*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 6*
*DOI: https://doi.org/10.17762/ijritcc.v11i6.7787*
***Article Received**: 08 April 2023 **Revised**: 16 June 2023 **Accepted**: 02 July 2023*

# Feature based-Learning with Data Increasing for video Recommendation and Computing

**Ramakrishna Reddy K[1], Dr. B. Dhanasekaran[2], Prof. Dr V. K. Sharma[3]**
[1]Research Scholar, Department of Computer Science and Engineering
Bhagwant University,Rajasthan-India,
Assistant Proffesor Department of Computer Science and Engineering
KG Reddy College of Engineering and technology,hyderabad,telangana,india
Krkreddy20@gmail.com
[2]Research Supervisor,Department of Computer Science and Engineering
Bhagwant University,Rajasthan-India.
drbdhanasekaran@gmail.com
[3]ResearchCo-Supervisor,DepartmentofEEE
Bhagwant University, Rajasthan-India.
Viren_krec@yahoo.com

**Abstract** :Image content analysis is crucial for determining the reliability of a link between two videos. Video characteristics are increasingly being used in image and video representation as custom pre-trained picture and video convolutional neural networks become generally available. People also have limited access to video editing tools for a variety of reasons, such as ownership and privacy concerns. You don't need to go back to the source video data to use the refined features again. An affine transformation, for instance, can be used to map a well-studied function onto an unfamiliar domain. To do this, we use a unique triplet failure in conjunction with the re-learning strategy. We propose a contemporary data augmentation method that may be applied to functionality on various frames for videos as an alternative to employing specific motion data. Extensive testing on the well-known Hulu content-based Video Relevance challenge demonstrates the process's efficacy and provides solid evidence of state-of-the-art performance.

**Index Terms**—Feature re-learning, ranking loss, data augmentation, content-based video recommendation.

## I. INTRODUCTION

Several recommendations have been issued in feature based learning such as [1, 2, 6] recognition, image segmentation, video copy identification, retrieval behind the very beginning background of film, of video. The algorithm tries to help users find something that they would like to see, whether it is interesting or not [3,4,5]. Instead of words, a video's images are more dynamic, bringing a fresh dynamic to several videos. For example, it may be created for the purpose of increasing traffic, but not necessarily material of the video is fine. The kinds of approaches dependent on the text's subject matter may vary shoddy workmanship. According to Yang et al., [7], consumers would use video suggestion even more than text. Darkness, black and white, or colorful which motion features are accessible, and how often should they be used? A work of recent vintage extracted using Deldjoo and his team predicting converted Convolution Neural Network (CNN) in comparison with other similar models carrying out vision-related tasks [8,9,10,11,12]. In this paper, it concentrates on the graphic/graphic design and visuals. Off-the-self visual characteristics are not unusual. Techniques in general should be used to increase the weight of video importance. We are seeing the beginning of an attempt to introduce a new video feature room, thereby determining the importance of the video. It emphasized on increasing the meaning, semantic vectors efficient retrieval or dimension reduction through hashing full with several meanings. A consequence of this is that guided learning is essential. There is an absolute need for function learning.

It is essential to make the most of the features in a video. The more features, the more important the video, the less importance there is to be attached to it. It's like trying to make a video more interesting by adding more features to a video that is already popular. It works by making the video more and more interesting and visually engaging. It makes the video look more like a novel. It creates a novel way of showing the video in a new and different way. It gives the user the chance to experience the video for the first time in a different way, in a unique and novel way. The video is more interesting than the video before. We augment an existing data augmentations shoot explicitly for name recognition and video-level purposes, without cutting something else. We test video relevance by using a content-based approach context-driven approach. However, taking lots of videos that describe

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 6*
*DOI: https://doi.org/10.17762/ijritcc.v11i6.7787*
***Article Received**: 08 April 2023 **Revised**: 16 June 2023 **Accepted**: 02 July 2023*

the same topic and finding common patterns among them is a laborious and monetary intensive process. The need in viewing original footage in this article, we test video relevancy. In order to meet these expectations, they'll have to replay the initial video. It is customary for movies and TV shows [13] to be covered by copyright, but not required.

There is the difficulty of estimating how interesting seed videos would be and a short video as well. No ability to read initial source files Instead, they facilitate the two visual elements, drawn from the amount of single-frame images and predefined Inception-3 [14] tagsC3D versions. An appropriate methodology to test our approach on philanthropic activities

There are the following contributions towards this model:

- It proposes a retrieval-based function numerical relationship. It uses a short-term maximum distance loss function for video feature predictions. Overused triplet rating. This new theory provides improved results, even though it loses less performance. Yet has a greater pace of convergence.
- We propose an improved feature: "feature re-learning "manipulation the overall game plan. This may be extended at the video or frame level functions. Without reference to prior knowledge or theory. Due to our commitment to videos, our security policy is helpful to us filtering camera info.
- Accompanied by the function set lack of NERT as well figures representing data augmentation, one of the computer's planned video features on real data sets generous.

A finely-crafted digital app for Hulu was one of the three winning entries in the video relevance competition. It has created a new loss role to oversee feature re-relearning. We are (partially) understood between candidate videos advocate a modern formula for determining video significance. The goal is to help the customers pinpoint areas in which we need to improve our services. It equality regards to the function architecture the precision of the suggested approach. This feature enables the user to add a qualified model for video relevance.

## II. RELATED WORK

Relevance is determined by importance of media as set in time or dimension spaced by normal unit of measurement. The papers by Bhalgat et al. present [15] present use state-of-the techniques. It's advantageous to be progressive. The benefit of our research often consists of learning new [16,17,18,19]. Reinvention for feature relearning is a good way to learn how to use video function space classification. It works well when shown as part of a series of videos creative inputs instead of looking for innovative ways to create value in the video. The video was hopelessly badly scored. However, at runtime, the model is still considered to be a model. Triple start/triplet loss is popular for quantitative ranking tasks To increase the visual impact of two-dimensional video, Karpathy et al. [4] have applied generative algorithms in their model. It's possible to develop a distance constraint fostering the hostility of a more distant negative in the diagram.

The problem of how to download and handwritten character recognition image relevance has not been substantiated. This issue has been definitively resolved. It uses extra sorting in place of grouping additional ground reality data. It considers both our score and our rank identical of look and style as well as our planned loss to rank our performance. We believe that the new method is a more effective way of ranking score. It doesn't need extra sorting. It uses extra data. It is a new method of grouping data. We use pixel-level data augmentation to make our frames look like that. Instead of segments, we collected segments inconsistent. A cluster is made up of equally or homogenously down sampled parts. As the frame rate of video images gets faster, so does the quality. These videos tend to be an equal creative level a simple scheme that causes noise to be injected into conditional generative models. We have full control of the initial recordings; it naturally guarantees protection and privacy problems about video data. It's the most significant gap generative model to generate the noise. Using the network would help us to produce different types of samples, when we're engaged in this project fuzzy signals are used to boost the system's discrimination. The system would then use fuzzy signals to boost its discrimination. The video-level data increase emerges to be comparable to GAN [20] and CVAE [21]. Based on the above related work, it proposed the video relevance framework as next section.

## III. VIDEO RELEVANCE FRAMEWORK

We focus on supplying relevant video content based on 2D video when discussing d-dimensional attributes a graphic representation of the topic. When considering two images aimed towards both content creators and those who search for new content resembling, then we can determined the content based relevance as per following equation.

$$\mathrm{Cos}(v,v') = \frac{v.v'}{||v||.||v'||} \qquad (1)$$

Where v and v′ are two videos based on d-dimension. As mentioned in Section 1, it is built as a component pre-trained by a CNN model which is not so effective task. Therefore, it suggest a new video features represent as Φ(v) where cos(Φ(v),Φ(v′)) create better video relevance with certain characters based on ground truth data. Thus relevant video

*International Journal on Recent and Innovation Trends in Computing and Communication*
***ISSN: 2321-8169 Volume: 11 Issue: 6***
***DOI: https://doi.org/10.17762/ijritcc.v11i6.7787***
***Article Received**: 08 April 2023 **Revised**: 16 June 2023 **Accepted**: 02 July 2023*

pairs represent as D = $\{v,v^+\}$. It begins on ranking-oriented function with use off-the-shelf computer learning approaches that create a new function and then consider this choice for implementation into another feature space tactical multi-level augmentation which also calls the overall strategy interface elements such as framing and captioning. In the end, it represent the strategies for video relevance prediction as per video feature space.

*A. Ranking- based Feature Learning*

It consider rank-dependent with feature re-relearning algorithm to redesign an old video element into a new/use feature space which do a great job of representing key features. The amount of frame features can change over the course of images, depending on needs feature data. It uses pooling which has been shown to be easy and reliably useful capable of using a wide variety of media for different creative activities on various activities task [1,14,22,23]. The more sophisticated features are used in addition to create new feature space. Thus the fresh function can be defined with feature vector as presented as new:

$$\Phi(v) = Wv + b \qquad (2)$$

Where W Є $R^d$ trained matrix, b Є $R^p$ is bais term. Their resemblance is calculated according to the cosine similarity between them. The affine transition is often referred to as a single-layer, integrated usage i.e., fully connected network (FCN). In theory, for example, a multilayer FCN, like a 2-layer FNC,we can use a two-layer filter-class network. We the network design will examine which one best suits. In order to facilitate learning, we apply a negative feedback (or gradient) to the model. When it comes to numerical indexing, the triplet ranking loss(TRL) is extremely popular and is very effective operational functions the equivalent of three losers to complete a pattern for practice, we use triplets Inspiration from a related video set; Inspired by: positive videos and counter-relevant. A slight pause is desirable before making up lost time by an irreplaceable miscalculation (or loss) which significantly influences its success influencing the content of videos by training videos relevant to the target user. By using this principle, we increase the risk tolerances of the model. Adding a negative pair restriction to the TRL makes it easier. We apply the concept by using max(0, Cos Φ) as the ideal observation.

*B. Increasing feature sets on many levels*

One strategy to boost the effectiveness of the analysis is to use a larger data set. Using such models produces remarkable outcomes. Not enough data from training has been collected. The lack of original, on-demand video programming without robotic camera work is a major drawback of Hulu's model. Traditional instance-level enhancement as a foundation for video production is becoming increasingly problematic. There is no use for any of these methods (which include spinning, turning, zooming in, and out). On top of that feature, we are currently building a multidimensional improvement. One strategy that works well with video is constantly adding new content. Multimedia pieces, as opposed to videos of original content, are the focus of our policy.

Enhancing features in a given frame getting ideas from real life We introduce these few frames as the basis for the rest of the video, so that viewers may get the idea even if they didn't see the whole thing. Methodology of the Survey Using a non-random sample Locate a video that contains n still images. We like to prank each other. When drum samples overlap, the tempo changes by more than one beat at a time. As so, fresh sequences of design elements are generated. which means that once we discovered the necessary means pooling, we were able to actualize the v. Both Stride and the full series allow for sampling, which provides more data for testing and refining the appropriate video-level interest training. Fine-tuning is the process of making slight but noticeable adjustments to improve the quality of a video's characteristics. It is not possible to perceive some pixels with the naked eye since they are so small. Our video rating framework, likewise, must not be distracted by trivial details. When features are extracted, we use that information to make more targeted adjustments to each video frame's function based on the image's content. A d-dimensional video level of detail in a 2-dimensional format. However, noise is generated through purely random processes. The function v has been perturbed, leading to these findings.

The minimal perturbation has a probability of 0.5 for Bernoulli random variables (p). Based on the values of the other variables, a noise vector with the label "e=1" was generated from a three-dimensional Gaussian distribution. Symbol means multiplication of all elements thorough multi-level function synthesis pixel-level features obtains preparation data genius creative assimilation awareness are essential to photographic success. Innovation should be individually applied to improve and expand the capabilities of man power. It is only capable of producing video-level additional examples that have further instruction. If the option is not present at the frame level of a video, disable it. The feature-oriented rating re-caller feature is much better due to additional function augmentation training commonly referred to as "deanonymized or de-identified details", or otherwise neutral task-based truth with respect to specific information. More information can be expressed in a modern visual space seeks is from tailoring to the material Relevance and use with online video.

*International Journal on Recent and Innovation Trends in Computing and Communication*
***ISSN: 2321-8169 Volume: 11 Issue: 6***
***DOI: https://doi.org/10.17762/ijritcc.v11i6.7787***
***Article Received**: 08 April 2023 **Revised**: 16 June 2023 **Accepted**: 02 July 2023*

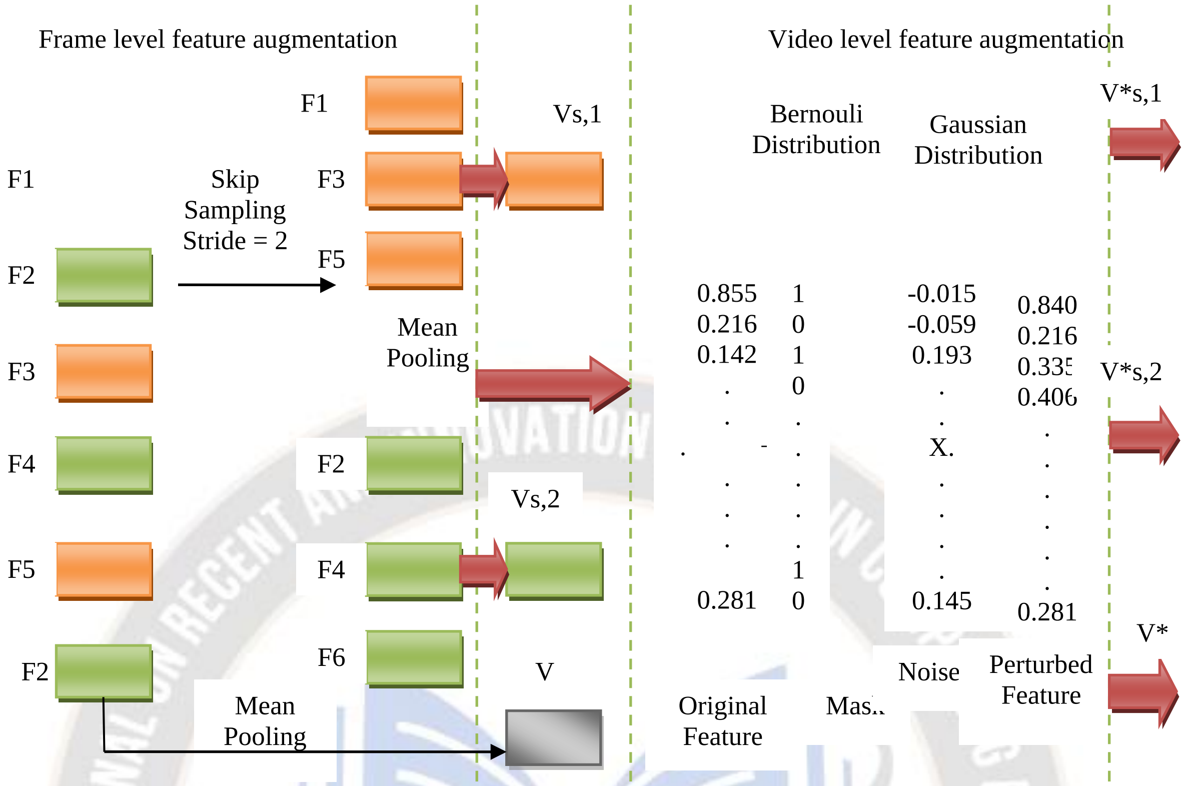



Fig. 1. Two steps: Frame-level feature augmentation and video-level feature augmentation. for features

## IV. EVALUATIONS

### *A. Experimental Setup*

In experiment parts, it considered Hulu Dataset for evaluation. This part points to evaluate the feasibility of the concept Fictional solution: suggested function training, we use the TV shows Pictures and movie data set given by HULU in the appropriate sense Semantic Video Relevance Analysis. Each part of the dataset has been subdivided into distinct parts for training measurement component of3,000video testing. It collected 4,500 movies and a dataset of 1,188 videos dataset. Both videos are complete dramas instead of clips from TV shows or movies. For Movies, feature-length films, there is a video for each recording in the training and test data sets. For a collection of related videos, the sourced information is recommended for it. Videos that are important to make interests relevant video that is labeled as R is said to be half of video resolution. The correct result would be reflected in the ground reality of the test the collection is unaudited. Regarding the category of recommended videos, we adhere to the union of recordings must always be employed when measuring the performance of achievements validation set, as a test.

We used videos when we were taking the exam on the programmed set. Regarding the above problem, the HULU challenge does not make provision fascinating concept rather than two previously computed functions. For example, they are analogous to functionality, such as frame-specific content and video-specific attributes. For higher-level functionality, such as videos coded to handle data at one-bit-per-second. Inception converts the decoded frames constructed on the

Image Net dataset of 2,048 layers of ReLU activations, the final secret layer “Perspective" is one of the style layer features. For color fidelity, this feature set was used to create the Sports C3D model that was trained on the Sports1 dataset. They’re worked to the full. Each video is encoded and decoded at 8 frames per second. The 512-dimensional activations of the pool5 layer are used one of the last video clips. To facilitate rapid comprehension, we will use for the sake of convenience. The iterative approaches use an evaluation procedure that is in line with standard creative practice. Two rank-based measures, recall@k where (k=50, 100, or 200) and hit@k where (k=5, 10, or 20), are used to calculate the success score. A digital computation seed v is generated for a movie in the following way:

$$\text{recall@k}=\frac{|R_v \cap \widehat{Rvk}|}{|R_v|} \qquad (3)$$

*International Journal on Recent and Innovation Trends in Computing and Communication*
***ISSN: 2321-8169 Volume: 11 Issue: 6***
***DOI: https://doi.org/10.17762/ijritcc.v11i6.7787***
***Article Received**: 08 April 2023 **Revised**: 16 June 2023 **Accepted**: 02 July 2023*

$$\text{hit@k} = \begin{cases} 1, & if\ recall@k > 0 \\ 0, & otherwise \end{cases} \quad (4)$$

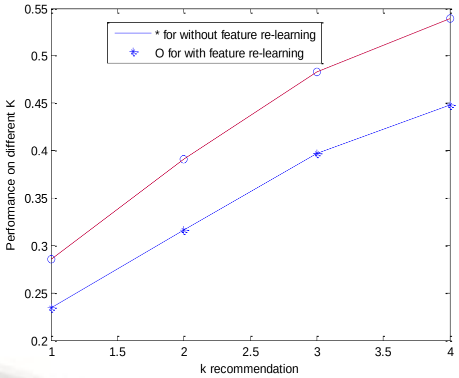
where $\widetilde{R_{vk}}$ is considered k recommended videos from individual video set.The creativity is rated by counting recall and applying it to the total ranking. Both the videos are handled using state-of-the-art tools make improvements. Further, PyTorch is considered in training for reinforcement learning. We also installed a deep learning environment empirically, we placed the margin effect at the values marg1 and marg2.Equivalent result got as in 0.2 and 0.05 in order to allow. We model our approach with the stochastic gradient descent algorithm with Adam Empirically begin at a value of 0.001 and apply the learning function to that value have a maximum batch size of 32. Once the error rate remains constant, it changes the learning rate to half of its previous value measure up to expectations.

The aim of this experiment is to see how powerful learning features are corresponding to creative development video relevance prediction. We examine the interrelationship from the overall results to the complexities of the dimensions. Specifically, by way of example we equate the effects of Inception v3 and C3D functions. Here the width is multiplied by 2, such that it reaches a size in the region as shown in table 1. Here, the movies dataset often did not use the proposed data augmentation approach in this instance. The remainder of the tests take place in 512 locations for next evaluation.

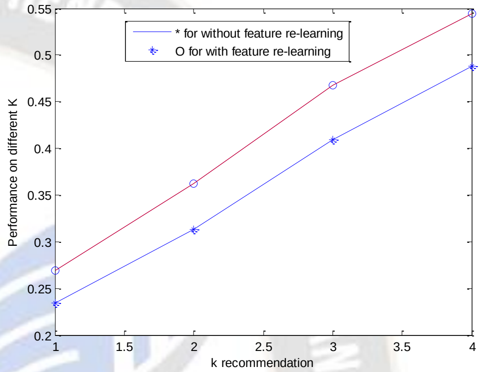
Table 1: The performance on (a) TV-shows and (b) Movies datasets

| | TV shows | | Movies | |
|---|---|---|---|---|
| Dimensionality | Inception v3 | C3D | Inception v3 | C3D |
| 32 | 2.61 | 2.5 | 1.95 | 1.89 |
| 256 | 2.68 | 2.55 | 2.005 | 1.92 |
| 512 | 2.7 | 2.62 | 2.031 | 1.95 |
| 1024 | 2.7 | 2.6 | 2.024 | 1.94 |
| 2048 | 2.69 | 2.56 | 2.010 | 1.92 |

Figures 2 and 3 demonstrate the models' performance regardless of whether or not they have this feature. It is unable to relearn a predefined measurement standard because it already knows the underlying features. An enormous productivity boost can be achieved by re-learning the interpretive meaning of all datasets.

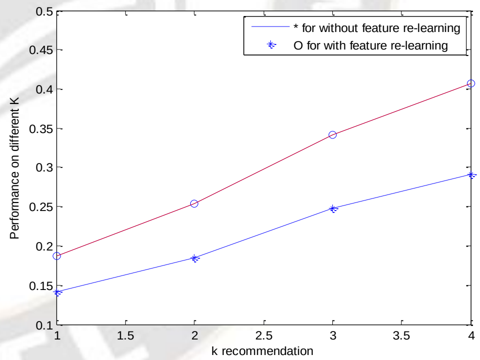



(a)Inception v3

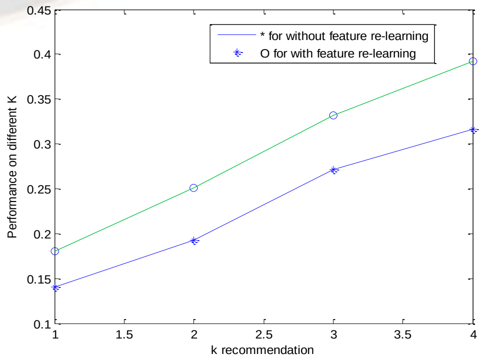



(b)C3D

Fig 2: The result of (a)Inception v3 and (b)C3D TV shows




(a)Inception v3




(b)C3D

Fig 3: The result of (a)Inception v3 and (b) C3D Movies

Figures 2 and 3 display hit@k's efficiency at recommending hits throughout a range of k values, from 5 to 30. It evaluated its performance in both the presence and absence of feature re-learning. The model's ability to learn new information is useful in many different contexts. The one where we never have to teach the model anything new but instead just train it to consistently outperform its competitor using the features we care about. In addition, you can choose between Inception v3 and C3D. User expectations for the value of taught features are typically met by findings explained creatively. Both Inception-v3 and C3D features perform similarly well with and without feature relearning in this model. The importance of feature relearning in video relevance prediction is demonstrated by these results.

## V. CONCLUSIONS

To identify the task-specific importance of videos, this research suggests feature re-learning model with a deeper dive into the data. An updated version of the negative-enhanced triple rating failure (NETRL) learns the model in the context of a predetermined task. HULU's Content-based Relevance and Video Relevance Prediction Challenge data sets were used to conduct extensive experiments that corroborate the findings. NETRL is superior to traditional triplet rating defeats in terms of both consistency and speed of convergence. It is possible to improve video importance prediction by taking into account the candidate videos' interconnectedness. There is medium between the model's efficiency and Gaussian noises to the multi-level increase in data. As per HULU model, the presented approach could be useful for other projects that require predicting the relevance of videos based on their content.

## REFERENCES

[1] M. Liu, X. Xie, and H. Zhou, “Content-based video relevanceprediction challenge: Data, protocol, and baseline,” 2018, arXiv:1806.00737.

[2] X. He, Z. He, J. Song, Z. Liu, Y.-G. Jiang, and T.-S. Chua, “NAIS:Neural attentive item similarity model for recommendation,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 12, pp. 2354–2366, Dec. 2018.

[3] J. Song, L. Gao, F. Nie, H. T. Shen, Y. Yan, and N. Sebe,“Optimized graph learning using partial tags and multiple features for image and video annotation,” IEEE Trans. Image Process.,vol. 25, no. 11, pp. 4999–5011, Nov. 2016.

[4] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, “Eventdriven web video summarization by tag localization and key-shot identification,” IEEE Trans. Multimedia, vol. 14, no. 4, pp. 975–985,Aug. 2012.

[5] X. Liu, L. Zhao, D. Ding, and Y. Dong, “Deep hashing with categorymask for fast video retrieval,” 2017, arXiv:1712.08315.

[6] H. Liu, H. Lu, and X. Xue, “A segmentation and graph-basedvideo sequence matching method for video copy detection,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 8, pp. 1706–1718, Aug. 2013.

[7] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, “Onlinevideo recommendation based on multimodal fusion and relevancefeedback,” in Proc. ACM Int. Conf. Image Video Retrieval,2007, pp. 73–80.

[8] H. Xie, S. Fang, Z.-J. Zha, Y. Yang, Y. Li, and Y. Zhang,“Convolutional attention networks for scene text recognition,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 15, no. 1s, 2019,Art. no. 3.

[9] C. Xu, X. Zhu, W. He, Y. Lu, X. He, Z. Shang, J. Wu, K. Zhang,Y. Zhang, X. Rong, Z. Zhao, L. Cai, D. Ding, and X. Li, “Fully deep learning for slit-lamp photo based nuclear cataract grading,”in Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervention,2019, pp. 513–521.

[10] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, “Automated pulmonarynodule detection in CT images using deep convolutional neural networks,” Pattern Recognit., vol. 85, pp. 109–119, 2019.

[11] Z. Chen, S. Ai, and C. Jia, “Structure-aware deep learning forproduct image classification,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 15, no. 1s, 2019, Art. no. 4.

[12] Annam, S. ., & Singla, A. . (2023). Estimating the Concentration of Soil Heavy Metals in Agricultural Areas from AVIRIS Hyperspectral Imagery. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 156 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2519

[13] M. Wang, C. Luo, B. Ni, J. Yuan, J.Wang, and S. Yan, “First-persondaily activity recognition with manipulated object proposals and non-linear feature fusion,” IEEE Trans. Circuits Syst. Video Technol.,vol. 28, no. 10, pp. 2946–2955, Oct. 2018.

[14] M. Larson, A. Zito, B. Loni, and P. Cremonesi, “Towards minimalnecessary data: The case for analyzing training data requirements of recommender algorithms,” in Proc. FATREC Workshop ResponsibleRecommendation, 2017, pp. 1–6.

[15] Smith, J., Ivanov, G., Petrović, M., Silva, J., & García, A. Detecting Fake News: A Machine Learning Approach. Kuwait Journal of Machine Learning, 1(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/142

[16] M. Mazloom, X. Li, and C. G. Snoek, “TagBook: A semantic videorepresentation without supervision for event detection,” IEEE Trans. Multimedia, vol. 18, no. 7, pp. 1378–1388, Jul. 2016.

[17] Y. Bhalgat, “FusedLSTM: Fusing frame-level and video-level featuresfor content-based video relevance prediction,” 2018, arXiv:1810.00136.

[18] Muhammad Rahman, Automated Machine Learning for Model Selection and Hyperparameter Optimization , Machine Learning Applications Conference Proceedings, Vol 2 2022.

[19] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris,“Near-duplicate video retrieval with deep metric learning,” in Proc.IEEE Int. Conf. Comput. Vis., 2017, pp. 347–356.

[20] J. Lee and S. Abu-El-Haija, “Large-scale content-only video recommendation,”in Proc. IEEE Int. Conf. Comput. Vis. Workshop, 2017,pp. 987–995.

[21] Y. Dong and J. Li, "Video retrieval based on deep convolutionalneural network," in Proc. Int. Conf. Multimedia Syst. Signal Process., 2018, pp. 12–16.
[22] J. Lee, S. Abu-El-Haija, B. Varadarajan, and A. P. Natsev,"Collaborative deep metric learning for video understanding," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2018,pp. 481–490.
[23] Prof. Shweta Jain. (2017). Design and Analysis of Low Power Hybrid Braun Multiplier using Ladner Fischer Adder. International Journal of New Practices in Management and Engineering, 6(03), 07 - 12. https://doi.org/10.17762/ijnpme.v6i03.59
[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets,"in Proc. Int. Conf. Neural Inf. Process. Syst., 2014, pp. 2672–2680.
[25] K. Sohn, H. Lee, and X. Yan, "Learning structured output representationusing deep conditional generative models," in Proc. Int. Conf. Neural Inf. Process. Syst., 2015, pp. 3483–3491.
[26] J. Dong, X. Li, W. Lan, Y. Huo, and C. G. M. Snoek, "Early embeddingand late reranking for video captioning," in Proc. ACM Int. Conf. Multimedia, 2016, pp. 1082–1086.
[27] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embeddingand translation to bridge video and language," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4594–4602.