ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Video Annotation and Retrieval System using SIFT and HOG Features

Sowmya P, Naveen Kumar B

M. tech Student, Dept. of Computer Science Engineering, UBDTC, Davanagere, India

Assistant Professor, Dept. of Computer Science Engineering, UBDTCE, Davanagere, India

ABSTRACT: As the amount of information uploaded in internet is increasing day by day, in order to get the required information the huge database of information are needed to be analyzed properly. One of the widely used methods to analyze and retrieve these huge video data is video annotation. Video annotation process requires a large amount of processing to analyze the contents in the video as the process is complicated. This paper introduces a video annotation and retrieval system using SIFT and HOG features, in that SIFT given for training the classifiers. The SIFT and HOG features of the images are extracted and used for training the classifiers for doing a comparison on performance of the classifiers. An analysis of the results is done to find which feature is better to train the classifier for getting more prominent annotated video database. Based on objects from the annotated video database, retrieval of the videos is also done.

KEYWORDS: SIFT and HOG features, SVM classifier.

I. INTRODUCTION

As the multimedia content over the internet is increasing day by day, efficient methods for retrieval of these huge amount of data is required. The World Wide Web is the most important source of information now a day. The knowledge and information which we obtain from internet is in the form of textual data, images, videos etc. There are many resources on the internet which people can use to create process and store videos. This has created the need for a means to manage and search these videos. Image retrieval and annotation method is a technique for searching and retrieving videos from a large database of several videos. A large collection of these videos is referred to as video database. An video database is a system where video data are integrated and stored. During the analysis of video database effective effort is required for the correct analysis and retrieval of the videos. Efficient browsing of the video database depends on the correctly retrieved or indexed videos from the database. Content based video search and retrieval is becoming more important as it helps for retrieval of videos based on the semantics in the video but it is a challenging problem as it needs to represent the semantics of the video. The content based video retrieval systems suffered from the problem of semantic gap-a gap that exists between low level features and high level features. In order to bridge the gap, the solution is to define a large number of semantic detectors which detect the semantic concepts. These detectors try to learn the mapping between the low level visual features with the high level concepts from the video. The detection of the semantic concepts is still very difficult, so their detection in videos is a challenging problem. Annotation of the video can be used for analysis and retrieval of the video database using HOG and SIFT features extraction. Annotation is an explanatory note or body of notes added to a text, diagram, document, image or video. Annotation files are generally extensible mark-up language (XML). These can be referred for the retrieval of desired image or video. Video annotation aims to assign several semantic & visual features to the contents of video. It is done to help semantic retrieval of videos from a large video database. Here an approach for video annotation and retrieval system based on HOG and SIFT features has been introduced to perform more complete semantic annotation.

The HOG feature extraction is very simple to understand. HOG will take a dense sampling strategy. To train the classifiers machine learning technique is adopted. i.e., first training the object detectors and then testing. Here five object classifiers are trained for HOG features. Now HoG features for some test images are calculated and are given to the corresponding trained classifiers. The object is then recognized. Support Vector Machine (SVM) is used as the classifier. SIFT (Scale Invariant Feature Transform) features are used in video annotation are of very high dimension. They are

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

invariant to changes in scale, 2D translation and rotation transformations. The large computational effort associated with matching all the SIFT features for recognition tasks. Considering that the videos have annotation, to access it only the annotation file have to be consulted in-order to determine whether the video is pertinent to the search, reducing overall access time.

II. LITERATURE SURVEY

Altadmri, A et.al [1] proposed a semantic based approach for video annotation with the aim to bridge the semantic gap. On receiving a new video input to be annotated, this framework uses a pre-annotated video dataset to identify similar videos. The matched annotations are then semantically analyzed and the best description for this new video is obtained using commonsense knowledge base. Commonsense is the term referred to identify information and facts that are expected to be known by ordinary people. Using these results, the new video is annotated.D. Chen, J [2] proposed a segmentation method based on Markov random field to extract more accurate text characters. This methodology allows handling background gray-scale multimodality and unknown text gray-scale values. Support vector machine (SVM) is used for text verification followed by traditional OCR algorithm.Shih-Wei Sun, Yu-Chiang Frank Wang in [3] proposed a robust moving foreground object detection method followed by the integration of features collected from heterogeneous domains. More focus is on annotating rigid moving objects & considers videos with only one foreground object present.J. W. Jeong Et.al [4] has proposed automatic video annotation technique based on ontology is proposed. MPEG-7 visual descriptors are used as features which are mapped to semi concepts and finally to objects. Ontologies are exploited to hierarchically describe the contents among heterogeneous users and devices. They have applied the technique in the smart TV environment. Video concept detection approaches using semantic inference rules and SVM classifier for efficient video search, sharing and browsing is proposed. Bogdan Vrusias, Dimitrios Makris, et.al., have presented a framework for semantic annotation of CCTV video which combines computer vision algorithm that extract visual semantics, along with Natural Language Processing that builds the domain ontology from unstructured text annotations.

III. METHODOLOGY

Figure 1 represents the proposed architecture. The algorithm consists of major two parts testing and training phase respectively. In training phase the system is trained to create a knowledge base. First stage of training phase is to read the videos stored in the database. Then videos are pre-processed. When the video is pre-processed feature extraction is achieved. For feature extraction HOG and SIFT features are used. Using SVM training algorithm a knowledge base will be created which is used in testing phase.

Testing consists of video is the system input to the top layer of the architecture, and then the video is converted into video frames. Then for these videos key frames features are extracted by using Gradient calculation. Next, video is pre-processing is used to change the video frame into gray scale conversion and resizing. For feature extraction HOG and SIFT features are used. SVM classifier will compare the result of feature extraction in testing phase with the database present in the knowledge base which we have stored at the training phase. Finally the similar type of videos are retrieved and displayed.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

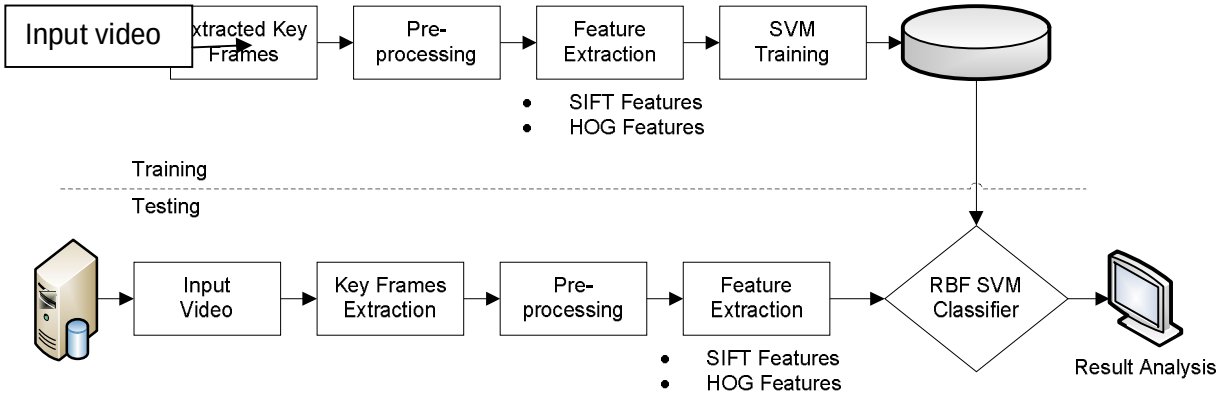

Figure 1:Block Diagram Of Proposed System.

A. PRE-PROCESSING

Pre-processing is mainly used to enhance the contrast and to adjust the size of the video and removal of noise. Pre-processing is any form of signal processing for which the output is an image or video, the output can be either an image or a set of characteristics or parameters related to image or videos to improve or change some quality of the input. This process will help to improve the video or image such that it increases the chance for success of other processes. In this paper we considered videos as input and those videos are subjected to pre-processing this will resulting in color conversion and resizing.

B. FEATURE EXTRACTION

First stage of training phase is to read the character images stored in the database. Then images are pre-processed. When the image is pre-processed feature extraction is achieved. For feature extraction HOG (Histograms of Oriented Gradients) and SIFT.

I. HOG FEATURES

HOG stands for Histograms of Oriented Gradients. HOG is a feature descriptor used to make the classification task easier under different conditions, main intension of feature descriptor is to generalize the object in such a way that the same object produces as close as possible to the same feature descriptor. The creators of this approach trained a Support Vector Machine (a type of machine learning algorithm for classification), to recognize HOG descriptors of videos. The HOG feature extraction is very simple to understand. HOG will take a dense sampling strategy. To train the classifiers machine learning technique is adopted. i.e., first training the object detectors and then testing. Here five object classifiers are trained for HOG features. Now HoG features for some test images are calculated and are given to the corresponding trained classifiers. The object is then recognized. Support Vector Machine (SVM) is used as the classifier. It is a supervised technique which classifies into two classes. i.e., object present or absent. The HOG will not use a collection of local features rather than it uses a global feature to describe features. Object/image is represented by a single feature vector, as opposed to many feature vectors representing smaller parts of the object. HOG features for some test images are calculated and are given to the corresponding trained classifiers. The video is then recognized. Support Vector Machine (SVM) is used as the classifier.

II. SCALE INVARIANT FEATURE TRANSFORM (SIFT) FEATURE

By using SIFT feature we can do effective retrieval of videos. In this method Video is divided into frames, and frames are divided into images. The object is separated from the segmented image. The segmented object is referred as a part of image. From the segmented image needful features are extracted. This method the features are extracted by using the Scale Invariant Feature Transform (SIFT) and are used to find the key points from the images because they are invariant to image. In this method first the video is pre-processed and converted into images. These images are then segmented using the segmentation algorithm to get the object image. Features are retrieved from the object image using the SIFT algorithm.

In extracting the SIFT features, key points and descriptors from the images are found and then histogram of visual words is constructed from a set of images. These features are given as input to the object classifiers for training. Creating of a vocabulary of visual words is done in pre-processing step. We have to consider a set of images as a input in order to

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

create the vocabulary of visual words The features from each of these images are extracted and saved. The visual words from the extracted features are constructed by performing k-means algorithm. In order to find the input to the classifier for training based on SIFT features, the vocabulary of visual words constructed in the pre-processing step is used. The input images to be trained along with the vocabulary are used to find the histogram for each image. These histograms are saved as mat file for later use. Also the labels are to be calculated based on the positive and negative images to be given as input to the classifier. These inputs are given to the linear SVM classifier and are trained.

SIFT (Scale Invariant Feature Transform) features are used in object recognition are of very high dimension. They are invariant to changes in scale, 2D translation and rotation transformations. The large computational effort associated with matching all the SIFT features for recognition tasks, limits its application to object recognition problems.

C. KEY FRAME EXTRACTION

Large amount of frames will be present in a video. Video annotation and analysis based on these large amounts of frames is a complicated task. As analyzing all the frames in the video is a time consuming task it is better to analyze only the frames which represent the best visual features of the scene. So key-frame detection can be used to reduce the frames. There are several key frame extraction algorithms that can be used. Key frame extraction algorithms select a subset of the most informative frames from videos. An algorithm which was proposed in that uses edge difference to find the similarity between two frames is used to find the key frames. Firstly, we construct a temporally maximum occurrence frame which considers the spatial and temporal distribution of the pixels throughout the video shot. Then, a weighted distance is computed between frames in the shot and the constructed reference frame. The key frames are extracted at the peaks of the distance curve and can achieve high compression ratio and high fidelity.

D. SVM CLASSIFIER

SVM stands for Function Support Vector Machines. SVM first maps the input vector into a higher dimensional feature space in order to achieve the optimal separating hyper-plane in the higher dimensional feature space,

The kernel function on two samples X and X', represented as feature vectors in some input space, is defined as

$$K(X, X') = \exp\left(\frac{\|X-X'\|^2}{2\sigma^2}\right) \quad (3)$$

$\|X - X'\|^2$ is the squared Euclidean distance between the two feature vectors. σ is a free parameter.

Furthermore, a decision boundary, i.e. the separating hyper-plane, is determined by support vectors rather than the whole training samples and thus is extremely robust to outliers. Exactly an SVM classifier is designed for binary classification. That is, to separate a set of training vectors which belong to two different classes. Note that a decision boundary similar to the support vector i.e. training samples. To provide the required mapping to ice-water labels a soft-margin SVM classifier is used. An SVM works by computing a linear decision boundary in a high dimensional space using the subset of labeled training samples near the decision boundary (called the support vectors). The SVM decision boundary equation is

$$f(x) = \sum_{\forall i} y_i \alpha_i K(x_i, x) \quad (4)$$

Where $K(x_i, x)$ Kernel function is defined by

$$K(x_i, x) = \exp(-\gamma |x_i - x|^2) \quad (5)$$

Then according to the result which we stored in the knowledge base at training phase SVM classifier will classify the videos by comparing it with SVM trained knowledge base database.

E. EXPERIMENTAL RESULT

Figure 2 represent the experimental result of the proposed system. Firstly will consider the video frames as input image, which is shown in Figure 2(a) after considering video frames from the input video those frames are pre-processed, in that gray scale converted images shown in Figure 2(b), from those gray scale images will apply feature extraction methods SIFT and HOG. Extract features of those gray scale converted images as shown in Figure 2(c) next SVM classifier will classify the taken input video as which category by video annotation and retrieval method as shown in Figure 2(d).

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

(a) **(b)** **(c)** **(d)**

Figure 2: (a) Video Frames; (b) Gray Scale Images from Input videos Image; (c) Feature Extracted Images; (d) Output.

F. CONCLUSION

As there is a more importance in video annotation and retrieval of video system has increased, because it helps for retrieval of videos based on the semantics of the video, but it is a challenging problem as it needs to consider the semantics of the video. Here a system proposed for video annotation based on HOG and SIFT features for training the SVM classifiers is proposed. Experimental results show that the performance of classifiers trained based on HOG features are best as compared to SIFT features. Since, only the key frames are analyzed for the object identification, much time is saved when compared to analyzing all the frames in the video. On the basis of HOG and SIFT feature extraction the annotation files are created, the retrieval of the videos from these annotated databases is much easier as compared to the huge database of videos.

REFERENCES

[1] Altadmri, A. and Ahmed, A, "Automatic Semantic Video Annotation in Wide Domain Videos Based on Similarity and Commonsense Knowledge bases" IEEE International Conference on Signal and Image Processing Applications, UK, Pp. 74 – 79, 2015.
[2] D. Chen, J. Odobez, H. Bourlard, "Text detection and recognition in images and video frames, Pattern Recognition 37" pp. 595 – 608, 2004.
[3] Shih-Wei Sun,Yu-Chiang Frank Wang, "Automatic annotation of web videos" IEEE 2011.
[4] J. W. Jeong, H. K. Hong, D. H. Lee, "Ontology-Based Automatic Video Annotation Technique in Smart TV Environment", IEEE Transactions on Consumer Electronics, Vol. 57, No. 4, pp. 1830-1836, 2011.
[5] N. Magesh, P. Thangaraj, "Semantic Image Retrieval Based on Ontology and SPARQL Query" In proceedings of International Journal of Computer Applications (IJCA) – ICACT, Number 1, pp.12-16, August 2011.
[6] K.Khurana, M.B.Chandak, "Key Frame Extraction Methodology for Video Annotation" International Journal of Computer Engineering and Technology, Volume 4, issue 2, pp.221-228, 2013.
[7] S Zhang, T, "A Generic Framework for Video Annotation via Semi-Supervised Learning" IEEE Transactions on Multimedia, Vol. 14, Issue 4, Pp. 1206 – 1219, 2012.
[8] R Datta, D Joshi, J Li, and J. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, Vol. 40, No. 2, 2008.
[9] Troncy, R, "Integrating Structure and Semantics into Audio-visual Documents" In Proceedings of the 2nd International Semantic Web Conference (ISWC), Sanibel Island, Florida, USA. Lecture Notes in Computer Science, Volume 2870, pp. 566 –58, 2003.