An Effective Video Annotation of Semantic and Visual Context Using Spatiotemporal Fuzzy K-Means Clustering

Akileshwari R[#1], Grace Selvarani A[*2]

[#]*Department of Computer Science and Engineering,*
[#] *Sri Ramakrishna Engineering College,*
[#]*Tamilnadu, India*
[1]akilarathinam@gmail.com

[*]*Department of Computer Science and Engineering,*
[*] *Sri Ramakrishna Engineering College,*
[*]*Tamilnadu, India*
[2]*hod-mecse@srec.ac.in*

Abstract — **With the rapid increasing of video content and multimedia applications on the internet, there is a huge demand for video content analysis utilizing effective technologies. Earlier method used semantic and visual context of video for video annotation. Semantic context mining indicates the human understanding of video from labels and plays an important role in annotation. On the other hand, visual context mining reflects the natural property of video and could be considered for further refinement of annotation, but it could not be perfectly modeled. To deal with the multi-feature present in the videos and to improve the performance of annotation, a new clustering algorithm namely spatiotemporal fuzzy K-Means clustering algorithm is proposed. This method clusters the spatial and temporal features to yield the better annotated result of video annotation.**

***Keywords*— Video annotation, Context mining, Semantic context, Visual context, Fuzzy k-means clustering**

I. INTRODUCTION

With the rapidly increasing multimedia technologies and publicly available video data on the internet , the video analyzing and processing methods such as rating, indexing ,labeling for searching and retrieval purposes. One most popular technique is to annotate the videos using some automatic video annotation tools or using some technologies [2]. Due to the existence of semantic gap between low-level visual features of video and high level human understanding, many semantic concepts are still difficult to be accurately detected. Therefore, effective semantic concept detection in video remains a challenging problem. So far, content-based video retrieval [9] has achieved some good results. Although the work in [9] only proposed the content-based video clip retrieval, but the concept-based video retrieval is absent. To achieve better Video annotation performance, we proposed a new context mining of video using spatiotemporal fuzzy k-means clustering approach. Here, we combine the spatial and temporal features of the videos for clustering in semantic context mining followed by visual context mining. Firstly, semantic context mainly describes the relationship exists among the concepts present in video by using the concept labels of the training set are given by the people and also the semantic context is close to human understanding and is helpful for improving the accuracy of video annotation. Secondly, visual context means that shots with similar visual features share some common concept labels which reflect the natural property of the video shots using KNN search which could be used for further improvement of video annotation.

Based on the above analysis and consideration, we propose a new method to refine the result of video annotation by exploiting the semantic and visual context of video. Spatiotemporal fuzzy K-Means clustering algorithm is proposed for grouping the spatial and temporal information of the video shots. This approach tackles the meaningful annotation of non-domain-specific videos, whereas most of the previous work has been designed for domain-specific videos and annotation can be improved.

II. RELATED WORK

Compared with data mining, multimedia mining reaches much higher complexity resulting from: (a) The huge volume of data, (b) The variability and heterogeneity

of the multimedia data (e.g. diversity of sensors, time or conditions of acquisition etc.) and (c) the multimedia content's meaning is subjective and it is shown in fig 1.The high dimensionality of the feature spaces and the size of the multimedia datasets make the feature extraction a challenging problem. The main issue in video annotation is the existence of semantic gap. [2] Proposed an automatic semantic video annotation by utilizing two layers. Firstly,

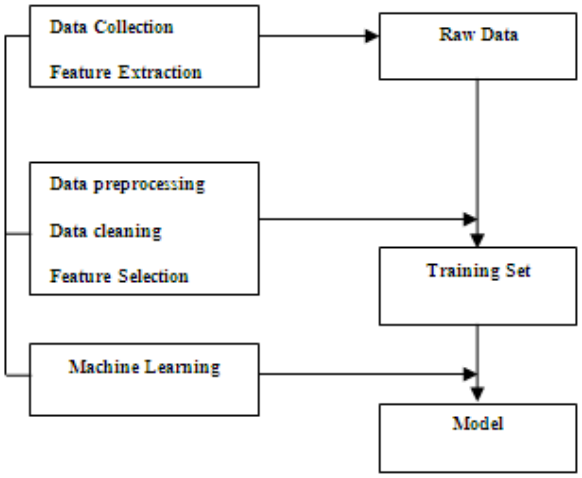

Fig .1 Multimedia Mining Process

The extraction of spatiotemporal features, suitable for efficient matching of objects, actions and scenes present in the videos; and secondly, the representation and use of semantic relationships between objects, actions and scenes to validate annotation. But this technique did not use any context information of videos. Our proposed method solves this problem. As our previous method [1] for context mining mainly contains two parts: semantic and visual context mining. Semantic context mining plays the main role for performance improvement, and visual context mining is considered as compensation and it could achieve better performance than existing approaches. K-means algorithm is a typical iterative clustering algorithm. K-Means algorithm aims at classifying data items into a fixed number of classes starting [7, 8]. To improve the video annotation result, we turn to the context mining in video, which is the main work of this paper.

III. PROPOSED WORK

Many methods used content based multimedia information retrieval [9], but the concept mining is absent in such methods. To improve the annotation using context mining by apply our new clustering approach. Here, we proposed the spatiotemporal fuzzy k-means clustering algorithm in semantic context mining followed by visual context mining as shown in fig 1.

A. Semantic Context Mining

In semantic context mining, we extract the spatiotemporal features and annotate using spatiotemporal fuzzy k-means clustering algorithm after predicting the concepts. The concepts and to measure the strength of relationship between concepts are predicted using Pearson product moment correlation is given in equation [1].

1) Feature extraction: In this module, Low level features of shots are considered as follows:

Observed feature:

Here, the low level features are extraction of color moment, wavelet texture key point from the video shots.

$X^j =< x_i^j, x_i^k >$, $(1 \le i \le n, c_k \in REL_j)$,is the observed feature for the concepts in the shot, where x_i^j is the observed feature for concept c_j and x_i^k is the observed feature for concept c_k in shot s_i[1].

2) Concept prediction: In semantic context mining, there is no need to model relationship among all concepts. Some of the concepts are weakly correlated. So, avoiding the relationship among these concepts could bring some advantages: Firstly, the constructed models are more efficient and the complexity of relationship modeling could be reduced; Secondly, the relationship modeling could be more accurate by omitting the weakly related concept, which may leads to noises because of the inaccurate concept detectors, and requires more data for effective training.

Pearson product moment correlation from [1] is used to measure strength of relationship between the concepts c_j and c_k , whose definition is given as follows:

$$PM(c_j, c_k) = \frac{\sum_{i=1}^{|S_{trn}-1|} (y_i^j - \mu_j)(y_i^k - \mu_k)}{|S_{trn} - 1|\sigma_j\sigma_k} \quad (1)$$

Where, S_{trn} is the training set μ_j and σ_j are sample mean and standard deviation of observing concept c_j in training set S_{trn} respectively.

B. Spatiotemporal Fuzzy K-Means Clustering

In clustering spatiotemporal data, we assume that there are n data $x_1, x_2 \ldots x_n$, each has its spatial and temporal components. The i^{th} data x_i is represented as a combination of its spatial and temporal components, namely, $x_i = [x_i(s) | x_i(t)]$ T,

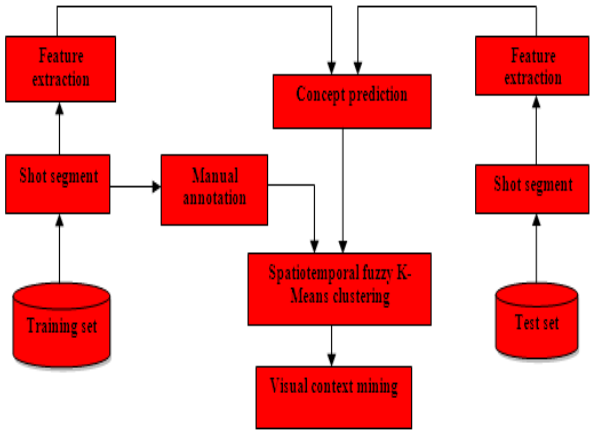

Fig .2 Spatiotemporal fuzzy K-Means clustering

where $x_i(s)$ is the spatial part of x_i, while x_i (t) denotes the temporal part of the same data point. By considering r features in the spatial part and q features in the temporal one, we have

$x_i =$
$[x_i(s)|x_i(t)]T = [x_{i1}(s) \ldots x_{ir}(s) |x_i1(t). \ldots x_{iq}(t)] T$ (2)

The fuzzy k-means clustering algorithm partitioning the data points into k clusters S_1 (l = 1, 2… k) and clusters S_l are associated with cluster center C_l. The relationship exists between a data point and cluster representative is fuzzy. That is, a membership $\mathcal{U}_{ij} \in [0, 1]$ is used to represent the degree of belongingness of each data point X_i and cluster center C_l. We denote the set of data points as S = {X_i}. The FKM algorithm is based on minimizing the following distortion,

$$\sum_{j=1}^{K}\sum_{i=1}^{N} u_{i,j}^{m} d_{ij} \quad (3)$$

With respect to the cluster representatives C_j and memberships $\mathcal{U}_{i,j}$ where N is the number of data points; m is the fuzzy parameter; k is the number of clusters; and d_{ij} is the squared Euclidean distance between data point X_i and cluster representative C_j. It is noted that ui,j should satisfy the following constraint:

$$\sum_{j=1}^{K} \mathcal{U}_{i,j} = 1, \text{ for i=1 to N}$$

The major process of FKM is mapping. It begins with a set of initial cluster centers and repeats this mapping process until it satisfies the entire stopping criterion. It has constraint that no two clusters have the same cluster representative.

C. Visual Context Mining

To achieve better performance, we further refine the annotation using the visual context information which uses KNN search for graph construction. Let f_i be the visual features extracted from $shots_i$. We construct the graph from [1] G= (V, E) for visual context mining. G mainly consists of two parts: node set V and edge set E. In $\mathcal{V}$= {v_i}, (1≤ i ≤ n), v_i models shot s_i where,s_i is a shot, and n is the number of the video shots. In $\mathcal{E}$= {$\mathcal{E}_{i,i'}$},(1≤ $i, i' \leq n, i \neq i'$) , $\mathcal{E}_{i,i'}$ is the edge connecting nodes v_i and v_i', whose weight is $w_{i,i'}$. $w_{i,i'}$. is computed by the normalized cross-correlation using $f_i$and f_i'.
Let $\rho^j = < \rho_1^j, \rho_2^j, \ldots . \rho_n^j >$be the previously obtained probabilities of the shots forc_j. The refined probabilities $\rho^{-j} = <\rho_1^{-j}, \rho_2^{-j}, \ldots . \rho_n^{-j}>$ using visual context mining could be obtained by

$$\rho^{-j} = (I - \mathcal{K} \times \mathcal{S})^{-1} \rho^{j} \quad (4)$$

$$\mathcal{S} = \mathcal{D}iag^{-1/2} \times \mathcal{W} \times \mathcal{D}iag^{-1/2} \quad (5)$$

Where $\mathcal{W} = \{w_{i,i'}\}, (1 \leq i, i' \leq n, w_{i,i} = 0)$ is the weights of edges, $\mathcal{D}iag$ a diagonal matrix with its (i,i) element equal to the sum of the ith row of $\mathcal{W}$, and $\mathcal{K}$ is a parameter.

Hence the video annotation is refined further according to the visual property of the videos.

IV.EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed method, we mainly conduct our experiment on UCF or TRECVID 2006 dataset, Here we choose the UCF Action dataset for two reasons: Firstly, it has been widely used in many recent works of context mining for video annotation [2] and [5] ;Secondly, this dataset contains various video content and it is better enough to evaluate our method.UCF dataset contains many types of manmade objects designed for information retrieval which can be downloaded from [5]. This dataset contains 1600 video clips related to many activities including: basketball shooting, biking, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and dog walking.

A thumbnail sample of this dataset is presented in Fig. 2. The focus of our work is mainly on context mining, so the experiments carried on the UCF action dataset are enough to show the effectiveness of our proposed method. Table I describes this method. As shown in this table, our existing and proposed method using classifiers i.e., SVM based on three different kinds of visual features for each

concept. For each shot to be processed, each classifier could predict a score value of the concept occurring in this shot. Then, the three scores are averaged finally as the predicted probability for this concept. By applying our methods to the dataset, we can evaluate the performance of our methods for context mining. The annotated results for each concept are results as a shot list in the descendent rank according to the refined probability of the video shots for the concept, so that the performance could be measured by the taken metrics as mean inferred average precision (infAP) [10] in information retrieval.

InfAP is an approximation of the average precision (AP) with the incomplete ground truth and AP approximates the area under the precision-recall curve. To find the performance over multiple semantic concepts, mean inferred AP is used. Larger value of mean InfAP gives better performance.

A. Results:

To show the effectiveness of the proposed approach for semantic context mining, we compare our method. Our previous method mainly focused on the semantic context mining and our method focused on refinement of context mining. So, we used clustering approach for spatial and temporal features followed by visual context mining. As shown in Table II, our method outperforms the existing method with a 52.6 % performance gain on the UCF dataset actions. The main reasons for these improvements are as follows. Here, we perform the clustering approach using fuzzy k-means clustering to form the clusters with the relevant concepts based on the spatial and temporal feature values and then we used the visual context mining for further refinement.

Here, we compared the average precision (InfAP) of two methods for the data sets with its dataset. Concepts from [10]. The fig 4 compares the performance gain of two methods with its InfAPs for 10 concepts in the official evaluation of UCF benchmark, using our method for the choosen datasets and the table 2 shows the InfAP values for the each video in the dataset.

TABLE I

SUMMARY OF OVERALL PERFORMANCE GAINS ON BASELINES FOR 10 OFFICIAL EVALUATED CONCEPTS IN UCF ACTION DATASET BENCHMARK BY APPLYING SPATIOTEMPORAL FUZZY K-MEANS CLUSTERING IN CONTEXT MINING. SYMBOL '+' MEANS PERFORMANCE GAIN (RELATIVE IMPROVEMENT MEASURED BY MEAN INFAP [10] COMPARED WITH BASELINE.

Baseline	UCF Action Dataset
Mean InfAP	0.1948
Previous Method[1]	+30.2%
Our Method	+45.7%

Fig .3 A UCF actions videos dataset snapshot. Each row illustrates eight thumbnails from one of the 11 categories.

Fig .4 InfAPs for 10 concepts in the official evaluation of benchmark UCF Action Dataset, using our method for combining spatial and temporal context using Fuzzy k-means clustering approach

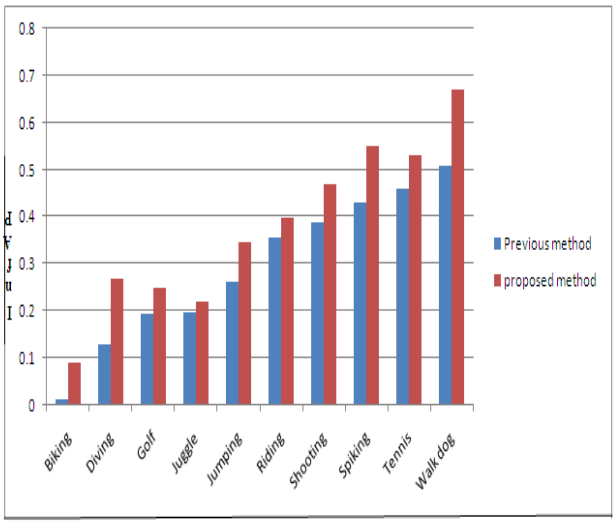

TABLE II

PERFORMANCE GAIN: INFAPS FOR 10 CONCEPTS IN THE OFFICIAL EVALUATION OF BENCHMARK UCF ACTION DATASET, COMPARISON OF INFAP OF TWO METHODS.

Dataset	Mean InfAP	
	Previous method[1]	Proposed method
Biking	0.023	0.095
Diving	0.136	0.282
Golf	0.195	0.250
Juggle	0.198	0.212
Jumping	0.261	0.342
Riding	0.371	0.393
Shooting	0.385	0.472
Spiking	0.444	0.561
Tennis	0.450	0.553
Walk Dog	0.501	0.682

V.CONCLUSIONS

Here, we proposed two steps to refine the context mining. In the present work, video annotation by mining the video context has been presented by using proposed spatiotemporal fuzzy K-Means clustering algorithm. Initially semantic context mining followed by visual context mining has been done. Semantic context mining indicates the human understanding of video from the labels and plays the main role in annotation. On the other hand, visual context mining reflects the natural property of the video shots which is considered as the further refinement process, which generally could not be perfect modeled. In semantic context mining, we model the spatial and temporal context information in video is clustered by using proposed fuzzy K-Means clustering. This clusters the information of spatial and temporal behavior of images.

It was shown that, for different nature of spatial and temporal components of the data, different treatment has to done to control the influence of temporal and spatial components. Comparing with existing methods of semantic context mining for video annotation, our method could more accurately capture concept relationship in video and more effectively improve the video annotation performance, by describing the impacts among concepts present in spatial and temporal context with the learned parameters. In visual context mining, the performance of context mining needs further improvement in future.

ACKNOWLEDGMENT

The authors would like to thank the Management, Director, and Principal of Sri Ramakrishna Engineering College for providing laboratory resources and valuable support.

REFERENCES

[1] Jian Yi, Yuxin Peng, and Jianguo Xiao," Exploiting Semantic and Visual Context for Effective Video Annotation", In: IEEE Transactions on Multimedia, Vol. 15, 2013.

[2] Amjad Altadmri and Amr Ahmed," A framework for automatic semantic video annotation Utilizing similarity and commonsense knowledge bases", In: Multimed Tools Appl, 2013.

[3] Y. Liu, T. Mei, X. Hua, X.Wu, and S. Li, "Multi-graph-based query-independent learning for video search," In;IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 12, pp. 1841–1850, 2009.

[4] Y. Peng et al., "PKU-ICST at TRECVID 2009: High level feature extraction and search," In Proc. TRECVID, Gaithersburg, MD, USA, Nov. 16–17, 2009.

[5] UCF_Computer_Vision_lab (2011) Ucf action dataset (11–11–2011). http://www.cs.ucf.edu/~liujg/YouTube_Action_dataset.html

[6] Y. Li, Y. Tian, L. Duan, J. Yang, T. Huang, and W.Gao, "Sequence Multi-Labeling: A unified video annotation scheme with spatial and temporal contexcontext,"In:IEEE Trans. Multimedia, vol. 12, no. 8, pp. 814–828, Dec. 2010.

[7] S. Eschrich, J. KE, L. O. Hall, and D. B. Goldgof. Fast accurate fuzzy clustering through data reduction. IEEE Transactions on Fuzzy Systems, 11(2):262–270, 2003.

[8] A. Liew, S. H. Leung, and W. H. Lau. Segmentation of color lip images by spatial fuzzy clustering. IEEE Transactions on Fuzzy Systems, 11(4):542–549, 2003

[9] M. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State-of-the-art and challenges," ACM Trans. Multimedia Comput., Commun., Appl., vol. 2, no. 1, Feb. 2006.

[10] E.Yilmaz and J.Aslam, "Estimating average precision with incomplete and imperfect judgments," in Proc. CIKM, 2006, pp. 102–111.

[11] Y. Peng and C. Ngo, "Clip-Based similarity measure for query-dependent clip retrieval and video summarization,"In: IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 5, pp. 612–627, 2006

[12] Zhao WL, Wu X, Ngo CW (2010) On the annotation of Web videos by efficient near-duplicate search. In: IEEE Trans Multimedia 12(5):448–461

[13] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, "An empirical study of context in object detection," in Proc. CVPR, 2009, pp.1271–1278

[14] D. Zhou, O. Bousquet, T. N. lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in Proc. Nips, 2004.

[15] X. Wu, A. Hauptmann, and C. Ngo, "Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts," in Proc. ACM Multimedia, 2007.

BIOGRAPHIES

Akileshwari.R, she is currently pursuing M.E. in computer science and engineering from Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India. Her research interests are Data Mining, Image and Video Processing.

Prof.A.GraceSelvarani, she is currently working as a Professor and Head for the Department of computer science and engineering (PG) in Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India. Her research interests are Image and Video Processing, Data mining, Cloud Computing.