ISSN (Print) : 0974–6846
ISSN (Online) : 0974–5645

Indian Journal of Science and Technology, *Vol 9(5), DOI: 10.17485/ijst/2016/v9i5/72065, February 2016*

Spatio-temporal based Approaches for Human Action Recognition in Static and Dynamic Background: A Survey

K. Anuradha[1]* and N. Sairam[2]

[1]School of Electrical and Electronics Engineering, SASTRA University, Tirumalaisamudram, Thanjavur - 613401, Tamil Nadu, India; kanukalyan79@gmail.com
[2]School of Computing, SASTRA University, Tirumalaisamudram, Thanjavur – 613401, Tamil Nadu, India; indsai@gmail.com

Abstract

The objective of this review article is to study the spatio-temporal approaches for addressing the key issues such as multi-view, cluttering, jitter and occlusion in recognition of human action. Based on high-level action units, a new sparse model was developed for recognition of human action in static background. Relevant to multi-camera view, a negative space approach for identifying actions taken from different viewing angles was proposed. An approach was based on space-time quantities was proposed to acquire the changes of the action instead of camera motion. This space-time based approach has handled both cluttering and camera jitter. In static background, a sparse model presented for recognition of human action acquires the fact that actions from the same class share same units. The presented method was assessed on numerous public data sets. This method has achieved a recognition rate of 95.49% in KTH dataset and 89% in UCF datasets. Based on negative space, a region based method was offered. This approach has managed the issue of long shadows in human action recognition. The approach was assessed by most common datasets and has attained higher precision than contemporary techniques. An approach based on space-time quantities was proposed to manage cluttering. This approach achieves a recognition rate of 93.18% in KTH dataset and 81.5% in UCF dataset. To handle occlusion, a model was presented with spatial and temporal consistency. The algorithm was appraised on an outdoor dataset with background clutter and a standard indoor dataset (HumanEva-I). Results were matched with advanced pose estimation algorithms.

Keywords: Action Recognition, Camera Jitter, Clutter, Multi-view, Occlusion, Segmentation

1. Introduction

Human activity recognition is about identifying or recognizing human action in different scenarios. Because of its requirements in various applications such as surveillance, entertainment environments, recognition of human activity has gained momentum in video analysis.

Relevant to research, human action recognition has its own part in active research. A 3-level classification of this recognition is shown in Figure1. Low level category identifies the core concepts such as segmentation of objects, extraction and detection of features. Mid-level category designates detecting or recognizing human activity in various scenarios such as single person, multiple persons, interactions and abnormal behaviors. High level category provides different applications of human activity recognition such as surveillance, entertainment environments, and healthcare.

Human activities can be categorized in to various types. Based on the intricacy, a classification of human activities such as interactions, gestures, group activities and actions is done. “Stretching an arm” and “raising a leg” can be specified as gestures. Actions which include many gestures are arranged temporally, viz., “waving,” “walking” and “punching.” “Two persons talking”, “a person stealing a suitcase from another” are interactions between two humans. Activities done by a group of persons or objects are referred as group activities. “A group of persons marching” and “two groups are fighting” are classic group activities.

*Author for correspondence

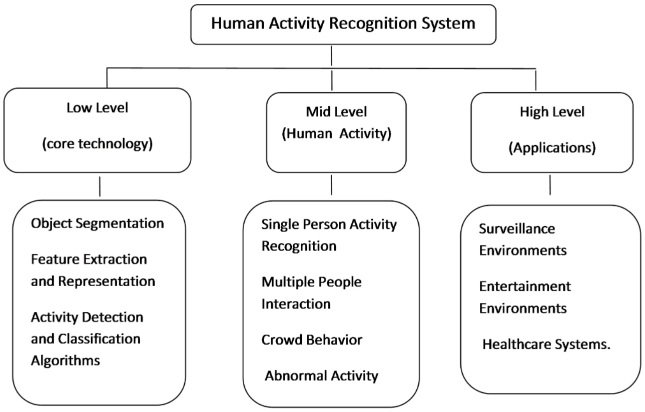
Human Activity Recognition System
Low Level
(core technology)
Mid Level
(Human Activity)
High Level
(Applications)
Object Segmentation
Feature Extraction and Representation
Activity Detection and Classification Algorithms
Single Person Activity Recognition
Multiple People Interaction
Crowd Behavior
Abnormal Activity
Surveillance Environments
Entertainment Environments
Healthcare Systems.

Figure 1. An outline sketch of human activity recognition system.

In conventional methods for recognition of human action, models were built using patterns of low level features such as appearance patterns, optical flow, space-time templates, 2D shape matching, trajectory-based representation and Bag-of-Visual-Words (BoVW). From the video sequences, STIPs are detected by detectors. Nearby these spatio-temporal interest points, local feature descriptors of cuboids extracted. Representations rely on these extracted descriptors. In trajectory-based approaches, human actions are designed as a set of spatio-temporal trajectories. By the arrangement of the trajectories of the important parts' movement, human activity is recognized. Recently, methods based on trajectories have been widely studied. Spatio-temporal volume-based approaches are also referred as holistic representations. Instead of applying sparse sampling using STIPs detectors or extracting trajectories, spatio-temporal volume-based approaches consider a video sequence as a whole.

Dynamic scenes are normally faced in indoor and outdoor situations. In such situations, objects such as swaying trees, spouting fountains, rippling water, moving curtains etc., are to be detected. In dynamic scenarios, many issues such as cluttering, occlusion are to be addressed. Cluttering means images or video sequences with a disordered or messy state.

1.1 Weizmann Event-based Analysis Dataset

Weizmaan Event-Based Analysis dataset is used for studying clustering algorithms and temporally segmenting videos with certain numerical metrics. The dataset has distinctive sequence of around 6000 frames, with different people, wearing different clothes, and carrying out different activities such as walking, running, and waving. A few frames are depicted in Figure 2.

1.2 KTH Recognition of Human Actions

KTH dataset is a video database. This dataset is formed with sequences of human actions with various scenarios. With a static camera, all the sequences were taken in the same background. The database consists of six types of human actions (walking, running, jogging, boxing, hand clapping, and hand waving) in four dissimilar scenarios. Figure 3 presents example frames conforming to various types of actions and scenarios. An actor, context: Static Homogenous Background (SHB), SHB with different clothes, SHB with scale variations, SHB with lighting variations and action type labels an action.

1.3 UCF Sports Action Dataset

University of Central Florida (UCF) has developed many human action datasets, which consists of actions collected from diverse sporting events. The dataset consists of actions featured in a extensive range of scenes and

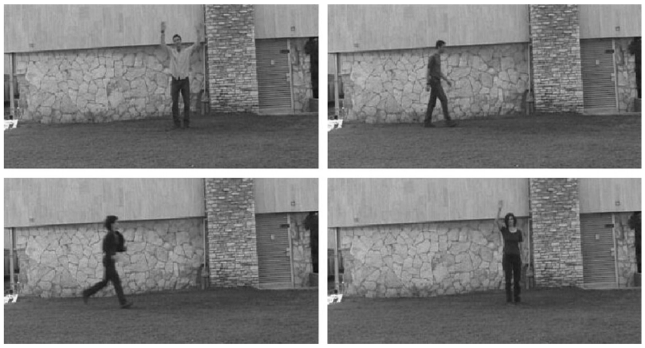

Figure 2. Weizmann event dataset with four sample frames.

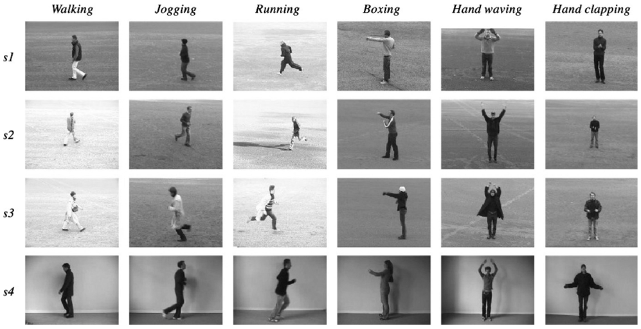

Figure 3. Dataset with six types of human actions in four dissimilar scenarios.

viewpoints. Figure 4 shows the UCF sports action dataset, which depicts various actions such as kicking, diving, lifting, golf swinging, horse-back riding, skating, running, walking and swinging.

2. Categorization of Various Methods

2.1 Static Background

To represent human actions in videos Wang et al.[1] have presented a technique based on high-level action units. A new sparse model was developed for human action recognition. In this approach, three interconnected components were used. They were locally weighted word context, learning of action units and a sparse model. Results and investigations evidently exhibit the efficiency of the approach.

Based on history of information about movements, Thakur et al.[2] have offered an approach for identifying human actions from compressed videos. From the compressed MPEG stream, the coded motion details were applied to build the coarse Motion History Image (MHI) and the Motion Flow History (MFH). The spatio-temporal and motion vector information are portrayed by the features extracted from the static MHI and MFH. Computation is reduced to the maximum extent, by using the features taken from the partly deciphered sparse motion data.

A method named as activity components, was offered by Yuan et. al [3]. Using a set of mid-level components, this method aimed at detecting human activities. This method was developed on the basis of activity components. In this method, inter dependencies of activity components were established and a Spatio-Temporal Context Kernel (STCK) was built.

An innovative method for multi-action recognition was offered by Carvajal et al[4]. By modeling each action with a Gaussian mixture, this method has performed joint segmentation and classification. This modeling is done by applying resilient low-dimensional action features. Then, segmentation was accomplished by classifying on overlapping temporal windows. Earlier methods have used dynamic programming or Hidden Markov Models (HMMs), which has resulted in intrinsic computations. Results have proved that, this method was less complex than the earlier methods.

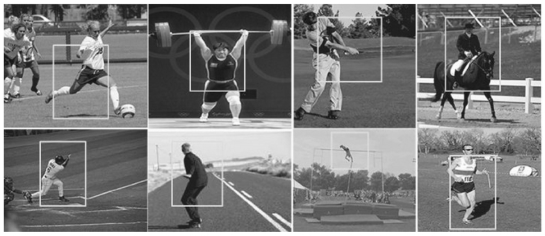

Figure 4. Screenshots of UCF sport action dataset.

By modeling the spatial-temporal structures of human actions, Wang et al.[5] have proposed a method for action recognition in videos. This work has attempted to improve a contemporary method for assessing human joint locations from videos. For selecting the best human joint location, this method has used the K-best approximations generated by the prevailing techniques. On these approximations, this method has included segmentation cues and temporal constraints. This method has merits such as easily interpretable, compact and robust to changes in joints.

2.2 Dynamic Back Ground

To segment prominent objects in videos Mahapatra et al.[6] have presented a technique referred as Coherency based Saliency Maps (CSM). CSM uses spatial and temporal coherency data. Spatial coherency map recognizes regions belonging to regular objects, whereas temporal coherency maps finds regions with high coherent motion. This method has combined the spatial and temporal coherency maps for getting the final spatio-temporal map. This final map finds the striking regions in the video. Experiments done on public datasets demonstrates that this technique outclasses two competing methods in this context.

Zhou et al.[7] have presented a joint learning framework to identify the spatial and temporal scope of the action of interest in training videos. By means of dense trajectories, this technique has taken videos as local features to represent actions. Dense Trajectories were used to get pixel-level localization results.

Luo et al.[8] had proposed a solution for the problem of human action recognition. This solution was developed by mingling the global temporal dynamics and local visual spatio-temporal appearance features. Using the model parameters as motion descriptors, a model has been developed to handle the motion dynamics with robust linear dynamical systems (LDSs). To quantify the similarity between LDSs, a distance based shift invariant subspace angles was presented. Finally, a classification was done by applying the maximum margin distance learning method.

Relying on Radial Basis Function (RBF) artificial neural networks, Huang & Do[9] have presented a new motion detection mechanism. A RBF neural network has the robust nonlinear mapping capacity and the local synaptic plasticity of neurons with a nominal network structure. Because of this nature, RBF approach was appropriate for motion recognition applications with static or dynamic scenes.

A new framework for long-duration complex activity prediction was presented by Li and Fu[10]. By mining temporal sequence patterns, this framework has predicted the complex activities. Results show that, the approach has achieved superior performance with regard to foretelling global activity classes and local action units.

Wang et al.[11] have presented a video sequence. Authors have focused on efficient and accurate action detection. This video sequence was a collection of feature trajectories. To assess the mutual information of feature trajectories, this method started with the generation of a random forest. Then, a one-order Markov model was applied to recursively infer the action regions at successive frames. This method finds the time-continuity property of feature trajectories. By finding the time-continuity property, the action region is effectively inferred even at large temporal intervals. Finally, an ST- tube was formed by adding the successive action regions bounding the human bodies.

2.3 Methods Dealing Cluttering

An innovative compact local descriptor of video dynamics with regard to action perceiving and recognition was presented by Derpanis et al.[12]. This method was based visual space-time oriented energy measurements. From coarse image intensity data, this descriptor solves the issues related to flow-based features. Regardless of spatial appearance viz., changes persuaded by resilience to clutter and with clothing, the descriptor compares the dynamics of two space-time video segments. A relevant similarity measure is presented that declares effective thorough exploration for an action template, resulting from a single prototype video, across a series of videos.

To enforce spatial coherence on the foreground regions, Gao et al.[13] have presented a technique. This technique has two-pass Robust Principal Component Analysis, to dynamically estimate the support of the foreground regions through a motion saliency estimation step. This technique has obtained precisely defined foreground regions. This technique deals with large dynamic background motion in a better way. Also, this method has handled camera jitter with an image alignment. Experiments on standard datasets illustrate that our technique works efficiently on par with contemporary approaches in complex scenarios.

2.4 Methods Dealing Occlusion

Lin et al.[14] have presented an efficient background subtraction technique. This technique has attempted to solve the problems relevant to background subtraction in video surveillance. This technique is based on learning and keeping an array of dynamic texture models with the spatio-temporal representations. This method has extracted a sequence of regular video bricks (i.e.) video volumes covering over both spatial and temporal domain. While adjusting the scene variations, a background modeling modeled as pursuing subspaces within the video bricks. To manage the variations in foreground objects and scene changes, subspaces were incrementally updated during online processing. Validation of the proposed method in complex scenarios has proved superior performances over other advanced approaches for background subtraction.

A novel probabilistic 3D scene model was proposed by Wojek et al.[15]. This work has integrated object tracking, multiclass object detection and scene labeling along with geometric 3D reasoning. This 3D-model could signify complicated interactions such as physical exclusion between objects, inter-object occlusion and geometric context. By applying only monocular video as input, implication in this model has enabled the recovery of 3D scene context, make 3D multi-object tracking from a moving observer and for objects of many categories. Dissimilar to other approaches, this model has done explicit occlusion reasoning. Hence, this model has the ability to track objects that were partially occluded.

An occlusion aware algorithm for identifying human gesture in a video was offered by Ramakrishna et al.[16]. This algorithm has resolved the issue of double counting. Human body is a grouping of singleton parts (viz., head and neck) and pairs of parts (viz., knees, shoulders, and feet). To avoid double counting by reasoning about occlusion, symmetric body parts were identified with mutual exclusion constraints.

2.5 Methods Dealing Multi-view

To solve the multi-view action recognition problem, Zhu et al.[17] have proposed a method. This new multi-sensor fusion method was built, by using a local segment similarity voting method. On the random forests classifier, this method has applied local segments of binary silhouettes. To label the testing actions, a new voting approach was applied. This approach was extended from single camera to multi-camera fusion. Results were assessed on camera fusion set-ups in the IXMAS dataset.

By altering the computation of motion feature and modeling the data from different viewing angles, Rahman et al.[18] have presented the idea of negative space for identifying actions taken from different viewing angles. The problem of long shadows in recognition of human action was managed by this approach. By knowing the fact that, the negative space feature can balance the positive space method, the approach has attempted at constructing a better action recognizer.

A new method presented by Wu and Shao[19] for human action recognition, has joined the gains of local and global representations. This model referred as, bag of correlated poses, was presented to encode temporally local features of actions. A soft-assignment strategy was adopted to lessen the dimensionality of the model and evade the penalty of computational complexity and quantization error. An extended motion template (i.e.) extensions of the motion history image (MHI) was presented to acquire the holistic structural characteristics. By combining the merits of local and global features, this method has offered an eminent depiction for human actions.

Spatio-Temporal Laplacian Pyramid Coding (STLPC), for holistic depiction of human actions was contributed by Shao et al.[20] The authors have presented the spatio-temporal Gaussian/Laplacian pyramids for multi-resolution video analysis. In the pyramid model, a sequence with action was disintegrated into a series of band-pass-filtered parts. Then, spatio-temporal prominent features with several sizes could be well-localized and enriched in these parts. This technique has offered a proficient possibility for holistic human action representation.

2.6 Methods Dealing Interaction-level Human Activities

Ryoo and Matthies[21] have presented a work, which has focused on two features of first-person activity recognition. This work has attempted at finding answers to two queries. This approach has taken the global motion descriptors and local motion descriptors. Then, it has combined both to describe multi-channel kernels for recognition. Apart from this, a new kernel-based activity recognition approach was presented. This kernel-based activity recognition approach explicitly studies structures of human actions from training videos. This work has learnt sub-events consisting of an action and their temporal arrangement. This approach has attained superior performance in first-person activity recognition. Experiments have shown that, this approach could consistently detect activities from continuous videos.

Choi et al.[22] have presented a framework, which recognizes multiple, interacting, people from a mobile platform. To fix all the trajectories of a 3D coordinate system, this approach has assessed the camera's 3D-motion and the people's tracks within a single comprehensible framework. By applying the reversible jump Markov chain Monte Carlo (RJ-MCMC) particle filtering method, the tracking problem was resolved by getting the MAP solution of a posterior probability. This framework was approximated on perplexing datasets with scenes from moving cameras, indoor and outdoor street scenes

2.7 Method Dealing with Multiple Scenes

For video analysis and processing tasks, effective video representation models were critical. Relevant to this context, Moghadam et al.[23] have presented a framework. This framework was built on obtaining a sparsest solution to model video frames. For modeling the spatio- temporal information, one scene was divided into two, a common frame, and a set of innovative frames. To evaluate the innovative frames and common frame for each video, this method was developed on results in the field of compressed sensing. CIV (Common and Innovative Visuals) model may be applied to get scene change boundaries and be extended to videos from many scenes.

2.8 Methods for Human Actor Detection

Kumar and Sangaiah[24] have presented a method for detecting abnormal traffic actions like pedestrians crossing the junction. This method was based on petri nets.

Kheirkhah and Tabatabaie[25] offered a Hybrid Face Detection System. This method detects face of persons in in color images and intricate background. This system has enhanced precision and pace.

3. Various Methodologies

3.1 Static Background

Human actions in videos were represented using high-level action units, in the approach presented in[1]. In this approach, three interconnected components were used. First is the locally weighted word context, a novel context-aware spatial- temporal descriptor, to increase the discriminability of the conventionally used local spatial-temporal descriptors. Second, learning of action units was done with the graph regularized nonnegative matrix factorization. Semantic gap in action recognition was filled by efficiently linking the units. Third, a sparse model relied on a joint l2,1-norm was offered, to maintain the representative items and suppress noise in the action units. Sparse model acquires the fact that actions from the same class share same units. The presented method was assessed on numerous public data sets. Experiments were done to certify that the action unit based representation is acute for modeling intricate activities from videos.

Relying on compressed videos, an idea for human action recognition was presented in[2]. The coarse MHI and MFH was built using the motion information present in the compressed MPEG stream. A set of features taken from static MHI and MFH neatly describe the spatio-temporal and motion vector information. For identifying a set of human actions, the features mined from MFH were employed to train Support Vector Machine, K-Nearest Neighbor, and Bayes and Probalistic Neural Network. With relevance to different classifiers, the outcome was evaluated for each set of features.

In [3], an approach for human action representation was proposed. In this method, activity components (mid-level components) were taken hierarchically from videos. To portray various features of activity components, the approach has focused on information such as appearance, shape and motion. Second, the dependencies between activity components were used. Finally, a spatio-temporal context kernel was built. STCK, has acquired feature's local properties and their spatio-temporal context information. Assessments done on inspiring datasets for activity recognition proved that, the method outpaces standard spatio-temporal features and STCK context kernel and has improved results. However, this approach did not deal with extraction of activity components.

A new method for multi-action recognition presented in[4], needs certain parameters. For modeling each action, this method has used Gaussian Mixture Model (GMM) with the characteristics of actions. This was used as training data for single action videos. A resilient low dimensional action features with image gradient information and optical flow was used. Those features with high spatial frequency only were considered as part of an action. On a temporal sliding window, GMM classifier was applied for segmenting actions. This has made the approach to handle the temporal misalignment in an improved way and resulted in enhanced performance. Evaluations done on the KTH dataset illustrate that, the mechanism has realized superior performance over a recent HMM-based approach. Apart from its merits, this approach has not handled spatial irregularities. Also, this method has not separated irrelevant movements from action motion.

A technique for action recognition in videos was proposed in[5]. For estimating human joint locations from videos, this method has tried to improve an existing, advanced method. This technique has got the inputs from the prevailing method, and includes the segmentation cues and temporal constraints. A set of approximated joints was identified, and then grouped the joints in five body parts. By employing the data mining techniques, a depiction for the spatial-temporal structures of human actions was attained. This representation has acquired the spatial configurations of body parts in a single frame and also the motion of body parts. Assessments have illustrated that this approach was able to confine body joints more precisely than prevailing methods. This approach also has outclassed advance action recognition methods on the UCF sport, the Keck Gesture and the MSR-Action3D datasets. This method has not addressed some issues such as occlusion, missing parts, and recovery of poses in three-dimensions.

3.2 Dynamic Background

A new approach named as, CSM was proposed in[6]. CSM aimed at segmenting the significant objects in a video. A prominent region in a video has the anticipated features: 1) it belongs to a regular object; 2) has coherent motion over time, which is dissimilar to its neighbors. Spatial coherency and temporal coherency data from videos finds salient objects. Spatial coherency finds image regions, which belongs to a regular object. For the sake of finding patches on homogeneous objects, entropy of image patches is applied. Lower entropy values for patches on homogeneous objects.

In CSM, motion information was taken out in the form of motion vectors between consecutive frames, temporal coherency maps were built. Higher entropy values for a pixel over time designate higher motion. A motion center-surround map indicates regions with higher movement magnitude than their neighbors. Likewise, the direction center-surround map indicates regions with different movement direction than the neighbors. A direction entropy map indicates regions with regular motion direction and rejects regions with random motion. Because, regions with random motion have higher entropy.

CSM has combined these four maps, which yields the temporal coherency map. By combining spatial and temporal coherency maps, a saliency map was produced. Experiments done for moving object extraction on public datasets demonstrated that, this method has produced lower false positives and higher segmentation accuracy when compared with two other techniques for calculating motion saliency. By extensive analytical use of motion information, this method has attained higher accuracy. This method could not segment the objects, when several similar objects are moving close to each other.

A joint learning framework proposed in[7] identifies the spatial and temporal extents of the action of interest in training videos. This technique has used dense trajectories, which has taken videos as local features to represent actions. Dense Trajectories were used to get pixel-level localization results. A trajectory split-and-merge algorithm was proposed to segment a video into the background and split foreground moving objects. To enable segmentation, the inherent temporal smoothness of human actions was used. With the results of latent SVM framework on segmentation, spatial and temporal extents of the action of interest are considered as latent variables. Latent variables were inferred concurrently with action recognition.

Equated with the prevailing action detection methods that do not depend on bounding box annotations, this approach has two gains. First, by using dense trajectories creates fine-grained pixel-level localization results. Second, this method has learnt both temporal and spatial extents of the action of interest in positive examples. Experiments on two challenging datasets validate that action detection with this approach was superior than contemporary techniques.

The approach presented in[8] focused on solving the issues in human action recognition, by joining Linear Dynamical Systems (LDSs) and cuboids recognition in a maximum margin distance learning framework. For identifying human motion, this approach has used global dynamic information. The authors have studied and found that, no other work has taken this idea for recognizing human actions. A strong and reliable LDS learning algorithm was presented to define action sequence dynamics. A suboptimal solution was accomplished by repeatedly checking stability conditions with new constraints. This approach has attained a stable solution, by repeating the process. Since robust LDS that share the same dynamic features as the trasining data was important for shift invariant distance metric, this work has proved that sequences were developed in those lines. This work has done classification of action sequences independent of the frame, which contemporary methods did not achieve. The approach was assessed on five short clips data sets, called as Weizmann, KTH, UCF sports, Hollywood2 and UCF50, a three long continuous data sets, called as VIRAT, ADL and CRIM13. These assessment shows that; this work has produced competitive results as compared on par with current advanced methods. Since LDSs could not describe the non-linear dynamics, this method has not handled long-time sequences due to the long-scale temporal variations.

The approach developed in[9] has two important modules. Multi Background Generation (MBG) module makes a workable probabilistic model. This model was built by computing the Euclidean distance between the input pixels with the reference background. This information is transmitted to the network in the form of hidden layer neuron centers. By this way, the probabilistic background model builds a hidden layer in the RBF network structure. Then, the Moving Object Detection (MOD) module achieves complete and accurate recognition of moving objects by using two procedures, a block alarm procedure and an object extraction procedure. The block alarm procedure rejects unwanted checking of the dynamic and static background region. Then, the object extraction procedure manipulates those blocks with high likelihood of moving objects. Based on detailed range natural video sequences, this approach was evaluated. Results depicted that the approach outclasses the contemporary approaches in terms of similarity ratio.

The framework proposed in[10] comprises of, a general framework; Probabilistic Suffix Tree (PST); Sequential Pattern Mining (SPM); a Predictive Accumulative Function (PAF). General framework was developed with the intention of solving a problem of complex activity prediction, by mining temporal sequence patterns. Probabilistic Suffix Tree (PST) was presented to associate casual relationships among constituent actions. Sequential Pattern Mining (SPM) designed for interactive objects information using the context-cue, where a series of action and object co-occurrence were encoded as a complex symbolic sequence.

PAF was developed to portray the predictability of each type of activity. For action-only prediction and context-aware prediction, two experiments on two data sets were done. Results demonstrated that, the approach has attained superior performance with regard to foretelling global activity classes and local action units. This model did not work for actionlet sequences with noise.

The work presented in[11] has focused on (1) Feature trajectories for envisaging the evolvement of the action region at large temporal intervals. (2) Accurately localizing human actions. This approach finds the ST-tube, which consists of bounding boxes of human bodies in successive frames. By including the background clutters, the approach could retrieve human actions from surveillance video databases. (3) Hierarchical Branch-and-Bound Search (H-BBS), which updated the action region in the ST- tube. This was done for handling high-resolution videos. (4) Processing of sequential data. When matched with the traditional frame-by- frame tracking method, this approach has gained maximum savings in computational time. By tracking the action region through feature trajectories. It has exposed that the approach fits well for online applications, such as human computer interface and video surveillance. This method could not detect complex actions, which shares the common motion patterns.

3.3 Methods Dealing Cluttering

A work suggested in[12] reliant on space-time quantities acquire the changes of the action instead of camera motion. Camera is kept immobile. For action detection in a big search video, the video is skimmed with regard to space-time positions. By gliding a 3D template on all space-time positions, skimming was done. At these positions, calculation of similarity between corresponding positions of the template, search volumes and the histograms was done. For the purpose of action recognition, a database is formed with an uncropped video. This uncut video has a set of labeled video parts comprising spatio-temporally localized actions and the query video. For slight camera movements such as camera jitter, the proposed features were proved to be robust. On matching the query video with each action in the database, action is labeled with the maximum similarity value was produced as the class. The main restriction of this approach is, the usage of a single monolithic template for a specific action.

To solve issues such as large dynamic background motion and camera jitter, this work was presented in[13]. This is a hierarchical two-pass process. In this two-pass process, the first-pass RPCA quickly finds the possible regions of foreground in a sub-sampled image. A simple motion consistency scheme is applied to evaluate the motion saliency of the foreground regions. In the second pass, a block- sparse RPCA enforces the spatial coherence of foreground objects in the outlier matrix S, with the λ value set based on the motion saliency estimated in the first pass. By considering the motion saliency and the block-sparse structure of the outlier matrix S, this approach makes the foreground detection robust even with the clutter produced by the background motion. Comprehensive experiments on challenging videos and standard datasets validate that this technique outpaces advanced methods and works efficiently in complex scenarios.

3.4 Methods Dealing Occlusion

Based on the ideas such as Spatio - Temporal Representations, Pursuing Dynamic Subspaces and Maintaining Dynamic Subspaces Online, a framework for background modeling was presented in[14].

In spatio-temporal representations, observed scene was represented by video bricks. Video volumes spanning over both spatial and temporal domains were used for modeling joint spatial and temporal data. At each location of the scene, a sequence of video bricks was taken as observations. In each observation, learning and updating the background models were done. To precisely encode the video bricks against illumination variations, a brick based descriptor, referred as Center Symmetric Spatio-Temporal Local Ternary Pattern (CS-STLTP) was designed.

In pursuing dynamic subspaces, each sequence of video bricks at a certain location was considered as a consecutive signal, and produces the subspace within these video bricks. The linear dynamic system was implemented to characterize the spatiotemporal statistics of the subspace. A data matrix represents the observed video bricks, where each column has the feature of a video brick. The proposed background model has the information of appearance and motion, which were extracted over both spatial and temporal domains.

In maintaining dynamic subspaces online, this model has done a segmentation of moving foreground objects. This segmentation was done for the given newly appearing video bricks. At the same time, this model has kept the background models which has accounted for scene changes. The raising problem is to update parameters of the subspaces incrementally against disturbance from foreground objects and background noise. A new approach was presented in our model to overcome the problem of, one video brick be partially occluded by foreground objects. Out model has replaced the pixels labeled as non-background by the generated pixels to make new observations.

A new probabilistic 3D model was proposed in[15]. This model has done a monocular 3D scene geometry assessment in real-time traffic scenes. This assessment results in more consistent identification of objects, viz., trucks, pedestrians and cars. Even with a single camera, a reliable 3D description of a perceived scene was obtained with the information from object identification and low-level scene labeling. The approach was appraised for various types of challenging onboard sequences. These assessments on the approach have proved a remarkable progress in 3D multi-people tracking.

The work presented in[16] has attempted at tracking human pose, using an occlusion-aware model. This proposed model has imposed both spatial and temporal consistency and has made some attempts to avoid double counting. Enthused by ideal designs for multi-target tracking, the model has jointly tracked symmetric parts. The algorithm was appraised on an outdoor dataset with background clutter and a standard indoor dataset (HumanEva-I). Results were matched with advanced pose estimation algorithms.

3.5 Methods Dealing Multi-view

To solve the multi-view action recognition problem, an innovative method was proposed in[17]. This approach of multi-sensor fusion technique was developed to manage the problem of unequal classification capabilities. To attain this, this approach depends on each camera prediction histogram. Here, each voting segment is biased with regard to outcome of classification in the random forests. Results were matched with the Naive-Bayes Nearest Neighbor (NBNN) and baseline Bag of Words (BoW) methods. Detailed investigations illustrated that this algorithm outpaces the baseline BoW and NBNN methods.

From video sequences, a region based method was offered in[18] for recognition of human action. This method has worked with the negative space. Negative space is adjacent regions of the human silhouette. The proposed approach has handled the issue of long shadows in human action recognition. Most of the approaches tried to remove the shadows during the segmentation process. Since this approach has not removed the shadows, it did not rely on segmentation. To improve recognition, the method can match the positive space based techniques. The approach has a hierarchical processing. This processing started with histogram examination on input image which is segmented. It was followed by shape and motion feature extraction, pose sequence analysis by applying Dynamic Time Warping. Finally, categorization was done with Nearest Neighbor classifier. The approach was assessed by most common datasets and has attained higher precision than contemporary techniques. This method could not detect objects with multiple shadows. When the image plane and action performing plane are orthogonal to one another, this method could not recognize objects.

A method contributed in[19], for recognition of human action, which was termed as bag of correlated poses. In this method, the authors have contributed three key ideas. First, Correlogram of human poses in an action sequence was presented to encode temporal structural information. This model has considered the silhouette in each frame as a feature. An explicit model was developed to encode its temporal-structural information, by creating a correlogram. Second, a soft-assignment method was designed for avoiding the quantization error penalty. Finally, a holistic representation was presented descriptor for local representation. This method combines the temporally local descriptor with an extension of the holistic descriptor, motion history image (MHI). A joint model was presented to fuse two different descriptors and attain further progress over the separate methods. Experimental results have demonstrated that, the approach outclasses the advanced methods on the IXMAS action recognition dataset.

A pyramid model presented in[20], an action sequence was divided into a series of band-pass-filtered components. A set of spatio-temporal prominent characteristics with various sizes could be well-localized and improved in the band-pass-filtered parts. To get discriminative and invariant spatio-temporal characteristics, a set of 3-D Gabor filters and max pooling were continuously applied. This method has preserved motion and structural information. The method has accomplished excellent recognition rates on the KTH, the multiview IXMAS, UCF Sports, and HMDB51 datasets. This technique also outdoes advanced techniques in action recognition.

3.6 Methods Dealing Interaction-level Human Activities

A work presented in[21] has focused on two features of first-person activity recognition. This work has attempted at finding answers to two queries. This approach has taken the global motion descriptors and local motion descriptors. Then, it has combined both to describe multi-channel kernels for recognition. Apart from this, a new kernel-based activity recognition approach was presented. This kernel-based activity recognition approach explicitly studies structures of human actions from training videos. This work has learnt sub-events consisting of an action and their temporal arrangement. This approach has attained superior performance in first-person activity recognition. Experiments have shown that, this approach could consistently detect activities from continuous videos.

A person tracking system was proposed in[22]. This work was employed in two environments. One for tracking people from a mobile, ground-level camera. Another for tracking people indoors. from a robot platform. Authors have proposed an upright approach for tracking many people and assessing the movement of camera simultaneously. First, an innovative model that dealt with the process of producing videos from a mobile camera. As a second step, a motion model was presented that captures the communications between targets. Third, to construct a more robust and adaptable tracker, an upright approach for multiple person detection was used. This approach was proved to be effective with the application of reversible jump-Markov chain Monte Carlo (RJ-MCMC) particle filtering. Results of experiments illustrated that the presented approach has robustly estimated the movements of camera from scenes which change dynamically and also track people who are interacting or independently moving.

3.7 Method Dealing Multiple Scenes

Based on the theory of sparse coding, Common and Innovative Visuals (CIV) was presented in[23]. By applying temporal and spatial correlations, frames from one scene would be divided to a sum of a common part and relevant innovative components. The common part accounts for visual information, which did not change among frames, where as innovative, parts root from changes specific to each frame. Using some results from sparse coding and the theory of compressed sensing, this method has identified these parts. This approach has proved that, it could identify scene change boundaries and results can be extended to videos from many scenes. This method was resilient to noise. This method has proved efficiency and performance in terms of object tracking, video editing and scene change detections. Scene change boundaries could be detected by CIV. CIV was differentiated from other approaches in the sense that, no motion estimation/detection, image segmentation, feature selection/tracking, foreground/background separation or object tracking was needed in CIV. Without motion estimation, detection and image segmentation, simulation results have validated that common video tasks have been done effectively. When background motion is rapid, this method could not detect objects in a long video sequence. This method was not able to identify objects, when multiple large objects moving in different directions. CIV did not detect objects, when they are occluded by other moving objects.

3.8 Methods for Human Actor Detection

An event detection method suggested in[24] was capable of detecting strange events viz., needless crossing of pedestrians in traffic junctions. A Petri net model was developed for a particular event in traffic situation, with the aid of domain expert. Then, a video with such events is was applied to the Petri net models. This method was tested with actual traffic videos. Results exhibit that, the method has accomplished a detection rate of 88%.

Hybrid Face Detection System (HFDS) presented in[25] comprise two stages Face Candidates Extraction stage and Face Detection stage. Face candidate extraction stage takes out the potential regions in the image. In this stage, background elimination is performed. In the second stage, the segments generated in the first stage were processed. This method has achieved a detection rate of 98.88% greater than that of Viola-Jones method (98.66%).

4. Comparison of Results Presented by Wang et al.[1], Rahman et al.[18] and Wang et al.[11]

This comparison is done with the intention of finding an approach with superior performance in recognizing human actions. Human actions were taken from extensive datasets like KTH and UCF Sports action. Comparisons shown in Figures 5 and 6 has revealed that, each approach outperforms the other in recognizing specific actions.

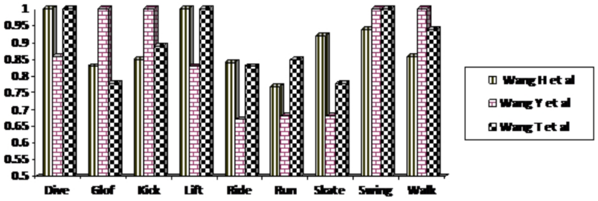

Figure 5. Comparision of results attained with UCF sports action dataset.

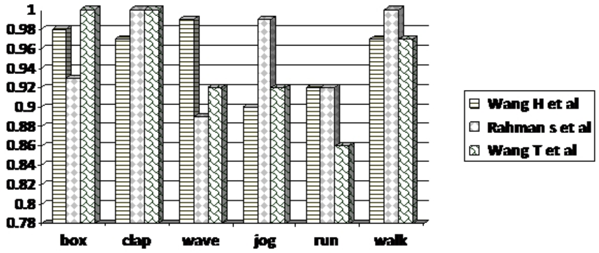

Figure 6. Comparision of results attained with KTH dataset.

Results also revealed that, a single did not perform well in all cases. This comparison opens lot of ideas in human action recognition, which can be taken up for research.

5. Conclusion

A survey is done on the spatio-temporal approaches with the intention of finding a good solution for important problems in human action recognition with static and dynamic background. Important issues that were addressed by these approaches are occlusion, multi-camera view, cluttering, dynamic background variations, camera jitter, and intensity variations. Since human action recognition has lot of applications, this survey may generate interest and opens new avenues of research in this area.

6. References

1. Wang H, Yuan C, Hu W, Ling H, Yang W, Sun C. Action recognition using nonnegative action component representation and sparse basis selection. IEEE Transactions on Image Processing. 2014;23(2):570–81.
2. Thakur R, Mehan N, Namitakakkark. Recognition of human actions using motion history information extracted from the compressed. International Journal of Advanced Research in Computer Science and Software Engineering. 2013;3(7):973–7.
3. Yuan F, Xia G, Sahbi H, Prinet V. Mid-level features and spatio-temporal context for activity recognition. Pattern Recognition. 2012;45(12):4182–91.
4. Carvajal J, Sanderson C, McCool C, Lovell B. Multi-action recognition via stochastic modelling of optical flow and gradients. Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis - MLSDA'14; 2014. p. 19.
5. Wang C, Wang Y, Yuille A. An approach to pose-based action recognition. 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013. p. 915–22.
6. Mahapatra D, Gilani S, Saini M. Coherency based spatio-temporal saliency detection for video object segmentation. IEEE Journal of Selected Topics in Signal Processing. 2014;8(3):454–62.
7. Zhou Z, Shi F, Wu W. Learning spatial and temporal extents of human actions for action detection. IEEE Transactions on Multimedia. 2015;17(4):512–25.
8. Luo G, Yang S, Tian G, Yuan C, Hu W, Maybank S. Learning human actions by combining global dynamics and local appearance. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(12):2466–82.
9. Huang SC, Do B-H. Radial basis function based neural network for motion detection in dynamic scenes. IEEE Transactions on Cybernetics. 2014;44(1):114–25.
10. Li K, Fu Y. Prediction of human activity by discovering temporal sequence patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(8):1644–57.
11. Wang T, Wang S, Ding X. Detecting human action as the spatio-temporal tube of maximum mutual information. IEEE Transactions on Circuits and Systems for Video Technology. 2014;24(2):277–90.
12. Derpanis KG, Sizintsev M, Cannons K, Wildes RP. Action spotting and recognition based on a spatiotemporal orientation analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(3):527–40.
13. Gao Z, Cheong L, Wang Y. Block-sparse RPCA for salient motion detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(10):1975–87.
14. Lin L, Xu Y, Liang X, Lai J. Complex background subtraction by pursuing dynamic spatio-temporal models. IEEE Transaction on Image Process. 2014;23(7):3191–202.
15. Wojek C, Walk S, Roth S, Schindler K, Schiele B. Monocular visual scene understanding: understanding multi-object traffic scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(4):882–97.
16. Ramakrishna V, Kanade T, Sheikh Y. Tracking human pose by tracking symmetric parts. 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 3728–35.

17. Zhu F, Shao L, Lin M. Multi-view action recognition using local similarity random forests and sensor fusion. Pattern Recognition Letters. 2013;34(1):20–4.
18. Rahman SA, Leung MKH, Cho S-Y. Human action recognition employing negative space features. Journal of Visual Communication and Image Representation. 2013;24(3):217–31.
19. Wu D, Shao L. Silhouette analysis-based action recognition via exploiting human poses. IEEE Transactions on Circuits and Systems for Video Technology. 2013;23(2):236–43.
20. Shao L, Zhen X, Tao X, Li X. Spatio-temporal laplacian pyramid coding for action recognition. IEEE Transactions on Cybernetics. 2014;44(6):817–27.
21. Ryoo MS, Matthies L. First-person activity recognition: what are they doing to me?. 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 2730–7.
22. Choi W, Pantofaru C, Savarese S. A general framework for tracking multiple people from a moving camera. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(7):1577–91.
23. Moghadam AA, Kumar M, Radha H. Common and innovative visuals: a sparsity modeling framework for video. IEEE Transaction on Image Processing. 2014; 23(9):4055–69.
24. Kumar PMA, Sangaiah AK. A petri net-based approach for event detection in pedestrian crossing sequence. Indian Journal of Science and Technology. 2014;7(4):439–46.
25. Kheirkhah E, Tabatabaie ZS. A hybrid face detection approach in color images with complex background. Indian Journal of Science and Technology. 2015;8(1):49–60.