science), vol. 7872. Berlin, Germany: Springer, 2013, doi: 10.1007/978-3-642-38067-9_8.

[24] A. Mehrabian, "Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression," *J. Psychopathol. Behav. Assessment*, vol. 19, no. 4, pp. 331–357, Dec. 1997, doi: 10.1007/BF02229025.

[25] M. Moolchandani, S. Dwivedi, S. Nigam, and K. Gupta, "A survey on: Facial emotion recognition and classification," in *Proc. 5th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Apr. 2021, pp. 1677–1686, doi: 10.1109/ICCMC51019.2021.9418349.

[26] D. S. Moschona, "An affective service based on multi-modal emotion recognition, using EEG enabled emotion tracking and speech emotion recognition," in *Proc. IEEE Int. Conf. Consum. Electron. Asia (ICCE-Asia)*, Nov. 2020, pp. 1–3, doi: 10.1109/ICCE-Asia49877.2020.9277291.

[27] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300738, doi: 10.1016/j.inffus.2017.02.003.

[28] A. Radoi, A. Birhala, N.-C. Ristea, and L.-C. Dutu, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, pp. 135559–135570, 2021, doi: 10.1109/ACCESS.2021.3116530.

[29] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 524–543, Apr. 2021.

[30] K. R. Scherer, "Towards a prediction and data driven computational process model of emotion," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 279–292, Apr. 2021, doi: 10.1109/TAFFC.2019.2905209.

[31] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005, doi: 10.1177/0539018405058216.

[32] H. Schlosberg, "Three dimensions of emotion," *Psychol. Rev.*, vol. 61, no. 2, pp. 81–88, 1954.

[33] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with AFFDEX and FACET: A validation study," *Behav. Res. Methods*, vol. 50, no. 4, pp. 1446–1460, 2018, doi: 10.3758/s13428-017-0996-1.

[34] B. R. Steunebrink, M. Dastani, and J.-J. C. Meyer, "The OCC model revisited," in *Proc. 4th Workshop Emotion Comput. Current Res. Future Impact, Appl. Comput. Sci. Inf. Technol., Baden-Württemberg Cooperat. State Univ. Stuttgart (DHBW)*, Stuttgart, Germany, 2009.

[35] Y. R. Veeranki, H. Kumar, N. Ganapathy, B. Natarajan, and R. Swaminathan, "A systematic review of sensing and differentiating dichotomous emotional states using audio-visual stimuli," *IEEE Access*, vol. 9, pp. 124434–124451, 2021, doi: 10.1109/ACCESS.2021.3110773.

[36] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 206–218, Jun. 2011, doi: 10.1109/T-AFFC.2011.12.

[37] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels (extended abstract)," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 477–483, doi: 10.1109/ACII.2015.7344613.

[38] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *J. Nonverbal Behav.*, vol. 37, no. 4, pp. 217–230, Dec. 2013, doi: 10.1007/s10919-013-0159-8.

[39] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 16–35, Jan. 2021, doi: 10.1109/TAFFC.2018.2879512.

AGNIESZKA LANDOWSKA is currently an Associate Professor at the Gdańsk University of Technology, Poland. She is the Leader of emotions at the Human–Computer Interactions Group (EMORG). She leads a project that develops mobile applications dedicated to autism therapy (Friendly Apps). She has supervised research in project AFFITS (Methods and tools for affect-aware intelligent tutoring systems) and AUTMON (Automated therapy monitoring for children with autism spectrum disorder) and taken part in UE COST LUDI (Play for children with disabilities) and SHELD-ON (Smart habitat for the elderly) projects. She is a Manager in a project EMBOA affective loop in socially assistive robotics as an intervention tool for children with autism. Her research interests include making technology more humane, including topics of human–computer and human–robot interaction, accessibility and adoption of technology, user experience, and affective computing.

TERESA ZAWADZKA received the master's degree in the specialty of engineering systems and databases from the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, and the Ph.D. degree in computer science from the Faculty of Electronics, Telecommunications and Informatics, in 2009. She is currently an Assistant Professor with the Department of Software Engineering. The degree was awarded by the Faculty Council, which has also decided to distinguish the work. The specialty of the doctorate, knowledge management, is also the subject of her current research interests include data processing and data mining and big data. For the last two years, she has joined her passion for knowledge management with affective computing.

MICHAŁ ZAWADZKI received the Ph.D. degree in knowledge management from the Gdańsk University of Technology, in 2009. He is currently a Specialist in the field of computer science. Since graduation, he has been taking part in several projects, both EU (PIPS—Personal Information Platform in Life and Health Services) and national (SYNAT). For over a decade, he has been connected with business, successfully assuming the roles of a Software Engineer, an Architect, and a Scrum Master. Despite achieving all of this, he still cooperates with the academic environment by providing his practical experience.