International Journal of Computer Applications Technology and Research
Volume 3– Issue 10, 612 - 616, 2014, ISSN: 2319–8656

Tagging based Efficient Web Video Event Categorization

A.P.V.Raghavendra
V.S.B Engineering College
Karur, India

Abstract: Web video categorization is one of the emerging research fields in the computer vision domain due to its massive volume growth in the internet which demands to discover events. Due to insufficient, noisy information and large intra class disparity makes it more daunting task to recognize the events. Most of the recent works focus on constrained (fixed camera, known environment) videos with supervised labelling to categorize the web videos. In this paper, we propose the subject based Part-Of- Speech (POS) Tagging technique with the assist of Named Entity Recognition (NER) and WordNet is applied on YouTube video titles to discover the events based on the subject, not on the objects visualized in the videos. Unsupervised learning method is used on high level features (titles) because of incoming videos are not known and large intra-class variations. For the experiment, we have chosen topics from Google Zeitgeist and downloaded the related videos from YouTube. A novel conclusion is derived from the experimental result that use of low level features will lead to a poor classification in discovering intra class events based on the subject of the videos.

Keywords: video categorization, Natural Language Processing, Parts Of Speech Tagging, Named Entity Recognition, WordNet

1. INTRODUCTION

Computer vision is one of the vast and important domain in which video event classification becomes more important than ever in nowadays. Video event classification is important because of the number of volumes growing at exponential rate in the internet and moreover replication of the identical video exist in different video sites calls a need to research and mine the efficient and effective events. According to YouTube [1], "100 hours of video are uploaded to YouTube every minute" shows the importance level of event recognition. In this paper, we have mined only the high level features to find subject based events and also proved that low level features will not be useful for classifying the intra-class events.

Lemmatization is the algorithmic process of determining lemma for a given word. In linguistics [3], "lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single term". Lemmatisation is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech.

Parts-Of-Speech (POS) Tagging is useful for identifying a word used in the corpus corresponding to a part of speech. In this work, we have used 36 types of POS tags to determine to each and every word. Named Entity Recognition (NER)[4] is the boost up for the POS tagging process because it is not able to identify the named words (names, places, songs, organizations, countries, etc). NER is used to tag the word with named entities. WordNet [5] is used to minimize the number of repeated terms to get an efficient mining result.
There are two novel conclusion found out in this work. First, the low level features are not efficient for classifying the intra class events for that SIFT (Scalable Invariant Feature Transformation) is used. Second, two level hierarchy events are represented such as for example Level 1 "blackberry" Level 2 "blackberry torch" which clearly shows that level 2 is the specified part of level 1. The rest of the paper is organized as follows. Section 2 gives a brief overview of related works. Section 3 gives the detail of the proposed framework for unconstrained video classification. Section 4 presents experimental results. Finally section 5 concludes the work.

2. RELATED WORKS

Supervised learning method is used to classify the videos by means of training set. Some of the works are either limited to some specific domains (e.g. movies [12, 13], TV videos [14, 15, 16] etc.) or focus on certain predefined content such as human face [17, 18] and human activities [19]. Nowadays, the challenging task is to categorize the videos by using unsupervised labelling techniques.

In recent trends, web video event classifications are done by using high level and low level features. In [6], association Rule mining is applied on the titles and descriptions of the video to mine the events and the find events are used as the label for classifying the videos by using statistics and distribution characteristics. Video taxonomic classification systems are presented in [7, 8], with more than one thousand categories in consideration. However, in [7, 8], the taxonomic-structured category labels are predefined by domain experts. Also those categories can include anything and do not necessarily correspond to events.

3. PROPSED FRAMEWORK

In the proposed framework, POS tagging is mainly applied on the YouTube titles by considering descriptions as a noisy data based on the analysis. The following diagram depicts the architecture of the proposed framework.

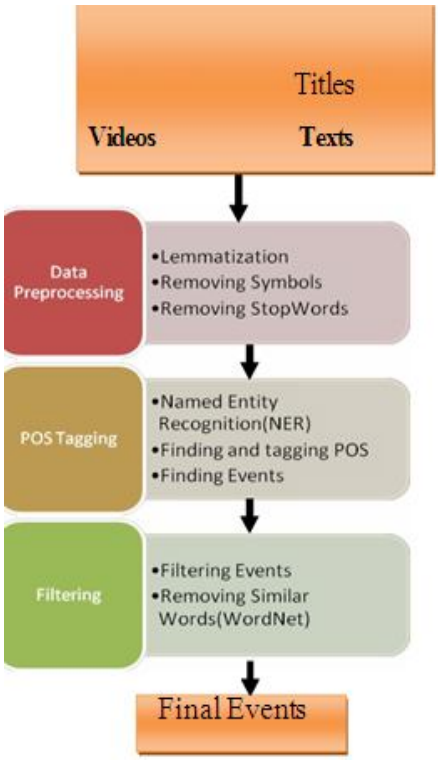

Figure 1. Architecture of proposed framework

3.1 Data Preprocessing

3.1.1 Lemmatization

Lemmatization or Stemming is the process of reducing inflectional forms and sometimes derivationally related forms of a word to a common base form. In stemming, the effective and popular algorithm is Porter's algorithm. But, there is small different between lemmatization and stemming is that stemming refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. *Lemmatization* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma.*
Examples:

sinking -> sink+ ing -> sink
Received -> receive
Went -> go

3.1.2 Removing Symbols

In the event recognition, symbols are not important to mine the events. Hence, we remove all the symbols (`!@#$%^&*()-_=+;:"<>',./\[]{}) by using the tokenization concept.

3.1.3 Removing stopwords

Stopwords is a set of words which is not constructive in the process of mining the events from the corpus text. The list of words that are not to be added is called a stop list. Stop list is used as a training set to remove all the stopwords. In order to save both space and time, these words are dropped at indexing time and then ignored at search time. We have used the stop list from [9] which is recommended and we have added own stopwords based on the analysis

3.2 POS tagging

3.2.1 Trained Entity Recognitition

Named Entity Recognition (NER) is typically viewed as a sequential problem. Named Entities is the single term in the text corpus belonging to predefined classes such as person, location, nationalities, organizations, etc. The effective NER can be built by using heavy trained data. We have collected maximum data to make it as well trained set. The list of training data is given below as:

Table 1. Training set used in NER

List of topics	Number of words
Numbers, temporal words, currencies, measurements	1103
Names	649
Places	20226
Jobs	10,020 1,064
Titles	915
Location	952,674
Artworks	128,193
Competition	129,388
Films	146,129
Object name	148,125
Songs	212,851
People	2,319,335
Total	4,814,852

3.2.2 Finding and tagging POS

Parts Of Speech (POS) tagging is the problem of assigning each word in a sentence the part of speech that it assumes in that sentence. In our work, there are 36 tags are covered to find and tag the words [2,10]. The SNOW (Sparse Network Of Linear Separators) [11] utilizes the Winnow learning algorithm is used to tag the each word in the sentences.

Training of the SNOW tagger network proceeds as

follows: Each word in a sentence produces an example. Given a sentence, features are computed with respect to each word thereby producing positive examples for the part of speech the word labelled with, negative examples for the other parts of speech. In testing, this process is repeated, producing a test example for each word in a sentence.

Once an example is produced, it is then presented to the networks. Each of the sub-networks is evaluated and we select the one with the highest level of activation among the separators corresponding to the possible tags for the current tags. After every prediction, the tag output by the SNOW tagger for a word is used for labelling the word in the test data. Therefore, the features of the following words will depend on the output tags of the preceding words.

3.2.3 Finding events

Out of all the tags, we are mainly concentrating on noun tags such as NN (Noun Singular), NNP (Noun Plural), and NNPS (Proper Noun Singular) to find the events based on the subjects.

Examples:
Emerging **India** vs sinking **Pakistan**
Blackberry torch 9800 for sale

3.3 Filtering

3.3.1 Filtering events

It is the process of removing similar events for finding at the first instances. It not only filters the event but also gather the events which are similar to each other and represent that event with all the related videos under particular event.

3.3.2 Removing words

It is the intelligent process of filtering the events. WordNet is a lexical database for the English language. It groups English words (nouns, verbs, adverbs, adjectives) into set of synonyms called as synsets, provides short, general definitions and records the various semantic relations between these synonym sets. Thus the final events will b displayed along with the related vedios

4. EXPERIMENTAL RESULTS

To evaluate the effectiveness of proposed work, we have collected and used our own dataset. Since all the existing dataset is using the videos which is constrained and only used to find particular set of events. Thus, we have collected the top ten events happened in India, 2013 by getting the result from Google Zeitgeist.

Table 2. Dataset used

Events	Number of vedios
PM candidature of Narendra Modi	200
Blackberry sale	200
Dravid retirement	200
Air india news	200
Indian economy	200
Laptop distribution scheme	200
Karnataka election results	200

The above events are searched in the YouTube and gathered all the live videos along with its title. This live dataset is useful to test the level of effectiveness in the categorization process.

The outcome of the result consists of list of events in which the hierarchy based on the number of occurrence occurred in the final event result. The higher occurrence event is considered as a Level 1 hierarchy and level 2 hierarchy i.e sub event will be the lower number of occurrence.

In the proposed work, low level features are not included due to the poor classification result in the intra class event categorization process. The derived conclusion is represented as below:

Table 3. Derived conclusion

Parameters	Subject based classification	Object based classification
Events	Yes	Yes
Intraclass events	**No**	Yes

Yes – low level features can be used
No- Low level features cannot be used

The above table shows the clear cut picture of the usage of low level features for classification. In order to prove the above conclusion, the sample proof is given below as:

Event name= Arsenal under 18

Event name = arsenal

Figure 2: Contradiction in classification for intra class event classification using SIFT

The above diagrams shows that the both features are merely same to each other. Features are identified by using SIFT (Scalable Invariant Feature Transform) technique which is most preferred technique used in the low level feature extraction. Thus, when the events are finding out by basing on the subjects, intra event classification contradiction will occur and lead to the poor classification if the low level features are used for classification.

5. CONCLUSION

The fast growth of the volume of videos in the internet becomes exponential which needs an urgent call to research in the video categorization to find out the effective major events. Web video categorization becomes more challenging by considering the unconstrained videos with the situation of not using low level features during the intra class events makes it more complicated. In the proposed work, three eyes namely POS Tagging, NER and WordNet enhance process and detect the useful events from the titles and also the novel conclusion is derived that low level feature is not useful for classifying the intra class events based on the subject.

6. REFERENCES

[1]http://www.youtube.com/yt/press/statistics.html

[2]http://nlp.stanford.edu/software/corenlp.shtml

[3]Collins English Dictionary, entry for "lemmatise"

[4]L. Ratinov and D. Roth, Design Challenges and Misconceptions in Named Entity Recognition. CoNLL (2009)

[5]G. A. Miller.Wordnet: A lexical database for english. (11):39-41.

[6]Chengde Zhang, Xiao Wu, Mei-Ling Shyu and QiangPeng, " Adaptive Association Rule Mining for Web Video Event Classification", 2013 IEEE 14th International Conference on Information Reuse and Integration (IRI), page 618-625.

[7] Y. Song, M. Zhao, J. Yagnik, and X. Wu.Taxonomic classification for web-based videos.In CVPR, 2010.

[8] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtube-cat: Learning to categorize wild web videos. In CVPR, 2010.

[9] http://www.ranks.nl/resources/stopwords.html

[10]http://cs.nyu.edu/grishman/jet/guide/PennPOS.html

[11]Roth and D. Zelenko, Part of Speech Tagging Using a Network of Linear Separators. Coling-Acl, The 17th International Conference on Computational Linguistics (1998) pp. 1136—1142

[12]O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Au- tomatic annotation of human actions in video. In Proc. of ICCV, 2009.

[13]Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld.Learning realistic human actions from movies. In Proc. of CVPR, 2008

[14]M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy automatic naming of characters in tv video. In Proc. of BMVC, 2006.

[15]F. Smeaton, P. Over, and W. Kraaij.Evaluation campaigns and trecvid. In Proc. of ACM Workshop on Multimedia In- formation Retrieval, 2006

[16]J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In Proc. of ACM MM, 2007.

[17] M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao. Au- diovisual celebrity recognition in unconstrained web videos. In Proc. of ICASSP, 2009.

[18] J. Liu, J. Luo, and M. Shah.Recognizing realistic actions from videos.In Proc. of CVPR, 2009.

[19] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569-571, Nov. 1999