iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51753

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Unsupervised Video Segmentation using Quantization

Shreyas KS[1], Sushmitha Y V[2], Vakamalla Keerthi Priya[3], Gudapati Sai Ramakrishna[4], Prameetha Pai[5]
[1, 2, 3, 4]*Undergraduate Student,* [5]*Assistant Professor, Department of Computer Science Engineering BMS College of Engineering Bangalore, India*

Abstract: ***Unsupervised video segmentation is a challenging task in computer vision that involves dividing a video into meaningful segments without any labeled data or prior knowledge of the video content. One approach to achieving this is the use of quantization, which involves clustering similar image patches or video frames into discrete groups based on their visual features.Beta-VAE is a type of variational autoencoder (VAE) that is capable of learning disentangled representations of data. It can be applied to video segmentation using quantization, allowing for more effective segmentation and analysis of complex video datasets.The use of Beta-VAE and quantization is essential for several reasons. Firstly, it allows for the automatic analysis of large video datasets without the need for manual annotation, which can be time-consuming and expensive. Secondly, it enables the detection and tracking of objects and events in videos, which has applications in surveillance, robotics, and autonomous driving. Finally, it can be used for content-based video indexing and retrieval, which is crucial for video search and recommendation systems.The advantage of using Beta-VAE in video segmentation is that it can learn disentangled representations of video frames, which separates the data into meaningful factors of variation. This leads to more accurate segmentation, as the model can distinguish between different objects and events in the video, even if they have similar visual features.In conclusion, the use of Beta-VAE and quantization is necessary for advancing the field of computer vision and improving the analysis and understanding of videos. It has numerous applications in various industries and can contribute to the development of more effective and efficient video analysis tools.***

Keywords: Video Segmentation, Deep Learning, Quantization

I. INTRODUCTION

Imagine you are watching the movie 3 Idiots by Aamir Khan: In the climax scene Aamir Khan is seen operating on a Remote-control plane which glides over lake Ladakh. Suddenly, the story jumps into an emotional scene where Aamir Khan is shocked to see Kareena Kapoor approaching him. Such a dramatic change of scenes plays an important role in the movie's storytelling. Generally speaking, a movie is composed of a well-designed series of fascinating scenes with transitions, where the underlying storyline determines the order of the scenes being presented. Therefore, recognizing movie scenes, including the detection of scene boundaries and the understanding of the scene content, facilitates a wide-range of movies understanding tasks such as scene classification, cross movie scene retrieval, human interaction graph and human-centric storyline construction.

It is worth noting that shots and scenes are essentially different. In general, a shot Is captured by a camera that operates for an uninterrupted period of time and thus is visually continuous; while a scene is a semantic unit at a higher level. A scene comprises a sequence of shots to represent a semantically coherent part of the story. Therefore, whereas a movie can be readily divided into shots based on simple visual cues using existing tools like Temporal video segmentation, the task of identifying those sub-sequences of shots that constitute scenes is challenging, as it requires semantic understanding in order to discover the association between those shots that are semantically consistent but visually dissimilar.

There has been extensive research on video understanding. Despite the progress in this area, the currently available works focus on recognizing the categories of certain activities from short videos like violence etc. More importantly, these works assume a list of predefined categories that are visually distinguishable. However, for movie scene segmentation, it is impossible to have such a list of categories. Additionally, shots are grouped into scenes according to their semantic coherence rather than just visual cues. Hence, a new method needs to be developed for this purpose.

To associate visually dissimilar shots, we need semantic understanding. The key question here is “How can we learn semantics without category labels?”. Our idea to tackle this issue consists of two aspects 1) Instead of attempting to categorize the content, we focus on scene boundaries. We can learn what constitute a boundary between scenes in a unsupervised way, and thus get the capability to differentiate between within-scene and cross-scene transitions.

2) We can leverage the cues contained in multiple semantic elements, including place, cast, action and audio, to identify the associations across shots. By integrating these aspects, we can move beyond visual observations and establish the semantic connections more effectively. 3) We can also explore the top-down guidance from the overall understanding of the movie, which may bring further performance gains.

Based on these ideas, we are proposing to develop a framework that performs scene segmentation through three stages: 1) Extracting shot representation from multiple aspects 2) Making local predictions based on the integrated information, and finally 3) Optimizing the grouping of shots

Scene boundary detection and segmentation. The earliest works exploit a variety of unsupervised methods.clusters shots according to shot color similarity. In the author plots a shot response curve from low-level visual features and sets a threshold to cut scenes.further group shots using spectral clustering with a fast global k-means algorithm. predict scene boundaries with dynamic programming by optimizing a predefined optimizing objective.

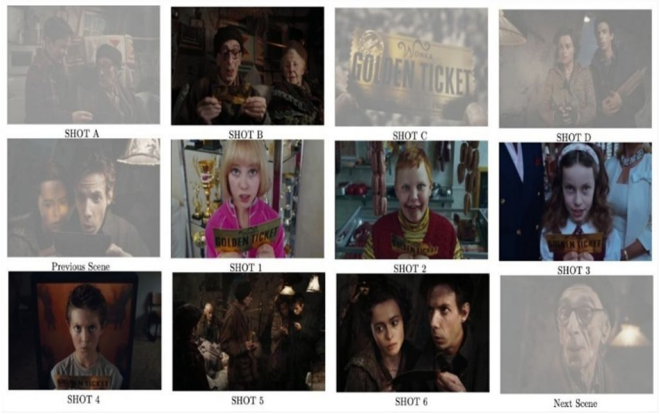


Figure 1: When we look at any single shot from above, e.g. By just looking at SHOT B, we cannot infer what the current event is. Only when we consider all the shots from 1 to 6, we can recognize that "All the characters are being introduced"

II. LITERATURE SURVEY

A. Self-supervised Visual Feature Learning

The proposed novel approach for feature extraction and composition using denoising autoencoders [1] involves training neural networks to reconstruct input data that has been corrupted by adding noise, leading to more robust features that can better handle input perturbations. The experimental results presented in the paper demonstrate that the denoising autoencoder approach outperforms other unsupervised feature learning methods, making it a promising technique for various applications. Besides, a new method for unsupervised feature learning using image inpainting [2] was considered, which involves using a neural network to predict the missing parts of an image conditioned on the rest of the image, leading to features that capture spatial relationships between different parts of the image. The experimental results presented in the paper show that this context encoder approach leads to better performance than other unsupervised feature learning methods, further demonstrating its potential for various applications.

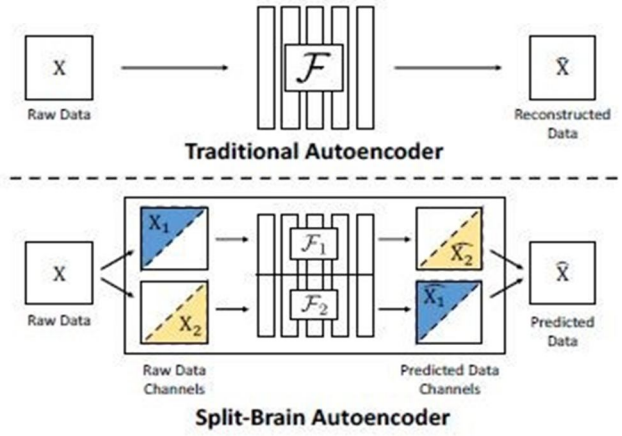


Figure 2: Traditional vs Split-Brain Autoencoder architectures.

Split-brainautoencoders[3] (top) Autoencoders learn feature representationF by learning to reconstruct input data X. (bottom)The proposed split-brain autoencoder is composed of twodisjoint sub-networks F1,F2, each trained to predict one data subset from another, changing the problem from reconstruction to prediction. The split-brain representation F is formed by concatenating the two sub-networks which involve using neural networks to predict one channel of the input data from another channel, were also discussed. The experimental results presented in the paper demonstrate that this approach is effective in capturing useful features that can be used for various tasks such as image classification and object detection. Furthermore, a new approach to unsupervised feature learning using image rotations [4] was proposed, which involves using a neural network architecture to predict the angle by which an image has been rotated, leading to features that capture spatial relationships between different parts of the image. The experimental results presented show that the proposed approach outperforms other unsupervised feature learning methods on benchmark tasks such as image classification and object detection, making it a promising technique for various applications. Similarly, a method for unsupervised feature learning using mutual information estimation and maximization [5] was discussed, which involves using a neural network architecture to map inputs to a lower-dimensional representation, and then estimating mutual information between the input and the representation using a predictor. The network is trained to maximize the mutual information estimate, leading to representations that capture useful information about the input. The experimental results presented in the paper demonstrate that this approach can be effective for various applications, making it a promising area of research. Moreover, a method for semi-supervised learning using large-scale self-supervised pre-training [6] was proposed, which involves augmenting labeled data using data augmentation techniques and mixup regularization, and then using the resulting augmented samples to train a large self-supervised model on an unlabeled dataset. The pre-trained model is then fine-tuned on the labeled data using supervised and unsupervised loss functions, leading to improved performance on various tasks. This method has the potential to be used for a wide range of applications, making it a promising approach for semi-supervised learning. Finally, Contrastive Predictive Coding (CPC) [7], a self-supervised learning method that predicts future samples from a sequence of past samples, was considered. The model is trained to distinguish between positive and negative samples using a contrastive loss function, leading to representations that capture useful information about the input. The experimental results presented in the paper demonstrate that CPC can be effective for various applications, making it a promising area of research. Additionally, an approach proposes Non-Parametric Instance Discrimination (NPID) [8] , an unsupervised feature learning method that learns discriminative features by differentiating between similar and dissimilar instances of input data. The method involves formulating the problem as a clustering problem and learning representations that capture intra-class variation and inter-class diversity. The model is trained with a contrastive loss function and showed impressive performance in various benchmarks on unsupervised and semi-supervised settings, making it a promising approach for various applications.

B. Unsupervised Semantic Segmentation

The field of unsupervised learning for image classification and segmentation is rapidly evolving, with several new methods achieving state-of-the-art results on benchmark datasets. [9]Invariant Information Clustering (IIC) and Contrastive Clustering are two such methods. IIC learns image representations by maximizing mutual information between pairs of transformed versions of the same input image, while [10]Contrastive Clustering maximizes agreement between positive pairs and minimizes agreement between negative pairs in the representation space. Both methods outperform other state-of-the-art unsupervised methods on various benchmarks. The DeepCluster v2 extension further improves clustering performance with a dynamic sampling strategy.

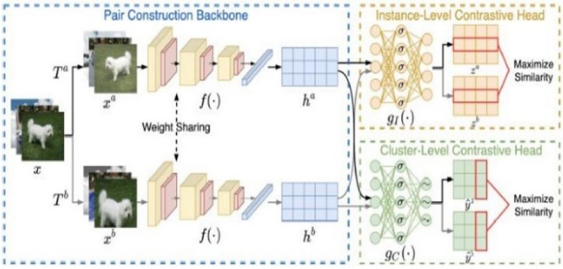


Figure 3: The framework of Contrastive Clustering. We construct data pairs using two data augmentations. Given data pairs,one shared deep neural network is used to extract features from different augmentations.

Two separate MLPs (_ denotes the ReLU activation and _ denotes the Softmax operation to produce soft labels) are used to project the features into the row and column space wherein the instance and cluster evel constrstive learning are conducted respectively
Self-supervised learning is another area of research where recent advances have led to promising results.[11] SCAN is a self-supervised learning method that learns representations by clustering semantically similar images and combining their features to form cluster-level representations. It outperforms other self-supervised methods when combined with fine-tuning on small labeled datasets. [12]Another novel unsupervised semantic segmentation approach based on clustering techniques is PICIE. It utilizes both invariance and equivariance properties and uses an encoder-decoder architecture and K-means clustering to generate feature embeddings. An ablation study shows the effectiveness of various components of PICIE, including using multi-level features and incorporating attention mechanisms. PICIE shows promise in the field of unsupervised semantic segmentation.
[13]In addition to these unsupervised methods, there are also supervised approaches that can improve segmentation performance. SegSort is a method for semantic image segmentation that employs discriminative sorting of image segments. It takes a bottom-up approach, first segmenting the image into smaller regions and then sorting them into different classes based on visual similarity. SegSort is based on a fully-convolutional network architecture and is trained in a supervised manner. A new dataset, COCO-SS, is introduced. Segment sorting as a pre-processing step can improve the performance of existing segmentation models.
[14]Finally, Deep feature factorization is a method for discovering high-level concepts in deep neural network representations. It aims to factorize a set of features into a small number of common factors that correspond to underlying concepts. The method is unsupervised and does not require prior knowledge about the concepts to be discovered. It can discover meaningful concepts such as texture, color, object shape, and object category and can be used for unsupervised image classification and semantic segmentation. The method can be used for a wide range of applications in computer vision where unsupervised concept discovery is needed. These recent advances in unsupervised and supervised learning methods offer exciting opportunities for improving image classification andsegmentation performance.

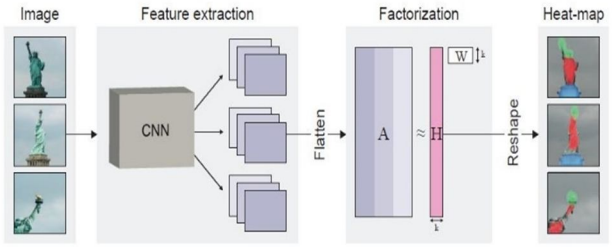


Fig. 4: An illustration of Deep Feature Factorization. We extract features from a deep CNN and view them as a matrix. We apply NMF to the feature matrix and reshape the resulting k factors into k heat maps. See section 3 for a detailed explanation Shown

C. Visual Transformers

In the field of computer vision, various approaches have been proposed to improve the performance of visual tasks. One such approach is the use of self-supervised learning in vision transformers, which has shown promise in learning complex visual tasks with state-of-the-art performance.
Recent works have proposed novel architectures that adapt transformer networks for dense prediction tasks such as object detection and semantic segmentation. [15]The ViP architecture, for example, combines a CNN with a transformer network to capture spatial information and achieve state-of-the-art performance with fewer computations compared to existing CNN-based models. Similarly, Non-local Neural Networks (NLNet) capture long-range dependencies in images by computing self-attention between all pairs of positions in an input feature map. This approach has shown effectiveness in various tasks, including medical image analysis and video understanding. These architectures demonstrate how transformers can be adapted for different visual tasks.
[16]Self-attention mechanisms have been instrumental in improving the performance of various deep learning models. The Self-Attention Generative Adversarial Network (SAGAN), for instance, uses self-attention mechanisms to learn long-range dependencies in images and generates high-quality and diverse images compared to other state-of-the-art GAN models. [17]Transformers, on the other hand, were initially designed for modeling relationships between sequential inputs and have been used in various NLP applications. The Transformer's success led to its extension and adaptation to other domains, including computer vision. The success of these models demonstrates the versatility of self-attention mechanisms in various deep learning domains.

Recent works have also explored ways to train data-efficient transformer-based image recognition models. [18]One such work is the "Training Data-Efficient Image Transformers & Distillation through Attention," which proposes a method called "Attention Condensation" to compress attention maps, resulting in fewer parameters and faster training. They also introduce a novel distillation approach that leverages attention maps from a large teacher model to improve the performance of a smaller student model. [19]Another work, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," proposes a new pretraining approach called "training on large-scale unlabeled data" for the ViT architecture. The approach divides an input image into patches and linearly projects each patch into a lower dimensional space to obtain a sequence of vectors that are fed into the transformer encoder. The ViT architecture and the pretraining approach provide insights into how transformers can be used to improve the performance of image recognition tasks with fewer parameters and better generalization.

III. DATASET

A. *MNSIT*

The MNIST (Modified National Institute of Standards and Technology) dataset is a widely used benchmark dataset in the field of machine learning and computer vision. It is a collection of 70,000 handwritten digits (0-9) images, with 60,000 images for training and 10,000 for testing. The images are grayscale and have a resolution of 28x28 pixels. The MNIST dataset is often used for developing and testing image recognition algorithms, particularly for classification tasks. Due to its small size and simplicity, it has become a standard dataset for testing various machine learning models, and its performance is often used as a baseline for evaluating new models.

B. *ImageNet*

ImageNet is a large-scale visual database that contains more than 14 million images, annotated with over 21,000 categories. The dataset is used primarily for training and evaluating computer vision models, particularly deep convolutional neural networks (CNNs). ImageNet is widely recognized as a benchmark dataset for image classification, object detection, and other computer vision tasks, and has played a significant role in advancing the field of computer vision over the past decade. The dataset is challenging due to the large number of categories, the variability of images within each category, and the presence of many fine-grained categories.

The images in ImageNet come from a variety of sources, including Flickr and other image-sharing websites, as well as professional photographers and stock photo agencies. The dataset is organized into a hierarchy of categories, with each category represented by a set of images. The most popular subset of ImageNet is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which includes 1.2 million training images, 50,000 validation images, and 100,000 test images.

C. *NORB*

The NYU Object Recognition Benchmark (NORB) dataset is a small and standard vision dataset used for evaluating algorithms for object recognition, stereo, and human vision. It consists of a set of images of 50 toys from 5 categories captured under different lighting, azimuth and elevation angles.

The images are grayscale and have a resolution of 96x96 pixels. The dataset contains two parts: the small NORB (noisy ORB) dataset with about 10k images for training and testing, and the full NORB dataset with about 290k images. The dataset is often used in the development and evaluation of 3D object recognition algorithms, shape modeling, and shape retrieval.

D. *COCO*

The Common Objects in Context (COCO) dataset is a large-scale object detection, segmentation, and captioning dataset. It contains over 330,000 images with more than 2.5 million object instances labeled across 80 categories, making it one of the largest and most comprehensive datasets for computer vision.

The images are collected from various sources, including Microsoft Bing image search, Flickr, and others, and are taken in a wide range of environments, including indoor and outdoor scenes. The dataset is widely used for benchmarking object detection and segmentation models, as well as for training models for image captioning and visual question answering tasks. COCO also holds annual challenges where researchers can submit their models to compete with others on various tasks, and the dataset is continuously updated with new annotations and images.

E. INRIA

The INRIA Holidays dataset is a benchmark dataset for image retrieval tasks. It contains a total of 1,491 high-resolution images of various sizes with a resolution of 1024 x 768 pixels. The images are classified into 157 groups, with each group containing a minimum of four and a maximum of 50 images. These groups represent different types of scenes and objects, such as buildings, animals, landscapes, and people.The images in the dataset are manually annotated with ground truth correspondences, which makes it an ideal dataset for evaluating image retrieval algorithms. The annotations are provided in the form of text files that contain the file names and corresponding ground truth matches for each image.The dataset has been widely used for evaluating image retrieval algorithms, especially for testing the performance of feature extraction and matching methods. It has also been used for evaluating deep learning-based approaches for image retrieval tasks.

IV. EVALUATION MATRIX

Automatic evaluation of video scene segmentation methods is a challenging task due to the complexity of video data and the subjective nature of evaluating segmentation results. While the ideal evaluation method is through human judges, it is often slow and expensive. To address this issue, several automatic evaluation metrics have been proposed, including temporal Intersection over Union (tIoU) and F1 score. However, these metrics have limitations in capturing the visual quality of the segmented scenes, and they may not always align with human perception.

On the other hand, evaluating video captioning systems poses similar challenges, but there has been more progress in developing automated evaluation metrics. Metrics such as BLEU, ROUGE, METEOR, and CIDEr have been widely used to evaluate the quality of generated captions. However, these metrics have been criticized for their tendency to favor generic captions over more specific and detailed ones. Despite these challenges, evaluating video scene segmentation and captioning systems is critical to advancing research in these fields and improving their practical applications.

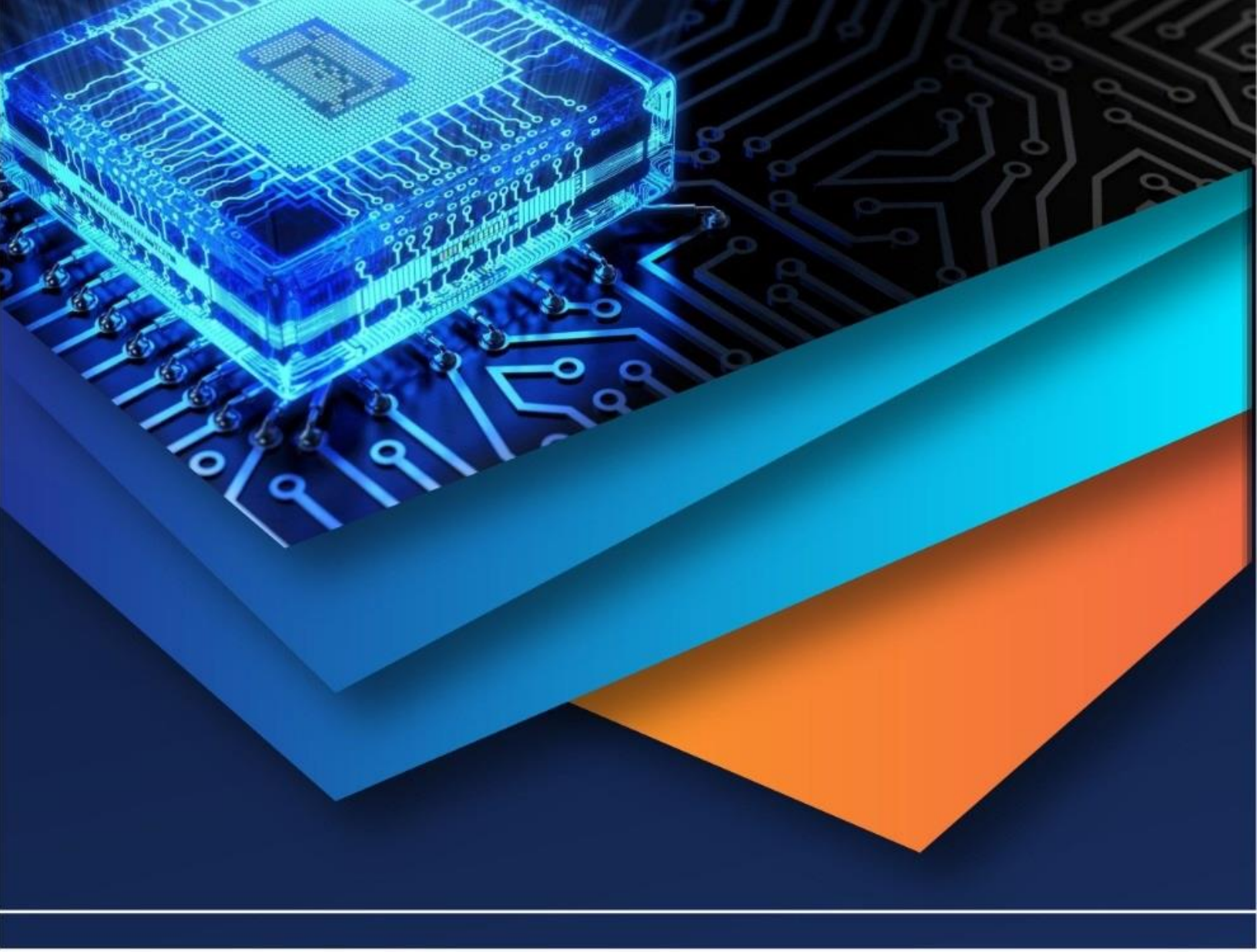
V. CONCLUSION

In conclusion, video scene segmentation is a challenging task that requires accurate detection and classification of objects and their interactions in dynamic environments. There are various automatic evaluation metrics available to measure the performance of video scene segmentation algorithms, but they have their limitations and cannot replace human judgment entirely. Despite these challenges, recent advances in deep learning and computer vision have shown promising results in video scene segmentation. Techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been effectively utilized to capture spatial and temporal features of videos and classify objects and actions. However, further research is required to improve the accuracy and efficiency of video scene segmentation systems, especially in complex and dynamic environments. Moreover, video scene segmentation systems have various potential applications in fields like surveillance, robotics, and autonomous vehicles. With the continued development of deep learning techniques and the availability of large-scale annotated datasets, it is expected that video scene segmentation systems will become more powerful and widely used in various domains.

REFERENCES

[1] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pp. 1096–1103, 2008.
[2] Deepak Pathak, Philipp Krahenb ¨ uhl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context en- ¨ coders: Feature learning by inpainting. In CVPR, 2016.
[3] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In CVPR, 2017.
[4] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728, 2018.
[5] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2018.
[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big selfsupervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029, 2020a
[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.
[8] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via nonparametric instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3733–3742, 2018.
[9] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised ˜ image classification and segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9865–9874, 2019
[10] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. arXiv preprint arXiv:2009.09687, 2020.
[11] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In European Conference on Computer Vision, pp. 268–285. Springer, 2020

[12] Jang Hyun Cho, U. Mall, K. Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. ArXiv, abs/2103.17070, 2021.
[13] Jyh-Jing Hwang, Stella X. Yu, Jianbo Shi, Maxwell D. Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. CoRR, abs/1910.06962, 2019. URL http://arxiv.org/abs/1910.06962.
[14] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 336–352, 2018.
[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803, 2018.
[16] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In International conference on machine learning, pp. 7354–7363. PMLR, 2019.
[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008, 2017.
[18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve J ´ egou. Training data-efficient image transformers & distillation through attention. In ´ International Conference on Machine Learning, pp. 10347–10357. PMLR, 2021.
[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY