Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 29, Num. 02 (2022) 48-62

RESEARCH ARTICLE

Towards Causal Effect Estimation of Emotional Labeling of Watched Videos

Rumo à Estimativa do Efeito Causal da Rotulação Emocional de Vídeos Assistidos

Eanes Torres Pereira [1]*, Geovane do Nascimento Silva[1]

Abstract: Emotions play a crucial role in human life, they are measured using many approaches. There are also many methodologies for emotion elicitation. Emotion elicitation through video watching is one important approach used to create emotion datasets. However, the causation link between video content and elicited emotions was not well explained by scientific research. In this article, we present an approach for computing the causal effect of video content on elicited emotion. The Do-Calculus theory was employed for computing causal inference, and a SCM (Structured Causal Model) was proposed considering the following variables: EEG signal, age, gender, video content, like/dislike, and emotional quadrant. To evaluate the approach, EEG data were collected from volunteers watching a sample of videos from the LIRIS-ACCEDE dataset. A total of 48 causal effects was statistically evaluated in order to check the causal impact of age, gender, and video content on liking and emotion. The results show that the approach can be generalized for any dataset that contains the variables of the proposed SCM. Furthermore, the proposed approach can be applied to any other similar dataset if an appropriate SCM is provided.
Keywords: Affective Computing — Causal Inference — Pattern Recognition — Multimedia

Resumo: As emoções desempenham um papel crucial na vida humana, elas são mensuradas por meio de várias abordagens. Existem também várias metodologias para elicitar emoções. A elicitação de emoções por meio de vídeos assistidos é uma abordagem importante usada para criar conjuntos de dados de emoções. Contudo, o link causal entre o conteúdo do vídeo e as emoções elicitadas não foi bem explicado pela pesquisa científica. Neste artigo, apresentamos uma abordagem para computar o efeito causal de conteúdos de vídeos em emoções elicitadas. A teoria do Do-Calculus foi empregada para computar a inferência causal e um SCM (Structured Causal Model) foi proposto considerando as seguintes variáveis: sinal de EEG, idade, gênero, conteúdo do vídeo, gostar/não-gostar e quadrante emocional. Para avaliar a abordagem, dados de EEG foram coletados de voluntários assistindo a uma amostra de vídeos do conjunto de dados LIRIS-ACCEDE. Um total de 48 efeitos causais foi avaliado estatisticamente de modo a checar o impacto causal de idade, gênero e conteúdo de vídeo em gostar/não-gostar e emoção. Os resultados mostram que a abordagem pode ser generalizada para qualquer conjunto de dados que contenha as variavéis do SCM proposto. Além disso, a abordagem proposta pode ser aplicada a qualquer outro conjunto de dados similar se um SCM apropriado for fornecido.
Palavras-Chave: Computação Afetiva — Inferência Causal — Reconhecimento de Padrões — Multimídia

[1] *Unidade Acadêmica de Sistemas e Computação, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande (UFCG), Campina Grande - Paraíba, Brazil*
***Corresponding author**: eanes@computacao.ufcg.edu.br
DOI: http://dx.doi.org/10.22456/2175-2745.111817 • **Received:** 01/03/2021 • **Accepted:** 02/03/2022

1. Introduction

Emotions play a crucial role in human life, they influence to a great extent our decisions and quality of life. Emotions are measured using many approaches, such as responses to questionnaires and physiological signals acquired by diverse types of sensors (electroencephalography, electrocardiography, body temperature, etc). There are also many methodologies for emotion elicitation, such as playing music and presenting images and videos. Applications of emotion understanding range from Health Sciences to Marketing and Economy.

There is a great quantity of research being executed on classification of affect based on emotion elicitation through video watching. Some of the well-known EEG (Electroencephalography) datasets [1], [2] [3] for emotion classification

were constructed using data from people who watched videos. There are some unanswered questions in the Affective Computing area, for example:

- How do we know which are the specific video contents responsible for certain affective and liking responses?
- Is there any causal relationship between subject age, gender, and his affective state after watching a specific video content?
- Is there any causal relationship between subject age, gender, and his liking label assigned to a video?

The research presented in this paper aims to answer those questions using Causal Inference, specifically the Do-Calculus methodology developed by Pearl [4]. To achieve the objective, videos of the LIRIS-ACCEDE dataset were presented for volunteers, who were wearing EEG sensors on their scalps. Each volunteer labeled their emotional states after watching each video in terms of valence, arousal and liking. A DAG (Directed Acyclic Graph) was created to model the causal relations among the studied variables and Do-Calculus was employed to measure the causal effects.

Do-Calculus allows to draw causal inference about experimental interventions from observational data. Therefore, by using Do-Calculus without having to perform the experiment, but from data collected from other observational studies it is possible to answer a question like the following: What would the valence of the emotion felt by the subjects be if they had watched a video containing dogs? Causal Inference is not new in the sciences and Philosophy, and Do-Calculus is not the only existing approach for causal inference. Another well-known approach, specially in Psychology, is the Propensity Scores [5]. The Do-Calculus was chosen to analyze the results obtained in this research because its usage of DAG (Directed Acyclic Graphs) makes it more intuitive and allows a graphical definition of causation [6].

Therefore, the goal of this research is to present a methodology to evaluate the causal effect of specific video contents on affective and liking states of subjects who watched videos. Furthermore, this article also presents the results of the proposed methodology applied to the evaluation of causal effects of age and gender on affective and liking states.

The outline of the paper is composed by 6 sections. Section 2 presents the related work on video concepts and on EEG-based age, gender, and emotion classification. The Do-Calculus basic theory is presented in Section 3. Section 4 presents the proposed approach. Experimental results are presented and discussed in Section 5. Finally, Section 6 presents the conclusions of the research.

2. Related Work

This section is composed by three subsections. Two of them present researches on EEG-based classification of emotion, age and gender. The last section presents research on video concepts. As video concepts extraction and classification are a fundamental topic for this research, it was decided to include a review of related work on that subject matter.

2.1 EEG-based Emotion Classification

Current researches have demonstrated that EEG signals carry information about age and gender [7] [8], The frequencies of EEG signals are commonly grouped in bands, such as: delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (14-30 Hz), gamma (over 30 Hz). There are many approaches proposed for EEG-based emotion classification. In the beginning (before the rising popularity of Deep Learning approaches/techniques), researchers used handcrafted techniques for extracting features like entropy, PSD (Power Spectrum Density) and HOC (Higher Order Crossings) [9] for training classifiers.

There are many EEG datasets proposed for emotion classification, for instance DEAP [1], MAHNOB-HCI-Tagging [2], and STEED [10]. Although there is a lack of consensus in some methodological aspects (e.g., quantity of channels, position of channels, frequency bands, signal duration, signal features, etc), one may say that is possible to classify emotions using EEG signals. This subsection presents a review on recent research on the topic of EEG-based emotion classification.

An important evidence for the possibility of EEG-based emotion classification is the existence of works evaluating all the features that are being used for the classification task. For instance, Jenke et al. [11] affirm that when they were writing their article, there was no systematic comparison of features for EEG-based emotion classification, therefore that was the objective of their article. According to Jenke et al. [11], there is no agreement in literature about which features are most appropriate for emotion classification. Of course, they are not considering the use of end-to-end (E2E) Deep Learning approach in which the same architecture extracts features and trains a classifier. In order to evaluate feature extraction and feature selection approaches, Jenke et al. [11] collected their own EEG dataset from 16 subjects (nine male) with ages between 21 and 32 years. The duration of the signals for each subject was 30 seconds and the signals were labeled in five emotions: happy, curious, angry, sad, and quiet. For emotion induction, authors used IAPS [12] pictures, and for classification they used Quadratic Discriminant Analysis (QDA). Results showed that multivariate feature selection techniques performed better than the univariate counterpart. One of the best feature extraction methods was the HOC (Higher Order Crossings), which performed better than spectral power bands approaches. Two main extensions of the Jenke et al. [11] research could be: To apply the methodology for emotion induction by videos, and to evaluate features across multiple datasets.

As there are multiple EEG channels available for processing, it is a valid idea to think that neighbor channels share information and, therefore, they use an approach as a DCNN

(Deep Convolutional Neural Network) to process EEG data as it was performed by Putten et al. [8]. Another crucial source of information in EEG is in the time dimensional. Therefore, another good idea would be to use a recurrent neural network to explore temporal information. Fourati et al. [13] proposed to deal with the matrix of EEG data as a multidimensional time series and process it for emotion classification using an Echo State Network (ESN), which is a type of recurrent neural network. ESNs, in general, are composed by three layers: input, hidden, and output. The hidden layer is called the reservoir. To deal with the problem of ESN's random initialization, Fourati et al. [13] performed unsupervised learning of the random reservoir before the output layer learning. Fourati et al. [13] used preprocessed version of DEAP dataset in their experiments. Four types of classification experiments were performed: valence, arousal, emotional stress, and states/calm. Two types of inputs were given to the proposed approach: raw signal and 4 bands filtered using daubechies db5 with 5 levels wavelet decomposition. Results of Fourati et al. [13] were compared with some state of art works. In all cases, except for stress/calm results, Fourati et al. [13] proposed approach achieved highest accuracy. Fourati et al. [13] did not mention the unbalance of DEAP dataset and only evaluated the experiments using accuracy. If unbalance is not dealt adequately it may compromise results, as it was exposed by Pereira et al. [14].

Although, there are many recent approaches for EEG-based emotion classification using some variation of Deep Learning (DL), SVM are still one of the most popular classifiers for that task. Liu et al. [15] constructed a standardized database of 16 emotional videos. As their main objective was not to propose a new classification approach, they employed the well-known feature extraction and SVM classification paradigm for EEG-based emotion classification. One important contribution of Liu et al. [15] work is that their database was constructed to evoke target-specific emotions. EEG signals were collected using Emotiv Epoc Headset. The dataset was constructed under supervision of nine research assistants who majored in Psychology, and three specialists in the area of emotion elicitation evaluated the film which were selected by the nine research assistants. Afterwards, 462 Chinese-speaking students (195 males) age ranging between 18 and 30 years. watched and rated the videos. The duration of the final selected videos was between 1 and 3 minutes. For EEG acquisition the number of volunteers was 30, and all volunteers were male with age ranging between 19 and 26 years. The standardized database was used to elicit seven emotions and neutrality. The features extracted from EEG for classification were: PSD, SLDA (Sparse Linear Discriminant Analysis), and asymmetry features. For eight-class classification, anger versus disgust versus fear versus sadness, the accuracies were low: 32.31% (0.46%) and 65.09% (0.66%). However for non-neutrality versus neutrality, positive versus negative, and amusement versus joy versus tenderness the accuracies were higher than 86%. The emotion elicitation video dataset proposed by Liu et al. [15] has the drawback of the language of videos be in Chinese, which does not allow replication of results by researchers in other countries. However, the EEG signals were collected from a high quality procedure and is well-balanced for the emotional labels, with 2 videos for each emotion: Joy, amusement, tenderness, anger, sadness, fear, disgust, and neutrality.

Song et al. [16] proposed to use a variation of GCNN (Graph Convolutional Neural Network) in order to exploit the relationships of information present in different EEG channels for improving emotion classification. They called their approach DGCNN (Dynamical Graph Convolutional Neural Network). A graph may be represented by an adjacency matrix, and the DGCNN may learn the adjacency matrix of EEG channel relationships dynamically. The network architecture is composed by four main layers: graph filtering layer, convolutional layer, Relu activation layer, and full connected layer. Five EEG features were extracted from signals and given as input to the DGCNN. Five types of features were evaluated: PSD (Power Spectrum Density), DE (Differential Entropy), DASM (Differential Asymmetry), RASM (Rational Asymmetry), and DACAU (Differential Causality).The proposed approach was evaluated using two datasets: SEED and DREAMER. Results for the SEED dataset were compared with other state of art approaches. For PSD features, the proposed approach achieved the highest accuracy. In subject independent (LOSO: Leave-one-subject-out) experiments, the proposed approached achieved the highest accuracy only using PSD features extracted from delta band. When using a combination of features from all bands (alpha, beta, gamma, theta, and delta) the best result was achieved for LOSO using DE features as input for DGCNN (79.95%).

An important question was addressed by Zheng et al. [17]: which EEG patterns are stable among different subjects and EEG collection sections? Zheng et al. [17] point out that *the stability of patterns and performance of models over time has not been fully exploited*. To allow a better evaluation of EEG pattern stability, the authors proposed a new dataset (called SEED) with an important methodological difference compared to existing datasets: the data were collected from the 15 volunteers, three times with intervals of one week or longer. EEG Feature extraction followed a procedure similar to that followed by Song et al. [16]. The extracted features were given as input to a GELM (Discriminative Graph Regularized Extreme Learning Machine) and, for comparison, to a SVM. As occurred in Song et al. [16] experiments, the highest results were obtained by the classifier trained with DE features. The dataset was also evaluated, subject-dependently, using training data for one of each section to test data from one of the three sections. There are many accuracy results higher than 80%, but for some subjects the accuracies were near 50% (subject 5, for example).

2.2 Age and Gender EEG-based Classification

One of our main assumptions in this research is that the EEG signals carry information about subject gender and age which could be used for train a classification algorithm. Therefore, in this section we present papers from research in neuroscience corroborating our assumption.

Nguyen et al. [7] proposed an approach for age and gender classification using features extracted from EEG signals obtained from the Australian EEG Database [18]. The Australian EEG Database is comprised of EEG recordings of 40 patients, 20 male and 20 female with ages between 19 and 69 years. From those signals two types of features were extracted: (i) Power Spectral Density, 11 coefficients of an 11th-order Auto-regressive model and 3 Hjorth parameters (activity, mobility and complexity); (ii) MFCC (Mel-frequency cepstral coefficients), LFBP (Log filter-bank powers), and LSP (Line spectral pairs). Nguyen et al. [7] called the second type of para-linguistic features, because that type of feature is employed for speech classification. The features were extracted from 8 EEG channels (F3, F4, C3, C4, P3, P4, O1, and O2) and were given as input to SVM (Support Vector Machine) classifier. For age, there were 3 classes (Young, Middle, and Elderly), and for gender two classes (Male and Female). All the accuracy results obtained in the experiments were higher than 96%.

The research presented by Nguyen et al. [7] is the traditional way of pattern recognition: extracting features and training a classifier. With the advancement of Deep Learning, more and more researchers are going directly to algorithms which are able of performing feature extraction and classifier training simultaneously. One of those algorithms was implemented as Deep Convolutional Neural Networks (DCNN) and was employed by Putten et al. [8] in order to classify people gender using EEG signals obtained from their scalps. Putten et al. [8] used EEG from 1308 subjects (mean age 43.38 years and 47% males) to train a DCNN which has the following architecture details:

- Input layer: 2 seconds of EEG from 24 channels, corresponding to 256×24, because sampling rate was 128Hz;
- Intermediate layers: convolutional layer of 50-300 3×4 filters, 2×2 pooling layer, and 25% dropout layer;
- Output: two classes: male or female.

The accuracy obtained by Putten et al. [8] were around 80%, which is 16% lower than the accuracy obtained by the approach proposed by Nguyen [7] which was published 5 years before. However, the main motive why Putten et al. [8] results are important is because they did not performed feature engineering. The same algorithm which selected the features was used to train a classifier. That is one of the main motives to use DCNN in this type of research. In a near past, researchers discussed which were the best EEG features for classifying a specific pattern [11]. Nowadays this discussion is changing.

Tomescu et al. [19] explain that the structural changes which occur in brain during aging are well documented, furthermore males and females differ in terms of brain volume, grey/white matter ratio, regional cerebral blood flow. However the differences in dynamics of mental activities changes in relation to gender and age are not yet well understood. In order to set some light on that problem, Tomescu et al. [19] proposed to use the concept of EEG micro-states (periods of about 100ms of stable scalp potential fields) to study gender and age changes in EEG. In their study, Tomescu et al. [19] used 204-channel EEG signals from 179 subjects (6-87 years old, 90 female). In order to select the periods for extracting the EEG micro-states, the local maxima of the Global Field Power (GFP) was computed, and the EEG signal was cut only at the corresponding region of GFP peaks. The GFP peaks were submitted to k-means clustering algorithm. Using a fitting process, Tomescu et al. [19] determined: the mean duration of micro-states, the occurrence of each micro-state, and the transition between micro-states. Micro-states were further grouped in four categories. ANOVA (Analysis of Variance) results showed a general tendency for increased duration of the micro-states with age, and for one of the four micro-states categories there was a difference between males and females.

Vandenbosch [20] studied whether it is possible to predict the age of children and adolescents using EEG recordings. Two datasets were used: Netherlands Twin Register dataset (NTR, 836 subjects), and Washington University in St. Louis dataset (WUSTL, 702 subjects). The subject ages varied from 5 to 18 years. Power Spectrum Density (PSD) was extracted from 12 EEG channels (F3, F4, F7, F8, C3, C4, P3, P4, P7, P8, O1, and O2) and it was given as input for three types of classifiers: random forest (RF), support vector machines (SVM), and relevance vector machine (RVM). As there were data of subjects from the same family in the datasets, the data were reduced by excluding duplicate subjects of the same family. A six-fold-cross validation was employed and repeated a number of time equal to the quantity of different subjects from the same family, selecting only one subject per family in each iteration. Two types of age classifications were performed: child-adolescence, ages (5, 7, 12, 14, 16, 18). For child versus adolescence classification the accuracy was higher than 93%, and for age categories the MAE (Mean Absolute Error) was 1.22 years for RF and 1.46 years for RVM, while SVM regression did not perform well. The experiments performed by Vandenbosch [20] showed that using popular EEG features and classifiers it is possible to achieve good age classification results.

Hu [21] performed some experiments to evaluate the feasibility of classifying human gender using EEG signals. Hu [21] used data collected from 28 healthy subjects (13 male) with ages between 18 and 30 years. The data were digitized at 1000Hz from 32 channels. For each subject, five minutes of signal were sectioned in one second periods. Therefore, the dataset was augmented to 8400 samples ($28 \times 5 \times 60$). Four types of entropy features were computed: fuzzy entropy (FE),

sample entropy (SE), approximate entropy (AE), and spectral entropy (PE). Six different classifiers and three different ensemble of classifiers were trained: K-Nearest Neighbors (KNN), Logistic Regression (LR), Quadratic Discriminant Analysis (QDA), Decision Tree (DT), Bagging, Boosting, and vote classifier. Although Hu [21] says that he used 10-fold-cross-validation, there is no information about any caution on the fact of samples from the same person occurring in the training and test sets. Perhaps, there is some correlation among samples coming from the same subject and it would be a good practice to avoid different samples of the same subject contained in the training and test sets simultaneously. When evaluated isolated, all classifiers trained with a single type of feature produced accuracies higher than 65%. There was also classifier training using all features as input, in this case all accuracies were higher than 95%. The accuracies for bagging and boosting were all higher than 98%. Hu [21] obtained high accuracy results for gender classification using EEG, the only criticism to his work is the absence of caution of dealing with signals coming from the same subjects in the training and testing datasets.

Kaur et al. [22] used wireless EEG device (Emotiv Epoc Plus) for collecting data from 60 subjects with ages between 6 and 55 years. The wireless EEG device records signal from 2 references and 14 channels: AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, O2. The used sampling frequency was 128Hz. Signals were smoothed using a least square filter (Savitzky-Golay filter) and they were processed by Discrete Wavelet Transform (DWT) to extract five frequency sub-bands: delta, alpha, beta, gamma, and theta. Three features were extracted from each frequency band: mean, energy, and root-mean-square. A random forest classifier used the features to train and testing. The dataset is well-balanced for quantity of subjects across age ranges, which were: 6-10, 12-15, 18-23, 25-29, 33-38, 42-55. But there was a small unbalance for gender classes (35 male and 25 female). For all band waves, the used classifiers had higher accuracy for gender classification than for age classification. Theta, Delta and Gamma bands had the highest accuracies. Kaur et al. [22] analysed the accuracies changing signal duration in steps of 1 second, from 1 to 10 seconds. The highest accuracies were between 7 and 8 seconds. A comparison with two other classifiers was also performed: SVM (Support Vector Machine) and ANN (Artificial Neural Networks). The results for Random Forest were higher than the results for SVM and ANN in all tested cases.

A well-known problem in training deep learning classifiers is their need for great amount of data. Putten et al. [8] used EEG data collected from 1308 subjects in order to train their DCNN approach. In order to explore the temporal information, a deep learning approach that may be more appropriate for EEG signal processing are the long short-term memory classifiers (LSTM) which are recurrent artificial neural networks. Kaushik et al. [23] used the dataset produced by Kaur et al. [22] which contains data from 60 subjects (35 males) with ages ranging from 6 to 55 years. Data were collected for 10 seconds using a well-known commercial EEG device: Emotive Epoch plus. Therefore, only 14 channels were used (due to the capacity of EEG device). Kaushik et al. [23] preprocessed the signals using the Daubechies Db-8 Wavelet transform, resulting in four wavelet coefficients to noise and five wavelet coefficients corresponding to alpha, beta, delta, gamma and theta brain waves. Three Deep Learning Architectures were investigated: Long Short Term Memory (LSTM), Bidirectional LSTM (BLSTM), and BLSTM-LSTM.

Two types of experiments were performed: Using raw signal, and using wave-bands obtained by wavelet processing. The higher accuracies were obtained for alpha and beta bands, and alpha accuracy was higher than beta accuracy. The alpha accuracy for age in the proposed model was very near the BLSTM result (91.31% against 91.96%, respectively). For gender, beta waves were better than alpha waves, with results of 97.5% and 95.5% for BLSTM and BLSTM-LSTM, respectively. The research of Kaushik et al. [23] shows the feasibility of age and gender classification using EEG. An important factor to emphasize here is that the better frequency bands for the results obtained by Kaushik et al. [23] were discrepant with the results obtained by Kaur et al. [22]. Kaur et al. [22] research pointed to better results in beta and theta for age classification and delta wave for gender classification, whilst Kaushik et al. [23] results pointed to alpha and beta bands having the highest results. The discrepancy must be further investigated.

2.3 Video Concepts

There are two main types of video temporal units: shots and scenes. According to Sidiropoulos et al. [24], the main feature in a shot is the sequence of images taken without camera interruption, and scenes are longer higher-level temporal segments. Shot detection, for example, may be performed by detecting video editing effects. However, scene detection is a challenger problem because it is based on semantic criteria. Sidiropoulos et al. [24] proposed an approach to decompose videos into scenes, which is based on a multi-modal scene segmentation technique called Generalized STG-based (GSTG). The technique proposed by Sidiropoulos et al. [24] exploits features from visual and auditory channels. The most import module of the approach is the STG (Scene Transition Graph), which computes similarities between key-frames and construct a connected graph. As the STG presents a high computational complexity due to compute similarities between every two shots, Sidiropoulos et al. [24] proposed an approximation algorithm to limit the number of shot pairs to be used in computing similarities. Sidiropoulos et al. [24] also proposed a variation of STG to low-level audio features. The STGs algorithms for audio and visual features were combined to form the GSTG. The approach was evaluated on two datasets: Netherlands Institute for Sound & Vision dataset, and the TRECVID dataset. The F-score of the experimental results were higher than 86%.

According to Sidiropoulos et al. [25], a video concept is a high-level video-feature, and concept detection means: *estimating for each concept a degree of confidence in the hypothesis that this concept is suitable for describing the contents of a given elementary piece of a video stream*. The video concept detection problem is challenging because it is an open set problem: the number of different possible concepts is unknown a priori for any video. Sidiropoulos et al. [25] affirms that the majority of approaches focus on extracting image features (like SIFT, SURF, DAISY, etc) from key-frames, and training classifiers such as SVM to create base detectors. In their work they proposed to use, as another source of data for training base detectors, the video tomographies [26], which are spatio-temporal slices with one axis representing time and another axis representing space. The video tomographies were used in conjunction with visual features extracted from key-frames. Furthermore, Sidiropoulos et al. [25] proposed to use genetic algorithms for selecting subsets of best-performing base detectors, they used 37 linear SVM classifiers as base detectors. The approach was tested on the 2011 and 2012 TRECVID SIN Tasks datasets, which contained 50 and 46 concepts for evaluation, respectively. The experimental results were evaluated in terms of Mean Extended Inferred Average Precision (MXinfAP), and using all concepts or motion-related concepts only. In all cases, the quantity of base detectors was lower for the selection among 37 base detectors, and in almost all cases the MXinfAP was superior for the approach.

Apostolids and Mezaris [27] proposed an algorithm for fast temporal video segmentation into shots, they affirm that there are two types of shots, abrupt and gradual, and their algorithm is able to detect the two types of shots. The approach is executed in three steps: (i) computation of video frames similarity to detect abrupt transitions; (ii) detection of gradual transitions; and (iii) filtering of shots detected wrongly due to object/camera movement or camera flash-lights. The video dataset was comprised of 15 videos obtained from German public broadcaster RBB, cultural heritage show of the Dutch public broadcaster AVRO, and videos from the archive of the Netherlands Institute for Sound and Vision. The ground-truth was created by human annotation. The results were evaluated in terms of precision, recall and F-Score. For the three metrics, results were higher than 88% and higher than other approaches selected for comparison.

Markatapoulou et al. [28] affirms that *semantic video concept detection in video aims to annotate video fragments with one or more concepts chosen from a predefined concept list*. A crucial step in video concept detection is segmenting the video in shots. A common practice in concept detection is to combine classifiers trained with different features (late fusion), but due to unbalance in the quantities of different concepts, the classifier combination may be challenging. To deal with the classifier combination problem and to better integrate handcrafted features and DCNN-based features, Markatapoulou et al. [28] proposed to use a cascade of classifiers. The classifier cascade approach became famous after Viola and Jones used it for object detection [29]. An important characteristic of cascade approaches is organizing classifiers in stages in such a way that the more computational consuming operations are in the latter stages. The DCNN architecture used was the 16-layer pre-trained DCNN proposed by Simonyan and Zisserman [30]. Authors used features extracted by the last hidden layer of the DCNN. The classifiers were evaluated in terms of MXinfAP. The classifier trained using only DCNN features obtained higher results than using handcrafted features (ORB, SIFT, SURF). However, the highest MXinfAP was obtained when the classifier used a combination of ORB, SIFT and DCNN features.

According to Markatopoulou et al. [31], there are two main categories of methods for considering the relationships between concepts: modelling the label relationships, and exploiting the task relationships. Markatopoulou et al. [31] proposed to use a variation of the ELLA (Efficient Lifelong Learning Algorithm) [32] algorithm for video concept detection because they affirms that MTL (Multi-task learning) algorithm might be appropriate for semantic concept detection. They called their ELLA approach of ELLA_LC (LC means Label Constraint). Experiments were performed using the TRECVID 2013 SIN dataset, which contains 800 hours of video for training and 200 hours of video for testing. Four DCNN classifiers were used for feature extraction: an 8-layer CaffeNet; a 16-layer ConvNet; a 22-layer GoogLeNet; and a DCNN trained by the authors with architecture similar to the 22-layer GoogLeNet. In the tests, the MXinfAP was evaluated for 38 concepts. In all cases, except one of them, the proposed approach achieved highest results.

Markatopoulou et al. [33] point out that the existing approaches for video concept annotation do not consider semantic relationships or inter-dependencies among concepts. Therefore, they proposed a DCNN architecture to explore both implicit and explicit concept relationships (visual-level and semantic-level). Implicit concept relationships were modelled using Multi-task Learning (MTL) to learn shared feature vectors encoded in DCNN layers. Explicit concept relations were modelled using a new cost function (CCE-LC: Cost Sigmoid Cross-entropy with Label Constraining) for a set of DCNN layers which exploits correlations between concepts. Experiments were performed on four datasets: TRECVID-SIN 2013, PASCAL-VOC 2007, PASCAL-VOC 2012, and the NUS-WIDE. The problem was defined as: given a concept, retrieve the 2000 video shots which are mostly related with it. The best MXinfAP results achieved by the proposed approach were 33.77%, 87.00%, 88.69% and 60.73% respectively for the following datasets: TRECVID-SIN, PASCAL-VOC 2007, PASCAL-VOC2012, and NUS-WIDE.

Vasileios Mezaris and associated researchers developed a web service called VideoAnalysis4ALL [1] which allows the upload of videos for shot and scene segmentation, and visual

[1] Available at ⟨http://multimedia2.iti.gr/onlinevideoanalysis/service/start.html⟩. Last access: April 16, 2020.

concept detection. The papers reviewed in this subsection were published by the team of Vasileios Mezaris, who is Senior Researcher in the Multimedia Knowledge and Social Media Analytics Laboratory of the Information Technologies Institute. The VideoAnalysis4ALL was used for detecting the concepts used in the research presented in this article.

3. Do-Calculus and Structured Causal Model

The Causal Inference approach used in this paper is mainly based in Structured Causal Model (SCM) and in Do-Calculus, as proposed by Pearl et al. [6] [4]. A SCM is a mathematical structure $S = \{U, V, F, G\}$, where:

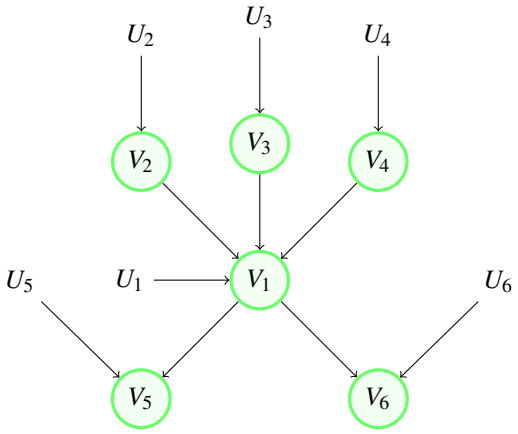
- U is a set of exogenous variables, i.e., variables that could not be measured or variables that are not directly present in the model;
- V is a set of endogenous variables, i.e., variables that are present in the model and which can be measured and manipulated directly;
- F is a set of functions ruling the relationships among endogenous and exogenous variables;
- G is a directed acyclic graph depicting the structural relationship of causation among variables of U and V.

The graph, G, may be obtained by two main ways: (i) by domain knowledge, a specialist may propose the cause and effect relationships among variables of a problem; (ii) by causal search, algorithms are employed to search for the causal relations among variables. In this research, domain knowledge was employed to propose a graph depicting cause-effect relationships among the following variables: age range, gender, video concept, liking, emotional quadrant. The modeled graph is presented in Figure 1, and the set of endogenous and exogenous variables are, respectively, $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ and $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$. For the purposes of this paper, it is not necessary to define the set of functions F because this research is not interested in the exact numeric functional influence of each causal variable in the effect variable.

Figure 1 represents the causal diagram of the studied problem. In this research, EEG signal was clustered in 4 groups, ages were divided in three ranges, video content was clustered in four groups, like was labeled in three categories (like, dislike, neutral), and emotions were grouped according to valence-arousal quadrants. In Figure 1, each graph's node represents:

- $V_1 = \{1, 2, 3, 4\}$. EEG signal feature category. Each number represents one of 4 possible clusters resulting from the application of K-means to the HOC features;

Figure 1. Proposed Causal Diagram. The set of functions F is depicted as the arrows.



- $V_2 = \{0, 1, 2\}$. Age of the subject, obeying the following distributions 0: 18-22, 1: 23-26, 2: 27-30;
- $V_3 = \{0, 1\}$**: Gender of the subject (0 female, 1 male)**;
- $V_4 = \{0, 1, 2, 3\}$. Video content category obtained by video-concept clustering as described in Subsection 4.3;
- $V_5 = \{1, 2, 3\}$. Like/Dislike value, 0: like, 1: neutral, 2: dislike;
- $V_6 = \{1, 2, 3, 4\}$. Emotional Quadrant. From quadrant 1 up to quadrant 4 we have the following valence-arousal relations, respectively: HV-HA, HV-LA, LV-LA, LV-LA. Where, the letters mean: HV = high valence; HA = high arousal; LV = low valence; LA = low arousal.

One of the most important contributions of Do-Calculus is the possibility of answering interventional questions using observational data. In order to answer such questions, Do-Calculus [4] provides, among other mathematical tools: The Back-Door Criterion, The Front-Door Criterion, and The Three Rules of Do-Calculus. The definitions and theorems on Do-Calculus presented in this section are written exactly as in [4].

Definition 3.1. (**Back-Door**) A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:

(i) no node in Z is a descendant of X_i;and

(ii) Z blocks every path between X_i and X_j that contains an arrow into X_i.

An explanation about item (ii) in Definition 3.1: When it says that Z blocks a path, it means that conditioning on one or more elements of the Z set blocks the causal dependence between the other variables. This blocking/unblocking is explained using D-separation theory. Due to textual space constraints, it is not possible to explain all details about D-separation, and the reader is advised to read one of Pearl books referred in this article.

Theorem 3.1 (**Back-Door Adjustment**). If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by Equation 1.

$$P(y|do(x)) = \sum_{z \in Z} P(y|x,z)P(z) \tag{1}$$

In Theorem 3.1 the identifiability refers to the possibility of obtaining a *consistent estimate of the probability of Y under the condition that X is set to x by external intervention, from data involving only observed variables* [34].

Definition 3.2. (**Front-Door**) A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if:

(i) Z intercepts all directed paths from X to Y;

(ii) there is no unblocked back-door path from X to Z; and

(iii) all back-door paths from Z to Y are blocked by X.

Theorem 3.2 (**Front-Door Adjustment**). If Z satisfies the front-door criterion relative to (X, Y) and if $P(x,z) > 0$, then the causal effect of X on Y is identifiable and is given by the Equation 2.

$$P(y|do(x)) = \sum_{z \in Z} P(z|x) \sum_{x'} P(y|x',z)P(x') \tag{2}$$

Both Back-Door and Front-Door adjustment equations are used to go from an interventional probability equation to an observational probability equation. But there are also more high level criteria for accomplishing the conversion from intervention to observation: The following Rules of Do-Calculus are used for that purpose.

Rules of Do-Calculus [35]. Let X, Y, Z, and W be arbitrary disjoint sets of nodes in a causal DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$. The following three rules are valid for every interventional distribution compatible with G.

1. Insertion/deletion of observations: Equation 3.
2. Action/observation exchange: Equation 4.
3. Insertion/deletion of actions: Equation 5.

$$P(y \mid do(x), z, w) = P(y \mid do(x), w) \quad \text{IF } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}} \tag{3}$$

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w) \quad \text{IF } (Y \perp\!\!\!\perp Z) \mid X, W)_{G_{\overline{X},\underline{Z}}} \tag{4}$$

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), w) \quad \text{IF } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X},\overline{Z(W)}}} \tag{5}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

By applying Do-Calculus rules in the proposed SCM of Figure 1, it is possible to answer interventional queries and to verify the causal impact among variables.

In Do-Calculus, an intervention in the variable X is represented by $do(X)$, and the graph G is modified to $G_{\overline{X}}$ by excluding all arrows that arrives into X. In this article, the causal effect was computed using the library PyAgrum [36].

4. Proposed Approach

This Section presents the proposed approach and the description of the data used, such as video dataset and EEG dataset. In this Section, the proposed research questions are presented in Do-Calculus notation, and the adjustment formula for one of the research questions is developed (Research Question 1). The adjustment formula for all research questions was not presented because they are very similar with the one presented.

4.1 Overall Presentation

The results presented in this paper were obtained by applying Causal Inference to a dataset which is composed by the following parts:

- 60 videos from the LIRIS-ACCEDE video dataset;
- EEG samples from volunteers who watched LIRIS-ACCEDE videos;
- Video valence, arousal and liking labels assigned by volunteers for the watched videos;
- Information about age and gender of the volunteers.

4.2 Data Description

4.2.1 Video Dataset

As one of the main questions of this research is to evaluate the causal impact of video content in human emotional states, a video dataset containing affective annotations is necessary. Among the well-known labeled video datasets publicly available, the LIRIS-ACCEDE [37] was chosen due to following the main reasons:

- The dataset consists of 9,800 good quality video excerpts;
- All video excerpts are shared under the Creative Commons license;

- There are affective annotations for all videos;
- The affective annotations were obtained from people from diverse cultural backgrounds;
- In the dataset, there are short videos with duration of less than 60 seconds.

After downloading the dataset, the next step was to select 60 videos which would be presented to volunteers. For each valence-arousal quadrant, 15 representative videos were selected using the scores available in the LIRIS-ACCEDE dataset. The names of the videos are presented in Figure 2, in Section 4.3.

4.2.2 EEG

The EEG signals used in our experiments were collected from volunteers in our laboratory while they were watching videos. The volunteers were undergraduate students from our university. Table 1 presents detailed information about volunteers gender and age. After watching each video, a self-assessment questionnaire was prompted to the volunteer asking three questions: how was valence perceived (in a scale from 1 to 9)?; how was arousal perceived (in a scale from 1 to 9)?; and, if they liked, did not liked or were indifferent to video content.

Table 1. Detailed information about volunteers, before and after removing problematic samples.

	Before	After
Min. Age	17	18
Max. Age	21	30
Avg. Age	20	22
Std. Age	1.41	3.87
Male	15	9
Female	11	7
Tot. Volunteers	26	16

EEG signals are affected by many types of noise sources, for instance: eye-blinking, eye-rolling, chewing, shaking arms and legs, etc. There is not a method for EEG signal filtering which has good results for all types of noise. Even deep learning has been employed for filtering EEG signals, but results still require improvements [38]. Although there is not a known study about the effect of filtering signals containing affective content, all filtering approaches may discard signal information that could be important for emotion classification. Therefore, in this research, the EEG signals were used without any type of filtering. This decision may affect the quality of EEG features and, consequently, the clustering. However, the combination of evidence allowed by probabilistic approaches as those used in this research could attenuate possible problems in one of the evidence sources.

Only channels FP1, FP2, F3 and F4 where used for extracting features in this research. This decision is in accordance with the common practice in the literature [14] [10]. The chosen feature was the HOC (Higher Order Crossings) [9] which is one of the best features for EEG-based emotion classification [11]. The feature vectors were given as input to K-means clustering algorithm to group them in 4 clusters. Those 4 clusters represent the levels of variation of variable V_1 in the causal diagram presented in Figure 1.

4.3 Concept-Based Video Clustering

The tool available in the website *VideoAnalysis4ALL* [2] was used to extract the concepts of all 60 videos. The video concept clusterization may be formalized as follows. Consider a set of videos, $V = \{v_1, v_2, \ldots, v_n\}$, and a function α representing the processing of *VideoAnalysis4ALL* tool.

$$\alpha : V \rightarrow C \times S \tag{6}$$

where V represents the input set of videos, C represents the set of concepts, and S represents the set of shots. Equation 6 shows that the output of α is a matrix containing all the concepts of the input video grouped by shots. However, not all concepts were used in this research, only the most important, which were selected by employing function β.

$$\beta : C \times S \rightarrow C_{10} \times S_{10} \tag{7}$$

In Equation 7, domain and counter-domain look similar, except by the fact that the output matrices have at most 10 concepts from at most 10 shots. The concept selection was performed by sorting the concept scores. After applying Equation 7, the videos were clustered by concept. For each video, a text document was created containing all the 10 selected concepts presented in the video sorted by scores from each video scene. In the best case, for each video there were 100 concepts. However, the same concept could occur in more than one scene indicating its importance for that video. Besides, not all video had 100 concepts because some of them did not had 10 scenes or 10 concepts for each scene. Some video concepts are presented in Table 2.

After clustering, the video concepts were grouped in 4 groups of video contents, and those groups represent the levels of variation of the variable V_4 in Figure 1. In order to allow the 2D visualization, the 4-dimensional dissimilarities were given as input to a multi-dimensional scaling (MDS) algorithm to project the data into 2 dimensions. Figure 2 shows two higher level clusters: indoor and outdoor. The indoor clusters contain people, in one of them appearing mainly entertainment situations, and in the other appearing mainly people faces. The outdoor clusters contain nature, one of them contains animals and the other contains mainly natural environment.

4.4 Proposed Research Questions in Do-Calculus Notation

4.4.1 Research Questions

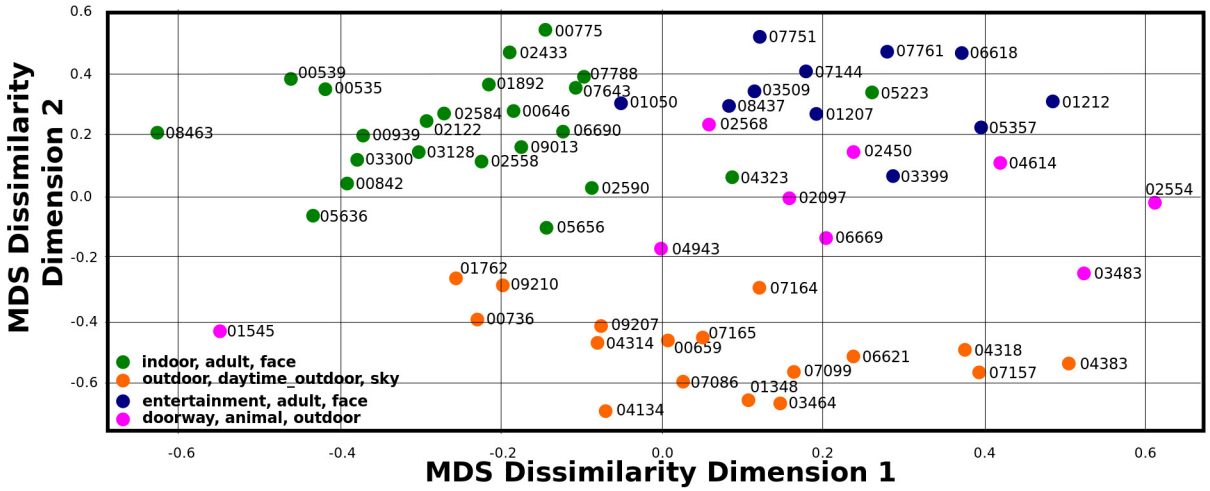
Section 5 presents results for the following causal queries:

[2] Available at: http://multimedia2.iti.gr/onlinevideoanalysis/service/start.html . Last access: March 07th, 2020.

Table 2. Sample of concepts for each video labeled as quadrant 2. In reference to Figure 2, these are the correspondences of cluster numbers and colors: 0 = green, 1 = orange, 2 = blue, 3 = pink.

Video	Cluster	Concept						
ACCEDE00539	0	Kitchen	Table	Room	Person	Eaters	Indoor	Adult
ACCEDE06621	1	Sunglasses	Outdoor	Sky	Sunny	Vehicle	Person	Clouds
ACCEDE00059	1	Vegetation	Plant	Daytime_Outdoor	Person	Child	Girl	Trees
ACCEDE05223	0	Person	Female_Person	Talking	Adult	Male_Person	Face	Civilian_Person
ACCEDE00535	0	Person	Eaters	Adult	Table	Kitchen	Female_Person	Indoor
ACCEDE00775	0	Person	Adult	Sitting_Down	Table	News_Studio	Talking	Eaters
ACCEDE01892	0	Eaters	Person	Adult	Face	Food	Joy	Flowers
ACCEDE01762	1	Female_Person	Face	Teenagers	Female_Human_Face	Person	Female_Human_Face_Closeup	Single_Person_Female
ACCEDE00842	0	Female_Person	Single_Person_Female	Teenagers	Female_Human_Face	Single_Person	Face	Female_News_Subject
ACCEDE01545	3	Domesticated_Animal	Cats	Quadruped	Animal	Mammal	Dogs	Explosion_Fire
ACCEDE04314	1	Cats	Domesticated_Animal	Dogs	Quadruped	Animal	Mammal	Person
ACCEDE04318	1	Waterscape_Waterfront	Islands	Outdoor	Lakes	Mountain	Oceans	Valleys
ACCEDE09210	1	Vegetation	Plant	Trees	Daytime_Outdoor	Landscape	Forest	Person
ACCEDE07099	1	Urban_Park	Tent	Trees	Ground_Vehicles	Outdoor	Vegetation	Vehicle
ACCEDE09207	1	Female_Person	Trees	Person	Female_News_Subject	Face	Teenagers	Female_Human_Face

Figure 2. Concept-Based Clustered Videos.



1. $P(V_5 \mid do(V_4))$: Which is the causal impact of a given video concept in the subject like/dislike state for that video?
2. $P(V_5 \mid do(V_3))$: Which is the causal impact of the subject's gender in the subject like/dislike state for a video?
3. $P(V_5 \mid do(V_2))$: Which is the causal impact of the subject's age in the subject like/dislike state for a video?
4. $P(V_6 \mid do(V_4))$: Which is the causal impact of a given video concept in the subject emotional state after watching a video?
5. $P(V_6 \mid do(V_3))$: Which is the causal impact of the subject's gender in the subject emotional state after watching a video?
6. $P(V_6 \mid do(V_2))$: Which is the causal impact of the subject's age in the subject emotional state after watching a video?

The causal formulae for all 6 causal queries have similar structures, for each one only the variables are changed. In order to calculate the causal impact of an intervention in V_i in the value of V_k, it is necessary to calculate $P(V_k \mid do(V_i))$, which may be obtained by successively applying Do-Calculus rules, to achieve Equation 8. The results of applying Equation 8 for answering the proposed causal questions are presented in Section 5.

4.4.2 The Adjustment Equation for $P(V_5|do(V_4))$

Let V_5 and V_4 be endogenous variables in Figure 1 without considering the exogenous variables. So, the first step is to verify if there is any change in the Graph of Figure 1 after excluding the arrows coming to V_5. As there is no change, we verify the possibility of applying Back-Door or Front-Door adjustment. As there are observations available on V_2, V_3, and V_6, one should try to apply Back-Door criterion with $Z = \{V_2, V_3, V_6\}$, but V_6 could not be used because does not attends item (ii) of Back-Door definition (V_6 does not block the path between V_5 and V_6). For Back-Door, Z would change to $Z = \{V_2, V_3\}$. But neither V_2 nor V_3 blocks a path between V_4 and V_5 which contains an arrow into X_4. Therefore, Back-Door may not be applied to this problem.

The next step is verifying the possibility of applying Front-Door criterion. Let's start with $Z = \{V_1, V_2, V_3, V_6\}$, and ex-

Table 3. Concept clusters for each used video. In reference to Figure 2, these are the correspondences of cluster numbers and colors: 0 = green, 1 = orange, 2 = blue, 3 = pink.

Video	Cluster	Video	Cluster
ACCEDE00059	1	ACCEDE02558	0
ACCEDE05656	0	ACCEDE00535	0
ACCEDE02568	3	ACCEDE06618	2
ACCEDE00539	0	ACCEDE02584	0
ACCEDE06621	1	ACCEDE00646	0
ACCEDE02590	0	ACCEDE06669	3
ACCEDE00736	1	ACCEDE03128	0
ACCEDE06690	0	ACCEDE00775	0
ACCEDE03300	0	ACCEDE07086	1
ACCEDE00842	0	ACCEDE03399	2
ACCEDE07099	1	ACCEDE00939	0
ACCEDE03464	1	ACCEDE07144	2
ACCEDE01050	2	ACCEDE03483	3
ACCEDE07157	1	ACCEDE01207	2
ACCEDE03509	2	ACCEDE07164	1
ACCEDE01212	2	ACCEDE04134	1
ACCEDE07165	1	ACCEDE01348	1
ACCEDE04314	1	ACCEDE07643	0
ACCEDE01545	3	ACCEDE04318	1
ACCEDE07751	2	ACCEDE01762	1
ACCEDE04323	0	ACCEDE07761	2
ACCEDE01892	0	ACCEDE04383	1
ACCEDE07788	0	ACCEDE02097	3
ACCEDE04614	3	ACCEDE08437	2
ACCEDE02122	0	ACCEDE04943	3
ACCEDE08463	0	ACCEDE02433	0
ACCEDE05223	0	ACCEDE09013	0
ACCEDE02450	3	ACCEDE05357	2
ACCEDE09207	1	ACCEDE02554	3
ACCEDE05636	0	ACCEDE09210	1

clude V_6 from Z because it does not block a path from V_4 to V_5. So, our Z would be $Z = \{V_1, V_2, V_3\}$, and criteria (i) and (ii) are satisfied, because Z intercepts all directed paths and there is no back-door path from V_4 to Z. As there is an arrow from V_4 to V_1, V_4 would block all back-door paths from Z to V_5, if they exist. Therefore, the Front-Door Adjustment can be applied to this problem, and we get Equation 8.

$$P(V_5|do(V_4)) = \sum_z P(z|do(V_4),z) \sum_{i=1}^{i=4} P(V_5|V_{4,i},z)P(V_{4,i}) \quad (8)$$

where $z \in \{V_1, V_2, V_3\}$. The equations for the probabilities of the other research questions are obtained similarly by applying Front-Door Adjustment.

5. Results

As it may be seen in Table 1, there were data of only 16 volunteers for experimentation, with a total of 960 samples (16 subjects times 60 videos). In order to statistically evaluate the results, the data were randomized and 100 random sets containing 600 samples each were created. For each set containing 600 samples, a causal effect evaluation was performed, and hypothesis tests were computed as described in this section. For each set of samples, 48 causal effect evaluations were performed as presented in Table 4.

The first step in analysing the collected data was to verify their normality. The normality test results are presented in Table 4. In order to verify the normality, the Shapiro-Wilk test for normality was employed using the causal effect estimations obtained for each parameter combination. The Shapiro-Wilk test was performed as hypothesis test considering the null hypothesis as the data was drawn from a normal distribution.

Table 4 presents the results for normality tests. A total of 48 tests were performed, and for all of them the null hypothesis was that the data follow a normal distribution. As it may be seen in the right column of Table 4, there was not any case in which the null hypothesis was rejected. Furthermore, in a great number of tests, the p-value was very high. Only in some cases, p-value was not so high. For example, only in experiments 23, 33, 37, and 41 the p-value was lower than 10%, but still higher than 5%.

After checking for normality, the next step was to verify the hypothesis of the effects of video content, age, and gender on the volunteer video-liking. For those tests, a T-test for the means of two independent samples was employed. Table 5 presents the 8 results for normality tests on the liking data. For all normality liking tests the null hypothesis was rejected, which means that not every pair of samples has similar average values.

As it was verified by results showed in Table 5, liking and disliking results have different averages. Therefore, the next step is to verify which are the highest causal effects on liking/disliking of variables video-cluster, age, and gender.

Table 4. Normality Tests. Meaning of variables is as follows: V_1 = EEG Cluster, V_2 = Age, V_3 = Gender, V_4 = Video Content Cluster, V_5 = Liking, and V_6 = Emotional Quadrant.

Exp. #	Cause	Effect	Stat.	p-value	Reject H_0?
1	$V_4 = 0$	$V_5 = 1$	0.9912	0.7672	False
2	$V_4 = 0$	$V_5 = 3$	0.9902	0.6901	False
3	$V_4 = 1$	$V_5 = 1$	0.9866	0.4161	False
4	$V_4 = 1$	$V_5 = 3$	0.9901	0.679	False
5	$V_4 = 2$	$V_5 = 1$	0.9907	0.7245	False
6	$V_4 = 2$	$V_5 = 3$	0.9881	0.521	False
7	$V_4 = 3$	$V_5 = 1$	0.9903	0.6963	False
8	$V_4 = 3$	$V_5 = 3$	0.9903	0.6954	False
9	$V_3 = 0$	$V_5 = 1$	0.9827	0.2195	False
10	$V_3 = 0$	$V_5 = 3$	0.986	0.3826	False
11	$V_3 = 1$	$V_5 = 1$	0.989	0.5912	False
12	$V_3 = 1$	$V_5 = 3$	0.9868	0.4325	False
13	$V_2 = 0$	$V_5 = 1$	0.9888	0.578	False
14	$V_2 = 0$	$V_5 = 3$	0.9917	0.806	False
15	$V_2 = 1$	$V_5 = 1$	0.9797	0.1289	False
16	$V_2 = 1$	$V_5 = 3$	0.9945	0.9612	False
17	$V_4 = 0$	$V_6 = 1$	0.9935	0.9209	False
18	$V_4 = 0$	$V_6 = 2$	0.9835	0.2532	False
19	$V_4 = 0$	$V_6 = 3$	0.9941	0.9468	False
20	$V_4 = 0$	$V_6 = 4$	0.9951	0.9774	False
21	$V_4 = 1$	$V_6 = 1$	0.9894	0.6209	False
22	$V_4 = 1$	$V_6 = 2$	0.9898	0.6559	False
23	$V_4 = 1$	$V_6 = 3$	0.9755	0.06128	False
24	$V_4 = 1$	$V_6 = 4$	0.9873	0.4625	False
25	$V_4 = 2$	$V_6 = 1$	0.9865	0.7619	False
26	$V_4 = 2$	$V_6 = 2$	0.9839	0.7456	False
27	$V_4 = 2$	$V_6 = 3$	0.9895	0.2377	False
28	$V_4 = 2$	$V_6 = 4$	0.996	0.234	False
29	$V_4 = 3$	$V_6 = 1$	0.9911	0.7619	False
30	$V_4 = 3$	$V_6 = 2$	0.9909	0.7456	False
31	$V_4 = 3$	$V_6 = 3$	0.9831	0.2377	False
32	$V_4 = 3$	$V_6 = 4$	0.9831	0.234	False
33	$V_3 = 0$	$V_6 = 1$	0.9755	0.06083	False
34	$V_3 = 0$	$V_6 = 2$	0.9895	0.6336	False
35	$V_3 = 0$	$V_6 = 3$	0.9884	0.5467	False
36	$V_3 = 0$	$V_6 = 4$	0.9841	0.2775	False
37	$V_3 = 1$	$V_6 = 1$	0.9778	0.09226	False
38	$V_3 = 1$	$V_6 = 2$	0.9787	0.1081	False
39	$V_3 = 1$	$V_6 = 3$	0.99	0.6744	False
40	$V_3 = 1$	$V_6 = 4$	0.9859	0.374	False
41	$V_2 = 0$	$V_6 = 1$	0.9747	0.05315	False
42	$V_2 = 0$	$V_6 = 2$	0.9821	0.1977	False
43	$V_2 = 0$	$V_6 = 3$	0.9923	0.8499	False
44	$V_2 = 0$	$V_6 = 4$	0.9908	0.7351	False
45	$V_2 = 1$	$V_6 = 1$	0.9789	0.1121	False
46	$V_2 = 1$	$V_6 = 2$	0.9884	0.5477	False
47	$V_2 = 1$	$V_6 = 3$	0.988	0.5139	False
48	$V_2 = 1$	$V_6 = 4$	0.9915	0.7863	False

Those causal effects are presented in Table 6. By analysing results from Table 6, the following conclusions are drawn:

- All clusters of concepts have a higher causal effect on not liking (LIKING=0) than on liking (LIKING=1);
- The video dataset had a negative causal impact on female volunteers (they did not liked the videos);
- The video dataset had a positive causal impact on male volunteers (they liked the videos);
- Youngest volunteers (AGE=0) were negatively impacted by videos (they did not liked the videos);
- Oldest volunteers (AGE=1) were positively impacted by videos (they liked the videos).
- The content of video cluster number 4 caused more liking than the other clusters.

Overall, the video dataset negatively impacted the liking of volunteers, and oldest and male volunteers were more positively impacted than youngest and female volunteers. Table 7 presents results of causal impact of cluster of concepts, gender, and age on emotional quadrants. As it may be seen, all video concept clusters had a negative causal impact on the emotional responses of volunteers. However, the video cluster number 4 caused more positive emotions than the other video clusters. This result for emotional quadrant is in accordance with the result for liking. The LV-LA quadrant had the highest causal impacts. This result is related to the fact already noted about results presented on Table 6 where it was seen that, in general, volunteers were negatively affected by video concepts. The valence-arousal analysis of causal effects on all ages and gender clusters had a highest causal effect on the LV-LA quadrant, which indicate a high propensity for attenuated (low arousal) negative emotions (low valence) for all the volunteers when watching the videos. However, in 3 of 4 cases, the second highest causal effects for AGE and GENDER occur in HV-LA quadrant, with the following values: 0.2020 (GENDER=1), 0.1858 (AGE=0), and 0.2020 (AGE=1).

One hypothesis we have to justify is the negative impact of the video content in liking and emotional responses is that the contents of the chosen sample were not attractive to the volunteers. Another hypothesis is related to the fact that the videos were not obtained from high-quality productions like Hollywood companies. A great number of videos were produced by novel companies or by amateur producers. During the EEG acquisition, some volunteers commented about the quality of the videos. Anyway, those hypotheses do not invalidate this research, because the proposed approach can be replicated by any researcher using other datasets and other volunteers. Our findings demonstrate the usefulness of the proposed approach to determine the causal effect of video content and quality in the liking and emotional responses of subjects.

Table 5. Normality Liking Tests

Exp. Number	CLUSTER_VID	GENDER	AGE	Stat.	p-value	Reject H_0?
P(LIKING — DO(CLUSTER_VID))						
1	0	-	-	8.515	4.353e-15	True
2	1	-	-	7.944	1.498e-13	True
3	2	-	-	9.966	3.471e-19	True
4	3	-	-	8.189	3.317e-14	True
P(LIKING — DO(GENDER))						
5	-	0	-	12.52	8.347e-27	True
6	-	1	-	5.03	1.107e-06	True
P(LIKING — DO(AGE))						
7	-	-	0	9.084	1.151e-16	True
8	-	-	1	8.006	1.028e-13	True

Table 6. Liking Tests. In this paper, V_5 = Liking, V_4 = Video Content Cluster, V_3 = Gender, and V_2 = Age.

	LIKING=0	LIKING=1
$P(V_5 \mid DO(V_4 = 0))$	0.3543	0.3162
$P(V_5 \mid DO(V_4 = 1))$	0.3547	0.3150
$P(V_5 \mid DO(V_4 = 2))$	0.3548	0.3131
$P(V_5 \mid DO(V_4 = 3))$	0.3555	0.3170
$P(V_5 \mid DO(V_3 = 0))$	0.3598	0.3361
$P(V_5 \mid DO(V_3 = 1))$	0.3040	0.3507
$P(V_5 \mid DO(V_2 = 0))$	0.3568	0.3271
$P(V_5 \mid DO(V_2 = 1))$	0.3161	0.3483

Table 7. Emotional Quadrant Tests. From quadrant 1 up to quadrant 4 we have the following valence-arousal relations, respectively: HV-HA, HV-LA, LV-LA, LV-LA. Where, the letters mean: HV = high valence; HA = high arousal; LV = low valence; LA = low arousal. The meaning of variables in the Table are: V_6 = Emotional Quadrant, V_4 = Video Content Cluster, V_3 = Gender, and V_2 = Age.

	HV-HA	**HV-LA**	**LV-LA**	**LV-HA**
$P(V_6 \mid DO(V_4 = 1))$	0.1229	0.1836	0.5173	0.1762
$P(V_6 \mid DO(V_4 = 2))$	0.1198	0.1830	0.5200	0.1772
$P(V_6 \mid DO(V_4 = 3))$	0.1186	0.1801	0.5209	0.1804
$P(V_6 \mid DO(V_4 = 4))$	0.1207	0.1867	0.5189	0.1737
$P(V_6 \mid DO(V_3 = 0))$	0.1360	0.1591	0.5005	0.2044
$P(V_6 \mid DO(V_3 = 1))$	0.1092	0.2020	0.5333	0.1555
$P(V_6 \mid DO(V_2 = 0))$	0.1207	0.1858	0.5181	0.1753
$P(V_6 \mid DO(V_2 = 1))$	0.1215	0.2020	0.5215	0.1814

6. Conclusion

This article presented an approach for drawing causal effect answers to the problem of emotion elicitation in videos. Two main theories were employed in this research: video concepts and causal inference (Do-Calculus). For the videos used in this research, the obtained results allow to conclude: Although in general volunteers did not like the videos, it is possible to say that, in general, men liked and women did not like the videos. A conclusion about age: youngest volunteers did not like the videos, but oldest volunteers liked the videos. The results for valence-arousal quadrants are in accordance with liking results: the videos may have induced attenuated (low arousal) negative emotions (low valence) for all volunteers.

One limitation of the proposed method is the necessity of the expert to design the causal graph. This limitation could be overcome using causal search methods. Other limitations of the proposed method are related to experimental conditions, especially the number of volunteers and EEG samples. Furthermore, the proposed method would be better evaluated using professional high-quality videos instead of amateur videos.

The proposed approach may be employed to evaluate liking and emotional elicitation for any video dataset if it has information about EEG, age, and gender of subjects who watched the videos. A future work would be to verify whether there is transportability [39] among EEG datasets. An application of transportability would be to verify if the results obtained using the proposed EEG dataset is transportable to other datasets like DEAP [1]

Acknowledgements

The authors would like to thank to "Programa Institucional de Bolsas de Iniciação Científica (PIBIC) do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) do Brasil."

This study was partially funded by CNPq (Brazilian National Council for Scientific and Technological Development), Grant N. 459763/2014-8.

Author contributions

Eanes Torres Pereira contributed to designing experiments, analyzing results, and in overall text writing.
Geovane do Nascimento Silva implemented software, conducted experiments, wrote the results section of the paper and part of the discussion.

References

[1] KOELSTRA, S. et al. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, IEEE, v. 3, n. 1, p. 18–31, 2012.

[2] SOLEYMANI, M. et al. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, IEEE, v. 3, n. 1, p. 1–14, 2012.

[3] ZHENG, W.-L.; LU, B.-L. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, v. 7, n. 3, p. 162–175, 2015.

[4] PEARL, J. *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press, 2009.

[5] LANZA, S. T.; MOORE, J. E.; BUTERA, N. M. Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American Journal of Community Psychology*, v. 52, n. 0, p. 380–392, 2013.

[6] PEARL, J. et al. *Causal Inference in Statistics: A primer*. United Kingdom: Wiley, 2016.

[7] NGUYEN, P. et al. Age and gender classification using eeg paralinguistic features. In: *6th Annual International IEEE EMBS Conference on Neural Engineering*. San Diego, CA, USA: IEEE, 2013. p. 1295–1298.

[8] PUTTEN, M. J. A. M. van; OLBRICH, S.; ARNS, M. Predicting sex from brain rhytms with deep learning. *Nature Scientific Reports*, v. 1, n. 8, p. 1–7, 2018.

[9] PETRANTONAKIS, P. C.; HADJILEONTIADIS, L. J. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine*, v. 14, n. 2, p. 186 – 197, 2010.

[10] PEREIRA, E. T. et al. Empirical evidence relating eeg signal duration to emotion classification performance. *IEEE Transactions on Affective Computing*, p. 1–12, 2018.

[11] JENKE, R.; PEER, A.; BUSS, M. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions of Affective Computing*, v. 5, n. 3, p. 327 – 339, 2014.

[12] BRADLEY, M. M.; LANG, P. J. International affective picture system. In: ZEIGLER-HILL, V.; SHACKELFORD, T. K. (Ed.). *Encyclopedia of Personality and Individual Differences*. Switzerland: Springer International Publishing, 2017. p. 1–4.

[13] FOURATI, R. et al. Unsupervised learning in reservoir computing for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, p. 1–1, 2020.

[14] PEREIRA, E. T.; GOMES, H. M. The role of data balancing for emotion classification using eeg signals. In: *Digital Signal Processing Conference*. Beijing, China: IEEE, 2016. p. 1–6.

[15] LIU, Y.-J. et al. Real-time movie-induced discrete emotion recognition from eeg signals. *IEEE Transactions on Affective Computing*, v. 9, n. 4, p. 550–562, 2017.

[16] SONG, T. et al. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, p. 1–10, 2018.

[17] ZHENG, W.-L.; ZHU, J.-Y.; LU, B.-L. Identifying stable patterns over time for emotion recognition from eeg. *IEEE Transactions of Affective Computing*, v. 10, n. 3, p. 417 – 429, 2019.

[18] HUNTER, M. et al. The australian eeg database. *Clinical EEG and neuroscience*, v. 36, n. 2, p. 76–81, 2005.

[19] TOMESCU, M. et al. From swing to cane: Sex differences of eeg resting-state temporal patterns during maturation and aging. *Developmental Cognitive Neuroscience*, n. 31, p. 58–66, 2018.

[20] VANDENBOSCH, M. M. L. J. Z. et al. Eeg-based age-prediction models as stable and heritable indicators of brain maturational level in children and adolescents. *Human Brain Mapping*, Wiley, n. 40, p. 1919–1926, 2018.

[21] HU, J. An approach to eeg-based gender recognition using entropy measurement methods. *Knowledge-Based Systems*, p. 1–8, 2018.

[22] KAUR, B.; SINGH, D.; ROY, P. P. Age and gender classification using brain-computer interface. *Neural Computing and Applications*, Springer, v. 31, p. 5887–5900, 2019.

[23] KAUSHIK, P. et al. Eeg-based age and gender prediction using deep blstm-lstm network model. *IEEE Sensors Journal*, p. 1–8, forthcoming.

[24] SIDIROPOULOS, P. et al. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 21, n. 8, p. 1163–1177, 2011.

[25] SIDIROPOULOS, P.; MEZARIS, V.; KOMPATSIARIS, I. Video tomographs and a base detector selection strategy for improving large-scale video concept detection. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 24, n. 7, p. 1251–1264, 2014.

[26] TONOMURA, Y.; AKUTSU, A. Video tomography: An efficient method for camerawork extraction and motion analysis. In: *International Conference on Multimedia*. San Francisco, California, USA: ACM, 1994. p. 349–356.

[27] APOSTOLIDS, E.; MEZARIS, V. Fast shot segmentation combining global and local visual descriptors. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, 2014. p. 1–5.

[28] MARKATOPOULOU, F.; MEZARIS, V.; PATRAS, I. Cascade of classifiers based on binary, non-binary and deep convolutional network descriptors for video concept detection. In: *International Conference on Image Processing (ICIP)*. Quebec City, QC, Canada: IEEE, 2015. p. 1–5.

[29] VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *Conference on Computer Vision and Pattern Recognition*. Kauai, HI, USA: IEEE Computer Society, 2001. p. 511–518.

[30] SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*. San Diego, CA, USA: ICLR, 2015. p. 1–15.

[31] MARKATOPOULOU, F.; MEZARIS, V.; PATRAS, I. Online multi-task learning for semantic concept detection in video. In: *International Conference on Image Processing (ICIP)*. Phoenix, AZ, USA: IEEE, 2016. p. 1–5.

[32] EATON, E.; RUVOLO, P. L. Ella: An efficient lifelong learning algorithm. In: *International Conference on Machine Learning*. Atlanta, Georgia, USA: PMLR, 2013. p. 507–515.

[33] MARKATOPOULOU, F.; MEZARIS, V.; PATRAS, I. Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. *Transactions on Circuits and Systems for Video Technology*, IEEE, v. 29, n. 6, p. 1631–1644, 2018.

[34] GALLES, D.; PEARL, J. Testing identifiability of causal effects. In: *Conference on Uncertainty in Artificial Intelligence*. Montreal, Quebec, Canada: ACM, 1995. p. 185–195.

[35] PEARL, J. *do*-calculus revisited. In: FREITAS, N. de; MURPHY, K. (Ed.). *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. Corvallis, OR: AUAI Press, 2012. p. 4–11.

[36] GONZALES, C.; TORTI, L.; WUILLEMIN, P.-H. aGrUM: a Graphical Universal Model framework. In: *International Conference on Industrial Engineering, Other Applications of Applied Intelligent Systems*. Arras, France: Springer, 2017.

[37] BAVEYE, Y. et al. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, v. 6, n. 1, p. 43–55, 2015.

[38] LEITE, N. M. N. et al. Deep convolutional autoencoder for eeg noise filtering. In: *International Conference on Bionformatics and Biomedicine (BIBM)*. Madrid, Spain: IEEE, 2018. p. 2605–2612.

[39] BAREINBOIM, E.; PEARL, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, v. 113, n. 27, p. 7345 – 7352, 2016.