International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-3, Issue-6, June 2017 Pages 04-08

Video Annotation Model Based on Multi-Label Classifier and Fuzzy Knowledge Representation Schemes

M.Sumithra, V.Mercy Rajaselvi

Abstract— **Video annotation is a promising approach to facilitate video retrieval and it can avoid the intensive labor costs of pure manual annotation. But it frequently encounters several difficulties, such as insufficiency of training data and the curse of dimensionality. Video annotation is processed by three steps, In the first step, in order to extract the key feature, video is taken as input and a single frame is extracted from the video by using the video cutting tool. From the selected frame GIST Descriptors for spatial structure and the feature vector of 8x 8 encoding samples are extracted. In the second step, classifier are trained and annotation is done. In the training phase, trained classifier is obtained by using SVM algorithm and in the classification phase, labels are given to the object using the trained classifier. In the third step, scenes are recognized by inference based algorithms which takes object labels as input. The inference based algorithms are used for annotation refinement and scene recognition. These algorithms use fuzzy knowledge representation scheme based on Fuzzy PetriNet and KRFPNs. KRFPNs is defined to enable reasoning with concepts which is useful for video annotation**

Index Terms— **Video annotation, Key feature extraction, Object-level recognition, Scene-level recognition.**

I. INTRODUCTION

Video surveillance has long been in use to monitor security sensitive areas such as banks, department stores, highways, crowded public places and borders. The advance in computing power, availability of large-capacity storage devices and high speed network infrastructure paved the way for cheaper, multi sensor video surveillance systems. The increase in the number of cameras in ordinary surveillance systems overloaded both the human operators and the storage devices with high volumes of data and made it infeasible to ensure proper monitoring of sensitive areas for long times. In order to filter out redundant information generated by an array of cameras, and increase the response time to forensic events, assisting the human operators with identification of important events in video by the use of "smart" video surveillance systems has become a critical requirement. The making of video surveillance systems "smart" requires fast, reliable and robust algorithms for moving object detection , classification, tracking and activity analysis.

M.Sumithra ME Student, Easwari Engineering College, Chennai, Tamil Nadu
V.Mercy Rajaselvi Assistant professor , Easwari Engineering College, Chennai, Tamil Nadu

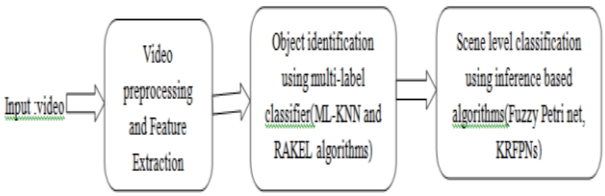
The procedure of moving object tracking is to decide whether there exist objects moving in video and to position the target basically and recognize it. A video sequence is made of basically a series of still images at a very small interval time between each capture. As video sequence consists of frame sequences which have certain temporal continuity, the detection for moving object in video is conducted in a way that frame sequences are extracted from the video sequence according to a definite cycle.

Therefore, moving object detecting has something similar to object detection in still images. Only moving object detecting is more relying on the motion characteristics of objects, i.e. the continuity of time, which is the difference between moving object and object detection in still images. The need of real-time object detection for video surveillance has spawned a huge amount of our daily life, especially in some domains where it has received considerable attention, for instance: criminology, sociology, statistic, traffic accident detection and military applications. Moving object detection is considered to be the most important task in automated video surveillance systems. It represents the low level image.

Key frame used for representing the main content of a video shot. Key Frame Extraction is the key technology for video retrieval, video query, video index, video browse and video abstraction. The algorithm for key frame extraction will affect the establishment and retrieval efficiency of video retrieval system. Key Frame Extraction-based video retrieval generally includes such steps as follows. Firstly, a video is divided into different shots, and key frames are extracted from these shots. Then the low-level visual features such as color, texture and shape are extracted from the key frames. These features are being used to build index and will be kept in database. After that, users can search videos from database by different search mechanisms.

II.RELATED WORKS

A visualization based batch mode sampling method to handle problem. An iso-contour based scatter plot is used to provide intuitive clues for representativeness and informativeness of samples and assist users in sample selection. A semi-supervised metric learning method is incorporated to generate an effective scatter plot reflecting the high-level semantic similarity for visual sample selection. Moreover, both quantitative and qualitative evaluations are provided to show that visualization based method can be effectively enhance sample selection in active learning.



III.PROPOSED SYSTEM

System Overview

Video annotation is the emerging approach it is used in the traffic process for analyzing the vehicle ,its colour and structure which are the key feature that are extracted using the GIST descriptor. The identical frames are selected and the objects in the frame are given label by using the svm algorithm. The label given to the objects are compared with the knowledge base which consist of database such as relation between object and scenes and relation between objects and then the scene is recognized using fuzzy knowledge representation schemes such as KRFPNs and fuzzy PETRINET thus the video is being annotated.

Color and spatial structure(key feature extraction)

In the Key Feature Extraction, the features are extracted from the frame. The frame is extracted from the input video by video cutting tool for frame identification. The feature set is made up of color and structure information of the object in the frame. The dominant color uses color histogram which is calculated by RGB color channel. Histogram bins with highest values for each channel were selected as dominant colors. After analyzing different number of different colors(3-36) per channel, only 12 of them in each image is selected as dominant color for our classification tasks. The information about the color layout of an image is preserved using 5 local RGB histogram. The size of DC vector is 180(12 DC X 3channel X 5image parts) DC1 to DC5. To calculate DC1, DC2, DC3 local features, a histogram is computed for each cell of a 3 X 1 grid applied to each image. The DC4 feature is computed in the central part of the image which has the size of ¼ diagonal size of the whole image and of the same proportions. The DC5 feature is the surrounding part of the image that is background of the image. GIST feature vector of 512 component is obtained using 8 X 8 encoding samples in the GIST descriptor within 8 orientations per scales of image components.

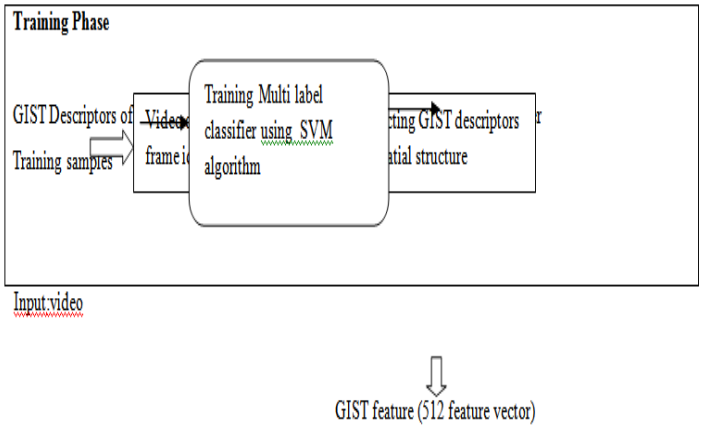


Fig 1 Key feature extraction

Object level recognition

The object level recognition is done by two phases training phase and classification phase. In the training phase, the classifier is trained by using SVM algorithm. The multi-label classification problem is expressed as

$$\Phi: E \rightarrow P(C)$$

Where **E** is a set of image examples.

P(C) is a power set of the set of class labels C.
And there is atleast one example ej ϵ E that is mapped into two or more classes.

The methods most commonly used to tackle a multi-label classification problem can be divided into two different approaches. In the first, the multi-label classification problem is transformed into more single-label classification problems, known as the data transformation approach. The aim is to transform the data so that any classification method designed for single-label classification can be applied. However, data transformation approaches usually do not exploit the patterns of labels that naturally occur in multi-label setting. In the second, adaptation methods extend specific learning algorithms in order to handle multi-label data directly and can sometimes utilise the inherent co-occurrence of labels.

For the multi-label classification task, the SVM algorithm is used which is an example of data adaptation methods.

Support vector machine (SVM) were originally designed for binary classification. Several methods have been proposed to construct a multi-class classifier [12] by combining one-against-one binary classifiers or one-against-all binary classifiers. The data sets can be linearly separable or nonlinearly separable. The nonlinearly separable cases require the use of kernel function in order to obtain linearly separable data sets. From N class in data sets, the one-against-one multiclass SVM method constructs N(N-1)/2 binary classifier where each one is trained on data from two classes.

In the classification phase, the labels are given to the object by using the trained classifier which is obtained as output in the training phase. The GIST descriptor test sample along with the trained classifier which has the many trained sample using this labels are given to the objects.

To design and extend SVM binary classifier into a one-against-one multiclass SVM, two groups of data examples are constructed from two classes. The obtained SVM binary classifier is trained to decide if the class is from the first class or it belongs to the second group of classes. This process is repeated for another couple of classes until finishing all the possible couples of the classes from data sets. So, by following this way, multiclass SVM is transformed to a multiple N(N-1)/2 SVM binary classifier.

Each SVM binary classifier is trained using a matrix of training data, where each row corresponds to the features extracted as an observation from a class. When classifying an object with an input features vector, each binary classifier from the multiclass SVM one-against-one model decides and votes for only one class.

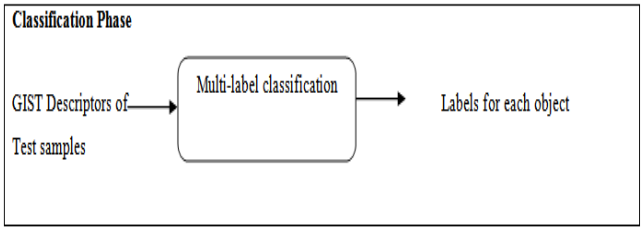


Fig 2 Object-level recognition

Scene level recognition

In the scene-level recognition, the output of object-level recognition that is object-labels are given as input and then applying fuzzy knowledge representation scheme based on FUZZY PETRINET and KRFPNs on the input, the labels for scene is obtained.

Each scene is treated as a composition of objects that are selected, based on the used training dataset. To enable making inferences about scenes, relation between objects and scenes and relation between objects in video are modelled and stored in a knowledge base. Facts about scenes and objects are collected automatically from the training dataset and organized in a knowledge representation scheme.

The elements of fuzzy knowledge used to annotate about scenes and relationships among objects, are presented using the KRFPNs scheme. The KRFPNs scheme concatenates elements of the FuzzyPetriNet(FPN) with a semantic interpretation and is defined as:

KRFPNs = (FPN; α; β; D; Σ)

FPN = (P; T; I; O; M; Ω; f; c) is a FuzzyPetriNet where P = {p1; p2; :::; pn};
n ϵ N is a finite set of places, T = {t1; t2; :::; tm}; m ϵ N is a finite set of transitions, I : T→P(P)/Ø is the input function and O : T→P(P)/Ø is the output function, where P(P) is the power set of places. The sets of places and transitions are disjunctive, P ∩ T = Ø, and the link between places and transitions is given with the input and output functions. Elements P; T; I; O are parts of an ordinary PetriNet(PN). M = {m1;m2;…;mr}; r ≥ 0 is a set of tokens used to define the state of a PN in discrete steps. The state of the net in step w is defined with distribution of tokens.

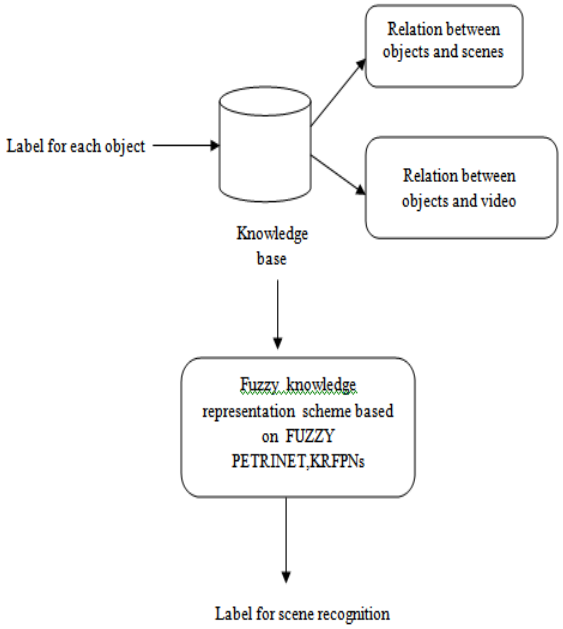


Fig 3 Scene-level recognition

IV.EXPERIMENTAL COLUMN

We compared the results of the proposed video annotation system at object and scene levels with previously published results. The performance of the annotation refers to multi-label classification at object level and is evaluated using different subsets of features.The obtained object labels are refined using the proposed annotation refinement algorithm,and the impact of that process is analysed at the scene level where the objects are treated as features for scene inference.The performance of the inference-based scene annotation in the is also compared with scene classification with the ordinary classification approach using the Naïve Bayes classifier with the same sets of features that were used in the first tier for object classification. The achieved results are averaged over 3 runs, since 3-fold cross validation was used.The classification performance at the object level was measured interms of instance- based and label-based accuracy, precision, recall and F1 score, while the classification performance at the scene level was measured using the label-based precision and recall metrics.

Evaluation measures

To define the evaluation measures,we assume that an image ej ϵ E; j = 1…..N should be classified into the set of true object labels Yj = {Cl; Cm;…; Cr }, Yj subset of C, where E is a set of images, C is a set of all class labels and N = Ej j corresponds to the number of images in the set E. For an example image ej, the set of labels that are predicted by a classifier is denoted as φj.

Instance-based accuracy is defined as the average ratio of correctly assigned labels and all labels assigned to each example by the classifier and the truelabels:

$$\text{Accuracy ins} = \frac{1}{N}\sum_{i=1}^{n}\frac{|Yi \cap \varphi i|}{|Yi \cup \varphi i|}$$

Instance-based precision is defined as the average ratio of correctly assigned labels and all labels assigned to each example by the classifier:

$$\text{Precision ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Yi \cap \varphi i|}{|\varphi i|}$$

Instance-based recall is defined as the average ratio of labels correctly assigned by the classifier and all labels in the ground truth for each example:

$$\text{Recall ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Yi \cap \varphi i|}{|\varphi i|}$$

The instance-based F-Measure is the harmonic mean of precision and recall:

$$\text{F1ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Yi \cap \varphi i|}{|Yi| + |\varphi i|}$$

These measures reach their best value at 1and the worst value at 0

The existing system with the scene-level label-based precision is 32.6% and recall is 27.5% . Where the proposed system for the scene-level label-based precision is 30% and recall is 25%.

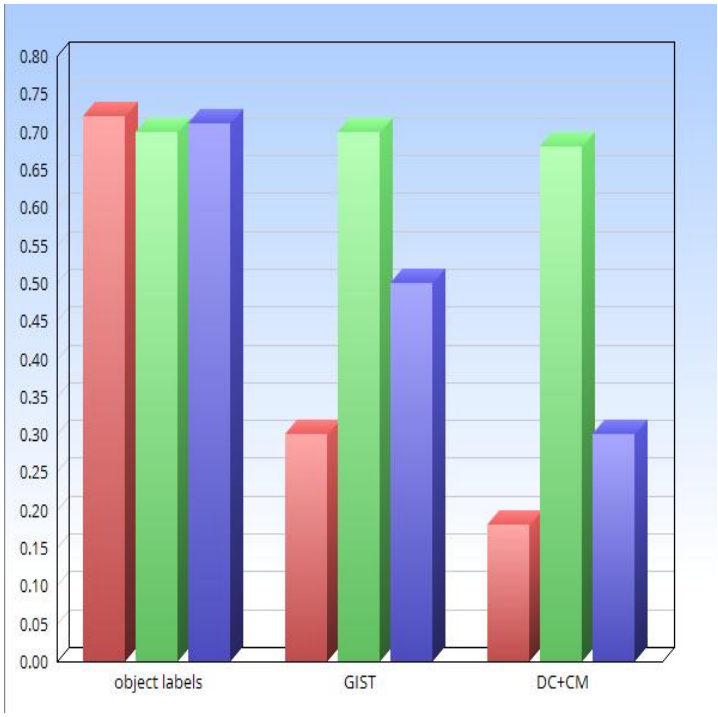


Fig 4 shows the precision, recall and f-score function values for scene-level annotation for the given key features

V.CONCLUSION

Video annotation using multi-label classification and fuzzy knowledge representation schemes is proposed in this paper. The hidden markov model used for labeling in the existing system does not provide correct results where as svm classifier trains the dataset and the trained labels are given to the objects in the frame. On separating object and scene level we gain more specified result compared to the existing system. The simplified data obtained for object label and scene label during existing system is 50% where as the proposed system is 30%. Future includes the video annotation for large size video which has more than 1000 frames.

REFERENCES

[1] Aaron O. Thomas*, Pavlo D. Antonenko, Robert Davis (2016), "Understanding meta comprehension accuracy within video annotation Systems", Computers in Human Behavior 58 269e277.

[2] Aftab Khan, David Windridge, and Josef Kittler (2014)," Multilevel Chinese Takeaway Process and Label-Based Processes for Rule Induction in the Context of Automated Sports Video Annotation", IEEE transactions on cybernetics, vol. 44, no. 10, october.

[3] Amir H. Shabani , John S. Zelek , David A. Clausi(2013),"multiple scale specific representations for human action recognition", Pattern Recognition Letters 34 (2013) 1771–1779

[4] hien-Li Chou, Student Member, IEEE, Hua-Tsung Chen, Member, IEEE, and Suh-Yin Lee, Senior Member (2016)," Multi-Modal Video-to-Near-Scene Annotation", IEEE 1520-9210 (c).

[5] Christoph Feichtenhofer, Axel Pinz Richard, P. Wildes (2016), "Dynamic Scene Recognition with Complementary Spatiotemporal Features", IEEE Transactions on Pattern Analysis and Machine Intelligence.

[6] Fatemeh Tabib Mahmoudi, Fellow, IEEE, Farhad Samadzadegan, and Peter Reinartz, Jr., Member (2015),"Object Recognition Based on the Context Aware Decision-Level Fusion in Multiviews Imagery", IEEE journal of selected topics in applied earth observations and remote sensing, vol. 8, no. 1, january.

[7] Hao Wang, Dit-Yan Yeung Senior Member (2015),"Towards Bayesian Deep Learning: A Framework and Some Existing Methods", IEEE 1041-4347 (c) 2016 IEEE.8. journal of latex class files, vol. 14, no. 8, august,

[8] Hongsen Liao, Li Chen, Yibo Song, Hao Ming (2013),"Visualization Based Active Learning for Video Annotation", IEEE .

[9] Iván González-Díaz , Tomás Martínez-Cortés, Ascensión Gallardo-Antolín, Fernando Díaz-de-María (2015)," Temporal segmentation and keyframe selection methods for user-generated video search-based annotation", Expert Systems with Applications 42 488–502.

[10] Jose M. Chaquet , Enrique J. Carmona , Antonio Fernández-Caballero," A survey of video datasets for human action and activity recognition ",Computer Vision and Image Understanding 117 (2013) 633–659.

[11] Kee-SungLee , Ahmad Nurzid Rosli , Ivan Ariesthea Supandi , Geun-SikJo (2014) ," Dynamic sampling-based interpolation algorithm for representation of clickable moving object in collaborative video annotation", Neurocomputing146(2014)291–300.

[12] M.Ravinder and T.Venugopal(2016),"Content-Based Video Indexing and Retrieval using Key frames Texture, Edge and Motion Features", International Journal of Current Engineering and Technology Vol.6, No.2 April

[13] Marina Ivasic-Kos , MiranPobar , SlobodanRibaric (2016),"Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme", , PatternRecognition52 287–305.

[14] Simon Jones, Ling Shao (2013), "Content-based retrieval of human actions from realistic video databases", Information Sciences 236 56–65.

[15] Thomas B. Moeslund , Adrian Hilton, Volker Kru¨ger," A survey of advances in vision-based human motion capture and analysis", Computer Vision and Image Understanding 104 (2006) 90–126.

[16] Wenjing Tong1, Li Song1,2, Xiaokang Yang1,2, Hui Qu1, Rong Xie1,2 (2008), "CNN-Based Shot Boundary Detection and Video Annotation".

[17] Yirui Wu, Palaiahnakote Shivakumara, Tong Lu, Member, IEEE, Chew Lim Tan, Michael Blumenstein, Senior Member, IEEE,

and Govindaraj Hemantha Kumar (2016), "Contour Restoration of Text Components for Recognition in Video/Scene Images", IEEE transactions on image processing, vol. 25, no. 12, december.

[18] Zengkai Wang, Junqing Yu, Member, IEEE, and Yunfeng He (2015)," Soccer video event annotation by synchronization of attack defense clips and match reports with coarse- grained Time information.", Neurocomputing.