An Adversarial Explainable Artificial Intelligence (XAI) Based Approach for Action Forecasting

Submitted: 6th April 2020; accepted: 21st August 2020

Vibekananda Dutta, Teresa Zielińska

DOI: 10.14313/JAMRIS/4-2020/38

Abstract:
Despite the growing popularity of machine learning technology, vision-based action recognition/forecasting systems are seen as black-boxes by the user. The effectiveness of such systems depends on the machine learning algorithms, it is difficult (or impossible) to explain the decisions making processes to the users. In this context, an approach that offers the user understanding of these reasoning models is significant. To do this, we present an Explainable Artificial Intelligence (XAI) based approach to action forecasting using structured database and object affordances definition. The structured database is supporting the prediction process. The method allows to visualize the components of the structured database. Later, the components of the base are used for forecasting the nominally possible motion goals. The object affordance explicated by the probability functions supports the selection of possible motion goals. The presented methodology allows satisfactory explanations of the reasoning behind the inference mechanism. Experimental evaluation was conducted using the WUT-18 dataset, the efficiency of the presented solution was compared to the other baseline algorithms.

Keywords: *Action prediction, Explainable artificial intelligence, Object affordances, Structured database, Motion trajectories*

1. Introduction

In the real-world scenarios forecasting the human action, before it is executed is a crucial problem. Such forecasting tool is needed for a wide range of applications in assistive, and social robotics.The recent events due to the global pandemic emphasized the role of service robots as health care assistants. Moreover, the robots, supporting the therapy of children with autism are employed to carry out social and assistive tasks, e.g., rewarding the person (by „musical dance" or „words of appreciation") if they performed the expected assignment (i.e., activities) without debacle. Such service robots are also useful for the therapy providing the guidance to the caretaker for avoiding any abnormal activities which can cause potential hazards. When developing the safe real-time human-robot interaction (HRI) it must be predicted what a person will do next [9]. Such ability requires tools and methods describing the temporal structure of human actions. For this purpose, several approaches such as the probabilistic methods, machine learning, or deep learning methods are widely used. Since the decision making is shifted from humans to machines, transparency and interpretability with reliable explanations are significant for getting a human trust in intelligent systems, the easiness of systems debugging and managing the ethical problems. With such capability (and with the transparency to the users) [15], the intelligent systems will be, able to plan ahead the responses with avoiding potential accidents or system faults.

Recent Machine Learning (ML) based intelligent systems are becoming increasingly complex, what makes difficult to the users to understand their actions [3]. Machine learning methods turn out to be un-interpretable „black boxes", which causes the problems with concluding about these systems robustness and reliability [1, 11]. Explainable Artificial Intelligence (XAI) is the method that is capable of explaining its own behaviour. XAI is known to have a positive influence on user trust in the understanding of the intelligent systems [14]. Fig. 1 illustrates the difference between traditional and XAI based reasoning. XAI thorough the explanations makes the underlying inference mechanism of an intelligent system transparent and interpretable for both: (a) the expert users (system developers) and (b) the non-expert users (end-users) [16,18,19]. It is worth mentioning that, the concept of XAI follows the workflow of the conventional machine learning approach in the „learning stage". However, the „application stage" offers interpretability of the learning mechanism, e.g., the significance of the applied features in the training stage, and how these features are mapped to the corresponding class label. Next, the explainability of the inference mechanism presents the influence of the decision system w.r.t. the selected features during classifications.

Forecasting human actions is a difficult problem that requires expertise in the area of robotics and artificial intelligence (AI). It involves the use of cognitive capabilities, e.g. perception, reasoning, prediction, learning, and planning, etc. and requires the semantics of the observed behaviour. The goal of this work is to create such capabilities for the robots to enhance their potentials to perform human service tasks and can help human beings with everyday activities. Having said that, such capabilities require human acceptance (i.e., trust, ethics, so on). Since such robotic platforms are lies at the interaction of human-robot interaction and machine intelligence. Therefore, the work concerning explainability and transparency of the prediction system is introduced to enhance human trust in the autonomous service robots.

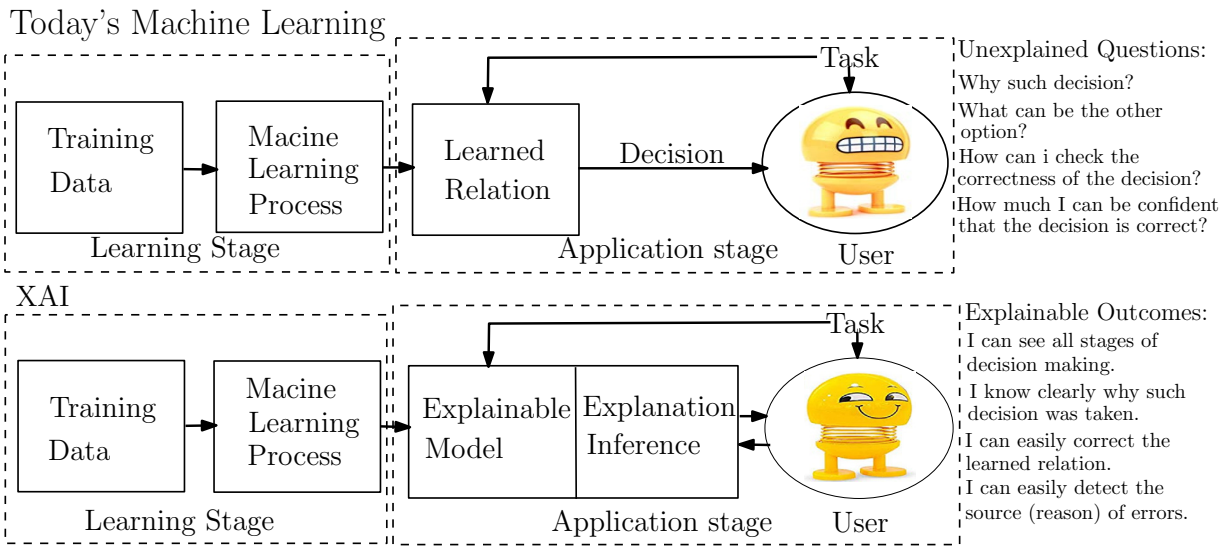


Fig. 1. The need for Explainable AI method in terms of transparency and interpretability [12]

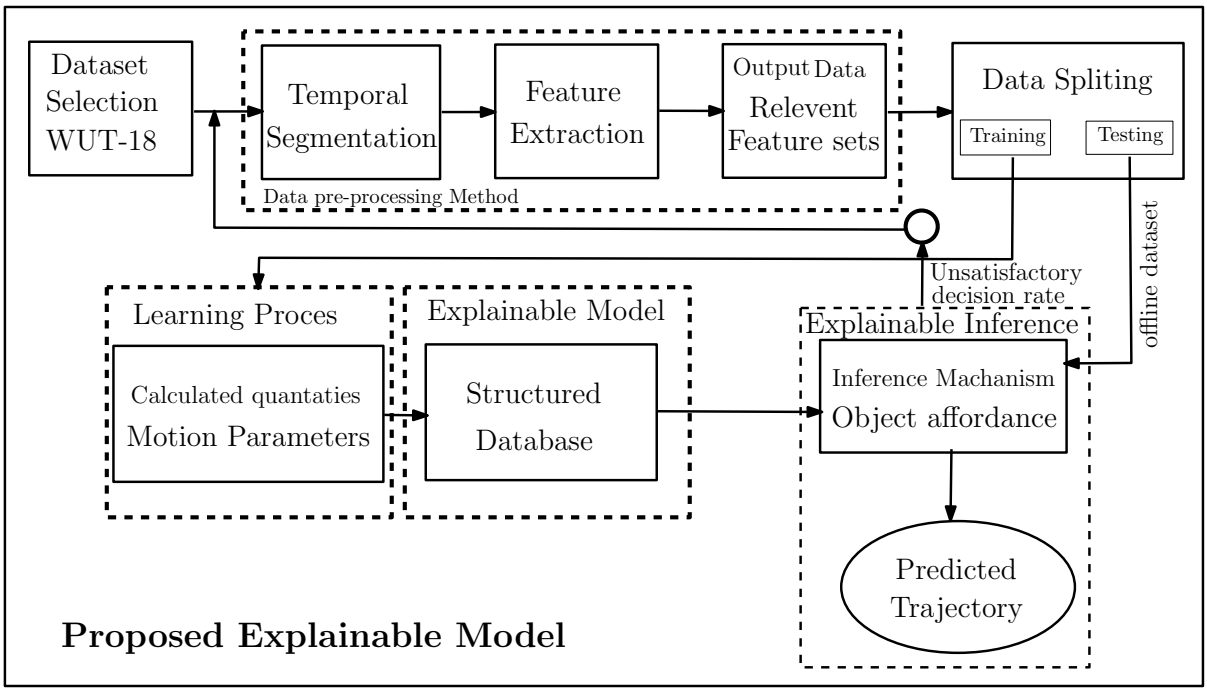


Fig. 2. General concept of proposed explainable method

2. Motivation and Objectives

Our work is motivated by the needs of trust system behaviour. The explainable systems are needed for actions forecasting in robotic autonomous servicing and care-giving. To address this challenge, we offer an adversarial Explainable Artificial Intelligence (XAI) approach for forecasting human actions using structured database and object affordances. The proposed approach was investigated in a supervised setting.

Comparing to our previous work [5], this paper focuses on the conceptual framework of explainability (i.e., XAI) on the following aspects. First, we give the short description of the formalization of the problem. Second, the definition of the structured database (explainable model) is proposed. Third, the object affordance explicated by the probability functions and reasoning (summarized in the graphs) is detailed. Finally, the comparison of the proposed method with the results obtained for other baseline methods is made.

The remaining part of the paper discusses this contribution in detail. Section 3 describes the proposed method focusing on the concept of the structured database and the object affordance. Section 4 presents the experimental results. The paper ends with the conclusions.

3. Proposed Method

3.1. General Definition

A human action is a state of doing. Since human action is a broader concept, for the sake of simplicity, only human actions involving objects are considered in our work. The proposed adversarial explainable ar-

tificial intelligence-based action forecasting consists of two phases: (a) the training phase (creating the actions model through gathering and processing the data with storing it in the database), and (b) the inference (prediction). The block diagram of the proposed method is depicted in Fig. 2.

Following [6], in this section, we give a general overview of the action prediction system. The scene is observed and the objects in the human vicinity are recognized. The objects are used as the discriminates for indicating which actions may be nominally taken by a human being. An action can be performed involving some objects. For example, a bottle, a cup, a box is placed on a table, therefore, for an action „reaching" involved object can be a bottle, a table, a cup, a box. Therefore, the objects are used as the discriminates for indicating which actions will be nominally taken by the human being.

First, we delineate the applied notations employed in this manuscript: (a) the capital letter (i.e. S, O) denotes a set, (b) the small letter (i.e. s, a, o) denotes an element of a set, (c) the upper script denotes the assignation, e.g., S^a means that the set S is assigned to a, (d) the lower script denotes the concrete element.

An action a_i is an elementary transformation of the human state. Therefore, potentially involved object o^{a_i} belongs to the set of all objects which can be involved in that action O^{a_i} ($o^{a_i} \in O^{a_i}$). The action is described by $a_i = a_i(s_{in}^{a_i}, s_{fin}^{a_i}, O^{a_i})$. If the specific object o_p is involved in action a_i, we denote it as a_i (o_p). Naturally, each action has its initial and final state, what is denoted by ($s_{in}^{a_i}, s_{fin}^{a_i}$), where $s_{in}^{a_i} \in S_{in}^{a_i}$, $s_{fin}^{a_i} \in S_{fin}^{a_i}$. It is to be noted that $S_{in}^{a_i}$ is a set of possible initial states and $S_{fin}^{a_i}$ is a set of possible final states. Introducing $S^{a_i} = S_{in}^{a_i} \cup S_{fin}^{a_i}$, the expression $a_i = a_i(s_{in}^{a_i}, s_{fin}^{a_i}, O^{a_i})$ can be rewritten as $a_i = a_i(S^{a_i}, O^{a_i})$.

When forecasting an action, we consider the scenario (observed scene). The scenario delivers the vocabulary. First, the objects are identified making the elements of vocabulary which is used in our database. Considering the set O of observed objects, based on the expression $a_i = a_i(S^{a_i}, O^{a_i})$, all possible actions are indicated. Lets $O = (o_p,\ o_w,\ o_z)$, where $o_p \in O^{a_i}$, $o_w \in O^{a_m}$, $o_w \in O^{a_j}$, $o_z \in O^{a_i}$, then the actions a_i, a_j and a_m will be indicated as possible.

3.2. Data Processing and Building the Structured Database (Explainable Model)

Referring to Fig. 2, the first step of the proposed method is the preprocessing of the recorded observations. The temporal segmentation and features extraction are made in this step. The goal of temporal segmentation of the video records is partitioned the recorded data for „discretized" and extracting of the relevant features from obtained data segments. We extract three groups of features: (a) human position h_p, (b) object position o_p, and (c) attributes describing human-object interaction: distance d, angle θ, and edge e. The temporal distance d and angle θ represent the distance and angular variable between human hand and the object of interest. Edge e which is the normalized distance obtained as the distance from the camera to the human hands normalized by the distance between the human hand and the object of interest. The description of the data processing phase is detailed in our previous work [4,10]. Here we are summarizing this step.

The mean value μ_i and variance σ_i^2 of temporal attributes (d, θ, e) is obtained during the preprocessing (training) phase. For each object which can be manipulated and for each action performed on it, the values of μ_d, μ_θ, μ_e, σ_d^2, σ_θ^2, σ_e^2 are calculated. Next, the database is created, the base is consisting of action lists taking into account the possible involved objects, as it is illustrated in Fig. 3a. As we can see, for each object, is given the list of possible actions. The database contains also the parameters (μ_i, σ_i^2) obtained from the recorded data segments [10]. With each addition of a new action or an object, the model must be additionally „trained" and the additional parameters (μ_i, σ_i^2) associated with the action (object) must be obtained. The quantities (μ_i, σ_i^2) are applied later as the parameters of the affordance (probability) functions used for prediction of a human hand motion trajectory and the motion target location. It brings clear interpretability of inference mechanism. During the application (or testing) phase once the objects are identified and recognized in the camera field of view, the database is accessed. Let us assume, that the objects $o_z = z$ and $o_w = w$ are noticed (to shorten the notation, the following abbreviations were introduced). The parameters of probability functions assigned to those objects are accessed (see Fig. 3a). Thereafter, considering the object affordance functions for o_w and o_z, the actions probability is obtained. The action with the highest probability is selected as the prediction. The possible future trajectories to the goal of interests are visualized by Bezier curves.

Fig. 3b illustrates the replicated representation of the database in graphical form reflecting the expressions $a_i = a_i(s_{in}^{a_i}, s_{fin}^{a_i}, O^{a_i})$ and $a_i = a_i(S^{a_i}, O^{a_i})$. The nodes represent the human states illustrated by initial states (i.e., s_{in}) or final states (i.e., s_{fin}) and the edges illustrate the transformations (actions - $a_{(.)}$). This graph built out of the sequence of consecutive actions a_i and a_{i+1}. Therefore, the final state s_{fin} of a previous action makes naturally an initial state s_{in} of a next action. The example graph is made for actions depicted in Fig. 3a. In this example possible sequences of actions w.r.t. the objects are: (a) $\{a_i(v), a_j(v)\}$ performed on an object v as it is shown in Fig. 3a, (b) $\{a_b(z), a_c(z), a_d(z)\}$ performed on one object z, (c) $\{a_n(w), a_k(w), a_d(w), a_p(w)\}$ performed on an object w respectively. Expanding the graph means the proper update of the database with properly feeding it with objects, actions and calculated quantities collected during the training phase [7].

3.3. Object Affordances Representation (Interpretability of the Inference Mechanism)

In this section, we discuss the inference mechanism of forecasting the human actions. The inputs are the depth information and the video data. Once the object is recognized and the features (d, θ, e) are

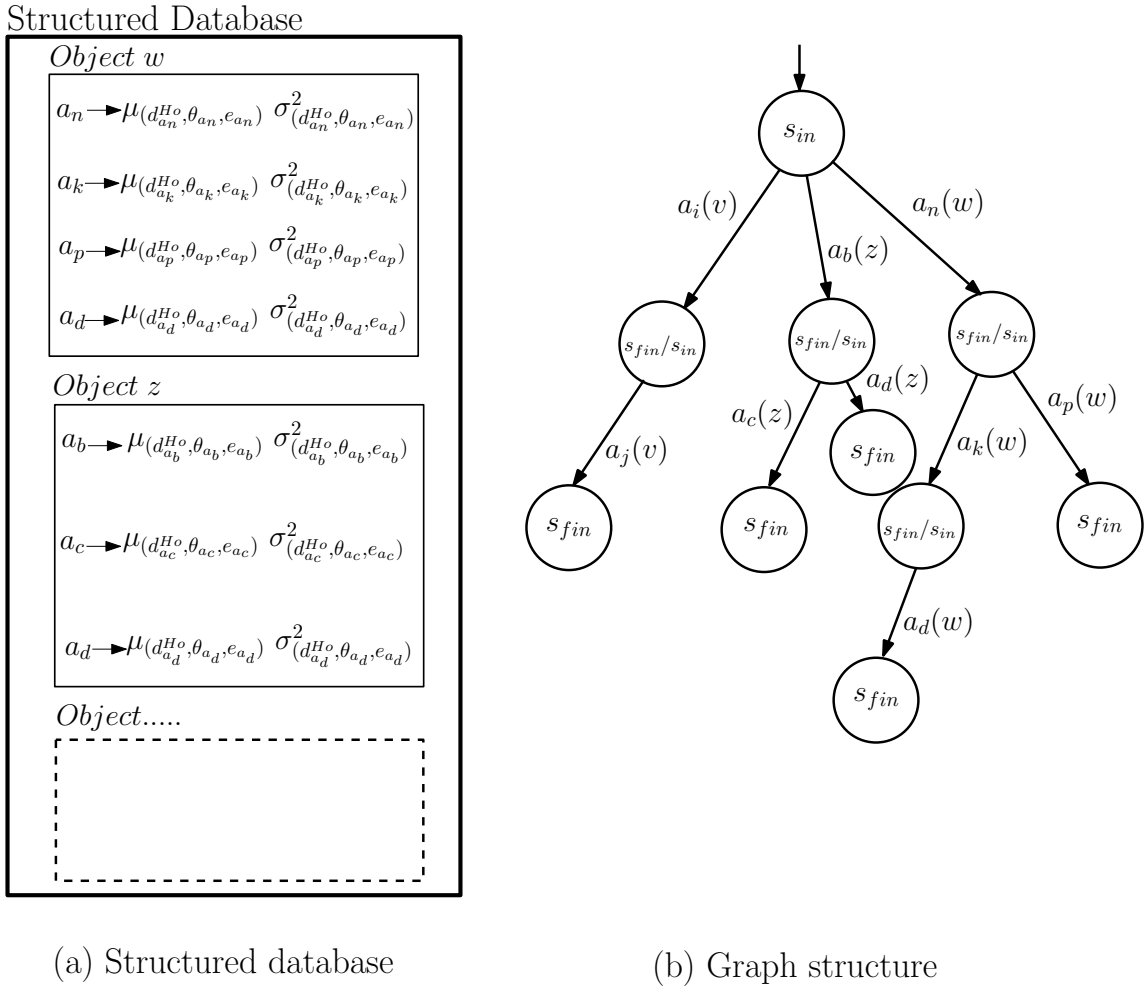


Fig. 3. Structured database and its replicated representation in the graphical form [7]

obtained, the set of actions associated with this object is considered (Fig. 3). Then the probability (value of affordance function) is calculated considering these actions. The affordance in our case results from the angular, distance, and edge preference considering the final state of an action (what in our case means the human hand position on the end of action). For the sake of simplicity, we can say that during the human hand motion as the most possible object to be manipulated (this is associated with the action) will be indicated. Such object to which the current distance d, angle θ, and the edge e are closest to (μ_d, μ_θ, μ_e). More precisely, applied probability functions will be delivering the probability of reaching each of the objects of interest providing for each of them probability created on the basis of current value of d, θ, e and the set of μ_d, μ_θ, μ_e, σ_d^2, σ_θ^2, σ_e^2. For each possible a_i for which the probability $p(a_k)$ is biggest is calculated using the Eq. 1. This is an action selection. Such action a_k is selected among all possible actions $a_i (i = 1, ..., n)$.

$$p(a_k) = \max_{a_i\{i=1,..,n\}} \begin{cases} (p^{a_i}(e) \cdot p^{a_i}(\theta)) & \text{for } d > 20cm \\ (p^{a_i}(d) \cdot p^{a_i}(\theta)) & \text{for } d \leq 20cm \end{cases} \quad (1)$$

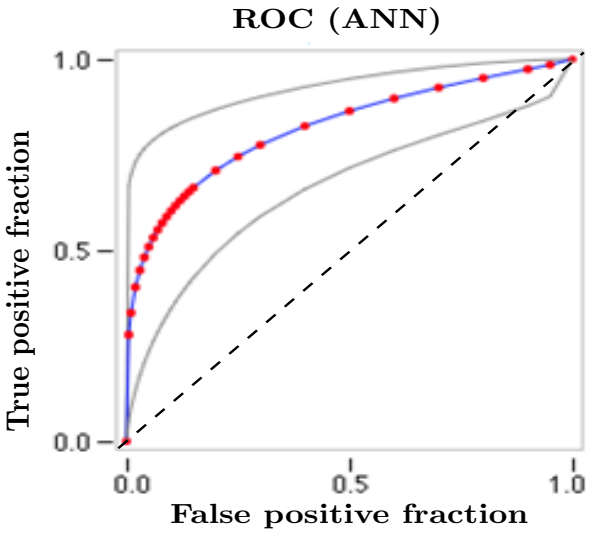
In our case, the threshold $20cm$ was selected heuristically noticing that for the hand being farther than $20cm$ from all the objects, any object can be targeted. Therefore, in this case the probability concerning the edge „preference" $p(e)$ (related to the easiness of motion) w.r.t. the action a_i is used. For the distance not bigger than 20cm the distance to the object is more relevant, therefore the probability considering the distance „preference" $p(d)$ instead of $p(e)$ is used, $p(\theta)$ takes into account the angular position towards the object (details are given in [7]) which represents the action a_i is contemplated. The forecasted trajectory is obtained using the parameterized cubic equation of the Bezier curve [7]. Detailed description of the above functions together with its validation is presented in our publication [7, 10].

4. Experimental Results

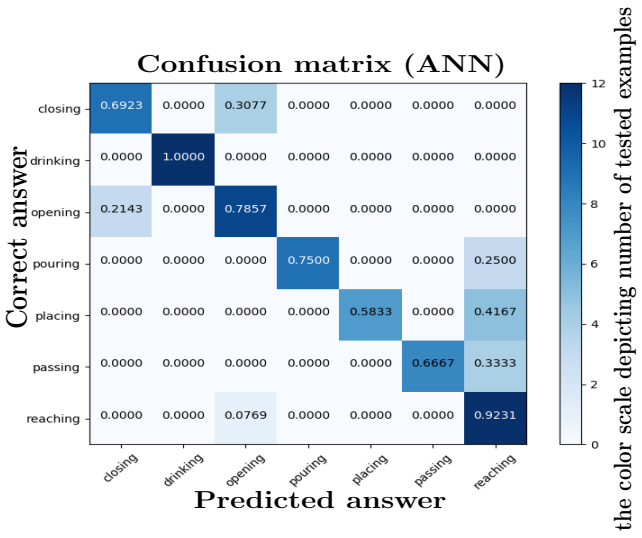
The proposed solution was evaluated using two methods: (a) a comparison with the state-of-the-art baseline algorithms (model test), (b) quality of prognosis depending on the amount of transparency of the decision system (explainability test).

The method was implemented using Intel Core i7 3.10GHz machine with 16 GB of RAM, with 64-bit Linux operating system. The C++ and python programming language (along with TensorFlow, Keras Packages) were used as a programming means.

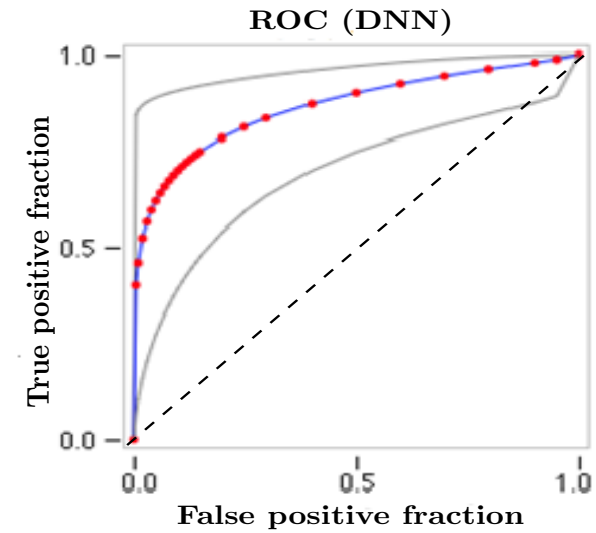
We created publicly available dataset (named as WUT-18) of the following daily activities: drinking wa-



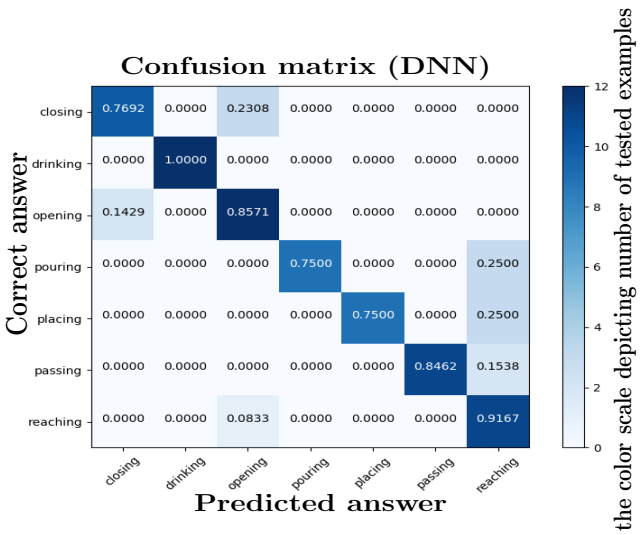
(a) Artificial Neural Network



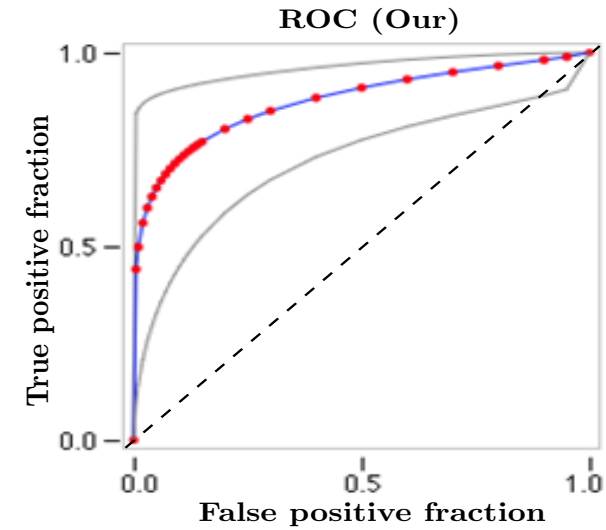
(b) ANN



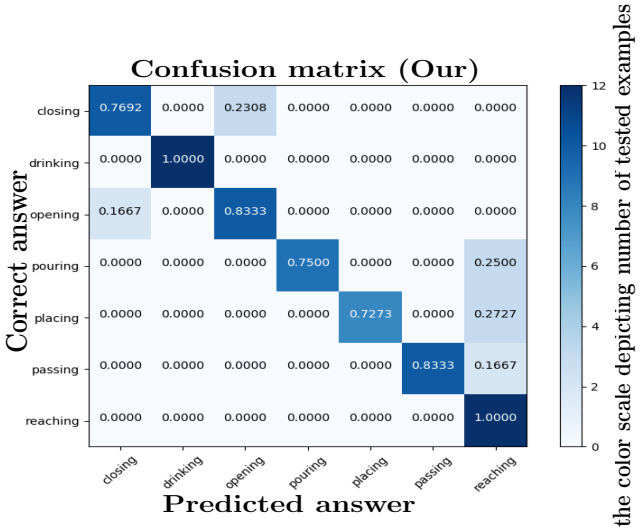
(c) Deep Neural Network



(d) DNN

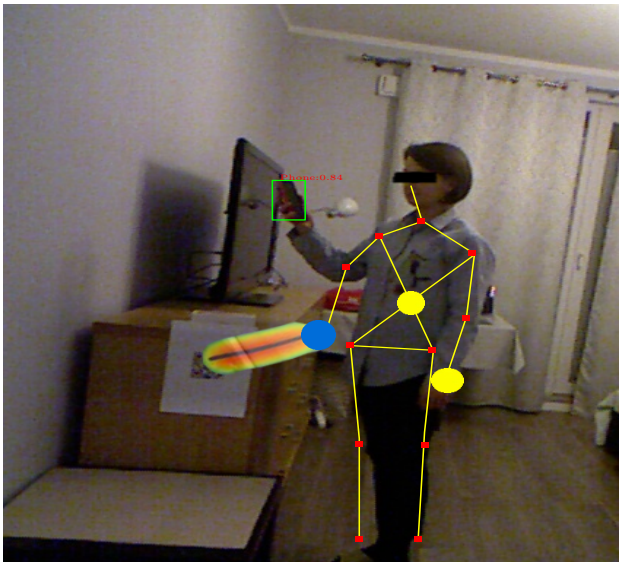


(e) Our Model

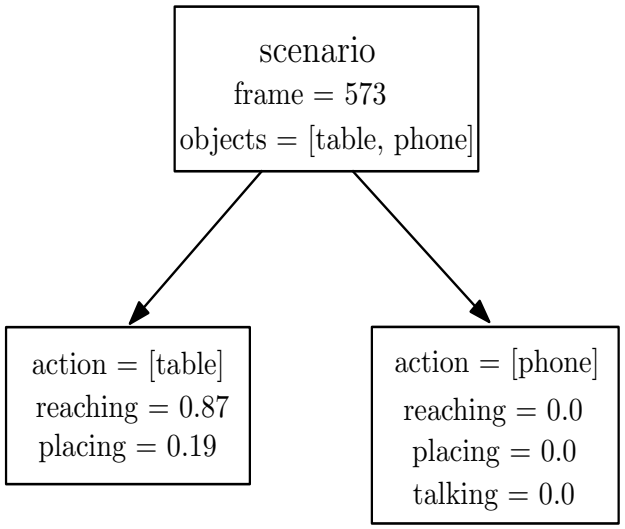


(f) Proposed Method

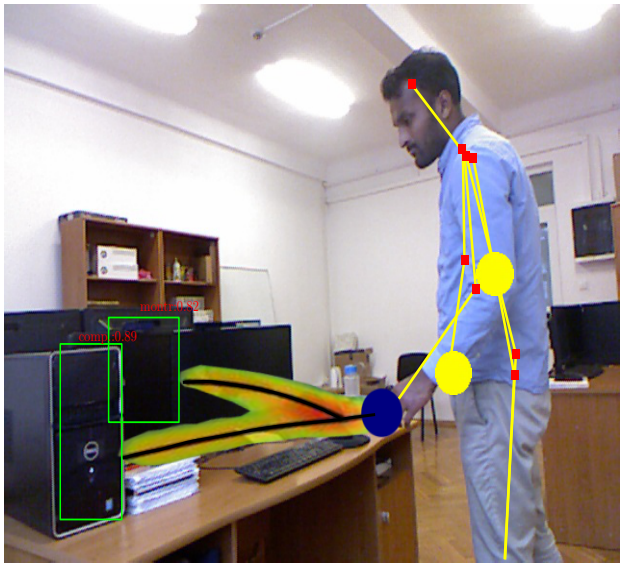
Fig. 4. The ROC curves (a, c, e) together with confidence intervals on the level of 95%. For better reference the straight dotted lines indicate the equal fraction of false and correct responses. The error matrices (b, d, f) for our method and the other baseline algorithms



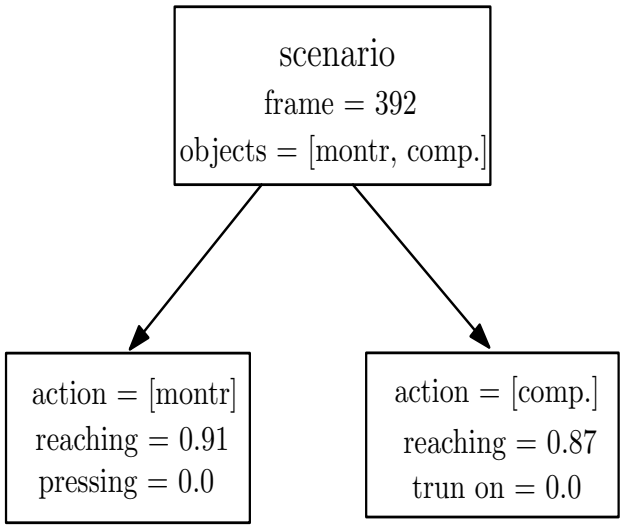
(a) Failed scenario



(b) Decision tree



(c) Successful scenario



(d) Decision tree

Fig. 5. Experimental results of different scenarios: (a) misclassified action, (c) correctly classified action. Results of the explanation process for inference mechanism using decision trees are depicted in b and d respectively

ter, activating computer, talking to phone, etc. These activities were performed by 6 participants in 3 different settings (a) an office, (b) a home, (c) a kitchen. The participants had neither prior knowledge of the purpose of the study nor instructions how to perform each activity. The data sets were collected under RGB-D settings, at the rate of 60fps. The cameras range for human observations was fixed and covered the relevant space.

Evaluating the proposed method we used the benchmark WUT-18 human activity dataset, and compared our method with two baseline algorithms: (a) Artificial Neural Network (ANN) based approach [13], and (b) Deep Neural Network (DNN) based approach [17]. Based on the trials and errors, we selected the parameters needed in the methods used for comparisons. The following parameters of ANN were defined: two hidden layers with 784 nodes each, activation function (ReLu, Softmax), optimizer (Adam), batch size (64), epochs (500), loss function (categorical-crossentropy). Similarly, in DNN the chosen parameters were: four hidden layers with following amount of nodes (560, 560, 256, 120), activation function (ReLu, Softmax), optimizer (Adam), batch size (64), epochs (500), loss function (categorical-crossentropy). The description of the parameters used in the baseline algorithms is given in [2, 8].

All these algorithms were tested using the same observation dataset. Evaluation results with respect to Receiver Operating Characteristic (ROC) curves and confusion matrices are shown in Fig. 4. *True positive fraction* means the number of correct responses normalised over the number of all samples(the decision is yes and the true response is yes as well), *false positive fraction* means the number of wrong responses normalised over the number of all samples (the decision is yes but should be not). As it is seen in Fig. 4a, 4c, 4e the amount of the correct decisions is significantly gre-

ater than the wrong decisions.

Fig. 4b, 4d, 4f is showing the so-called confusion matrix. The matrix contains the normalised numbers of predicted answers taking into account the correct answers. The color scale is visually depicting the number of samples for each result. The best accuracy was achieved for *drinking* action and the worst accuracy was achieved for *placing* action.

Fig. 5 shows the example images (left-hand side) together with visualization of the prediction process (right-hand side). This figure justifies the applied explainable approach which is visualizing the forecasting process. Fig. 5 illustrates the prediction for both: correct prediction and failed scenario. As it is seen the inference mechanism is commented displaying the correct decision. All objects considered as possible for performing an action are given (list – objects), moreover in each time instant the names of all possible actions are displayed (list – actions).

5. Conclusion

Due to the inability of explaining the decisions and actions, non-transparent machine learning algorithms should not be directly used in critical applications such as assistive robots and servicel robots. The wrong decisions of the system can result in harmful consequences. An explainability is needed when addressing such problems. To do this, an adversarial Explainable Artificial Intelligence (XAI) based method was proposed and discussed in this paper, the emphasis is laid on the explainability in the training and application stages.

The paper concerns very important and up-to-date problem of the broadly perceived artificial intelligence, namely the so-called explainable AI in which the reasoning process, analyses and actions undertaken are clearly visible and understandable for the human being. In such a way the models and procedures, as well as the results obtained, are trustworthy and hence much easier implementable. This paper can be viewed as proposing a conceptual framework and its proof, but not a complete final implementation.

The applied method was tested using a benchmark dataset WUT-18. Fig 5 delineates the use of proposed probabilistic approach in conjunction with explainability and interpretability. It offers enhancement in the transparency of the prediction system which makes our solution more comprehensive to the end-user. From the series of conducted experiments, it is also inferred that the proposed approach provides a significant improvement in terms of evaluation metrics when validated against pre-specified testing sets. Moreover, the following statistical significance tests are depicted in Fig. 4, we came to the verdict that the suggested approach outperforms the state-of-the-art baseline machine learning classifiers. Briefly, the proposed framework can eliminate the challenge of providing transparency of the decision system and offer an acceptable accuracy to forecast human actions. Obtained results were quantified and the method was validated as satisfactory. For selecting the possible actions we considered the probability functions based on the normal distributions. We expect to broaden the scope of applications focusing on the needs of a wide range of possible end-users.

AUTHORS

Vibekananda Dutta* – Institute of Micromechanics and Photonics, Faculty of Mechatronics, Warsaw University of Technology, ul. Św. Andrzeja Boboli 8, 02-525 Warsaw, Poland, e-mail: vibek@meil.pw.edu.pl, www: https://ztmir.meil.pw.edu.pl/web/Pracownicy/dr-Vibekananda-Dutta.

Teresa Zielińska – Institute of Aeronautics and Applied Mechanics, Faculty of Power and Aeronautical Engineering, Warsaw University of Technology, ul.Nowowiejska 24, 00-665 Warsaw, Poland, e-mail: teresaz@meil.pw.edu.pl, www: https://ztmir.meil.pw.edu.pl/web/Pracownicy/prof.-Teresa-Zielinska.

*Corresponding author

ACKNOWLEDGEMENTS

The research was supported by the funds of the Institute of Aeronautics and Applied Mechanics, Faculty of Power and Aeronautical Engineering, Warsaw University of Technology. The work on this manuscript and part of the research was also supported by the Preludium 11 (Grant No. 2016/21/ N/ST7/ 01614) funded by National Science Center (NCN), Poland.

REFERENCES

[1] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable Agents and Robots: Results from a Systematic Literature Review". In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, Richland, SC, 2019, 1078–1088, Montreal QC, Canada.

[2] F. Chollet, *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*, MITP Verlags GmbH: Frechen, 2018.

[3] M. G. Core, H. C. Lane, M. van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, "Building explainable artificial intelligence systems". In: *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI-06/IAAI-06*, 2006, 1766–1773.

[4] V. Dutta and T. Zielinska, "Action prediction based on physically grounded object affordances in human-object interactions". In: *2017 11th International Workshop on Robot Motion and Control (RoMoCo)*, 2017, 47–52, 10.1109/RoMoCo.2017.8003891.

[5] V. Dutta and T. Zielinska, "Action based activities prediction by considering human-object relation", *Prace Naukowe Politechniki Warszawskiej. Elektronika*, vol. 196, 2018.

[6] V. Dutta and T. Zielinska, "Activities Prediction Using Structured Data Base". In: *2019 12th International Workshop on Robot Motion and Control (RoMoCo)*, 2019, 80–85, 10.1109/RoMoCo.2019.8787354.

[7] V. Dutta and T. Zielinska, "Prognosing Human Activity Using Actions Forecast and Structured Database", *IEEE Access*, vol. 8, 2020, 6098–6116, 10.1109/ACCESS.2020.2963933.

[8] V. Dutta, M. Choraś, M. Pawlicki, and R. Kozik, "A Deep Learning Ensemble for Network Anomaly and Cyber-Attack Detection", *Sensors*, vol. 20, no. 16, 2020, 4583, 10.3390/s20164583.

[9] V. Dutta and T. Zielinska. "Predicting the Intention of Human Activities for Real-Time Human-Robot Interaction (HRI)". In: A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, eds., *Social Robotics*, vol. 9979, 723–734. Springer, Cham, 2016, 10.1007/978-3-319-47437-3_71.

[10] V. Dutta and T. Zielinska, "Predicting Human Actions Taking into Account Object Affordances", *Journal of Intelligent & Robotic Systems*, vol. 93, no. 3-4, 2019, 745–761, 10.1007/s10846-018-0815-7.

[11] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger. "Explainable AI: The New 42?". In: A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, eds., *Machine Learning and Knowledge Extraction*, volume 11015, 295–303. Springer, Cham, 2018.

[12] D. Gunning, "Explainable artificial intelligence (XAI)", *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, 2017.

[13] M. T. Hagan, H. B. Demuth, and M. Beale, *Neural Network Design*, PWS Publishing Co.: USA, 1997.

[14] M. Harbers, J. Broekens, K. van den Bosch, and J.-J. Meyer, "Guidelines for Developing Explainable Cognitive Models". In: *Proceedings of the 10th International Conference on Cognitive Modeling (ICCM 2010)*, 2010, 85–90.

[15] T. Lan, T.-C. Chen, and S. Savarese. "A Hierarchical Representation for Future Action Prediction". In: D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds., *Computer Vision – ECCV 2014*, volume 8691, 689–704. Springer, Cham, 2014, 10.1007/978-3-319-10578-9_45.

[16] Z. C. Lipton, "The mythos of model interpretability", *Queue*, vol. 16, no. 3, 2018, 31–57.

[17] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications", *Neurocomputing*, vol. 234, 2017, 11 – 26, https://doi.org/10.1016/j.neucom.2016.12.038.

[18] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences", *Artificial Intelligence*, vol. 267, 2019, 1–38, 10.1016/j.artint.2018.07.007.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, 1135–1144, 10.1145/2939672.2939778.