International Journal of Engineering & Technology, 7 (2.31) (2018) 89-93

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A survey on video classification using action recognition

Caleb Andrew[1*], Rex Fiona[2]

[1]Post Graduate Student, Department of Computer Sciences and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India.
[2]Assistant Professor, Department of Computer Sciences and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India.
**Corresponding author E-mail:calebandrew192@gmail.com*

Abstract

The growth in multimedia technology have resulted in producing a variety of videos every day. These videos should be classified in order to help people identify the correct video which they search for when needed. The video classification problem can be said as a probabilistic data classification problem which falls as a subcategory of the machine learning technique. Classification helps in indexing, analyzing, searching etc. A survey has been made on the present technologies that are used for video classification. Various techniques used for video classification such as Multiple Instance Learning (MIL), Conditional Random Field (CRFs) and classifying based on the action and gesture are studied.

Keywords: *Video classification, machine learning, multiple instance learning (MIL), conditional random field (CRFs), action recognition, gesture recognition.*

1. I. Introduction

In recent years the growth in portable multimedia devices has resulted in producing a multitude of videos in all genres every day [10]. These videos must be classified for easy accessing. The video classification method helps in indexing, analyzing, searching, etc. the videos based on categories. There are various methods available for classifying the videos. The existing traditional video classification method using MIL or CRFs considered as a single data instance (the entire video) from which the visual features is extracted and quantized using K-means and then by the concept of average pooling the quantized visual features is pooled to form a single featured vector. Further the feature vector is being used as inputs for a classifier and it is classified. Though these methods are efficient it has few drawbacks like bringing the noisy details from the background and sometimes fetching the non-related video frames. At times the key indicator of the action or an event is ignored as the temporal correlation between video frames is ignored both in training and testing. The present technique which combines the Multiple Instance Learning (MIL) and Conditional Random Field (CRFs) by a machine learning algorithm [12]. In which the whole video clip is taken as a bag and each segment is considered as an individual instance. The local patterns of the various video categories will be explored by the MIL and the CRF model exploits the intrinsic temporal dependencies between every instance. A novel conditional likelihood formulation is designed and that requires just the annotation on the video clips which is used during the training stage. During the testing stage the videos are classified by the learned CRF model. This algorithm performs well on synthetic data and realistic videos for action and gesture classification and also results in better performance.

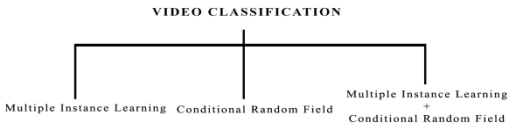

Fig. 1: Video classification methods

2. Video classifications

The video classification method helps in categorizing the videos and helps in indexing, analyzing, searching, etc. [10]. Video classification helps in classifying videos that are being produced everyday by various multimedia devices. The concept of video classification is a vast topic and it can also be termed as video interpreting. The term video represents a sequence of images that has both sound and images (sequence of images). In addition to the spatial features videos have an added property of temporal features providing more information and requiring larger computational models [18]. Classifying videos is one of the current challenges in machine learning. Machine learning is one among the present trending technique under artificial intelligence that automates analytical model by using algorithms that are iteratively learnt from the data [12]. The accuracy of the classification can be measured for the video-level and frame-level.

Action recognition

In computer vision, the term action recognition refers to the process of analysing an action that is present in a video sequence this helps in identifying the actions of interest in time and space. Videos have the information that can assist in analysing the actions that has been captured. Action recognition starts with the extraction of the features and then analysed to ensure whether the features can assist the task of action recognition then it is encoded. This encoded features is used as an input to classifier for

identifying the action class and to find the temporal and spatial locations.

Gesture recognition

Gesture recognition is the process of interpreting a human motion mathematically by the use of computer devices. Gestures originate from human body mostly from face and hands. Gestures can be used to control and interact with devices without the use of physical touch and also helps in human to communicate with the machine. Video classification using gestures can be challenging because identifying the right gesture or the accuracy in identifying the correct gesture when using hands is highly confusing.

Fig. 2: Taxonomy of action and gesture recognition

3. Video classification models

A survey has been made on the available algorithms that are used for the video classification method this helps in learning which among the all available technique is more accurate in predicting the action or gesture and can be used for classifying action class videos.

Multiple instance learning

Multiple instance learning falls under the supervised learning framework of the machine learning technique, every training instance has a label either real values or discrete. In a MIL the bags are named positive or negative based on the instance present within.

Thomas G. Dietterich et al. stated that the multiple instance problem begins when the training samples are not clear. A single object might contain other feature vector that describes it but amongst all only one of those feature vectors might be incharge of the observed classification of object [4]. Their proposal compares three different kinds of algorithms that studies axis-parallel rectangles to solve multiple instance problem. Algorithms that rejects the multiple instance problem performs poorly but the algorithm that confronts the multiple instance problem performs good and yields 89% accurate prediction on a dataset. In machine learning, problems occurs only when the learning system has partial knowledge about the training samples.

Annabella Astorino et al. deviced a model of Multiple Instance Learning (MIL) approach for classifying images. The model mainly emphasizes on a recent MIL method for binary classification in which the task is to differentiate between the positive and negative set of points [3]. These sets are termed as bags and the points within the bags are known as instances. In times of different classes of instances, a bag is termed positive if it has at least one positive instance and it is called negative if it has only the negative instances. For solving this problem there is two different approaches available: the bag-level approach and the instance level approach. During the starting stage the total entity of every bag is considered, in the latter stage a classifier is formed on the basis of the details of the instances, without checking the entire entity of every bag. The devised method uses the Lagrangian relaxation technique to a Support Vector Machine (SVM) model [3]. Experiments are carried out on the code MIL-RL by detaching, based on the MIL paradigm. The code has succeeded in classifying 110 pairs of images correctly. But this method is not tested for classification in large datasets.

Yu Zhou et al. presented a novel Semi-Supervised Multiple Instance Learning (Semi-MIL) approach. This approach uses a different mode of "bag of instances". The non-labelled data in the multiple instance learning problem can also be used because of this method. The method is formulated based on the Minimax kernel's graph model [6]. This algorithm also helps in visual tracking. Multiple Instance Learning (MIL) is a special studying that deals with the ambiguity of instance labels and used for the drug activity prediction issue. The training will contain bags and the bags has instances. Semi-supervised learning methods can be used for both the labelled and unlabeled data for achieving better classification accuracy even though other models stick to supervised learning [6]. The combination of semi-supervised learning and multiple instance learning called the Semi-Supervised Multiple Instance Learning (Semi-MIL) is developed, which uses the unlabeled bags in multiple instance classification problem. Based on the Semi-MIL algorithm, an effective tracking system is developed and tested on some videos which fetched better results.

Conditional random field

John Lafferty et al. presented conditional random field for creating probabilistic models to label the data. It is a sequence model which has the advantages of maximum entropy Markov models (MEMMs) and stochastic grammars that are used earlier for labeling the data [7]. Conditional random field also solves the label biasing problem in a proper way. The main difference between a CRFs and MEMMs is the CRFs undirected and the MEMMs directed graphical model. CRFs have a single exponential method for the group probability of a label sequence. Normalization is done globally not like other model which used to do it for each state individually [7]. CRFs assign proper probability distribution over possible labeling and it also normalize easily to analogues of stochastic context free grammars that is used for the natural language processing and prediction.

Abdallah Zeggada et al. In order to study the simultaneously spatial contextual information and cross-correlation between labels the multi labeling classification method of unmanned aerial vehicle (UAV) imagery within a conditional random field (CRF) model is developed [2]. This methodology consists of two main phases. First, the selected input UAV image is divided down into a grid of tiles and is processed. In the second phase, using a multi label CRF model the spatial correlation between consecutive tiles and the correlation between labels of the same tile are related. Unmanned aerial vehicles (UAVs) have several pros, extremely high resolution (EHR) is captured and the multi label classification can be used for analyzing these images. Structured random fields exhibit techniques used specially in EHR images, in which a single object may be consisting of thousands of pixels which is used in combining spatially neighboring information in the classification model [2]. The multi label CRF framework for EHR UAV images is used at a tile level under a CRF perspective. The main advantage is that the proposed CRF combines the spatial information within the class, and the cross-correlation information between different class labels.

Thomas Deselaers et al. proposed a conditional random field model for MIL. Multiple instances – conditional random field model considers bags as nodes and the instances as states. The model combines critical unary instance classifiers and pairwise non-similarity contents, these forces are joined together in an ordered manner by learning the parameters of the CRFs using the constraint generation [8]. Both forces improve the performance of classification. The MI-CRF consider all the bags together during the training as well as the testing stages, this helps in classifying the bags in a setup that classifies multiple test bags together and can possess other MIL algorithm as unary potentials helping in improving the performance.

4. Comparative study

This section gives a survey on various available techniques of video classification like using action and gesture recognition by Multiple Instance Learning (MIL) and Conditional Random Field (CRFs) algorithms.

Video classification

Chen Sun et al. emphasized the need for event recognition in computer vision research owing to its multifaceted applications. However, an enormous amount of work is performed on footage from a fixed camera, used environment and events. This method focuses on classification of unregulated web videos[15]. From the local feature descriptors the feature vectors of fixed lengths is built and SVM is used to classify. The important contribution is the study of the usage of Fisher Vector representation for improved results in comparison with the Bag-of-Words (BoW) method. This was useful only for the still image classification but not for moving image categorization in the past. Tests have been conducted on the NIST TRECVID Multimedia Event Detection (MED) dataset, which contains many hours of unregulated videos created by various user.

Andrej Karpathy et al. studied the performance of Convolutional Neural Network (CNN) in large scale video classification, in which the networks have access to complex temporal evolution in addition to appearance information present in static images [17]. since videos are comparatively tough to collect, interpret and store so practically there are no video classification benchmarks at present which match the scale and variety of actual image datasets. A new Sports-1M dataset consists of one million videos comprising of a diverse taxonomy of 487 categories of sport. CNNs require colossal periods of training time to effectively optimize the millions of parameters. Time complexity is further compounded because the network must process multiple frames of video at a time. An impressive approach to solve the above problems and boost the run time performance of CNNs is to rework the architecture to enclose two different streams of processing [17]. This shows various methods for supporting the connectivity of the CNN in time domain to take benefit of local spatio-temporal information and suggest a multiresolution architecture as an assured method to speed up the training. The generalization performance of the best model is studied by retraining the top layers on the UCF-101 Action Recognition dataset and significant performance improvement are observed compared to the UCF-101 baseline model.

Zuxuan Wu et al. stated that videos have a rich semantics and are mostly multimodal. The competitive task of classifying videos based on their high-level semantics like human actions. Various efforts have been taken to understand this problem, most available methods combine multiple features using simple fusion strategies and avoided the study of inter class semantic relationship [18]. Zuxuan Wu et al. proposed a unique framework that collectively understands feature relationships and maneuver the class relationship that improves the performance of video classification. These two relationships are learned and utilized in deep neural network (DNN). DNN can be efficiently initiated using a GPU implementation within a manageable cost [18]. By preparing the DNN with improved capability of analyzing both inter-feature and inter-class relationship, the presented consistent DNN is more applicable for identifying video semantics. Tests are conducted on the well-known Hollywood2 and Columbia Consumer video benchmarks exhibited superior performance.

Jingjing Liu et al. proposed a combination model that clubbed the Multiple Instance Learning (MIL) and Conditional Random Field (CRFs). The use of various multimedia devices has rapidly increased. The traditional method for classifying these videos take the entire video as a single data instance and the visual features are extracted and pooled together by K-means and then it is given as an input for the classifier and it is classified resulted in some drawbacks like fetching the noisy details from the background and non-relevant contents are fetched for classifying [12]. The proposed machine learning algorithm overcomes the drawbacks by taking the videos as bags and its segments as instances. The local patterns of the videos are explored by MIL and the temporal relation between instances is explored using CRF. At the training stage a conditional likelihood is formulated which requires just the annotation of videos and during the testing stage the videos are classified by the learned CRF model [12].

Video classification is one of the most needed task and it features variety of applications. Both static and motion information are contained in a video represented either by frames or optical flow. There is a probability of people encountering 82% video traffic by 2021. As a precaution for this problem there is a rapid need for video classification. Deep networks are being used for capturing the static and motion information, doing so have two major drawbacks [10]. First, the relationship between spatial and temporal attention is ignored. Second, the tight complementarity between static and motion information is ignored. The two-stream collaborative learning with spatial-temporal attention (TCLSTA) model proposed by Yuxin Peng et al. have overcome the above two imitations. The two models: (1) Spatial-temporal attention model: The salient regions in a frame is highlighted by the spatial level attention, and the discriminative frames of a video is studied by the temporal-level attention and both are developed to study the discriminative static and motion features together for good classification performance. (2) Static-motion collaborative model: In addition to the Spatial-temporal attention models work the model also adjust itself to learn the fusion weights of static and motion streams for a better classification [10]. This model achieves better results compared to the traditional model.

Nataliya Shapovalova et al. proposed an extension of the popular latent SVM. This new framework is applied to the action classification method. When a video is given as an input the system on its own identifies what action is taking place, this uses binary classification method which is very useful for many applications [19]. It is also required to know that where does the action occurs or what leads to the. The spatio-temporal position of the video is treated as latent variable. Building models that analyze such evidence requires hand labeling of regions of interest on training videos, but this is a time killing error-prone process and also using it in a large dataset will be expensive and further labeling where an action takes place is a nonrealistic task. To resolve this the authors main contribution is the development of the similarity constrained latent SVM. It adds the potential to motivate consistency of latent variables over all of the training data, accepting pairwise similarities of latent labels in a fashion related to Transductive SVM for semi supervised learning.

Gesture recognition

Hervé Jégou et al. address the problem of image searching in a very large scale, searching for the most relevant images in a large image repository. The author stress on the relative optimization of three constraints: the accuracy in search, the efficiency in search and the memory usage [16]. The last two can be related as the search efficiency can be measured by the amount of memory to be visited. Other models proposed an approx. closest neighbor search of BOF vectors. This model is cut short for small vocabulary sizes producing lower accuracy in searching compared to the others. A still image might require hundred bytes to produce low search accuracy. The efficiency problem was slightly covered by the min-Hash model. In contrast to the problem the proposed approach produces a higher accuracy for a 20-byte representation which is achieved by optimizing the representation.

Maryam Asadi-Aghbolaghi et al. stated that under the field of pattern recognition and computer vision the action and gesture recognition is studied more and improving results have been generated so far. Deep learning concept has achieved better results that outperform “non-deep” state of the art methods [5]. The temporal axis in a footage is makes the action and gesture recognition a competitive task in terms of the data to be tested and model complexity. The author proposes various techniques like frame sub sampling, collection of local frame-level features into mid-level video production or temporal sequence modeling (TSM). At present LSTM are an essential part of deep learning models for image sequence modelling for action and gesture recognition [5]. Along with the implicit method of Spatio-

temporal features using 3D convolutional nets, already computed motion-based feature and the linking of various visuals performs better in the state of the art model. The author also proposes a taxonomy that encapsulate important aspects of deep learning models.

Action recognition

Ivan Laptev et al. proposed a method that addresses the identification of human actions in distinct and real video settings. Due to the non-availability of real videos in the past this important and challenging subject has been ignored. The author addressed the limitation of the previous models and the usage of movie scripts to automatically define the actions in a video sequence and also the problem of action classification in a video samples [9]. The presented new idea includes local space-time features, space-time pyramids and multichannel non-linear SVMs. The tests made on challenging action classes in movies showed promising results and when tested on a standard KTH action dataset the proposed model resulted in achieving 91.8% accuracy.

Heng Wang et al. concluded that the local space-time features are the best suited representation for action recognition. The previous model introduced space-time interest points by adding Harris detector to videos. Feature descriptors vary from higher order derivatives, gradient information, optical flow and brightness information such as 3D-SIFT, HOG3D and extended SURF [11]. The 2D space domain and the 1D time domain in videos exhibit a different quality. The authors stated that tracking interest points through videos is a direct option. Heng Wang et al propose to sample feature nodes on a dense grid in every node and to track them using a state of the art dense optical flow algorithm. This improves the quality of trajectories over sparse tracking method, like KLT tracker. This model presents a video representation based on dense trajectories and motion boundary descriptors. In addition, a descriptor based on motion boundary histograms (MBH) has been introduced that rely on a differential optical flow. During the test the MBH reveals an improved performance especially on the real-world videos that has more camera motion.

The human-computer interactions, video surveillance and other video applications have increased and the actions involved in them are used for classifying these videos [1]. The use of dense trajectories to identify the human action is the most widely used model but the dense trajectories computing descriptors need to spend a lot of time and the bond between trajectories are ignored every time. So, to bring a solution for these issues Xiang Xiao et al. proposed the trajectories-based motion neighborhood feature (TMNF) model for action recognition. From the real video resolution, the trajectories of the central particular region are selected this helps in avoiding the unwanted background trajectories and the reduction of computation [1]. The orientation and the motion relationship between various trajectories is explored by the TMNF method. The improved vector of the locally aggregated descriptors (IVLAD) is being used for the video representation and the liner SVM algorithm is used for the classification of videos which performs well on several popular datasets.

Heng Wang et al. stated that the dense trajectories achieved state-of-the-art results on various action recognition datasets. To estimate the camera, motion the authors matched feature points between consecutive frames using SURF descriptor and dense optical flow. These matches are later used to sturdily analyze the homograph with RANSAC [13]. Human motion differs from the camera motions in general they generate conflicting matches and to improve this a human detector is assigned to remove these matches. The proposed model shows better improvements on motion-based descriptors like HOF and MBH.

Antonios Oikonomopoulos et al. in their survey addressed the problem of localization and identification of human actions in an unsegmented image sequences. The proposed framework permits us to work with multiple actions that occurs in a same scene. To choose characteristic ensembles per class the boosting technique is used. This directs to a group of class specific codebooks in which each code word is an ensemble of features. During the training stage the spatial positions are stored in relevance to set of reference points and the temporal positions with respect to the start and end of action instance [14]. While testing using the information stored during training each of the activated code words casts votes based on the spatiotemporal positions and extent of action. The author also proposed an extension in time implicit shape model which led to the invention of the spatiotemporal shape model, which helps us to localize both in time and space [14]. The use of class-specific codebooks and spatiotemporal models in a voting framework helps us to pact with the presence of dynamic background and with activities that happens frequently.

5. Conclusion

In this survey paper, various available techniques for video classification models using the action and gesture recognition is discussed. Though there are many traditional approaches for classifying the videos presently there are various new approaches that combines the traditional techniques to achieve better classification accuracy.

References

[1] Xiao X, Hu H & Wang W, "Trajectories-based motion neighbourhood feature for human action recognition", *IEEE International Conference on Image Processing (ICIP)*, (2017), pp.4147-4151.

[2] Zeggada A, Benbraika S, Melgani F & Mokhtari Z, "Multilabel Conditional Random Field Classification for UAV Images", *IEEE Geoscience and Remote Sensing Letters*, (2018), pp.399-403.

[3] Astorino A, Fuduli A, Veltri P & Vocaturo E, "On a recent algorithm for multiple instance learning. Preliminary applications in image classifications", *IEEE International conference on Bioinformatics and Biomedicine (BIBM)*, (2017), pp.1615-1619.

[4] Dietterich TG, Lathrop RH & Lozano-Pérez T, "Solving the multiple instance problem with axis-parallel rectangles", *Artificial. Intelligence*, (1997), pp.31–71.

[5] Asadi-Aghbolaghi M, Clapes A, Bellantonio M, Escalante HJ, Ponce-López V, Baró X, Guyon I, Kasaei S & Escalera S, "A survey on deep learning based approaches for action and gesture recognition in image sequences", *2th IEEE International Conference on Automatic Face & Gesture Recognition*, (2017), pp. 476-483.

[6] Zhou Y & Ming A, "Semi-Supervised Multiple Instance Learning and its application in visual tracking", *8th International Conference on Wireless Communications & Signal Processing*, (2016).

[7] Lafferty JD, McCallum A & Pereira FCN, "Conditional random fields: probabilistic models for segmenting and labelling sequence data", *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, (2001), pp.282–289.

[8] Deselaers T & Ferrari V, "A conditional random field for multiple-instance learning", *Proceedings of the Twenty-Seventh International Conference on Machine Learning (ICML)*, (2010), pp.287–294

[9] Laptev I, Marszalek M, Schmid C & Rozenfeld B, "Learning realistic human actions from movies", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, (2008), pp.1–8.

[10] Peng Y, Zhao Y & Zhang J, "Two-streams Collaborative Learning with Spatial-temporal Attention for Video Classification", *IEEE Transactions on Circuits and Systems for Video Technology*, (2018).

[11] Wang H, Kläser A, Schmid C & Liu CL, "Dense trajectories and motion boundary descriptors for action recognition", *Int. J. Comput. Vis.*, Vol.103, No.1, (2013), pp.60–79.

[12] Liu J & Chen C, "Video classification via weekly supervised sequence modeling", *computer vision and Image understanding*, Vol.152, (2016), pp.79-87.

[13] Wang H & Schmid C, "Action recognition with improved trajectories", *Proceedings of the IEEE International Conference on Computer Vision ICCV*, (2013), pp.3551–3558.

[14] Oikonomopoulos A, Patras I & Pantic M, "Spatiotemporal localization and categorization of human actions in unsegmented image sequences", *Trans. Image Process.*, Vol.20, No.4, (2011), pp.1126–1140.

[15] Sun C & Nevatia R, "Large-scale web video event classification by use of Fisher vectors", *Proceedings of the IEEE Workshop on Applications of Computer Vision WACV*, (2013), pp. 15–22.

[16] Jégou H, Douze M, Schmid C & Pérez P, "Aggregating local descriptors into a compact image representation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, (2010), pp. 3304– 3311.

[17] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R & Fei-Fei L, "Large-scale video classification with convolutional neural networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, (2014).

[18] Wu Z, Jiang YG, Wang J, Pu J & Xue X, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification", *Proceedings of the ACM International Conference on Multimedia MM*, (2014), pp.167–176.

[19] Shapovalova N, Vahdat A, Cannons K, Lan T & Mori G, "Similarity constrained latent support vector machine: an application to weakly supervised action classification", *Proceedings of the Twelfth European Conference on Computer Vision ECCV, Springer*, (2012).