Deep Learning Based Crime Investigation Framework

Lydia J Gnanasigamani, Seetha Hari

Abstract:— Deep learning has emerged as the best way to infer knowledge from data with more meaning and accuracy. The applications of Deep Neural Networks in a variety of domains have made it an important area of research. Crime analysis is the study of crime characteristics and their relationships. The huge volume of crime-related datasets and the various different types of crime and their different characteristics and the complex relationship between them make Deep Neural Networks an ideal choice for this domain. The knowledge gained from this analysis will enable law enforcement officers to process information rapidly and accurately. In the paper, we have proposed a crime framework for India keeping in mind the language and cultural differences across the country. We have proposed automatically creating keywords from the written complaints and tagging the cases using Natural Language Processing. We have also proposed an approach using Deep Neural Network to classify the crimes and match the crimes to offenders using the method of operation. We have also used Deep Neural Networks for prediction of crime rates and hotspots.

Index Terms:— Crime, Convolutional Neural Network, Deep Neural Network, Framework, Natural Language Processing, Recurrent Neural Network, Deep learning

1 INTRODUCTION

Crime data is derived from many sources like written receipts, a diary, emails, chats, digital documents stored on a computer, the complaint, the victim, offender and witness statements, photos, videos and more. In the USA they have the Uniform Crime Reporting (UCR) [1] system for keeping track of crimes nationwide. IN India, the crime statistics are maintained by the National Crime Record Bureau (NCRB) [2]. They release a yearly report containing the statistics of various crimes across all the states and union territories in India. The statistics are provided across many different crimes such as robbery, homicide, motor vehicle theft, the crimes against women, scheduled tribes etc. NCRB has been collecting the information from the states from 1986. All these statistics have been published [3]. But there is no common framework across India for crime analysis. Most of the data is collected in each state in their own manner and classified according to their state laws. Nation-wide crime analysis is still a challenge in India because of the inherent differences in the states like languages and different state laws. Most of the states in India have their own language, and most of the people working in the police force are only fluent in their own native languages. So the complaints, FIRs, written documents, receipts and any other pieces of evidence are bound to be in the language of the state. If a case spans countrywide the search and notification itself take a lot of time to reach the police stations in other states. Many states have their own laws which are written on top of central government laws and they generally pertain to the problems faced in the particular state. Like the law against eve teasing [4], or the ban of one-time use plastic [5] by the state of Tamil Nadu and the law prohibiting exorbitant interests [6] by Kerala and few other states in India. In this paper, we aim to propose a common integrated framework for India.

2 Literature Review

Over the years there have been a few frameworks that have been proposed for crime analysis. The Regional Crime Analysis Program (ReCAP)[7] was one such early framework that provided integrating data from multiple sources using data fusion and crime analysis using data mining.

This system gave importance to spatial data mining and had an integrated GIS to show the results of mining on a map. Another framework that was developed using data mining techniques was COPLINK[8]. It was developed for Arizona state police as a means of identifying links between gangs and the members of the gangs. It was a combination of two projects COPLINK CONNECT [9], [10] which used network analysis for finding the links and COPLINK DETECT which was used to detect the false identities and addresses supplied by the suspects. Another framework was proposed to predict crime rates using social indicators [11] such as unemployment rates, police expenditures and the dynamic change in population. They have used mathematical modelling and regression to predict the change in crime rates with the dynamic change in macro social indicators. A new framework for hotspot matrix [12] was proposed exclusively for identification and controlling of crime hotspots. Once a hotspot is identified then a matrix of solutions is proposed to the police for reducing crimes in the hotspot. It uses the aoristic approach to identify spatial-temporal hotspots. A framework for searching and matching similar crimes using data mining and neural networks was proposed in [13]. In this, Self-Organizing Map Neural Network for clustering data to match the crime to similar crimes in the past. In this paper, we are proposing a generic framework for the day to day crime handling and analysis for the police personnel.

3 PROPOSED FRAMEWORK

The new framework, shown in fig 1, has been proposed with the aim of using deep learning techniques for the intelligent and timely analysis of crime. It encompasses a systematic approach for using deep learning neural networks to analyze crime data. By proposing this new framework, we aim to be able to leverage the deep learning neural network for prediction of crime rates and possible hotspots and for proactive policing and prevention measures. Each component of the framework is discussed further in this section.

• *Lydia J Gnanasigamani, Assistant Professor, School of Computer Science and Engineering, VIT University, Vellore, Tamil Nadu, India, lydia.jane@gmail.com*

• *Dr. Seetha Hari, HOD, School of Computer Science and Engineering, VIT University, Amaravati, Andhra Pradesh, India, hariseetha@gmail.com*

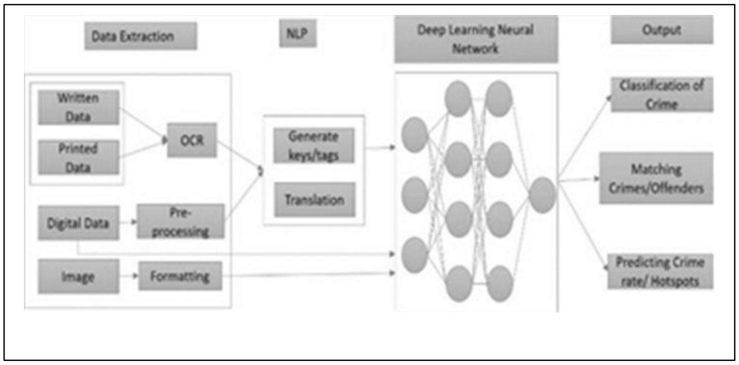

***Fig 1** Framework for Crime Analysis*

3.1 Automatic Data Extraction

Evidence related to a police case can be divided into text evidence, photo/videos and digital evidence like documents in the laptop or chats and messages in the smartphone. All of these need to be processed into a common storable format so that they can be used further for searching or matching crimes.The textual evidence comprises of the written complaint by the victim, the FIR (First Information Report), the narrative reports of the police officers conducting the case. The witnesses' statement of the crime, the evidence from the crime scene such as receipts, diary, letters, notebooks, post-its and so on. These documents are also data and they may play an important part in the identification of the offender. These documents, however, are not digital and have to be converted before they can be used for searching and other purposes. They are converted into digital format using Optical Character Recognition [14]. In India, these documents may be in different regional languages. The FIR is always registered in regional language. Now if the person is from another state and speaks another language, the evidence related to the case may be in another language. There are 22 official languages in India and all of them with their own script [16]. All Indian scripts are said to be derived from the Brahmi script [15]. All the languages scripts fall into three major families of scripts[17], the Devanagari script, the Dravidian script and Grantha script. A single OCR has not been developed yet for all the languages. Indian government sponsored OCR programs such as SanskritOCR, e-aksharayan[18] and Chitrakan. But still, these efforts have not resulted in an agile OCR. In many Indian languages, the base scripts may be combined to form compound scripts. These compound the problem of having to deal with varying scripts. Also, the handwritten documents are more complex because of the various different ways of writing the same letter.Other digital data may be in the form of documents or photos seized/recovered from laptops or PCs or smartphones. Some of these data may not be in the format that can be processed directly. Such data need to be formatted so that they can be stored along with the case data. One of the main difficulties in digital data is dealing with proprietary data. These require special programs to open them and most often they are licensed. Finding the right software that is required to open it is a challenge and storing the data in a format that is accessible for the case is a bigger challenge. Some of the documents may be encrypted or password protected or both. These may require subject experts to analyze and get the original document for investigation Photos and videos found during the investigation process can be preprocessed to extract important details from them. If it is a photo of person/persons, the facial features can be extracted and stored in the image database. It can also be matched with already existing images in the image database for known associates. If the image is a building then the building features can be extracted and its height and measurement can be compared to find the location. Similarly, we can use the extracted features of the vehicles to locate and intercept the offenders. If the picture is a natural scenery, then trees/plants and the animals/birds in the picture can tell the location as to where the picture was taken. Any drawings and paintings found during the investigation can also be preprocessed to extract the useful features out of them. The database that we have to build for the recognition system needs to be comprehensive so that we can get useful results In case of crimes within the home/apartment or office, we can use the crime scene pictures for reconstruction of crime scene thereby establishing the order of events. Every office and street today has a CCTV camera. They have also become part of households and a sense of security for most people. The videos that are recorded by CCTV cameras are accepted as evidence by all the courts today. We need advanced video and audio analytics to track and trace people and vehicles in real time. Facial recognition in real time video is an added advantage to the law enforcement officers.

3.2 Automatic Generation of Keys/Tags using NLP

Natural Language Processing is defined as the processing of language by computer. NLP is mostly associated with speech recognition, natural language generation and translation. Though the interest in NLP began as early as 1950 with Turing publishing his article on Intelligence [19], the field witnessed a huge growth once machine learning algorithms, both supervised and unsupervised were introduced for language processing. Hidden Markov Models and other statistical models are also increasingly being used in NLP. Nowadays deep neural networks called the neural machine translation (NMT) [20] are being used to directly learn transformation sequences eliminating the need for language modelling and word alignment requirement of the statistical methods. In our framework, the NLP component is used to generate automatic keywords from the case description and evidence. These keys are generated in a manner that they are easy to search, index and to translate. With 22 official languages, it is difficult to find a common way to represent the data. From the text data both the crime description and the accompanying evidence documents and the case notes by the police officers, using NLP we can generate keywords and fit the case into its appropriate category. The keyword generation is an important task as this is what would be used for matching by the next layer, the deep neural network. In India, we follow a common Criminal law across the country except for Jammu and Kashmir state. We will not deal with exceptions in this paper as our focus is to develop a common framework. The IPC (Indian Penal Code) [21] are followed almost uniformly across the country. But some acts like liquor ban is a law in few states like Gujarat when it is not the case in other states of India like Tamil Nadu, Andra Pradesh etc. Similarly, there are laws on cow slaughtering in states like Uttar Pradesh and Maharastra but in the rest of the country, there are no such laws. Also, there is the Goonda's act [22] which was enacted with the intention of curbing organized crime. This law is in effect in about 6 states in India including Tamil Nadu, Kerala, Karnataka and Maharastra. This law deals with gangsters, mafias and other organized crime. But sometimes it is also used against individuals to control the restlessness in the community or to stop a protest against the government. Similar types of crime also happen in states without this act. They may be registered under different IPC. We aim to overcome these differences by generating common keys and tagging the crimes under multiple sections so that we can find the similarity between them. Though most of the IPC remains same across India, the same is not the case for civil laws. The Civil laws are different in most states, meaning for the same offence the penalties may be different. By using keywords we can mitigate this issue to an extent. Apart from keyword generation and tagging, NLP is also used for translation. Translating one natural language to another while preserving the meaning of the input and following the grammatical rules of the target language is called machine translation. The advent of NLP made this task possible as it used huge datasets to infer rules. But even then the accuracy percentage was very low. Now deep learning is used for achieving more accuracy and customized solutions as per the requirements of the organization. All the cases are stored in regional languages and the evidence is stored in whatever language they are found. The search performed for any criteria, if it extends the boundary of state then it may require translation of the search query. The query can be performed in the regional language and all the other languages can also be translated to a regional language and the matching results can be produced. Or we can use a neutral language like English to store the keywords and do a common search across the database and produce the results. The results are then translated using NLP to a regional language and displayed accordingly If we use a neutral language like English, maintaining a common database for the entire country may be possible. But still, we have to think about the scalability and security of the database. Also, the government policies and rules that govern access to the database and data storage policies should be strictly followed.

3.3 Deep Neural Networks

Artificial Neural Network (ANN) is inspired by human biological neural networks, that is, the way in which neurons communicate with each other and the output which is the reaction to say heat or light. In the ANN [23], the signals are the activation functions and based on the output we get the results. They are many types of ANNs, from the simple perceptron to the advanced deep neural networks. In a feed-forward neural network, there is only one layer between the input and output layer. Deep Neural Networks (DNN)[24] is nothing but feedforward networks with more than one hidden layer between the input and output layer. The DNN uses sophisticated mathematical modelling to derive the output from the input. The relationship between input and output can either be linear or non-linear. Both recurrent neural networks [25], in which the data flows in different directions, and convolution neural networks [26], can be modelled as deep neural networks. Each layer models or transforms the input in a manner that is required for the output.

3.4 Classification using Deep Neural Networks

Classification is categorizing the data into already defined classes. It is a supervised learning methodology. In our proposed framework we use DNN for classification of crime. DNN is trained to classify the data according to the different inputs to an appropriate category or class. Once the training is complete, then the DNN is ready to classify the data into various categoriesMost of the crime data contain information like date, time, victim information, offender information and the type of offence. These parameters are crime characteristics. They describe a crime distinctively. The variables mentioned above are independent of crime type, that is, almost all the crimes have the spatial, temporal, victim, offender and crime specification. But there are also crime dependent variables, like, for homicide the additional variables will be like the weapon used, single/multiple victims, single/multiple offender and offender known to the victim or not. In case of burglary, the additional variables may be like the list of things stolen, entry method, exit method, type of the house, and method of burglary. Similarly, the additional variables will be different for vehicle theft and narcotics and for other crimes. As our aim is to develop a framework for the entire Indian country and we want to have to generic crime matching and prediction capabilities, we create our own defined categories, in-line with the IPC codes, to classify all the crimes. So the inputs to the DNN are both the generic variables and the crime-specific variables. Most of the data is expected to be text data and a few numerical data. Deep Convolution Neural Networks (DCN) [27] can be used effectively for text classification. Fig 2 shows the Deep CNN model. It has the convolution cells or the pooling cells and the kernel cells. The kernels process the input data and then it is simplified by the pooling cells.

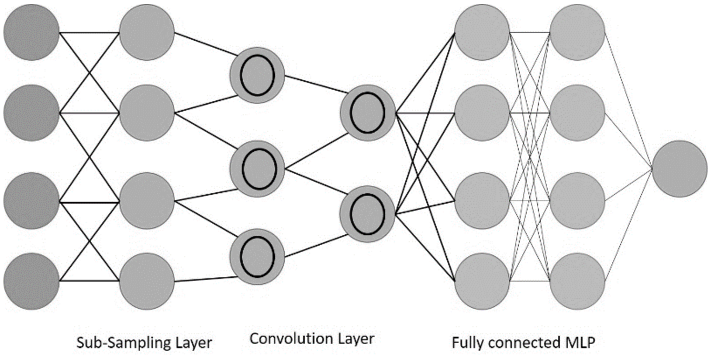

Fig 2 *Deep Convolutional Neural Network*

3.5 Clustering/Matching using Deep Neural Networks

Matching is matching either the criminal or the crime to similar incidences in the past. There are many unsolved cases in the police stations because of not enough evidence or not enough information. In our proposed system, we use clustering to match a crime to similar crimes in the past. For this, the input will the Method of Operation (MO). The idea is if the MOs are similar then the crimes may have been committed by the same

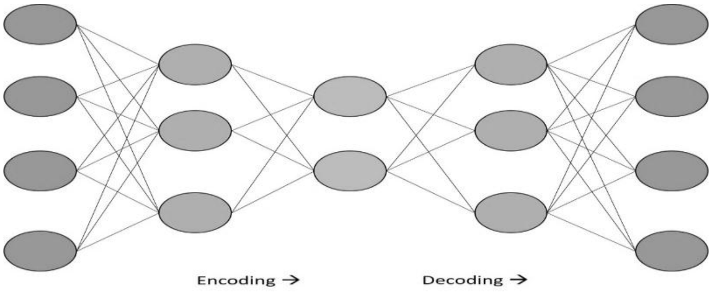

Fig 3 *Deep Auto Encoder*

person(s). This matching can work two ways. One is we have an already known MO and an offender who has the MO. So if in the future there is a crime with a similar MO, then it is possible that it was committed by the same person. Likewise, if we have a known offence and an MO, but we do not know the offender. If a person is arrested with a similar MO, then we can match the past cases too with the current case. The matching is done based on the clustering methodology. Clustering is grouping the data in a manner such that the members of a cluster are more similar to each other than the members of other clusters. For this operation the input to DNN will be the crime type and MO and if the offender information is known it is also provided. The result will the cases with similar MO. Autoencoders are the aptest choice for this. Autoencoders [28] are ANNs that compress data into a smaller representation or code, process it, and then decode into near original representation in the output in an unsupervised manner. So autoencoder by design automatically does dimensionality reduction. Deep autoencoders as shown in fig 3 have many layers between the input and output rather than the single layer as in a simple feed forward autoencoder.

3.6 Prediction using Deep Neural Networks

Prediction is forecasting the possible crime rate for the near future and the places that can become hotspots for a crime type. Crime rate prediction can be done for a type of crime or a place or both. Hotspot prediction is done for across a state or country and all the hotspots can be displayed. This analysis can be done by crime type. For both the analysis the historical data is important. We use temporal data such as year and month to calculate the crime rate using regression. Prediction is a mathematical model that tells us the future data by using past data. Mostly regression techniques are used for prediction. Regression is the method of modelling the relationship between data to analyze the way they contribute to the outcome together. Linear regression is used when the relationship between the variables is linear. If the relationship is non-linear than we can use polynomial regression. Logistic regression [29] also called the binomial regression can be used when the prediction has only two states. Overfitting is an issue with regression models. In Deep Neural Network we can use LSTM model as shown in fig 4 for prediction. Long Short Term Memory model or simply Recurrent Neural Networks [30] can remember the past states and makes use of the past information to make predictions.

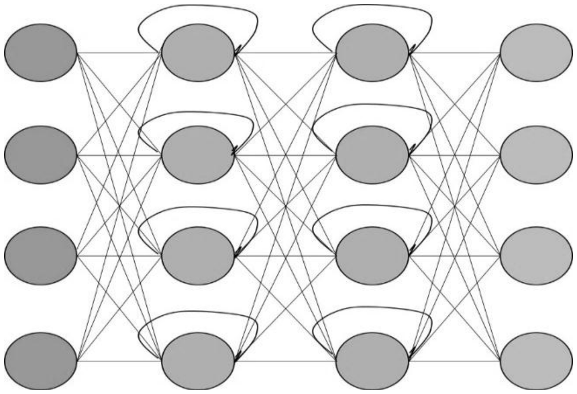

***Fig 4** Deep Recurrent Neural Network*

So, in the proposed framework, the Deep Neural Network layer will be a combination of both Convolutional Neural Network and Recurrent Neural Network in a manner that all of them work and the performance is good enough to get real time results. CNNs and RNNs are historically considered to be not compatible with each other. And in most cases where they have been used together, it is using the layered approach. CNN layers output has been used as input to the RNN layer. But there have been some cases in which CNN and RNN have been combined as a single neural network. These have been mostly in the domain of video analytics [31], weather prediction [32] and emotion detection [33]. Our proposed framework intends to extend this to the data analytics domain.

4 Conclusion

In this paper, we have seen about leveraging the power of Deep Neural Networks for crime data analysis. Through this approach, we think that crime fighting in India as a whole nation can be done better by improved response time and proactive policing in predicted crime hotspots. We also hope to clear cases at a faster rate by matching the crime details to offenders and past cases. We hope to expand this framework to include analysis of spatiotemporal data. We also hope to include known addresses and known associates of offenders in the database to identify the whole network. We also intend to add link analysis for social media data and include cybercrimes such as cyber terrorism and cyberstalking. We also intend to implement the components of the framework as an extensible software for the police to plug into their databases. We can also include cybercrimes and the tools for carrying out digital forensics in the framework.

References

[1] CA Sennewald, C Baillie, Effective security management, Butterworth-Heinemann; 2015 Aug 15.

[2] "National Crime Records Bureau", 2019. Ncrb.Gov.In. http://ncrb.gov.in/.

[3] "CRIME-IN-INDIA 2000-Till Date", 2019. Ncrb.Gov.In. http://ncrb.gov.in/StatPublications/CII/PrevPublications.htm.

[4] "Eve-teasing-act-1998,1998",2019. Draglc.Ac.In. http://www.draglc.ac.in/pdf/EveteasingAct1998.pdf.

[5] "Single-use Plastic Ban, 2018",2019. Tnpcb.Gov.In. http://www.tnpcb.gov.in/pdf_2018/G.O_84_BanPlastic3718.pdf.

[6] "Prohibition Of Charging Exorbitant Interest Act, 2013", 2019. Lawsofindia.Org. http://www.lawsofindia.org/pdf/kerala/2013/2013KERALA2.pdf.

[7] DE Brown, The regional crime analysis program (RECAP): a framework for mining data to catch criminals. InSMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218) 1998 Oct 11 (Vol. 3, pp. 2848-2853). IEEE.

[8] H Chen, W Chung, JJ Xu, G Wang, Y Qin, M Chau Crime data mining: a general framework and some examples. computer. 2004.

[9] H Chen, J Schroeder, RV Hauck, L Ridgeway, H Atabakhsh, H Gupta, C Boarman, K Rasmussen, AW Clements, COPLINK Connect: information and knowledge management for law enforcement. Decision support systems. 2003 Feb 1;34(3):271-85.

[10] H Chen, D Zeng, H Atabakhsh, W Wyzga, J Schroeder, COPLINK: managing law enforcement data and knowledge. Communications of the ACM. 2003 Jan;46(1):28-34.

[11] KC Land, M Felson, A general framework for building dynamic macro social indicator models: including an analysis of changes in crime rates and police expenditures. American Journal of Sociology. 1976 Nov 1;82(3):565-604.

[12] JH Ratcliffe, The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. Police practice and research. 2004 Mar 1;5(1):5-23.

[13] MR Keyvanpour, M Javideh, MR Ebrahimi, "Detecting and investigating crime by means of data mining: a general crime matching framework". Procedia Computer Science. Jan 2011 Jan 1;3:872-80.

[14] S Mori, H Nishida, H Yamada, Optical character recognition. John Wiley & Sons, Inc.; 1999 Jan 1.

[15] RM Sinha, A journey from Indian scripts processing to Indian language processing. IEEE Annals of the History of Computing. 2009 Jan;31(1):8-31.

[16] "Profile - The Union - Official Language - Know India: National Portal Of India", 2019. Knowindia.Gov.In. http://knowindia.gov.in/profile/the-union/official-language.php.

[17] India: Languages And Scripts" 2019. Cs.Colostate.Edu. https://www.cs.colostate.edu/~malaiya/scripts.html.

[18] "e-aksharayan",2019. https://tdil-dc.in/index.php?option=com_content&view=article&id=155:e-aksharayan&catid=78:main-areas&Itemid=435&lang=en.

[19] AM Turing, "Computing machinery and intelligence". In Parsing the Turing Test 2009 (pp. 23-65). Springer, Dordrecht.

[20] D Bahdanau, K Cho, Y Bengio, "Neural machine translation by jointly learning to align and translate". arXiv preprint arXiv:1409.0473. 2014 Sep 1.

[21] V Dhagamwar, Law, power and justice: Protection of personal rights under the Indian Penal Code. Sage Publications; 1992.

[22] "Goondas Act,1982". 2019. Lawsofindia.Org. http://www.lawsofindia.org/pdf/tamil_nadu/1982/1982TN14.pdf.

[23] X Yao, "Evolving artificial neural networks", Proceedings of the IEEE. 1999 Sep;87(9):1423-47.

[24] J Schmidhuber, Deep learning in neural networks: An overview. Neural networks. 2015 Jan 1;61:85-117.

[25] R Pascanu, C Gulcehre, K Cho, Y Bengio, How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026. 2013 Dec 20.

[26] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. InAdvances in neural information processing systems 2012 (pp. 1097-1105).

[27] Y Kim, "Convolutional neural networks for sentence classification". arXiv preprint arXiv:1408.5882. 2014 Aug 25.

[28] GE Hinton, RR Salakhutdinov, "Reducing the dimensionality of data with neural networks". science. 2006 Jul 28;313(5786):504-7.

[29] D Pregibon, Logistic regression diagnostics. The Annals of Statistics. 1981;9(4):705-24.

[30] C Huang, J Zhang, Y Zheng, NV Chawla, DeepCrime: Attentive Hierarchical Recurrent Networks for Crime Prediction. InProceedings of the 27th ACM International Conference on Information and Knowledge Management 2018 Oct 17 (pp. 1423-1432). ACM.

[31] C Streiffer, R Raghavendra, T Benson, M Srivatsa. "Darnet: a deep learning solution for distracted driving detection". In Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial Track 2017 Dec 11 (pp. 22-28). ACM.

[32] AG Salman, B Kanigoro, Y Heryadi. "Weather forecasting using deep learning techniques". In 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS) 2015 Oct 10 (pp. 281-285). IEEE.

[33] Y Fan, X Lu, D Li, Y Liu. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks". In Proceedings of the 18th ACM International Conference on Multimodal Interaction 2016 Oct 31 (pp. 445-450). ACM.