Overview of Video Concept Detection Using (CNN) Convolutional Neural Network

Ritika Dilip Sangale

Department Of Computer Engineering, Datta Meghe College of Engineering, Airoli, Navi Mumbai, University Of Mumbai.

Abstract - *Video Concept detection is the task of assigning an input video one or multiple labels. That indicating the presence of one or multiple semantic concepts in the video sequence. Such semantic concepts can be anything of users' interest that are visually observable. We Conducted survey of various techniques of Concept Detection. Concept Detection has various Steps. Shot Boundary Detection is first step in pre-processing Key-Frame Selection. Various features are extracted from these keyframes and stored as feature vector. These features are given to classifier which predicts presence or absence of the particular concept in given shot based on the previously trained model. The basic approach of concept detection is to use classification algorithms, typically SVM and/or CNN to build concept classifiers which predict the relevance between images or video shots and a given concept. We propose to use CNN for concept detection in video. We aim at developing a high-performance semantic Concept Detection using Deep Convolutional Neural Networks (CNNs) .we introduced Deep CNNs pre-trained on the ImageNET dataset to our system at TRECVID 2013, which uses GMM supervectors corresponding to seven types of audio and visual features.*

Key Words: Shot Boundary Detection, Keyframe Extraction, Feature Extraction, Deep CNN, GMM Supervectors, SVM, Score Fusion.

1. INTRODUCTION

With rapid advances in storage devices, networks, and compression techniques, large-scale video data become available to more and more average users. To facilitate browsing and searching these data, it has been a common theme to develop automatic analysis techniques for deriving metadata from videos which describe the video content at both syntactic and semantic levels. With the help of these metadata, tools and systems for video summarization, retrieval, search, delivery, and manipulation can be effectively created. The main challenge is to understand video content by bridging the semantic gap between the video signals on the one hand and the visual content interpretation by humans on the other. Since multimedia videos comprise of unstructured information that usually is not described by associated text keywords, semantic concept detection (also called semantic-level indexing or high-level feature extraction in some literatures) is needed to extract high-level information from video data.

Video Semantic concept detection is defined as the task of assigning an input video one or multiple labels indicating the presence of one or multiple semantic concepts in the video sequence. Such semantic concepts can be anything of users' interest that are visually observable, such as objects (e.g., "car" and "people"), activities (e.g., "running" and "laughing"), scenes (e.g., "sunset" and "beach"), events (e.g."birthday" and "graduation"), etc. Semantic concept detection systems enable automatic indexing and organization of massive multimedia information, which provide important supports for a broad range of applications are as Follows:

1. Computer Vision
2. Content or keyword-based multimedia search
3. Content-based video summarization
4. Robotic vision
5. Crime detection ,Natural disaster Retrieval
6. Interesting event Recognition from Games etc.

Video Semantic concept detection also shows that detecting the concepts that presence in the video shots. One of the technique that can be used for the powerful retrieval or filtering systems for multimedia is semantic concept detection. Semantic concept detection also provides a semantic filter to help analysis and retrieve a multimedia content. It also determines whether the element or video shot is relevant to a given semantic concept. For developing the models for annotated concepts we can use annotated training datasets. The goal of concept detection, or high-level feature extraction, is to build mapping functions from the low-level features to the high-level concepts with machine learning techniques.

Paring video into shots is the first processing step in analysis of video content for indexing, browsing and searching. A shot is defined as an unbroken sequence of frames taken from on camera. Shot is further divided into keyframe. Keyframes are representative frames from the entire shot. Various features are extracted from these keyframes and stored as feature vector. These features are given to classifier which predicts presence of absence of the particular concept in given shot based on the previously trained model. The basic approach of concept detection is to use classification algorithms, typically support vector machines (SVMs), to build concept classifiers which predict the relevance between images or video shots and a given concept. This paper is organized as follows: Section 2 presents Review of Literature .In Section 3, we will present our Proposed Methodology .At last, and Conclusion of this paper will be drawn in Section 4.

2. LITERATURE SURVEY

Mohini Deokar and Ruhi Kabra [2014] proposed a paper on "Video Shot Detection Techniques Brief Overview". In this paper the different techniques are discussed to detect a shot boundary depending upon the video contents and the change in that video content. As the key frames needs to be processed for annotation purpose, the important information must not be missed.

JingweiXu and LiSong, and RongXie [2016] proposed a paper on "Shot Boundary Detection Using Convolutional Neural Networks". In this paper a novel SBD framework based on representative features extracted from CNN. The proposed scheme is suitable for detection of both CT and GT boundaries.

DivyaSrivastava, RajeshWadhvani and ManasiGyanchandani [2015] proposed a paper on "A Review: Color Feature Extraction Methods for Content Based Image Retrieval". Here three basic low level features, namely color, texture and shape, describe and color is the most important various methods used for color feature extraction are illustrated.

Priyanka U. Patil, Dr. Krishna K. Warhade [2016] proposed a paper on "Analysis of Various Keyframe Extraction Methods". Here they explain key frame is a simple but effective form of summarizing a long video sequence. Keyframe will provide more compact and meaningful video summary. So with the help of keyframes it will become easy to browse video. This paper gives the brief review on the different keyframe extraction methods their features, merits and demerits.

Samira Pouyanfar and Shu-Ching Chen [2017] proposed a paper on "Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning". In the paper, a novel ensemble deep classifier is proposed which fuses the results from several weak learners and different deep feature sets. The proposed framework is designed to handle the imbalanced data problem in multimedia systems, which is very common and unavoidable in current real world applications.

Alex Krizhevsky, IlyaSutskever,Geoffrey E. Hinton [2012] Proposed paper on "ImageNet Classification with Deep Convolutional Neural Networks".In this paper author trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art.

Ashwin Bhandare, Maithili Bhide, Pranav Gokhale, Rohan Chandavarkar [2016] proposed a paper on "Applications of Convolutional Neural Networks". In this paper, author give a comprehensive summary of the applications of CNN in computer vision and natural language processing.Also,describe how CNN is used in the field of speech recognition and text classification for natural language processing.

Nakamasa Inoue and Koichi Shinoda and Zhang Xuefeng and Kazuya Ueki [2014] proposed a paper on "Semantic Indexing Using Deep CNN and GMM Supervectors". Paper proposed a high-performance semantic indexing system using a deep CNN and GMM supervectors with the six audio and visual features.

Kumar Rajpurohit, Rutu Shah, Sandeep Baranwal, Shashank Mutgi [2014], proposed a paper on "Analysis of Image and Video Using Color, Texture and Shape Features for Object Identification". This paper discussed and studied a few techniques and described the working of those techniques in relation to Content Based Video and Image Retrieval systems. Also presented different algorithms and their working in brief and the different low level features which can be used in order to extract and identify objects from image and video sequences.

3. PROPOSED SYSTEM

The proposed framework starts from collecting the data derived from videos. Each modality requires the corresponding pre-processing step. For instance, shot boundary detection and key frame detection are applied to obtain the basic video elements, e.g., shots and keyframes, respectively. Then, low-level visual features and audio features can be extracted from them.

3.1 Deep Convolutional Neural Networks

We use deep Convolutional Neural Network (CNN) trained on the ImageNET LSVRC 2012 dataset to extract features from video shots. A 4096-dimensional feature vector is extracted from the key-frame of each video shot by using the CNN [6].

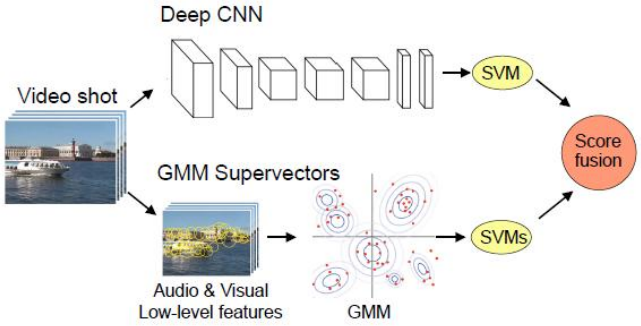

Fig -1: Proposed Architecture

The first to fifth layers are convolutional layers, in which the first, second, and fifth layers have max-pooling procedure. The sixth and seventh layers are fully connected. The parameters of the CNN is trained on the ImageNET LSVRC

2012 dataset with 1,000 object categories. Finally, from each keyframe, we extract a 4096-dimensional feature at the sixth layer to train an SVM for each concept in the Semantic Concept Detection [6].

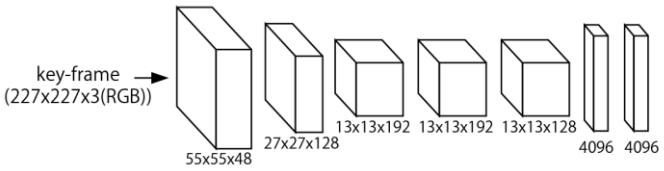

Fig -2: Deep Convolutional Neural Network

3.2 Low-Level Feature Extraction

1. SIFT features with Harris-Affine detector (SIFT-Har):

It is widely used for object categorization since it is invariant to image scaling and changing illumination. The Harris-Affine detector improves the robustness against affine transform of local regions. SIFT features are extracted from every other frame, and principal component analysis (PCA) is applied to reduce their dimensions from 128 to 32[8].

2. SIFT features with Hessian-Affine detector (SIFT-Hes): SIFT features are extracted with the Hessian-Affine detector, which is complementary to the Harris-Affine detector, it improve the robustness against noise. SIFT features are extracted from every other frame, and PCA is applied to reduce their dimensions from 128 to 32[8].

3. SIFT and hue histogram with dense sampling (SIFTH-Dense):

SIFT features and 36-dimensional hue histograms are combined to capture color information.SIFT+Hue features are extracted from key-frames by using dense sampling. PCA is applied to reduce dimensions from 164 to 32[8].

4. HOG with dense sampling (HOG-Dense):

32-dimensional histogram of oriented gradients (HOG) are extracted from up to 100 frames per shot by using dense sampling with 2x2 blocks. PCA is applied but dimensions of the HOG features are kept to 32[8].

5. LBP with dense sampling (LBP-Dense);

Local Binary Patterns (LBPs) are extracted from up to 100 frames per shot by using dense sampling with 2x2 blocks to capture texture information. PCA is applied to reduce dimensions from 228 to 32[8].

6. MFCC audio features (MFCC);

Mel-frequency cepstral coefficients (MFCCs), which describe the short-time spectral shape of audio frames, are extracted to capture audio information. MFCCs are widely used not only for speech recognition but also for generic audio classification. The dimension of the audio feature is 38, including 12-dimensional MFCCs [8].

7. SPECTROGRAM audio features:

Creating a spectrogram using the FFT is a digital process. Digitally sampled data, in the time domain, is broken up into chunks, which usually overlap, and Fourier transformed to calculate the magnitude of the frequency spectrum for each chunk. Each chunk then corresponds to a vertical line in the image; a measurement of magnitude versus frequency for a specific moment in time (the midpoint of the chunk). These spectrums or time plots are then "laid side by side" to form the image or a three-dimensional surface, or slightly overlapped in various ways, i.e. windowing.

3.3 GMM Supervector SVM

This Represent the distribution of each feature like each clip is modeled by a GMM (Gaussian Mixture Model), Derive a supervector from the GMM parameters and Train SVM (Support Vector Machine) of the supervectors[8]. GMM Supervector is the combination of the mean vectors. Finally Combine GS-SVM and Audio Feature (mfcc and spectrogram) and CNN-SVM and Calculate Score Fusion. Score Fusion Shows Concept Detected from the Video.

4. CONCLUSIONS

In this paper we discussed and studied a few audio & visual Local Feature Extraction Techniques and described the working of those techniques in relation to Concept Detection. We proposed a high-performance semantic Concept Detection using a deep CNN and GMM super vectors with the audio and visual features. We train Concept models from training dataset. We create distance matrix for the training and the test videos, and get SVM score of each video for each feature. The fusion weights of features were decided by 2-fold cross validation. The threshold is determined by the averaged threshold from the 2-fold cross validation. This is an Expected Proposed Work as per the basis of literature.

REFERENCES

[1] MohiniDeokar, RuhiKabra "Video Shot Detection Techniques Brief Overview". International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014

[2] JingweiXu , Li Song , RongXie "Shot Boundary Detection Using Convolutional Neural Networks" in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting,Ghent, 2016.

[3] DivyaSrivastava, Rajesh Wadhvani and ManasiGyanchandani "A Review: Color Feature Extraction Methods for Content Based Image Retrieval" IJCEM International Journal of Computational Engineering & Management, Vol. 18 Issue 3, May 2015.

[4] Priyanka U. Patil, Dr. Krishna K. Warhade "Analysis of Various Keyframe Extraction Methods" International Journal of Electrical and Electronics Research Vol. 4, Issue 2, April - June 2016.

[5] Samira Pouyanfar and Shu-Ching Chen "Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning" School of Computing and Information Sciences Florida International University, USA, March 27, 2017

[6] Alex Krizhevsky,IlyaSutskever,Geoffrey E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks"University of Toronto,2012

[7] AshwinBhandare, Maithili Bhide, PranavGokhale, RohanChandavarkar "Applications of Convolutional Neural Networks "AshwinBhandare et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (5) , 2016.

[8] Nakamasa Inoue and Koichi Shinoda and Zhang Xuefeng and Kazuya Ueki "Semantic Indexing Using Deep CNN and GMM Supervectors", Tokyo Institute of Technology, Waseda University,2014

[9] Kumar Rajpurohit, Rutu Shah, Sandeep Baranwal, Shashank Mutgi, "Analysis of Image and Video Using Color, Texture and Shape Features for Object Identification", Computer Department, Vishwakarma Institute of Information Technology, Pune University, India, Volume 16, Issue 6, Ver. VI (Nov – Dec. 2014)

[10] M. Hassaballah, Aly Amin Abdelmgeid and Hammam A. Alshazly, "Image Features Detection, Description and Matching", 2016.