Available online at: https://ijact.in

Date of Submission	09/04/2020
Date of Acceptance	11/05/2020
Date of Publication	31/05/2020
Page numbers	3698-3704 (7 Pages)

Cite This Paper: Khamael R Raheem, Israa H Ali, Multimodal content-based recommender system using three-dimension convolution neural network, 9(5), COMPUSOFT, An International Journal of Advanced Computer Technology. PP. 3698-3704.

ISSN:2320-0790

An International Journal of Advanced Computer Technology

MULTIMODAL CONTENT-BASED RECOMMENDER SYSTEM USING THREE-DIMENSION CONVOLUTION NEURAL NETWORK

Khamael Raqim Raheem[1], Israa Hadi Ali[2]

[1,2]College of Information Technology, Department of Software, University of Babylon, Iraq
[1]khmrak@itnet.uobabylon.edu.iq, [2]israa_hadi@itnet.uobabylon.edu.iq

Abstract: Research on Recommender Systems has grown tremendously over the past few years; however, the quest to date for how user emotions can be used as implicit feedback to supplement these systems is sparse. Recommender Systems should take advantage of the high availability of digital data to collect input data of various types that allow the system to enhance its accuracy implicitly or explicitly. In this study, a Multimodal Content-Based Recommender System for image recommendation is proposed which is based on Implicit and Explicit Feedbacks. In order to obtain the Implicit Feedbacks, a Convolution Neural Network with Three-Dimensions is constructed to predict the emotion of the user's face if it is positive or negative. The Convolution Neural Network making a mixture of spatial and temporal data in Three-Dimension Convolution in order to learn about a transition in consecutive frames. The results of predictions of Neural Network are used as Implicit Feedback for the recommendation algorithm. The Multimodal Recommender System is built by combining the output of two Content-based Recommender Systems using a binary Logistic Regression algorithm. Content-based Recommender System is built by training the Support Vector Machine classifier on features of item profile and Explicit or Implicit feedback. The performance measures are computed based on predicted and ground truth feedbacks. The result shows that the Three-Dimension Convolution Neural Network contributes to Implicit Feedbacks prediction in the Recommender System. Also, the combination of the results of two Recommender Systems with different feedback techniques can enhance the performance of the proposed system.

Keywords: Multimodal Recommender System; Content-Based Recommendation (CBR) ;Three-Dimension Convolution Neural Network (3D CNN); Support Vector Machine (SVM) ;Emotion; Implicit Feedbacks; Explicit Feedbacks

I. INTRODUCTION

Recommender systems (RSs) are software tools or techniques that support the user in the decision-making process by offering the opportunities the system forecasts the user would like [1]. The Recommender System applied to give recommendations for images [2], movies [3], purchasing [4] and traveling [5]. There are several algorithms applied in the RS field such as the Content-Based Recommendation (CBR) algorithm, Collaborative Filtering (CF) algorithm and knowledge-based recommendation (KBR) algorithm [6]. In movies recommendation, the items in the CBR algorithm are represented using features such as actor name, genre of movie, and subject matter that are stored in a structure called the item profile. The preferences for each user will be stored in the user profile. The CBR algorithm compares any unused item with those that establishing inside the user's profile to present the recommendations.

In each RS the user gives feedback to the observed item. There are two types of techniques to gather feedbacks which are Explicit Feedback technique and Implicit Feedback technique [7]. The explicit technique includes the users for relegating either numeric or score ratings for assessing the item. The precision of this type is higher than the implicit technique. On the other hand, the Implicit Feedback can be found in various applications, such as history browsing, site

usage history, and mouse or search pattern movement. This technique is less accurate compared to the Explicit Feedback technique and it is difficult to translate. Recent recommendation work has moved from Explicit Feedback[8]to Implicit Feedback, such as purchases, and watches [9]. In the proposed study, Implicit Feedbacks will obtain by tracking the user's face.

Recent researches on Recommender Systems using the user's affective state have appeared in different applications. One of the limited surveys in affective Recommender Systems is the one presented in [10]. In this work, the authors reviewed the literature on the topic until 2016 and presented classifications of the research that has been done. The group of researchers proved in [11] that the emotions detected through facial expression improved the performance of an RS. They explored the impact of affective metadata (metadata representing the emotions of the user) on the efficiency of the IAPS dataset sub-set CBR system[12]. They evaluated the performance of an RS by using four machine learning classifiers which are Support Vector Machine (SVM), Naïve Bayes (NB), Adaptive Boosting (AdaBoost), and C4.5. The performance of the system was computed by using Precision, Recall, and F-measure metrics.

The researchers in [13] presented that, when the user is looking for entertainment is because there is a want to maintain or strengthen positive states and reduce or direct negative ones. They conducted an experiment to detect what emotions users experience while conducting different video search tasks. They used facial expression recognition and biometrical signals to detect emotions and they made the participants rate the watched videos. They found out that facial expression is the best method for the detection of emotions. [14] Presented a system that records facial feature points as well as a viewer's feelings and proposing videos accordingly. [15]Analyzed a user's facial expressions and physiological parameters when viewing a video. This framework might consider the emotions of a viewer (happiness, sadness, angry, surprised, scared, disgusted and neutral) and select rational real-time video clips. This work claimed that the analysis of the video of user's facial expressions and physiological parameters can suggest future offers to users for video clips they currently like.

The researchers in [16]proposed an automated clothing recommendation system using in-store videos in real-time this can boost shoppers 'apparel experiences and increase sales of items. This approach focuses on findings that retail shoppers tend to seek to assess themselves in front of in-store mirrors on clothing. The machine uses a camera to record the behavior of a shopper opposite the mirror to draw inferences based on their facial expressions and the section of the clothing they inspect at each point of time. They used a CF algorithm to enforce the recommendation. [17] Developed an approach that utilized the emotional reactions of the audience as the basis for recommending new material. They conducted a study where the subjects' facial expressions and skin-estimated pulse were monitored while watching videos. The works discussed above not used a deep neural network for the Recommender System. [18] Stated that the Convolution Neural Network (CNN) can be used in Recommender Systems. The CNN with different architectures has been used in the literature for action classification [19, 20] and emotion classification in to positive or negative [21].

The Three-Dimension Convolution Neural Network (3D CNN) is a form of deep neural learning and feed-forward networks that have been tremendously used in many types of research as it provides better accuracy. We will design a 3D CNN to extract the changing of facial features in consecutive frames in order to predict the emotions of the user's face if it is positive and negative. Extracting the emotion from the video is a more complicated task where most conventional methods consider frames autonomously while disregarding the temporal relationships of the sequential frames in a series that is central to the identification of unpretentious changes in facial frame appearance. Figure 1 demonstrates a typical 3D CNN structure, this structure includes multiple Convolution layers followed by Pooling layers and one Fully Connected (FC) layer followed by an Output layer at the top. The input rectangle represents a stack of consecutive networked video frames.

In fact, 3D convolution is an extension of 2D convolution. The spatial data of input is considered for final features in the 2D convolutional layer while the mixture of spatial and temporal data is used in 3D convolution in order to learn about a transition in consecutive frames. The 3D convolution kernel (filter) is a 3D cube that considers neighboring frames of the local spatial area. For example, if the kernel size is 2x2x2x1 the dot product considers 2x2 in two consecutive frames to be receptive field. The fourth dimension is the color channel (which is 1 due to the use of the grayscale image in this case).

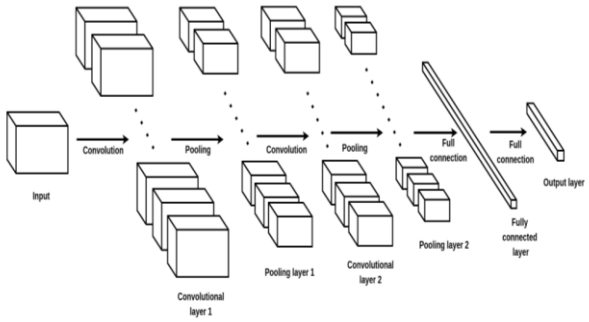

Figure 1: Typical structure of the 3D CNN

The proposed work employs the strength of Deep Neural Networks (DNN) for computer vision problems with the Recommender System field to improve the RS performance. The emotions of the user while he\she consuming the item will translate into feedbacks where the 3D CNN model will design to translate the user's emotions if it is positive or negative. Also, an offline Multimodal RS using the CBR algorithm for image recommendation will build. There are limited researches that constructed RSs based on multiple feedback techniques. The proposed system will combine the output results of two RSs; the first system uses Explicit Feedback as input to the recommendation algorithm while the second RS uses Explicit Feedbacks. The contribution of this study is to build a 3D CNN model to translate the user's face emotions into Implicit Feedbacks and build a Multimodal Recommender System based on Explicit

Feedbacks and the Implicit Feedback to enhance performance. The rest portion of the paper is organized as takes after. Section II presents research method. Section III clarifies the results and discussion and finally the conclusion presented in Section IV.

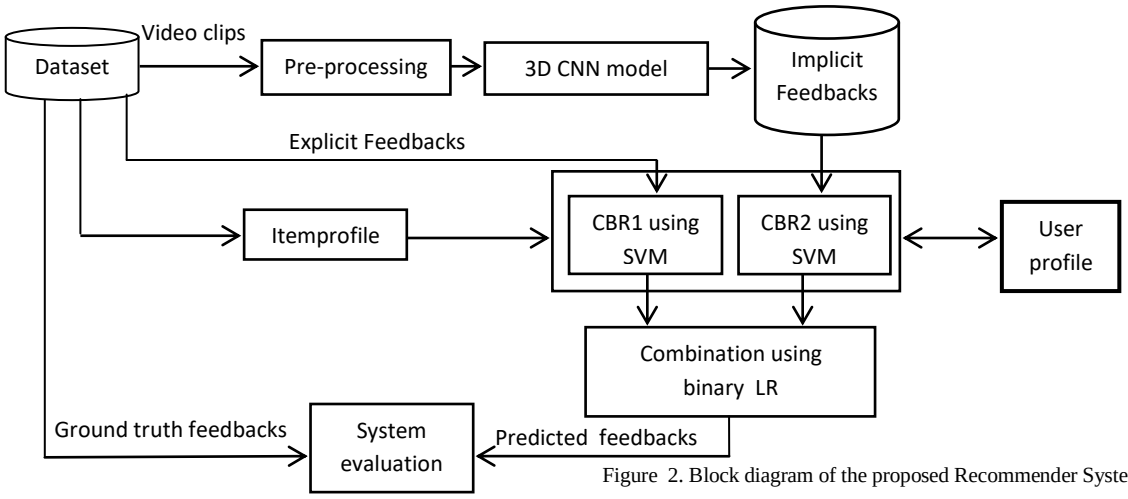

Figure 2. Block diagram of the proposed Recommender System

II. RESEARCH METHOD

In this section, important details about the dataset are explained and the main stages for the proposed RS will be illustrated. These stages are: pre-processing, the Implicit Feedbacks prediction and the recommendation algorithm stage. The block diagram in figure 2 illustrates the proposed RS design. In the pre-processing stage, the video clips in the dataset are processed to be contained the same number of frames. According to the results reported in [22], succinct frame fragments are used instead of a whole video. The study showed the concise 1-7 frame snippets are adequate to identify human action as expected by biological vision systems observation. In the proposed study, each video clip (sample) will contain eight frames to identify the emotion of the user. The proposed system builds supervised machine learning models; these models are a 3D CNN model, and a Multimodal RS model. Each supervised machine learning model has two phases, the training, and testing phase. The first model will build by training the 3D CNN on input samples. Each sample annotated with the class "0" or "1"; the class represents the emotional status (positive or negative emotion) of the user during viewing an item. In the training and testing phase, the convolution process will apply in a spatial and temporal direction where spatiotemporal features will extract from consecutive frames.

In the testing phase, the 3D CNN model will predict the emotional state for the user during viewing the item. The result of predictions will be stored in a database to be used as Implicit Feedback for the CBR algorithm. The Multimodal model will be built by combining the output results of two CBR systems. The first CBR system built the user profile for each user by training the SVM on the item profiles that belong to this user and Explicit Feedbacks obtained from the dataset, while the second CBR system will build the user profile by training the SVM on item profiles that belongs to this user and Implicit Feedbacks obtained from 3D CNN prediction.

The final results will be obtained by training the binary Logistic Regression (LR) on the outputs of the first and second CBR systems. The LR algorithm [23]measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. Mathematically, a binary logistic model has a dependent variable with two possible values, such as like/dislike which is represented by an indicator variable, where the two values are labelled "0" and "1". The evaluation will accomplish using the feedbacks predicted by the LR algorithm and ground truth feedbacks in the dataset.

A. Dataset

The dataset used in this work is LDOS-PerAff-1 [24]. The dataset contains videos for 52 users. For every video file, the frame rate is 15 frames per second. The video separated into 70 video clips where each video clip contains the emotional reaction (sequence of frames reflect the affective response of the user during viewing a specific item) of the user to 70 different items. The items are taken from the repository at the IAPS database of images. The total number of video clips is 3640 for 52 users responding to 70 varied visual stimuli items. Each video clip annotated with affective metadata and ground truth feedback. These annotations are stored in excel format in the dataset.

B. Pre-processing

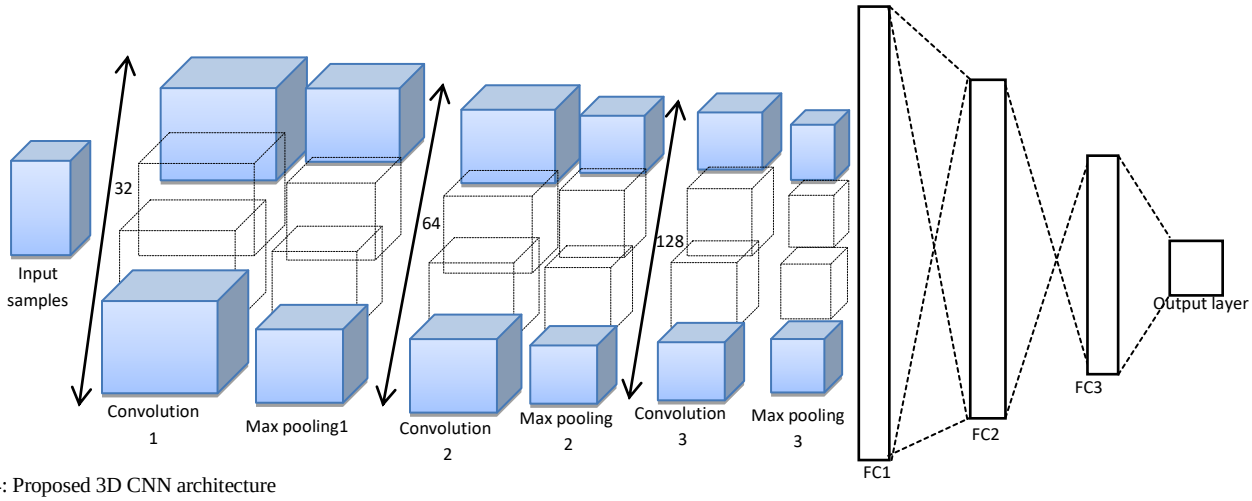

Figure 4: Proposed 3D CNN architecture

The video clips are processed to contain the same number of frames. We will call each video clip a sample term. The study in [25]stated that the "Positive affective responses result in positive evaluations of the focal product, whereas negative affective responses result in negative evaluations". The above study stated that each user gives positive feedback if he\she has positive emotion about the item or gives negative feedback if he\she has negative emotion about the viewing item. Therefore each sample will annotate with "0" or "1" class, the classes are taken from ground truth feedbacks. Figure 3 illustrates samples of the user with identifier 18 during watching the images with identifiers 8117 and 8499. Subfigure A shows the negative emotional reaction of the user when giving "0" as a feedback to the consuming item while subfigure B shows the positive emotional reaction of the user when giving "1" as a feedback to the consuming item. The frames inside each sample are resized to 32x32 pixels and converted to a grayscale form. After that, the face and the eye will extract from each frame in the sample.

We implemented the pre-processing over each frame in the sample using the open-source digital library (DLib) and Haar Cascade classifier. After that, the augmentation of the data is applied to increase the number of training samples because big datasets are needed in deep learning to increase system performance. Data augmentation is the process of collecting and processing samples that are already in a training dataset to create many modified versions of the same samples. To generate new samples and adding them to the training samples, we will make vertical and horizontal flipping, blurring and sharpening augmentation.

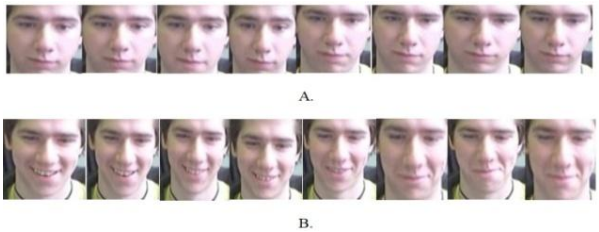

Figure 3: A. User 18 while viewing item 8117 and giving a "0" feedback , B. User 18 while viewing item 8499 and giving a "1" feedback

C. Implicit Feedback Prediction

The proposed work is using 3D CNN to model the emotion of the user's face where the emotions of the user to the consuming item can be predicted as positive (class 1) or as negative (class 0). Each machine learning system has training and testing stages. In both stages, spatiotemporal features will extract from consecutive frames that reflect the emotional reaction of the user to the consuming item. In the training stage, a model for emotion prediction will build using training samples with their classes, part of the training samples will be used as a validation set to tune the hyper parameters during the training stage. In the testing stage, the classes for each unseen sample will predict using the 3D CNN model.

The 3D CNN predictions represent the Implicit Feedback for the algorithm for a recommendation. It is important to remember that we model the users' visual emotions (for example, the observed image made me happy) and not the inherent material movements found in the image characters (for example, the image shows a happy person). The architecture of the proposed 3D CNN is illustrated in Figure 4. The Spatiotemporal features at this stage are extracted from each sample using 3D CNN. The proposed 3D CNN contains multiple layers with different purposes such as 3D Convolution, 3D Max pooling, Batch Normalization, Fully Connected, Dropout, Flatten and Output layer. The 3D convolution layer will apply the convolution mechanism using the 3D kernel to obtain the spatiotemporal features. In contrast to the 2D CNN, where the kernels are used in spatial directions, 3D CNN uses the kernel in both spatial and temporal directions. The Rectified Linear Units (ReLU) activation function will represent non-linearity in Convolution layers. Using the ReLU helps us to avoid the problem of the vanishing gradient induced by some other activation functions. The 3D Max pooling layer continually decreases the 3D Convolutional layer's dimensional output while it retains the important features. In a small Spatiotemporal window, the 3D pooling layer selects the finest representation of features. Batch Normalization is a technique that helps boost speed, performance, and reliability of CNN. It is used

to adjust and scale the activation in order to normalize the input data. Using Dropout in the network eliminates network overfitting of overtraining samples. So it is used to integrate the potential for regularization. Overfitting and long training time in particular are two fundamental problems in multilayered learning of the Neural Network and in deep learning. Dropout and Batch Normalization are two well-known approaches to such challenges [26].

The Flatten layer is the extension from multidimensional inputs to a One-Dimensional array necessary for the fully connected layer. The Fully Connected (Dense) layers are needed for the form of hierarchical feature extraction to provide more non-linearity within the network. The last Fully Connected layer represents the Output layer for the binary class results, with the sigmoid activation.

D. Recommendation Algorithm

A basic CBR scenario consists of a multimedia repository of items and a set of users. The profile of the item is represented using a feature set. The IAPS set of images we selected as the source for our content items is a compilation of stimuli collected in a controlled experiment. The study proposed in [11] used affective metadata for the representation of the image profile. In the proposed study, we will use this affective metadata for item profile representation. The affective metadata for each item profile is represented using the first two statistical moments (mean and standard deviation) for dimensional emotion (valence, Arousal, and Dominance values). The mean and standard deviation are computed for many users that watching a specific image. Table I illustrates an example of an item profile contains affective metadata to the item with the identifier 6910.

TABLE I. Example ofan Item Profile

Metadata Field	*Metadata Value*
Image id	6910
Valence mean	5.31
Valence standard deviation	2.28
Arousal mean	5.62
Arousal standard deviation	2.46
Dominance mean	5.10
Dominance standard deviation	2.46

In order to build the CBR algorithm, a supervised SVM classifier for binary classification will be used. The user profile is the result of training the SVM classifier algorithm. The user profile is a data structure based on past feedbacks and it may have different shapes, but generally, the shape reflects the metadata used in the item profile. The SVM classifier takes several records containing vectors of affective metadata as a training set and Explicit or Implicit Feedback as class values. The Implicit Feedbacks are the result of 3D CNN predictions. SVM learns and stores the relationships between the affective metadata and the classes in a data structure that represents the learned knowledge then it uses the data structure of the classifier to classify the unseen feature vectors. The proposed Multimodal CBR System will be build from two separate Recommender Systems and the outputs of these systems will be combined by using binary Logistic Regression (LR) algorithm.

III. RESULTS AND DISCUSSION

The proposed system designs 3D CNN model for Implicit Feedbacks prediction. The loss function, mean squared error (MSE) function and accuracy metric will be used to evaluate the 3D CNN. The Precision (P), the Recall (R) and F-measure (F) metrics will be used in RS evaluation. SVM classifier and LR algorithm are used in the proposed RS to classify the state of an item if it is relevant (class 1) or non-relevant (class 0). The output of the classification process is a confusion matrix, in binary classification the items identified as relevant or non-relevant where the item is correctly classified (True Positive (TP) or True Negative (TN)) or falsely classified (False Positives (FP) or False Negatives (FN)). Table II shows the confusion matrix of the binary classification. The Precision is the portion of items retrieved which is relevant to the search query, equation 1 illustrate the calculation process to this metric. The recall is the part of the items that are recovered from all the items in question, equation 2 displays the recall law. The F-measure reflects a trade-off between precision and recall, equation 3 outlines the F-measure computation method.

$$P = \frac{tp}{tp + fp} \quad (1)$$

$$R = \frac{tp}{tp + fn} \quad (2)$$

$$F = \frac{2 \text{ x } (P \text{ x } R)}{(P + R)} \quad (3)$$

TABLE II. Confusion Matrixfor Binary Classification

Predicted class	*Actual class*	
	Class 1	*Class 0*
Class 1	TP	FP
Class 0	FN	TN

The best videos for the users in the dataset will be selected in the proposed RS. The input samples are split into 80% training set and 20% test set. The training samples will be used to create 3D CNN model for Implicit Feedbacks prediction. The summary of layers in the proposed 3D CNN is shown in table III. The Binary Cross-Entropy loss function and Adaptive learning rate optimization algorithm (Adam) are used in the proposed network. The error functions will compute based on ground truth feedbacks and predicted feedbacks. The proposed network will apply using Keras library in python language. Moreover, the network will train for 30 epochs with a batch size of 32 sample per-epoch. Table IV illustrates the change of evaluation metrics through the training phase for the training and validation set (test set during model training). We can see that the error functions decreases when the epoch number increasing while the accuracy metric increases. The results on the validation set reached 0.49% for loss error, 0.16% for the MSE and 0.76% for the accuracy. The accuracy and loss function curves of the proposed 3D CNN illustrated in figure 5.

TABLE III. Summary ofthe 3D CNN Layers

Layer (type)	*Output shape*	*Parameters*
input_1 (Input Layer)	(None, 8, 32, 32, 1)	0
batch_normalization_1 (Batch)	(None, 8, 32, 32, 1)	3
conv3d_1 (Conv3D)	(None, 8, 32, 32, 32)	864
activation_1 (Activation)	(None, 8, 32, 32, 32)	0
max_pooling3d_1 (MaxPooling3)	(None, 8, 16, 16, 32)	0
dropout_1 (Dropout)	(None, 8, 16, 16, 32)	0
conv3d_2 (Conv3D)	(None, 8, 16, 16, 64)	55296
activation_2 (Activation)	(None, 8, 16, 16, 64)	0
max_pooling3d_2 (MaxPooling3)	(None, 8, 8, 8, 64)	0
dropout_2 (Dropout)	(None, 8, 8, 8, 64)	0
conv3d_3 (Conv3D)	(None, 8, 8, 8, 128)	221184
activation_3 (Activation)	(None, 8, 8, 8, 128)	0
max_pooling3d_3 (MaxPooling3)	(None, 8, 4, 4, 128)	0
dropout_3 (Dropout)	(None, 8, 4, 4, 128)	0
flatten_1 (Flatten)	(None, 16384)	0
dense_1 (Dense)	(None, 128)	2097152
activation_4 (Activation)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8192
activation_5 (Activation)	(None, 64)	0
dense_3 (Dense)	(None, 32)	2048
batch_normalization_2 (Batch)	(None, 32)	96
activation_6 (Activation)	(None, 32)	0
dropout_4 (Dropout)	(None, 32)	0
dense_4 (Dense)	(None, 1)	32
activation_7 (Activation)	(None, 1)	0

TABLE IV. Performance Measures Changing during 3D CNN Learning

Epoch number	*Training*			*Validation*		
	Loss %	*MSE %*	*Accuracy %*	*Loss %*	*MSE %*	*Accuracy %*
5	0.67	0.23	0.59	0.65	0.23	0.63
10	0.61	0.21	0.67	0.60	0.20	0.68
15	0.58	0.19	0.70	0.57	0.19	0.69
20	0.53	0.17	0.74	0.53	0.18	0.74
25	0.48	0.16	0.77	0.49	0.16	0.77
30	0.44	0.14	0.79	0.49	0.16	0.76

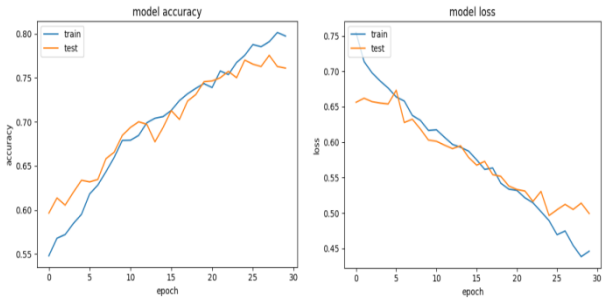

Figure. 5. Loss and accuracy metrics of the 3D CNN

The predictions of the proposed 3D CNN represent Implicit Feedbacks that will be used as input to the CBR algorithm. Three experiments conducted in the proposed Multimodal RS such that the first and second experiments will build two CBR Systems using a binary SVM classifier. In the first experiment, the affective metadata of items profiles for each user and Explicit Feedbacks will be used for SVM training. The second CBR will use affective metadata of items profiles for each user with Implicit Feedbacks that will be used for SVM training. The user profile will be constructed for each user by learning the SVM classifier on the user's item profiles and corresponding classes. Each CBR system will assess using a cross-validation test scheme[27]with 10-fold. The cross-validation will apply on each user profile then the P, R and F metrics will compute. The overall result of each CBR system is computed by calculating the average of P, R, and F for all users' profiles in the system. The third experiment will combine the output of the first and second experiments to improve system performance. Also, the cross-validation with the 10-fold will apply to compute the final results using the LR algorithm. Table V illustrates P, R and F metrics with each fold for this experiment such that the average of folds results represents the final prediction of the proposed Multimodal RS. The results of LR prediction are 0.69% for Precision, 0.72% for Recall and 0.70 for F-measure. Table VI outlines the performance measures for the three experiments in the proposed system.

We can note that the RS that relies on Explicit Feedbacks has higher performance than the one that depends on Implicit Feedback. We have combined the outputs of two RS with different feedback techniques in order to obtain higher performance for RS.

We made a comparison between CBR systems using two machine learning classifiers (NBand AdaBoost) and the proposed system. The results in table VII shows that the proposed Multimodal RS has performance higher than CBR systems based on Explicit Feedback technique.

TABLE V. Performance Measures with Each Fold for LR Algorithm

Fold Number	*P%*	*R%*	*F%*
1	0.78	0.61	0.68
2	0.68	0.81	0.74
3	0.58	0.61	0.59
4	0.89	0.71	0.79
5	0.60	0.75	0.67
6	0.67	0.78	0.72
7	0.67	0.71	0.69
8	0.80	0.76	0.78
9	0.63	0.80	0.71
10	0.67	0.67	0.67
	Average P= 0.69%	Average R= 0.72%	Average F=0.70%

TABLE VI.Results of the Experiments of the Proposed Recommender System

Feedback Technique	*Algorithm*	*P%*	*R%*	*F%*
Explicit	CBR using SVM	0.66	0.68	0.67
Implicit	CBR using SVM	0.62	0.61	0.61
Multimodal (Explicit with Implicit)	Proposed System	0.69	0.72	0.70

TABLE VII . Comparison Proposed RS withother CBR Systems

Feedback Technique	*Algorithm*	*P%*	*R%*	*F%*
Explicit	CBR using AdaBoost	0.60	0.61	0.61
Explicit	CBR using NB	0.44	0.45	0.44
Explicit & Implicit	Proposed system	0.69	0.72	0.70

IV. CONCLUSION

This paper proposes a Multimodal image Recommender System based on content. 3D CNN model is used in spatial and temporal directions for the combined convolution which leads to simultaneous spatiotemporal training. The proposed 3D CNN architecture is used to predict user's

emotions such that tracking is done for the user's face to determine the polarity of emotions if it is positive or negative. The predictions of the network are used as Implicit Feedbacks to the Content-based Recommender System. In the proposed CBR algorithms, the binary SVM classifier is used to build a user profile model and classify the user (relevant \ non-relevant) into the items that are consumed. Our experiments on the proposed Recommender System dataset shows that while combining the two RSs using different feedbacks techniques can improve the performance of the Recommender System. In the future work, we can extend this work by capturing the physiological parameters of the user to be used as Implicit Feedback in order to improve system performance.

V. REFERENCES

[1] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in Recommender systems handbook, ed: Springer, 2015, pp. 1-34.

[2] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic, "Affective labeling in a content-based recommender system for images," IEEE transactions on multimedia, vol. 15, pp. 391-400, 2012.

[3] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, "Content-based video recommendation system based on stylistic visual features," Journal on Data Semantics, vol. 5, pp. 99-113, 2016.

[4] X. W. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li, "We know what you want to buy: a demographic-based system for product recommendation on microblogs," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 1935-1944.

[5] T. D. Owusu and C. Hoffman, "The Personalization and Prediction Innovation of Mobile Recommender Systems," Issues in Information Systems, vol. 15, 2014.

[6] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer, and M. Stettinger, "Basic approaches in recommendation systems," in Recommendation Systems in Software Engineering, ed: Springer, 2014, pp. 15-37.

[7] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," Knowledge-based systems, vol. 46, pp. 109-132, 2013.

[8] K. Y. Collaborative, "filtering with temporal dynamics [J]," Communications of the ACM, vol. 53, pp. 89-97, 2010.

[9] I. Bayer, X. He, B. Kanagal, and S. Rendle, "A generic coordinate descent framework for learning from implicit feedback," in Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 1341-1350.

[10] R. Katarya and O. P. Verma, "Recent developments in affective recommender systems," Physica A: Statistical Mechanics and its Applications, vol. 461, pp. 182-190, 2016.

[11] M. Tkalčič, U. Burnik, and A. Košir, "Using affective parameters in a content-based recommender system for images," User Modeling and User-Adapted Interaction, vol. 20, pp. 279-311, 2010.

[12] P. J. Lang, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Technical report, 2005.

[13] Y. Moshfeghi and J. M. Jose, "An effective implicit relevance feedback technique using affective, physiological and behavioural features," in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 133-142.

[14] A. Mahata, N. Saini, S. Saharawat, and R. Tiwari, "Intelligent movie recommender system using machine learning," in International Conference on Intelligent Human Computer Interaction, 2016, pp. 94-110.

[15] A. Kaklauskas, R. Gudauskas, M. Kozlovas, L. Peciure, N. Lepkova, J. Cerkauskas, et al., "An Affect-Based Multimodal Video Recommendation System," Studies in Informatics and Control, vol. 25, p. 6, 2016.

[16] S. Lu, L. Xiao, and M. Ding, "A video-based automated recommender (VAR) system for garments," Marketing Science, vol. 35, pp. 484-510, 2016.

[17] Y. Diaz, C. O. Alm, I. Nwogu, and R. Bailey, "Towards an affective video recommendation system," in 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2018, pp. 137-142.

[18] J. Y. Liu, "A Survey of Deep Learning Approaches for Recommendation Systems," in Journal of Physics: Conference Series, 2018, p. 062022.

[19] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, "Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos," Procedia computer science, vol. 133, pp. 471-477, 2018.

[20] N. A. Rahmad, M. A. As'Ari, N. F. Ghazali, N. Shahar, and N. A. J. Sufri, "A survey of video based action recognition in sports," Indonesian Journal of Electrical Engineering and Computer Science, vol. 11, pp. 987-993, 2018.

[21] P. Barros, C. Weber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), 2015, pp. 582-587.

[22] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[23] J. Tolles and W. J. Meurer, "Logistic regression: relating patient characteristics to outcomes," Jama, vol. 316, pp. 533-534, 2016.

[24] M. Tkalčič, A. Košir, and J. Tasič, "The LDOS-PerAff-1 corpus of facial-expression video clips with affective, personality and user-interaction metadata," Journal on Multimodal User Interfaces, vol. 7, pp. 143-155, 2013.

[25] P. H. Bloch, "Seeking the ideal form: Product design and consumer response," Journal of marketing, vol. 59, pp. 16-29, 1995.

[26] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. batch normalization: an empirical study of their impact to deep learning," Multimedia Tools and Applications, pp. 1-39, 2020.

[27] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Ijcai, 1995, pp. 1137-1145.