OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

AUTOMATIC VISUAL CONCEPT DETECTION IN VIDEOS

Nilam B. Lonkar [1], Dinesh B. Hanchate [2]
Student of Computer Engineering, Pune University VPKBIET, Baramati, India
Computer Engineering, Pune University VPKBIET, Baramati, India

---

Abstract: The tasks like scene identification and object classification interrelated to the concept detection. It involves scene types as well as object categories which are relevant to the concepts. Visual concept has some negative aspects, while the visual concept related to event analysis is carried out significant improvement. Visual concept is defined by human and it has only one corresponding classifier in usual method are the negative aspects. This approach has proposed concept learning algorithm for handling all these issues for social event detection in videos. System imparts a powerful automatic concept mining algorithm with the help of N-gram internet services and flicker rather than defining visual concept manually. At the same time depending on the learned visual concept, system repetitively finds out the multiple classifiers for each and every concept. System gives a novel boosting concept learning algorithm, which increases the quality of being distinguishable.

Keywords *— Classification, Event analysis, Video recognition, Visual concept detection*.

I INTRODUCTION

To identify, manage and classify visual information, visual concept detection is a useful process. The task of visual concept detection is connected to the field of image and video analysis. Scene identification and object category recognition are strongly regarding to concept detection. Because scene types as well as object categories are related to concepts. The occurrence of the semantic idea (like objects, locations or activities) from the audiovisual content of the video stream is directed at automatically inferring concept recognition.

Two important things includes in concept detection, i.e mining of concepts and learning boosted concept. In first, ac-cumulate an auxiliary images with parallel textual descriptions of Flicker. Then, system naturally mine compact correct phrase segments as a concept based on the text related information. With the help of words closeness and detectable representativeness, phrase segments are discovered. Wikipedia and the Microsoft N-gram services are extracted segments which used to find out the word closeness of phrase segments. For checking visible closeness between pictures retrieved from Flicker is measured visible representativeness. Phrase phase is used as the search text. So, selection of phrase segments which have larger stickiness value and visible feature extraction which represents image are the two main tasks. Social media is one of the important stages of gathering and exposing the information in current stage. In the same manner the interaction degree of social site increases because of images and videos. This two fusion generates new category of content called as social multimedia. The today's expansion of effective phones, digital cameras and social media websites like Facebook, YouTube, Flicker. It is very easy for the group to discover and broadcast information online, which gives the functionality like information creation, allocation and transformation. So, media data are useful for efficient browse, search and observe social events over clients or government. It is very important for understanding social events from whole social media data.

The technology of concept finding is a powerful technique. It includes automatic detection and also a huge number of images management and classification. The system derived the images which are fulfilling some condition. To achieve the goal of analysis and judgment of the image with machines rather than of manual work system can realize the automatic image detection, identification and classification. In detection techniques of the visual concept detection system, the current re-search stage involves various approaches like, domain selection machine method, cross-domain learning method, extracting distinctive invariant feature on images learning method based on natural language for visual concept detection method

 ISSN (Online) 2456-3293

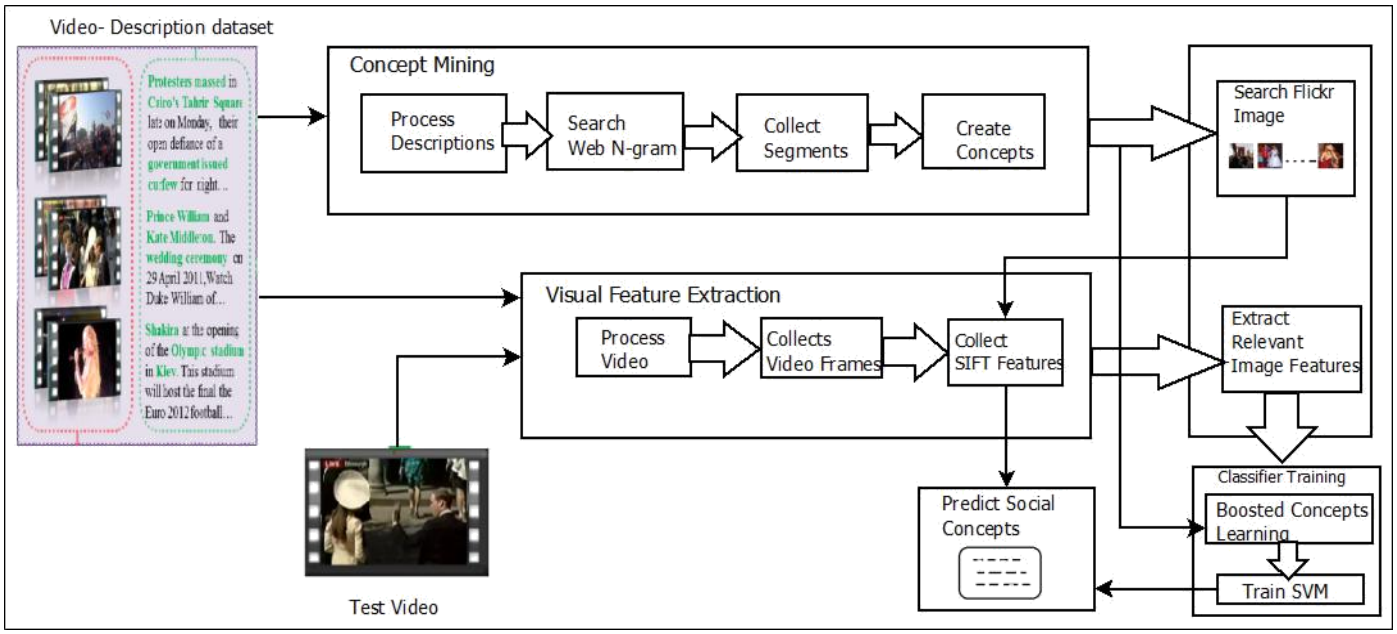

Figure 1 System Architecture

II RELATED WORK

H. Zhang and J. Guo [2] proposes a novel cross-domain learning mechanism which is used for delivering the correlation knowledge among different information sources used to divide condition of textual description missing in the image. In concept based representation method [3], to handle multimedia event, recounting approach plans a pilot analysis. It gives tips on, what basis this decision is made up and why this video is categorized in this event. The recounting covers all additional semantic declarations of the event classification. So, this approach is generally suitable for any supplement classifier. In heterogeneous features and model selection of event based media classification paper [5], it targets the basic problem handling of social media analysis. In knowledge adaptation method [7], the multimedia refers the infer knowledge for detecting event. It introduces various semantic concepts related to the Ad Hoc method of target videos. Firstly, this approach mines shared inconsistency and noise among the various video. This is very important to collect positive examples. J. Sivic and A. Zisserman explain object and scene collection method [9]. It finds user intended searched for those objects are found in the video. It finds user intended searched for those objects are found in the video.

III PROBLEM FORMULATION

The problem is to detect all meaningful event related concepts in the video, such that all video frames can be represented well. An instance of this problem consists of set of videos V

$V=\{V_0, V_{1, \ldots,} V_{n-1}\}$ where n is the total number of videos.

Let D be the set of description

$D = \{D_0, D_1, \ldots, D_n\text{-}_1\}$ where n is the total number of description.

Let G be the set of N-grams

$G = \{G_0, G_1, \ldots, G_{m-1}\}$where m is the total number of N-grams.

$$G_i = \begin{cases} True & if\ G_i \subseteq T\ . \\ False & Otherwise. \end{cases}$$

Let S be the set of segments

$S = \{S_0, S_1, \ldots, S_{n-1}\}$ where n is the total number of segments.

$$S_i = \begin{cases} True & if\ S \in G\ AND\ |S_i| \geq \theta. \\ False & Otherwise. \end{cases}$$

Θ = user defined threshold.

Let F be the set of features

$F = \{F_0, F_1, \ldots, F_n\text{-}_1\}$ where n is the total number of features.

Let C be the solution set of detected concepts

$C = \{C_0, C_1, \ldots, C_n\text{-}_1\}$ where n is the total number of concepts.

$$C_i = \begin{cases} True & if\ F_{ij} \cong F_{ik}\ . \\ False & Otherwise. \end{cases}$$

F_{ij} = feature of j^{th} image.

F_{ik} =feature of k^{th} image.

 ISSN (Online) 2456-3293

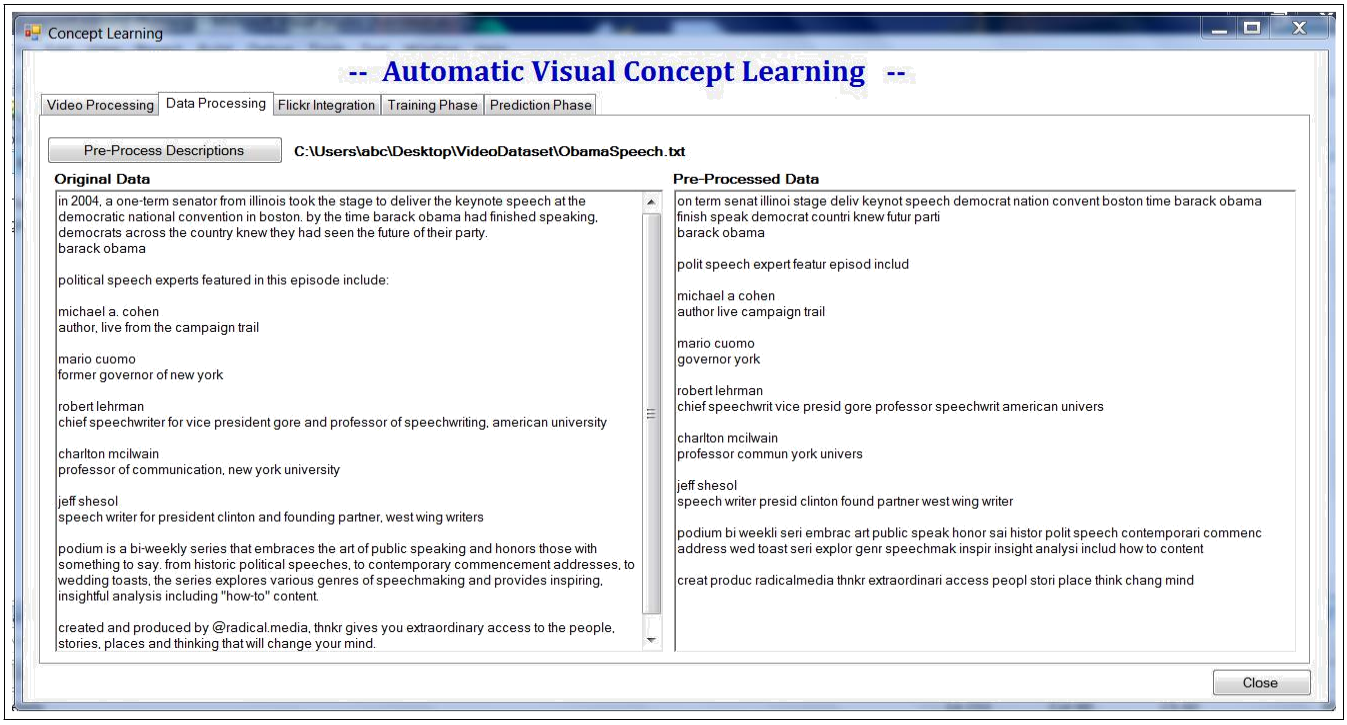

Figure 2 Main form shows window after data processing

IV SYSTEM ARCHITECTURE

A. Video description and datasets:

It contains the video and it's corresponding description which is used in the training phase.

B. Concept mining:

1) Process description: The preprocessing is done with the description of the video. The following steps are carried out: Remove the common stop words:

The words which are clarified in natural language may be before or after processing of words are called as stop words. The words like less function words the, is, at etc. are treated as the stop words.

- **Remove the common non words:**

In the non words remove the question marks, punctuation marks, etc., those which are the semantically irrelevant.

- **Stemming of the word:**

The method of cutting down derivational words from their roots forms generally a written word forms.

For example: removing and removal are stemmed to remove.

2) Web N-gram search: An n-gram is nothing but a nearby sequence of number of items from a given chain of text or speech. Depending on the application that things can be characterized, consonant, letters, words or common pairs. From a text or speech collection n-gram is generally gathered.

3) Collect segment: The valid parses or segments are collected from Web N-gram.

4) Create concept: By only seeing the text information, the phrase segment of the text description is obtained. To describe a specific event these segments are useful, but they are most likely being not able to use for visual information about event analysis. Both textual and visual information's need to be selected for considering the visual concepts from these segments. The chances of which segment is selected as concept is calculated by segment stickiness value and visual representative multiplication.

$$\textbf{Score(se)} = \textbf{Stc(se)}.\ \textbf{V}_{\textbf{flickr}}\textbf{(se)} \qquad (1)$$

Here, segment stickiness is represented by Stc(se), V_{flickr}(se) is the visual representation which is used as the effectiveness of segment by imparting the visual content of the videos. Specially, V_{flickr} (se) is figure out as the visual similarities of returning images. For search query we used I_{se} in which segment se retrieve from the Flickr. Fourier transformation is used for similarity measurement.

C. Visual feature extraction:

Here, at first it collects the video frames after that collects SIFT features of video frame image and result to retrieve from Flicker images. Finally, comparison of test video frames and result from SVM takes place and it gives the predicted concept.

 ISSN (Online) 2456-3293

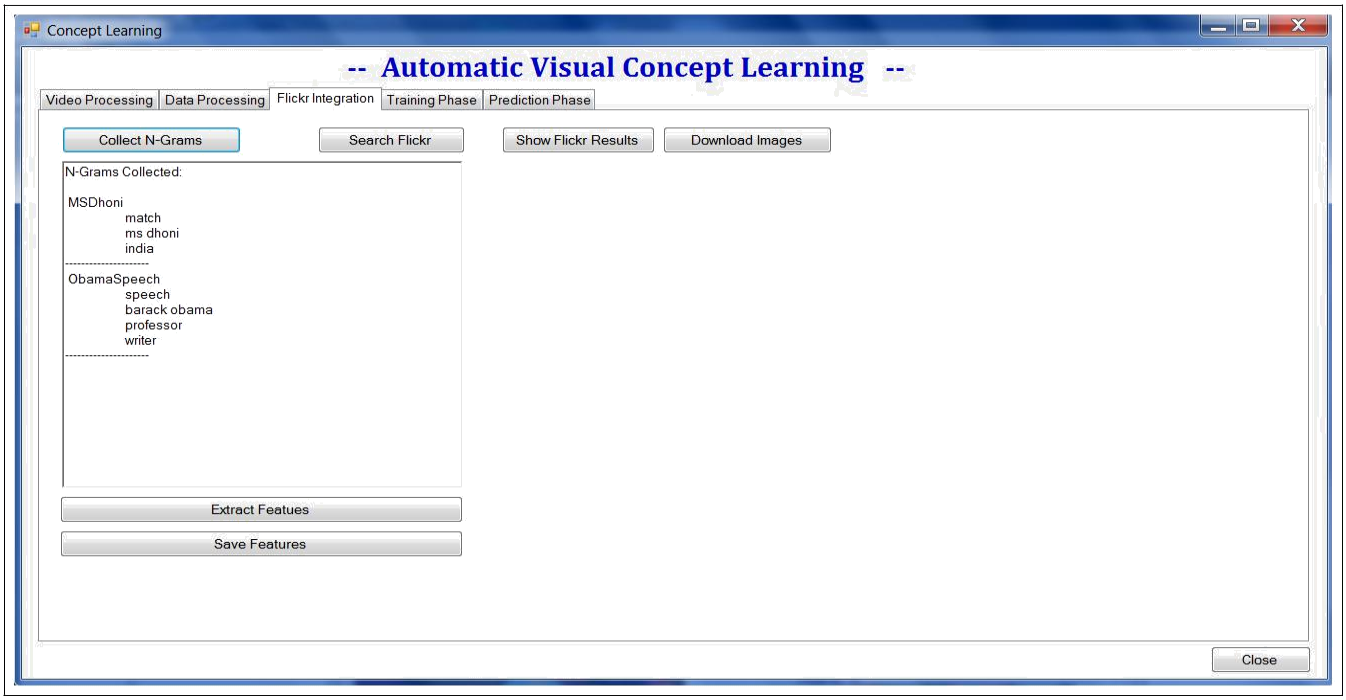

Figure 3. Diagram shows window of N-gram module

V IMPLEMENTATION

The input of this system is the video and its corresponding description and output is the concept presents in the video. Visual concept detection is having modules such as preprocessing, N-grams collection and feature extraction. In preprocessing module, video description is pre-processed first by removing stop words, non words. In the next step, N-grams are collected from segments which are derived from video description. This N-grams are information that is valid data retrieve from the description of the video. Then we annotate frames using N-grams. Features are extracted in the feature extraction module. The annotated frames are compared with frames retrieved in the testing phase.

VI RESULTS

In the pre-processing module, we remove the stop word, non word. After that find the root word by using stemming method. Figure [2] shows the window after processing video description. An n-gram is nothing but a nearby sequence of number of items from a given chain of text or speech. Figure [3] shows N-gram module for collecting valid segment from video description. The particular frame is annotated with the help of N-gram retrieving in the previous module. These annotated frames are used after comparison purpose. Figure [4] shows module for frames annotated by the N-grams, which is extracted previously. The features of the frames is calculated by using (Speeded up robust features) SURF method. Figure [5] shows features are extracted for collecting frames. Figure [6] shows the concept retrieve from video.

VII RESULT ANALYSIS

We use the video datasets, which contains videos and its corresponding description that crawl using You Tube. In table no I, we demonstrate the discussion about the data table which is made up from the dataset. If we take the 20 videos and its corresponding description. Here, the videos are relevant to each other. It gives the 14 concepts.
Sample:
https://www.youtube.com/watch?v=0SLfCkXDAf8

TABLE I
DATA TABLE DISCUSSION

Sr No.	#Videos	#Descriptions	#Concepts
1	20	20	14
2	20	25	16
3	25	25	18
4	30	35	23
5	38	38	32

 ISSN (Online) 2456-3293

Figure 4 Diagram shows window after frames annotated by the N-grams

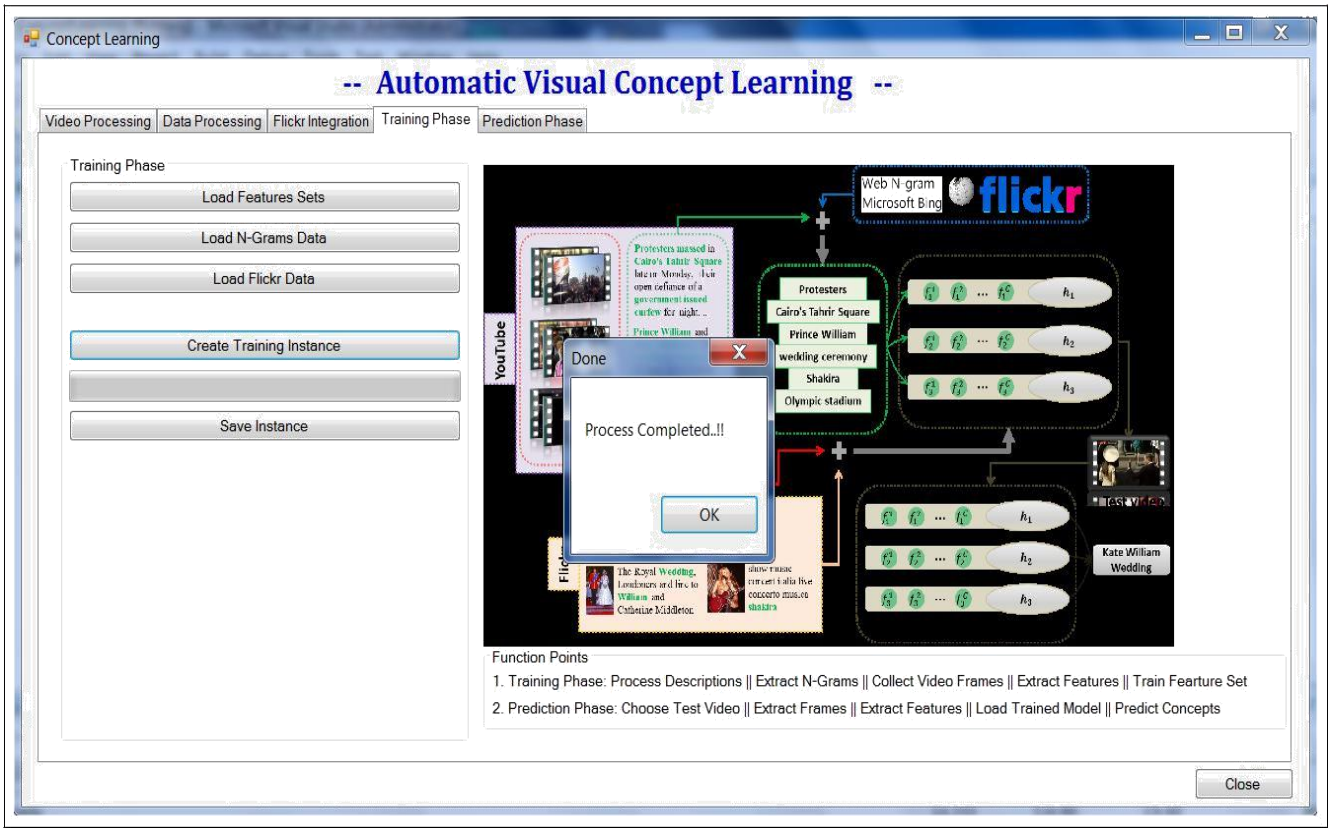

Figure 5. Diagram shows window after feature extraction

 ISSN (Online) 2456-3293

Table II Result table

SrNo	#videos	#frames	#derived Concepts	#Predicted Concepts	#Correctly PredictedCon.	Precision	recall
1	20	13760	12	10	9	0.9	0.83
2	20	14970	18	15	13	0.86	0.83
3	25	15750	17	15	12	0.8	0.87
4	30	20450	20	18	16	0.9	0.88
5	28	22500	25	22	20	0.9	0.88

Fig. 6. Diagram shows window after concept detection

Then, we take the 20 videos in which some video contains double description for better understanding. So the value of description is 25 and corresponding known concepts are 16. Similarly, We take 25 videos and its description, which has 18 concepts. The Table II is a discussion about result. Here if we consider the 20 relevant videos then we get approximately 13760 frames and the concepts retrieve from flicker which is derived concepts are 12. System predicted a concept is 10 out of 9 are correctly predicted. Performance measures are precision and recall are considered here for measuring performance. Similarly, all the values are calculated. Performance Measures:

$$1)\ \Pr ecision = \frac{Correctly\ \Pr edictedCon cepts}{\Pr edictedCon cepts}$$

$$2)\ \mathrm{Re}call = \frac{\Pr edictedCon cepts}{DerivedConcepts}$$

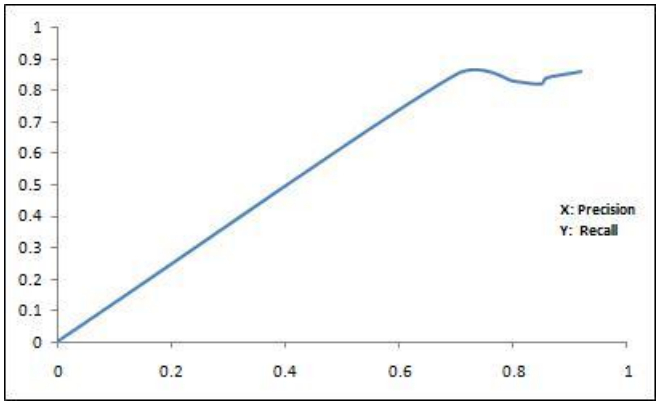

Figure 8 Result graph

VIII CONCLUSION

System models an automatically visual concept detection to find out the procedure for sociable occasion identification. To acquire this purpose, we firstly performs automatic concept mining. Then, extract the features from frames and then do the compression. Automatically mine visual concepts from the text, gives efficient and effective system which requires very less user interaction.

ACKNOWLEDGMENT

This paper would not have been written without the valuable advices and encouragement of Dr. D. B. Hanchate, guide of ME Dissertation work. Authors special thanks go to Prof. S. A. Shinde and Prof. S. S. Nandgaonkar, Head of Computer Department & principal Dr. M. G. Devamane, for their support and for giving an opportunityto work on concept detection in the videos.

REFERENCES

[1] L. Duan, D. Xu and S.-F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012.

[2] W. Lu, J. Li, T. Li, W. Guo, H. Zhang and J. Guo, "Web multimedia object classification using cross-domain correlation knowledge," IEEE Transactions on Multimedia, 2013.

[3] Q. Yu, J. Liu, H. Cheng, A. Divakaran and H. S. Sawhney, "Multimedia event recounting with concept based representation," in Proceedings of the ACM Conference on Multimedia, 2012.

[4] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng and H. S. Sawhney, "Video event recognition using concept attributes," in Proceedings of the IEEE Workshop on Applications of Computer Vision, 2013.

[5] X. Liu and B. Huet, "Heterogeneous features and model selection for event-based media classification," in Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, 2013.

[6] Z. Ma, Y. Yang, Y. Cai, N. Sebe and A. G. Hauptmann,"Knowledge adaptation for ad hoc multimedia event detection with few exemplars," in Proceedings of the ACM International Conference on Multimedia, 2012.

[7] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[8] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in Proceedings of the IEEE International Conference on Computer Vision, 2003.

[9] M. Zaharieva, M. Zeppelzauer and C. Breiteneder, "Automated social event detection in large photo collections," in Proceedings of the ACM International Conference on Multimedia Retrieval, 2013.

[10] Y. Yang, Z. Ma, Z. Xu, S. Yan and A. G. Hauptmann, "How related exemplars help complex event detection in web videos," in Proceedings of the IEEE International Conference on Computer Vision, 2013.

[11] G. Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of key points," in Proceedings of the European Conference on Computer Vision, Workshop, 2004.

[12] X. Yang, Q. Song, and Y. Wang, "A weighted support vector machine for data classification," International Journal of Pattern Recognition and Articial Intelligence, 2007.

[13] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in Proceedings of the International Conference on Machine Learning, 2011.

[14] S. Orlando, F. Pizzolon and G. Tolomei, "Seed: A framework for extracting social events from press news," in Proceedings of the International World Wide Web Conference,2013.