IJCTA, 9(26) 2016, pp. 137-143
© International Science Press

A Survey on Web Video Classification Based on Descriptor

R.Amsaveni[1] and R. Nedunchezhian[2]

ABSTRACT

Videos are a highly efficient and expressive media capable of capturing and presenting information. Every day, huge numbers of videos are uploaded online. In the field of computers, recognition of actions and scenes in videos based on web is one among the most critical problems. For the purpose of solving this issue in the identification and classification of videos, and the computing of the descriptors for videos is a significant job. It comprises of extraction of characteristics that represents the essential information present in the videos. This work yields a review of the research conducted recently in video analysis, inclusive of descriptor computation and recognition along with the classification of videos. Finally, an outline on the future scope on descriptor based video classification is also presented.

Keywords: Computer Vision, Recognizing Actions and Scenes, Descriptors, Classification of Videos.

1. INTRODUCTION

Nowadays with the technical advancements and the rapid increase in video capturing devices has led to video information grow up manifold. With the enormous progress made in multimedia communication, sharing and authoring techniques, multimedia, particularly video, services are going on to become highly speedy on the Web [1-2]. This, in turn, needs a rapid and accurate solution to be developed in the area of classification and action reorganization of web based videos. This is how videos classification comes into picture for different researchers [3]. Feature extraction and the descriptors computation are critical tasks relating to action recognition and videos classification [4-6].

The procedure of naming actions, generally as an action verb, applying sensory observations is referred to as action recognition [7, 8]. Action is a four-dimensional object that may be decomposed further into spatial and temporal components.. In order to attain that objective, the different approaches generally in this work chiefly are focused on a combination having vision and machine learning techniques [9]. Various kinds of classification have been studied in literature, a hierarchy that is used by Moeslund et al. [10]. Vision techniques try extracting action discriminative features/descriptors obtained from the video sequences, when also rendering suitable robustness towards cues that are distracting. Machine learning approaches attempts to learn mathematical models from those descriptors, and classify new descriptors based on the learned models [11]. For the cause of video classification and action recognition process, desirable descriptors are extracted at first and subsequently the video class is decided based on these descriptors. Because a video mostly consists of a sequence of frames, all the descriptors that are extractable from its frames can also be retrieved for accurate video classification and action recognition process.

In this Paper, the present day developments are reviewed and the open directions in a futuristic perspective for in descriptor based video classification and action recognition are analyzed. Initially the video descriptors

[1] Assistant Professor, Department of Information Technology, PSGR Krishnammal College for Women, Peelamedu, Coimbatore – 641004, Tamilnadu, India

[2] Director-Research and Vice Principal, KIT-Kalaignarkarunanidhi Institute of Technology, Kannampalayam,Coimbatore- 641 402, Tamilnadu, India

available for video classification and action recognition are studied, next the available video classification methods are discussed in detail and at last the inference obtained from the available work.

2. LITERATURE SURVEY

Multiple approaches have been introduced for the actions recognition and classification in different real-world videos, which needs the description and maintenance of videos in the order of billions. It is essential to define a video scene as compact as possible and to design an effective video classification process. For this purpose, the evaluation of the descriptors is conducted in the scene matching and recognition of the same scene or object that is viewed under diverse viewing conditions. In this work about a number of descriptors have been discussed that have previously exhibited a better performance and also the proposed descriptor for video classification and recognition process has shown good performance result in comparison to the existing approach in the section that follows.

2.1. Descriptors

In this section, the different kinds of descriptors utilized in literature for the purpose of video classification are presented. The descriptors are namely SIFT [12], Gradient Location and Orientation Histogram (GLOH), Shape Context [13], PCA-SIFT [14], GLOH is a new descriptor that is an extension of SIFT by varying the location grid and making use of PCA to minimize the size, and GIST descriptor [15].

SIFT descriptors

Lowe [12] rendered the code that are calculated for normalized scene patches with this code obtained from the videos in SIFT descriptors. It is a 3D histogram of gradient location and orientation, where the location is quantized to form a 4 × 4 location grid and the gradient angle is then quantized into eight different orientations. The resultant descriptor is of the dimension 128. Every orientation plane specifies the gradient magnitude related to an orientation given. In order to get the illumination invariance, the descriptor is then normalized by getting the square root of the sum of components that are squared.

Gradient location-orientation histogram (GLOH)

GLOH is an extension of the SIFT descriptor modeled to maximize its reliability and uniqueness. The SIFT descriptor computes for a log-polar location grid along with three bins in radial direction (the radius fixed to 6, 11, and 15) and 8 in angular direction, which gives result to 17 location bins. Especially the central bin is not segregated in angular directions. The gradient orientations are then quantized in 16 bins. This yields a 272 bin histogram. The size of this descriptor is later reduced with the help of PCA. The covariance matrix for PCA is further estimated on 47,000 scene patches that are gathered from different videos. The 128 largest eigenvectors are then employed for the purpose of description.

Shape context

Shape context is identical to the SIFT descriptor, though it is on the basis of the edges of the scenes. Shape context is usually a 3D histogram of edge point locations in addition to the orientations. Canny [16] detector is utilized for extracting the Edges from the scenes. Location is then quantized into nine bins of a log-polar coordinate system with the radius fixed at 6, 11, and 15 and then the orientation quantized into four bins (horizontal, vertical, and two diagonals). This way a 36 dimensional descriptor is obtained. Here, the research work used weight as a point contribution to the histogram along with the gradient magnitude. This has been indicated to yield better results compared to making use of the same weight for all the edge points, as said in [13]. Specifically, the actual shape context was calculated only not for orientations but only for edge point location.

PCA-SIFT descriptor

This descriptor is originally a vector of scene gradients in *x* and *y* direction that are computed within the support region. Then the sampling of the gradient region is done at 39 x 39 locations; hence, the vector has the dimension 3,042. The dimension is later reduced to 36 with the aid of PCA.

GIST

The GIST descriptor was first introduced in [17]. The concept is about developing a low dimensional representation of the scene, which does not need any kind of segmentation. The authors proposed a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that are indicative of the dominant spatial structure of a scene. It is shown that these dimensions may be estimated with reliability employing spectral and coarsely localized information. The scene is then divided into a 4×4 grid for which case the orientation histograms are extracted. It is to be noted that the descriptor is just as the same in spirit as the local SIFT descriptor [12].

Carneiro and Jepson [18] assessed the performance of point descriptors making use of ROC (Receiver Operating Characteristics). They exhibit that their phase-based descriptor outperforms other differential invariants. In the comparison made, detection of interest points is done by applying the Harris detector and the image transformations are then produced by artificial means. Ke and Sukthankar [19] have designed a descriptor that is identical to the SIFT descriptor. It applies the Principal Components Analysis (PCA) over the normalized image gradient patch and performs well in comparison to the SIFT descriptor on data generated artificially in recent times. In Li *et al.* [20] GIST is useful for retrieving an initial set of images of the same landmarks, for instance the statue of liberty, and afterwards, image point based matching is utilized for refining the results and for constructing a 3D model of the landmark. In Hayes and Efros [21] it is employed for completion of image. With a huge database of photographs collected from the web the algorithm patches up the holes observed in images by identifying same kind of image regions in the database on the basis of the GIST descriptor. Torralba *et al.* [22, 23] created multiple techniques for the compression of the GIST descriptor.

2.2. Video Classification Methods

Video classification is a significant means of improving the video retrieval efficiency. The task of Video classification [24], [25] is finding rules or information from videos making use of extracted features or mined results and then segment the videos into preset categories. In this paper, video genre classification techniques are reviewed in detail.

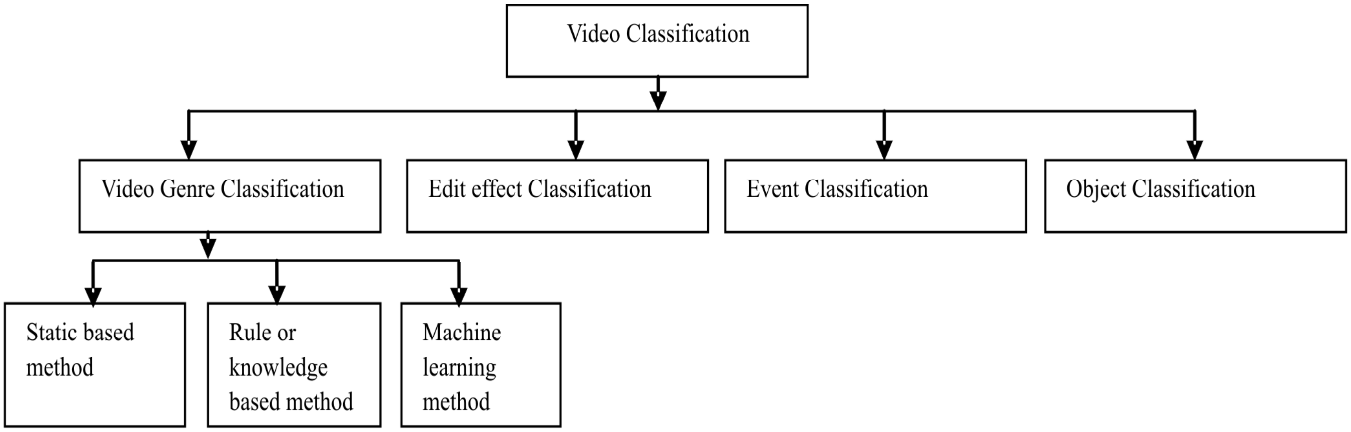

Figure 1: Different Video Classification Approaches

Video genre classification

Video genre classification is a technique of classifying videos into different kinds of genres like “movie,” “news,” “sports,” and “cartoon”. Strategies for classifying video genres can be categorized into statistic-based, rule- or knowledge based, and machine learning-based [26].

a) **Statistic-based Method:** This approach groups the videos by modeling diverse video genres statistically.

b) **Rule- or knowledge-based approach:** This approach uses heuristic rules from domain knowledge to low-level features for the classification of videos.

c) **Machine learning-based approach:** This approach utilizes labeled samples along with low-level features in order to train a classifier or a set of classifiers for grouping videos.

Author and Year	*Method*	*Classification Process*
Fisher *et al.* (1995) [27]	Statistic-based Method	It is used for classifying videos from television shows such as news, car race, tennis, animated cartoon, and commercials in this work. Initially, video syntactic properties including color statistics, cuts, camera motion, and object motion are examined. Thereafter the properties mentioned are employed for deriving more abstract film style attributes that includes camera panning and zooming, speech, and music. Finally, these style attributes that are detected, are mapped onto film genres.
Rasheed *et al.* (2005) [28]		Here, in this work, Rasheed et al employed statistic based technique for classifying films into comedies, actions, dramas, or horror films. According to the characteristics of films, this author thoug/ht of using only four visual features, such as average shot length, color variance, motion content, and lighting key. The classification is accomplished applying mean shift clustering approach.
Roach *et al.* (2001, a)[29]		This work introduces a cartoon video classification technique that makes use of motion features of foreground objects in order to differentiate between cartoons and non-cartoons.
Roach *et al.* (2001, b), [30]		Here, in this work, the classification of videos is based on the dynamic content present in short video sequences, where the foreground object motion and background camera motion are acquired from videos. The classified videos are inclusive of sports, cartoons, and news.
Chen and Wong (2001) [31]	Rule- or knowledge-based approach	The goal of this work is to design a knowledge-based video classification method, the related knowledge is coded as generative rules with confidences to create a rule-base. The Clip language is utilized to make a compilation employing the rule base.
Snoek *et al.* (2006) [33]		This work introduced a video classification and indexing technique, having a combination of video creation knowledge for the extraction of semantic ideas from videos by going through various paths through three sequential analysis steps namely the multimodal video content analysis step, the video style analysis step, and the context analysis step.
Zhou *et al.* (2000)[34]		This work presented a rule-based video classification technique that employs analysis of video content, feature extraction and clustering techniques for performing the semantic clustering of videos. Reports on experiments on basketball videos are also provided.

Zhou *et al.* (2002) [32]		This work proposed a Rule- or knowledge-based approach, which is a supervised rule-based video classification system, in which the higher semantics are extracted from using low-level features jointly along with classification rules which are obtained through a supervised learning process.
Fan *et al.* (2004) [37]	Machine learning-based approach	This method employs multiple degrees of concepts of video contents in order to accomplish hierarchical semantic classification of videos to facilitate access to video contents with good efficiency.
Mittal and Cheong (2004) [35]		In this newly introduced Machine learning-based approach, the Bayesian network is used for the classification of videos. The association existing between a continuous and nonparametric descriptor space and the classes is learned here and the minimum Bayes error classifier is then inferred.
Qi *et al.* (2006) [36]		This work made use of a video classification framework applying SVMs-based active learning. The outcomes of clustering every the videos in the dataset are provided as the input to the framework. The accuracy of the classifiers is then improved on a gradual scale during the active-learning process.
Truong *et al.* (2000) [38]		This method classifies the videos to belong to the genres of cartoons, commercials, music, news, and sports. The features that are utilized comprise of the average shot length, the percentage of every type of transition, etc. The C4.5 decision tree is helpful for building the classifier for the purpose of genre labeling.
Wu *et al.* (2004) [40]		In this work an online video semantic classification framework is proposed, where the local and global sets of optimized classification models are trained online by making the best use of local and global statistic characteristics of videos.
Yuan *et al.* (2006) [39]		The author introduced an automatic video genre classification mechanism which is based on a hierarchical ontology of genres of video. A group of SVM classifiers that are united in the form of a binary-tree allocate each video to its respective genre.

In accordance with the video genres classification approaches studied, few conclusions are discussed as given below [26].

1) These classification approaches can be utilize on stationary features, dynamic features and combination of both.
2) The approaches discussed above were use global statistical features. Such features are robust towards video diversity, rendering them suitable for video genre classification. Several algorithms try to add few semantic features based on these low-level features.
3) Previous domain knowledge is extensively applied in video genres classification. The usage of knowledge or rules can enhance the classification efficiency for few special domains, though the respective algorithms cannot be generalized for videos from other domains.

The abstraction of literature review states that in the video classification techniques, machine-learning approaches based on video genre classification yield better detection and classification of video and they either utilize only stationary features/descriptors, or only dynamic features, or a combination of both of them. Chiefly the machine learning method proposed along with GIST descriptor renders very good accuracy in the detection and classification of video. This approach preferentially employs global statistical low-level features/descriptors. Hence, such features/descriptors are reliable in terms of video diversity, yielding them suitable for video genre classification applying the machine learning method.

3. CONCLUSION AND FUTURE WORK

This paper is primarily focus on reviewing the video genre classification techniques under the broad coverage of video classification. Several researchers have demonstrated investigations to manage these problems related to action and scene detection through an extraction of descriptors and then the descriptor computation approach, which will make it possible for the classification the videos and it may also assist in real world video online for different purposes. Machine learning classifiers with GIST descriptors have been a desired topic for research for several years for web video classification with promising accuracies. However, these efficient classifiers also have their intrinsic setbacks and disadvantages. As said that, combination of the machine learning classifiers approaches will generally render superior performances rather than by utilizing them individually. Hence, these classification approaches will be advised to manage complicated issues in scene and action recognition and classification of videos.

REFERENCES

[1] Bughin, J., Corb, L., Manyika, J., Nottebohm, O., Chui, M., de Muller Barbat, B., & Said, R.,"The impact of Internet technologies: Search", *McKinsey & Company, High Tech Practice*, 2011.

[2] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H.,"Big data: The next frontier for innovation, competition, and productivity", 2011.

[3] Weinland, D., Ronfard, R., & Boyer, E .,"A survey of vision-based methods for action representation, segmentation and recognition", *Computer Vision and Image Understanding*, **115(2)**, 224-241, 2011.

[4] Solmaz, B., Assari, S. M., & Shah, M. .," Classifying web videos using a global video descriptor", *Machine vision and applications*, **24(7)**, 1473-1485, 2013.

[5] Kantorov, V., &Laptev. I. "Efficient feature extraction, encoding, and classification for action recognition In Computer Vision and Pattern Recognition (CVPR)", *IEEE Conference on* (pp. 2593-2600), IEEE. 2014.

[6] Hadid A..”Image and video descriptors In Image Processing Theory Tools and Applications (IPTA)", *2nd International Conference on* (pp. 11-12), IEEE. July 2010.

[7] Krüger, V.,Kragic, D., Ude, A., &Geib, C.,"The meaning of action: a review on action recognition and mapping", *Advanced Robotics*, **21(13)**, 1473-1501, 2007.

[8] Soomro, K., &Zamir, A. R.,"Action Recognition in Realistic Sports Videos. In Computer Vision in Sports", *Springer International Publishing*,(pp. 181-208), 2014.

[9] Gosselin, P. H., & Picard D.,"Machine learning and content-based multimedia retrieval", *In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* , 251-260, April 2013.

[10] Thomas B. Moeslund, Adrian Hilton, Volker Kruger.," A survey of advances in vision-based human motion capture and analysis", *Computer Vision and Image Understanding (CVIU)***104 (2–3)** ,90–126., 2006.

[11] Zabulis, X., Baltzakis, H., &ArgyrosA.,"Vision-based hand gesture recognition for human-computer iversal Access Handbook". LEA, **34-1,** 2009.

[12] D. Lowe., "Distinctive Image Features from Scale-Invariant Keypoints" *Int'l J. Computer Vision*, vol.2, no.60, pp. 91-110, 2004.

[13] S. Belongie, J. Malik, and J. Puzicha., "Shape Matching and Object Recognition Using Shape Contexts", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, no. 4, pp. 509-522, Apr. 2002.

[14] Y. Ke and R. Sukthankar., "PCA-SIFT: "A More Distinctive Representation for Local Image Descriptors", *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 511-517, 2004.

[15] A. Oliva and A. Torralba.,"Modeling the shape of the scene: a holistic representation of the spatial envelope", *IJCV*, **42(3)**:145–175, 2001.

[16] J. Canny., "A Computational Approach to Edge Detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, 1986.

[17] A.Oliva and A. Torralba., "Modeling the shape of the scene: a holistic representation of the spatial envelope", *IJCV*, **42(3)**: 145–175, 2001.

[18] G. Carneiro and A.D. Jepson., “Phase-Based Local Features”, *Proc. Seventh European Conf. Computer Vision*, 282-296, 2002.

[19] Y. Ke and R. Sukthankar., “PCA-SIFT: A More Distinctive Representation for Local Image Descriptors” , *Proc. Conf. Computer Vision and Pattern Recognition*, 511-517, 2004.

[20] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm.,”Modeling and recognition of landmark image collections using iconic scene graphs”, *In ECCV*, October 2008.

[21] J. Hayes and A. Efros.,”Scene completion using millions of photographs”, *In SIGGRAPH*, 2007.

[22] A. Torralba, R. Fergus, and Y.Weiss.,” Small codes and large databases for recognition”, *In CVPR*, 2008.

[23] Y. Weiss, A. Torralba, and R. Fergus.,”Spectral hashing In Advances in Neural Information Processing Systems”, 2009.

[24] D. Brezeale and D. J. Cook., “Automatic video classification: A survey of the literature” *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, **38(3)**, 416–430, May 2008.

[25] M. Roach, J. Mason, L.-Q. Xu, and F. Stentiford., “Recent trends in video analysis: A taxonomy of video classification problems” *In Proc. Int. Assoc. Sci. Technol. Develop. Int. Conf. Internet Multimedia Syst. Appl.*, Honolulu, HI, 348–354, Aug. 2002

[26] Y. Yuan., “Research on video classification and retrieval”, *Ph.D. dissertation, School Electron. Inf. Eng., Xi’an Jiaotong Univ., Xi’an, China*, 5–27, 2003.

[27] S. Fischer, R. Lienhart, and W. Effelsberg., “Automatic recognition of film genres” , *Proc. ACM Int. Conf. Multimedia*, , 367–368 , 1995

[28] Z. Rasheed, Y. Sheikh, and M. Shah., “On the use of computable features for film classification” , *IEEE Trans. Circuits Syst. Video Technol.*, **15(1)**, 52–64, Jan. 2005.

[29] M. J. Roach, J. S. D. Mason, and M. Pawlewski., “Motion-based classification of cartoons”, *Proc. Int. Symp. Intell. Multimedia*, 146–149, 2001

[30] M. J. Roach, J. D. Mason, and M. Pawlewski., “Video genre classification using dynamics”, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, **3**, 1557–1560,2001

[31] Y. Chen and E. K.Wong., “A knowledge-based approach to video content classification” , *Proc. SPIE Vol. 4315: Storage and Retrieval for Media Databases*, 292–300, Jan. 2001

[32] W. S. Zhou, S. Dao, and C. C. J. Kuo., “On-line knowledge- and rule based video classification system for video indexing and dissemination”, *Inform. Syst.*, **27(8)**, 559–586, Dec. 2002.

[33] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders., “The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing”, *IEEE Trans. Pattern Anal. Mach. Intell.*, **28(10)**, 1678–1689, Oct. 2006.

[34] W. S. Zhou, A. Vellaikal, and C.-C. J. Kuo., “Rule-based video classification system for basketball video indexing” , *Proc. ACM Workshops Multimedia*, 213–216, 2000.

[35] A. Mittal and L. F. Cheong., “Addressing the problems of Bayesian network classification of video using high dimensional features”, *IEEE Trans. Knowl. Data Eng.*, 16(2) 230–244, Feb. 2004.

[36] G.-J. Qi, Y. Song, X.-S. Hua, H.-J. Zhang, and L.-R. Dai., “Video annotation by active learning and cluster tuning”, *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshop*, 114–121, Jun. 2006

[37] J. P. Fan, A. K. Elmagarmid, X. Q. Zhu, W. G. Aref, and L.D. Wu., “ClassView: Hierarchical video shot classification, indexing and accessing”, *IEEE Trans. Multimedia*, **6(1)**, 70–86, Feb. 2004.

[38] B. T. Truong, C. Dorai, and S.Venkatesh., “Automatic genre identification for content-based video categorization”, *Proc. IEEE Int. Conf. PatternRecog.*, vol. 4, Barcelona, Spain, 230–233, 2000.

[39] X. Yuan, W. Lai, T. Mei, X.-S. Hua, and X.-Q. Wu., “Automatic video genre categorization using hierarchical SVM”, *Proc. IEEE Int. Conf. Image Process., Atlanta, GA*, 2905–2908, Oct. 2006.

[40] J. Wu, X.-S. Hua, and H.-J. Zhang., “An online-optimized incremental learning framework for video semantic classification”, *Proc. ACM Int. Conf. Multimedia, New York*, Oct. 2004, 320–323.