International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Video Segmentation & Retrieval

Brahanyaa Somasundaram, .S.Shridevi

Abstract— There is a tremendous growth in the fields of multimedia and web databases, and research has been stepping forward towards many computer vision applications. In many computer vision applications local features are needed. To address this specific issue, many large point descriptors and detectors have been invented throughout the years. Creation of effective descriptors is still a milestone. To combat the high computational cost and the hunger for training data, auto encoders are proposed for efficient image analysis and image retrieval. Based on the auto encoder concept, a novel descriptor has been introduced. The proposed descriptor reduces the size and complexity and hence reduces the time required by a database to produce and display the retrieval results.

Keywords—auto encoders, descriptors, detectors, computational costs, combat (key words)

I. INTRODUCTION

Due to advancements in the field of multimedia, storage databases, there is a huge amount of data collected in specific applications such as video recommendation, broadcasting and advertising websites. In recent years, research has evolved drastically in the field of CBIR(Content-Based Image Retrieval), Video retrieval , video to image retrieval and video to video retrieval. In this paper , we have proposed a method for retrieving top k most videos from the database , by using a query image. Numerous amount of effort has been put in the research community on image to video retrieval. Researcher de Araujo and Girod has achieved high accuracy by replacing an image by video clip for video-video retrieval. Thus, by doing so the search space reduces from a large set of database to small set of video clips. The proposed work in the paper aims to improve performance of a video retrieval by a query image in a large scale database. Visual similarity between query image and frame of a video is measured using various similarity functions and distance measures. Based on the existing proposed ideas , we have proposed a novel method , that not only improves the computational cost and time, but also helps in achieving greater accuracy. In this paper, we have utilized Convolutional Neural Network architecture for constructing the visual features of the dataset and for training the model. A novel index search and retrieval algorithm is proposed , which can be used to boost the search performance. Variational approaches have dominated many improvements , but the proposed methodology in the paper strives to achieve better efficiency and performance and can also be used in further Computer Vision Research areas.

Revised Manuscript Received on 22 May 2019.
***** Correspondence Author
Brahanyaa Somasundaram*, VIT University(Chennai Campus) (Tamil Nadu) India
Dr.S.Shridevi, VIT University(Chennai Campus) Tamil Nadu India

II. CONTRIBUTIONS

The main contributions of our work can be summarized asfollows:

- We introduce a formal definition of image-query problem and address the bottlenecks involved in video retrieval process
- A new architecture is proposed for video retrieval, which is quite different from the existing content based video retrieval methodologies.
- Accuracy has been improved and the loss of training data has been reduced by carefully training the data using CNN and Optical flow.

In this paper, we have conducted extensive research on the performance evaluation of the UCF101 dataset. Results demonstrate that the approach performs relatively well with respect to state-of-the art method.

III. RELATED WORK

Huge multimedia data storage capabilities and with the proliferation of the use of world wide web accessing, it becomes reasonable to imagine a world in which it would be possible to access any of the stored video information. Since much of this data will be in the form of video queries, it becomes very important to develop the technologies, necessary for segmentation, classification, indexing and retrieval. Researches on classification and retrieval of video are relatively new[3] and [2]. A key to the success of any multimedia content analysis algorithm is the type of features employed for the analysis. Generally, visual analysis will be conducted on low level visual properties extracted from individual video frames[5]. The system is designed so as to not only support the entire process of Video information management , but also to help the process of segmenting, indexing, retrieving and sequencing of video data. Motion analysis of data is a powerful data for studying dynamic behavior and determining sources of failures. In the case of failure analysis, the available video may be of poor quality, such as from surveillance cameras. To address such specific issues, we present an optical flow based video frame extraction algorithm incorporating physically based constraints to extract motion data from video. Optical flow concept came into existence , by the need to describe the visual stimulus given to the animals which are moving to the world. The American psychologist James J. Gibson was the first person who introduced the topic “OPTICAL FLOW” .He stressed on the importance of optic flow for affordance perception, the ability to discern possibilities like Lucas kanade, dense optical flow.

Autoencoders are effective unsupervised learning models that first encode an input into a lower dimensional representation. This representation, which constitutes the features in the input , can be particularly useful in the domain of image processing. Contents of a video can be encoded to a compressed representation, which can be used to build effective video retrieval applications.[7] This paper compares and evaluates several architectures, models and training approaches necessary to build such a retrieval system. Deep Neural Networks can learn many levels of non-linearity in the images to extract and represent features in an image. Autoencoders, a particular type of deep neural networks, learn to reconstruct images by first transforming an input into a hidden representation of its features. These representations can be used to compute semantic similarity in the images. Deep autoencoders are networks able to compress and decompress the input images, (i.e) the output is an approximation of the input reconstructed by the network after reducing the dimensionality in internal layers. Hinton and Salkhutdinov and Bengio, show this type of network benefits from a layer wise pre-training. The use of pretraining is shown to produce internal high-level abstractions of the input and results in effective training for arbitrary deep networks. Autoencoder models resemble in many ways single layer latent variable models. Autoencoders not only play an major role in data mining, but also can be used for large scale vieo-image retrieval[8]. In this paper, we have proposed a new architecture using Convolutional autoencoders and ANN algorithm for video-image retrieval.

Video retrieval continues to be one of the most exciting and fastest growing research areas in the field of multimedia technology. We have taken into consideration the existing constraints and limitations and have developed an application system that not only meets the users requirements but also is one of the novel methods in the field of computer vision[9]. This application system can be deployed in many fields such as Medical Image processing, Traffic monitoring system etc. To improve the current method of video retrieval, convolutional neural network (CNN) is used along with data augmentation. Data augmentation is used to improve the properties of images which leads to better feature extraction from foreground and background. The current approach of segmentation is proved to be computationally expensive. So the idea is improve the quality of video retrieval process along with the accuracy and make it computationally feasible. Hence the algorithm will help us to overcome the issues of graphics processing unit (GPU) issues. One of the main works of this paper include extracting frames from all the categories of the dataset using various efficient optical flow algorithms such as Lucas kanade , horn shock etc as described below. Optical flow cannot be computed locally, since only one independent measurement is available from the image sequence at a point, while the flow velocity has two components. A second constraint is needed. Therefore , a method for finding the optical flow pattern is presented which assumes that the apparent velocity of the brightness pattern varies smoothly almost everywhere in the image. As far as Lucas Kanade is concerned , although it is local method but still it is useful somewhat to find optical flow . It assumes that the flow is essentially constant in a local - neighbourhood of the pixel under consideration, By combining information from several nearby pixels, the Lucas–Kanade method can often resolve the inherent ambiguity of the optical flow equation. In this module we came to know, how Horn Schunck and lucaskanade algorithm will work. The outcome of this project is to improve the accuracy and suggest a hybrid algorithm by combining CNN with different machine learning algorithms and produce a novel “Video retrieval “ application useful for the users that helps in faster access of retrieved videos. The algorithms proposed are mainly for unsupervised learning which will helps in classification of images into their classes or clusters. CNN is studied with various combination of algorithms and determine the best out of it. Since convolutional nueral network itself comes with different types according to its dimensions such as convolutional 1D, convolutional 2D, convolutional 3D. 3D CNN is used with video segmentation for a certain datasets which contains 3D images which will determine length, breadth and height in the images. It is mostly used for medical purpose such as Magnetic resonance imaging MRI scans and for army camp. On the other hand convolution 1D and convolution 2D is used with long short term memory (LSTM) to predicting the next sequence of images from the given query of frames.

A.PROBLEM DEFINITION

In this section, we have introduced the definition of video retrieval by image query and related motions. One of the main goals of this paper involved developing an efficient CBVR (Content Based Video Retrieval System) . Initially, the dataset is pre-processed and frames from individual videos are extracted using Optical flow algorithm. CNN is used for 3D Video segmentation. The weights for Neural network are calculated as (n*m) , where ‘n’ is the number of inputs and ‘ m’ is the number of outputs. The main aim of this paper is to improve the existing algorithms by developing a fusion model from the existing works. The fusion model is developed with Deep Neural Network , that helps to increase more number of hidden layers which is used for extraction and classification of videos into their respective classes. Deep neural netwok improves after learning from the previous false negative classifications. An autoencoder is a Deep Neural Network is designed and implemented that learns to reconstruct its input. It typically has an encoder part which learns to represent the features in an input as a vector in latent space. This is followed by a decoder part which learns to reconstruct the input image from its hidden representation. Specifically, we have designed a neural network architecture such that we impose a bottleneck in the network which forces a compressed knowledge representation of original input. If the input features were each independent of one another , this compression and subsequent reconstruction would be a difficult task.

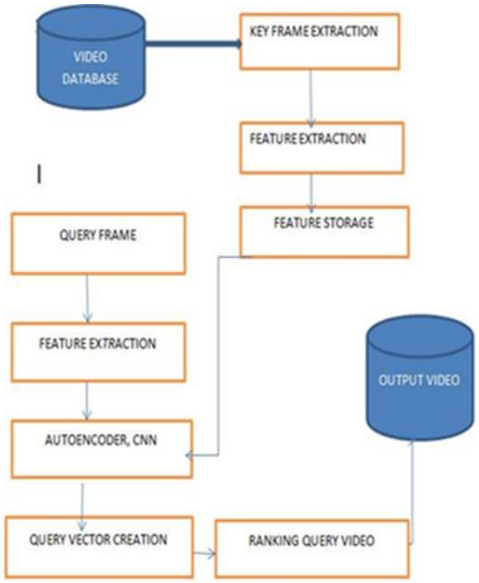
A Content Based Video Retrieval System in this paper is designed as follows: Multiple frames are extracted for the query video

A video is retrieved based on visual similarity match. Using , the feature vectors videos are represented from any one or a combination of more than one from different features extracted from numerous frames. Most similar videos with high precision value's are obtained based on mean square distance (or) cosine distance, euler's distance etc. between the feature vectors stored in the database and feature vectors of query image. As mentioned, classification and retrieval of multiple frames yields outstanding results without complexity in finding key frames to represent a single shot.

IV. PROPOSED MODEL

The proposed model describes how the video retrieval is different from existing traditional methods and the steps involved in the entire retrieval process. Individual key frames from multiple videos are extracted using optical flow and the dense layers are trained using CNN. Autoencoder model is built on top the CNN layers and search results are reformed based upon the query video. This model lowers the computational cost as well time.

Fig 1: Architecture of the CBVR System



The dataset has objects with varying degree of scale,translation, rotation and occlusion. This demands a sophisticated model with the ability to extract features from images effectively . The proposed architecture has 4 convolution layer blocks. Each block has 2 convolution layers with 3x3 filters, followed by a Max-Pooling layer with 2x2 filters. Dropout layers are placed after each convolution block to regularize the model. All the convolution layers have a stride of one and a padding to preserve initial size. At the end of 4th convolution block, we flatten out 2048 features extracted from the input image and feed it to 3 fully connected layers which progressively reduce the size of hidden representation to 128 (or) 64. This encoded vector can be used to index and retrieve images upon query.

Approach 1: This involves stacking up the layers of both encoder and decoder, and training all layers together . This is demonstrated in the Fig below.

Approach 2: Pre-train the Convolution layers for a different supervised task such as classification. Use these pre-trained Convolutional layers with fully connected layers from fig 3, and fine tune this fully connected autoencoder to reconstruct extracted features instead of input image.

The purpose of an autoencoder is to learn another representation of input data . An autoencoder is unsupervised learning method that sets the target values to be equal to the input values. Generally, an autoencoder contains one input layer, one or more hidden layers, and one output layer, which has exactly the same number of units as the input layer.

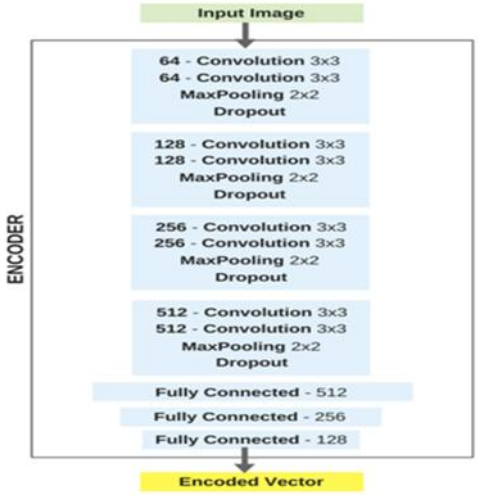


Fig 2: Encoder Network

Advantages of pre-trained Convolutional layers include:

a) Convolutional layers have already learnt how to extract features. They can be fine tuned only if necessary.

b) Fully connected layers learn to compress the extracted features instead of input image.

c) Allows higher order of compression.

V. PRE-TRAINING

The convolution layers were pre-trained with a classification task on the CIFAR4 dataset. This classification model consisted of four convolution blocks from (Fig 3) and three fully connected layers with a softmax loss function. The model was trained to reach state-of-the art accuracy of 76% on the CIFAR dataset. The following plots and tables show the progress of learning steps.

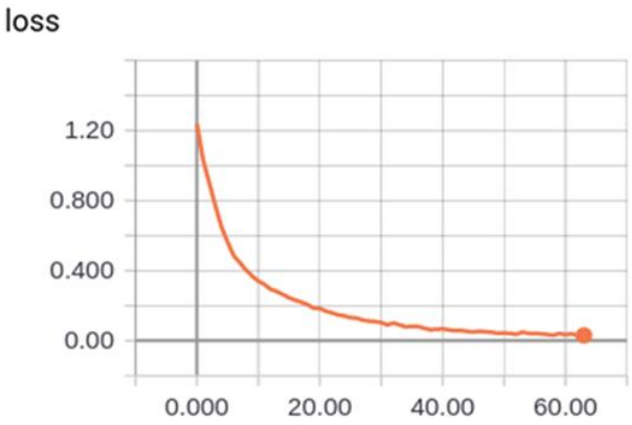


Fig 3: Pre-training Loss

The fully connected layers had 4096.4 hidden units. The weights were trained using Adam optimizer with a learning rate of le-4 and decay rate of le-6. Available GPU hardware permitted a batch size of 128 images per pipeline, with a training time of 28 mins, 80 secs per epoch. The model was regularized with Dropout and Batch Normalization layers as described previously. Dropout for first two Convolution blocks was set to 0.25 while the last two blocks had a dropout of 0.5. It should be noted that the model was trained to its full capacity and compares the state-of-the art accuracies of models similar in architecture

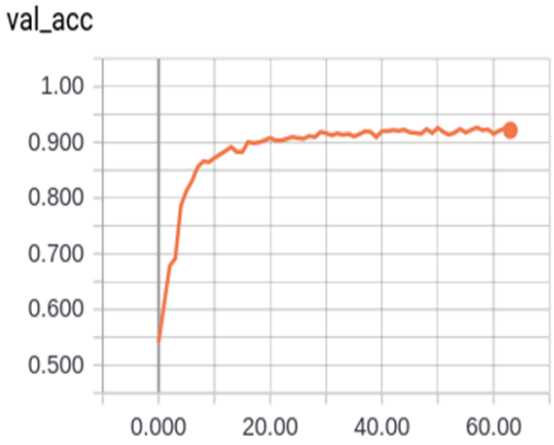


Fig 4: Validation vs Accuracy

V. INDEX CONSTRUCTION

We have proposed efficient models for each tasks .Despite that, if we were to store the compressed image vectors in a flat file and perform linear search, that would prove to be a bottleneck. A linear search involves:

- Loading all the images onto memory.
- Computing vector distance against query vector.
- Ranking (sorting) all vector distances to get the closest results.

Clearly this would not work in practical application with real-time user needs. An effective alternative approach would be ANN(Approximate Nearest Neighbor). Approximate Nearest Neighbor indexes each image and divides the search vector space into sub-trees. Each sub-tree may be, optionally, divided further to reduce search space. At the base of the tree, ANN performs an approximation of Nearest Neighbor search to retrieve search results. ANN is thus a minor trade-off of precise results in return for improved speed and memory utilization. This trade-off is perfectly acceptable in case of image data.

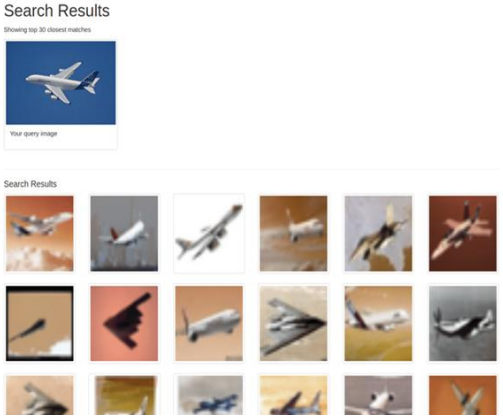


Fig 5: Search Results

VI. EVALUATION & RESULTS(HEADING 5)

Three models of varying complexity were built to fit the cifar dataset and evaluated to understand how they perform. Model A (Simple Auto encoder) constitutes only a single hidden layer per encoder and decoder. This was trained with Ad delta optimizer for 50 epochs over 50k dataset images.
Model B (Deep Auto encoder) constitutes three hidden layers per encoder and decoder. This model was found to perform best when trained with Ad delta optimizer for 60 epochs over the same training set.
Model C (Convolutional Auto encoder) consists of three convolutional layer blocks with pooling layers (maxpool)n per block. A complementing network performs the decoder part of the model. This model was trained with Adam optimizer for 80 epochs over the same training dataset.

Model	MSE Error
Model A	0.1266
Model B	0.1127
Model C	0.1015

Table 1: MSE Errors for 3 different Models

VII. RESULTS

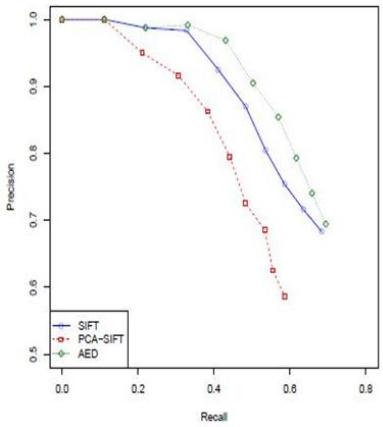
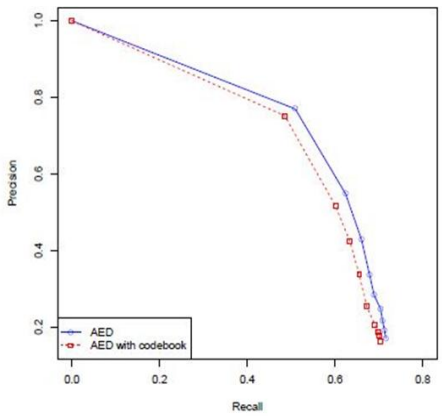


Fig 6: Precision vs Recall value for all methods involved in image analysis



.VIII. CONCLUSION

Three different autoencoder models were built and evaluated to quantify the effectiveness of Convolutional Autoencoders when compared to Simple and Deep autoencoders for image data. A classification model was built to pre-train Convolutional layers to extract features with an accuracy of 73% on the UCF101 dataset. The end-product of this project is not only user friendly, but also reduces time and manpower . This paper reduces manual computation for further research in the field of Multimedia and web databases. Accuracy for the overall model has been improved and the F1 score obtained is 0.78. As Deep learning approaches often require a huge amount of training data in order to solve a specific problem such as segmenting an object from a video, we have addressed such issues using GPU for training the DNN and performing computation using Parallel processors. We presented a new approach for learning to segment generic objects in video that achieves deeper synergy and appearance and also addresses practical challenges in training a deep network for Video retrieval and segmentation.

REFERENCES

1. A.Faktor and M.Irani . Video Segmentation by non-local consensus voting. In BMVC, 2014.5
2. A. Dososvitsiky, P.Fischer,C.Harzirbas and T.BroxFlownet: Learning optical flow with convolutional neural networks. In ICCV,December 2017.3
3. F.Galassso ,R.Cipolla and B.Schiele, Video segmentation using super pixels. In ACCV ,2012.
4. S.JiW.Xu, M.Yang and K.Yu.3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence,35(1): 221-231, 2013.3
5. S.Baker,D.Scharstein,J.Lewis, S.Roth ,M.J Black, and R.Szelski, A database and evaluation methodology for optical flow. International Journal of Computer Vision,92(1):1-31,2014.4
6. J.Dai, K,He,Y.Li, S.Ren and J.Sun. R-FCN:Object detection via region-based fully convolutional networks. In NIPS,2016-2
7. M.Grundman,V.Kwatra, M.Han and I.A Essa. Efficient hierarchial graph-based video segmentation.InCVPR ,2010.
8. F.Perazzi, J.Pont-Tuset, B. McWilliams ,L.Van God. A benchmark dataset and evaluation methodology for video object segmentation.In CVPR 2016.
9. A. Prest, Leistner, J.Civera, C.Schmid, and V.Ferari. Learning object class detectors from weekly annotated video. In CVPR,2012.
10. Y.J Lee, J.Kim and K.Grauman, Key segments for video object retrieval . In ICCV ,2011.