Exploring Behavior Analysis in Video Surveillance Applications

Ahmed Taha[1], Hala H. Zayed[2], M. E. Khalifa[3] and El-Sayed M. El-Horbaty[4]

[1, 2] Faculty of Computers & Informatics, Benha University, Benha, Egypt
[3, 4] Faculty of Computer & Information Sciences, Ain Shams University, Cairo, Egypt.

ABSTRACT

Video surveillance is recently one of the most active research topics in computer vision. It has a wide spectrum of promising public safety and security applications. As the number of cameras exceeds the capability of human operators to monitor them, the traditional passive video surveillance is proving ineffective. Hence, converting to intelligent visual surveillance is inevitable. Intelligent visual surveillance aims to detect, recognize and track certain objects from image sequences automatically, and more generally to understand and describe object behaviors. Many researchers have contributed to the field of automated video surveillance through detection, classification, and tracking algorithms. Despite recent progress in computer vision and other related areas, there are still major technical challenges to be overcome before reliable automated video surveillance can be realized. Recently, the problem of analyzing behavior in videos has been the focus of several researchers' efforts. It aims to analyze and interpret individual behaviors and interactions between different objects found in the scene. Hence, obtaining a description of what is happening in a monitored area, and then taking appropriate action based on that interpretation. In this paper, we give a survey of behavior analysis work in video surveillance and compare the performance of the state-of-the-art algorithms on different datasets. Moreover, useful datasets are analyzed in order to provide help for initiating research projects.

General Terms

Computer Vision, Video Surveillance, Object Tracking.

Keywords

Behavior Analysis, Action Recognition, Event Detection.

1. INTRODUCTION

Many large cities have crime and antisocial behavior problems, such as fights, vandalism, breaking and entering shop windows, etc. Often these cities have video cameras already installed, but what is lacking is automatic analysis of the video data. Such analysis could detect unusual events, such as patterns of running people, converging people, or stationary people, and then alert human security staff. As the amount of video data collected daily by surveillance cameras increases, the need for automatic systems to detect and recognize suspicious activities performed by people and objects is also increasing.

The back-bone of a general video surveillance system consists of six consecutive steps [1]: background model, foreground pixel extraction, object segmentation, object classification, object tracking, and action recognition. The first step is to build the background model. The purpose of the background model is to represent what the environment looks like without any foreground objects. In the literature, there are many methods that are proposed for constructing the background model [2-6]. The numerous approaches to this problem differ in the type of background model and the procedure used to update the background model. The methods generally fall into two main categories according to its capability to adapt with the environment variations: adaptive and non-adaptive [1]. That is, the ability of the background model to be updated to reflect the environment changes.

The second step in image understanding process is the foreground pixel extraction. It involves the extraction of pixels that are not part of the background model from an image that is being processed. These pixels serve as a basis for further analysis in the following steps. There are several approaches to foreground pixel extraction that researchers use (such as: temporal differencing, background subtraction, Gaussian BM subtraction, optical flow) [4, 7, 8].

The third step is object segmentation. It is the process of grouping similar foreground pixels into homogenous regions, better known as foreground objects or blobs. The similarity of the foreground pixels is determined by using a similarity metric. Similarity metrics are used to determine whether the pixels being compared belong to the same blob and to group the pixels into homogenous regions where all pixels have similar characteristics. Several similarity metrics, found in literature, are used for object segmentation, some of which are color based, proximity or location based, or mixture of characteristics [1, 9].

The fourth step is object classification. It is the process of identifying what kind of object is present in the environment. This is particularly useful when distinctly different types of objects exist in the environment and when a different tracking method is used for each type of objects. Therefore, it can be considered an optional step that is performed according to the application nature. There are two main categories of approaches for classifying moving objects: shape-based classification and motion-based classification [9, 25]. In addition, many metrics are found in literature for object classification: size metric [25], speed metric [11], and dispersedness [10].

The fifth step is object tracking. It is the process of locating a moving object (or multiple objects) over time. The objective of video tracking is to associate target objects in consecutive video frames. The association can be especially difficult when the video frame rate is slow relative to objects motion. Another situation that increases the complexity of the problem is when the tracked object changes orientation over time. The most critical issue in object tracking is to make sure that the same blob is being tracked in each subsequent frame by using object matching techniques (proximity-based techniques, prediction-based techniques, blob's characteristics based techniques, or blob model based techniques) [11, 12].

The sixth step of a general video surveillance system is the action recognition. It involves the analysis and the recognition of motion patterns to produce a high-level description of actions and interactions among objects. It is the process of recognizing the actions to understand what is happening in an environment [9]. In some circumstances, it is necessary to analyze the behaviors of people and determine whether their behaviors are normal or abnormal.

Behavior analysis using visual surveillance involves the most advanced and complex researches in image processing, computer vision, and artificial intelligence. The research in this area focuses mainly on the development of methods for analysis of visual data in order to extract and process information about the behavior of physical objects (e.g., humans & vehicles) in a scene. Behavior analysis is not restricted to only video surveillance systems but it can be extended to include interactive video games and many other applications. The behavior analysis in uncontrolled environments is critical to video-based surveillance systems, which is one of the extreme goals of vision technologies. The challenges can be summarized in two points:

a) The vast diversity of one event viewed from different view angles, at different scales, and with different degrees of partial occlusions.

b) The demand for efficient processing of huge amount of video data [1].

Extensive research has been reported on behavior analysis. This paper provides a survey of the various studies in this promising area. It presents an overview of current advances in the field.

The rest of the paper is organized as follows: in Section 2, different action representation methods are discussed showing the strengths and weaknesses of each method. Section 3 reviews the state-of-the-art methods for action recognition. Section 4 presents the datasets that are currently used by many of the action recognition approaches as a benchmark. Finally, we conclude the paper in Section 5.

2. BEHAVIOR REPRESENTATION

This section presents a review of representation methods used to discriminate actions from visual data. A first step in action recognition is the extraction of image features that are discriminative with respect to posture and motion of the objects.

Before recognizing any activity, the activity representation method must be determined. Activity representation concerns the extraction, selection, and transformation of low-level visual properties in video to construct intermediate input to an activity recognition model [9]. Activity representation should be expressive enough to describe a variety of activities yet sufficiently discriminative in distinguishing different individual activities. Various representations have been suggested. Some representations focus on maximizing the amount of high level information they could represent while others focus on maximizing the extraction efficiency [13]. Many challenges determine the choice of the action representation method. These challenges include intra- and inter-class variation, environment and recording settings, temporal variations, and the availability of training data and its labeling [25]. Different activity representations can be grouped into three categories as shown in Fig. 1: object based representations [9, 14], pixel based representations [15-17, 20, 21], and other feature representations [22, 24]. In the following subsections, we review the work presented in each category.

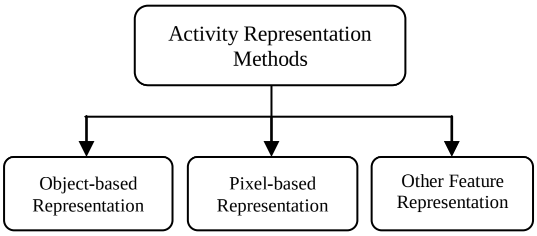

Activity Representation Methods
Object-based Representation
Pixel-based Representation
Other Feature Representation

Fig. 1. Activity representation methods

2.1 Object Based Representation

This type of representation depends on extracting a set of features for each object in the video. These features include trajectory or blob-level descriptors such as bounding box and shape. A trajectory-based feature is prevalently utilized to represent the motion history of an object in a scene [14]. Usually, a trajectory is formed by associating a set of attributes of detected object, such as appearance features and velocity over successive frames using motion tracking algorithms (Fig. 2). However, these attributes are highly dynamic and vary over time. So probabilistic frameworks such as Kalman Filter and Particle Filter are commonly adopted [9]. In addition, processing steps such as moving average smoothing, or trajectory merging are commonly employed to overcome noise problem or trajectory discontinuity problem to a certain degree.

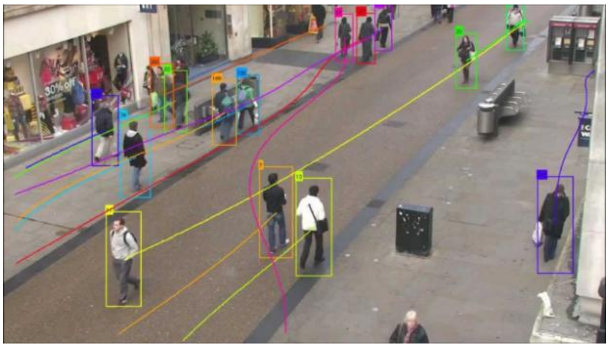


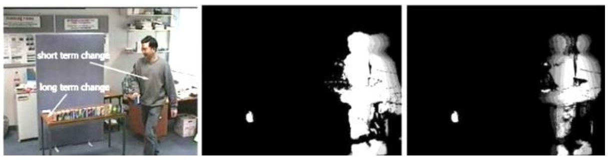
Fig. 2. Different trajectories for the movements of humans on the road

Object trajectories provide rich spatiotemporal information about an object’s activity. Therefore, trajectory information is typically employed to understand object’s behavior in the scene over time. However, using object-based representation in real-world surveillance can be challenging. Generally, object tracking depends on two important assumptions: the first one is that the object location can be determined reliably, and the second is that the spatial displacement of the same object between successive frames is small [11]. However, these assumptions are often invalid due to severe occlusions and low-frame rate surveillance videos. Specifically, the large number of objects with complex activities causes difficult and continuous inter-object occlusions (sometimes known as dynamic occlusions). Tracking of multiple objects in this environment is challenging since dynamic occlusions can cause ambiguities on the number and identities of targets, leading to temporal discontinuity in trajectories [12]. Also, given low-frame rate video, large spatial displacements of the object are detected between consecutive frames, causing severe fragmentation of object trajectories.

2.2 Pixel Based Representation

Pixel-based representation involves extracting pixel-level features such as color, texture, and gradient [9]. It does not gather features into blobs or objects like object-based representation. In the literature, the pixel-based representation methods can be categorized into three classes: foreground estimation, optical flow, and image appearance-based features.

The most common pixel-based representation is foreground pixels estimation through background subtraction. Despite its simplicity, it shows encouraging results in detecting unusual event by representing activity using both spatial and temporal distribution of foreground pixels. Many studies have shown the feasibility of this simple representation in human motion recognition [15] using Motion History Image (MHI) and in unusual event detection using Pixel Change History (PCH) or average behavior image (Fig. 3) [16]. In an MHI, pixel intensity is a function of the motion history at that location, where brighter values correspond to more recent motion. However, the MHI is a special case of the PCH. A PCH image is equivalent to an MHI image when the accumulation factor is set to 1. Foreground pixel-based representation is attracting lot of attention because it avoids explicit object segmentation and tracking and it is computationally feasible. Hence, It can be efficiently employed in representing activity in crowded scenes where tracking is a complicated problem.



(a) (b) (c)

Fig. 3. (a) A keyframe of a person shopping a can (b) MHI (c) PCH

Another common pixel-based method for activity representation is optical flow. It extracts the motion information (direction and speed) of individual pixels between successive frames (Fig. 4). An image space is usually divided into cells of a specific size (e.g., 10 × 10). For each cell, the average or median flow field is computed. Then, flow vectors are normally filtered based on a predefined threshold to reduce potential observation noise. Moreover, the extracted optical flow information is usually combined with a foreground mask so that only the vectors caused by foreground objects are considered, while all the flow vectors outside the foreground mask are set to zero [17]. Like foreground pixel-based representation, optical flow based representation avoids explicit tracking of individual objects. Hence, it is also used in highly crowded scenes with broad clutter and dynamic occlusions. However, optical flow has an additional advantage over foreground pixel-based representation. It already provides information about motion direction and speed, which are essential for understanding certain types of activity. In the other hand, most optical flow methods face problems in dealing with videos with very low frame rate and poor image quality. This is because they assume small object displacement and constant brightness for the computation of velocity field, which is invalid for these videos.

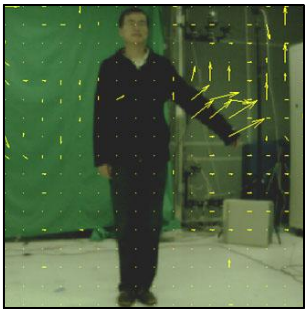


Fig. 4. Optical flow representation

A more recent direction in pixel-based representation method is utilizing image appearance-based features. In [18], Histogram of oriented gradient (HOG) features are utilized to detect unusual events in web-camera videos. Space time intensity gradients are applied as salient features to a nonparametric distance measure for learning the disparity between activities [19]. Also, spatiotemporal gradients of pixel intensities are extracted from video to characterize activities in extremely crowded scene [20]. Mixture of dynamic texture is utilized to represent activity patterns [21]. In general, studies in this direction show promising results. However, calculating such mixtures of texture or space-time gradients may be computationally expensive. Furthermore, the extraction of spatiotemporal gradient would definitely fail as a result of motion discontinuities given a low frame rate video.

2.3 Other Feature Representation

Some studies replace the low-level features representations (such as location, shape, and motion) with more complex ones for efficient modeling of complex behaviors. Kim et al. [22] apply a Mixture of Probabilistic Principal Component Analyzers (MPPCA) algorithm to learn a generative model for local optical flow patterns, which offers a compact representation by encoding the optical flow patterns as probabilistic words. Another relatively close work is presented in [23], the authors introduce an event-based abstraction that represents a behavior pattern using the probabilities of different classes of event occurring in each video. Different types of behavior patterns are either composed by different classes of events, or having different order of event occurrence. Also, Park et al. [24] presents a framework that switches between trajectory-based features (e.g. velocity and position) and blob-based features (e.g. aspect ratio of bounding box and height) based on the visual quality of detected objects. To sum up, Table 1 shows the advantages and disadvantages of different representation methods.

Table 1. Advantages and disadvantages of different representation methods

Representation method	**Advantages**	**Disadvantages**
Trajectory	▪ It provides rich spatiotemporal information about objects ▪ It does not require any appearance information of the individuals in the scene	▪ It is applicable in settings in which objects can be tracked with reasonable accuracy ▪ It fails with: o Low-frame rate surveillance videos o Severe occlusion o Large number of objects with complex activities
Foreground pixel extraction	▪ Simple ▪ Computationally feasible ▪ It avoids explicit object segmentation & tracking ▪ It represents activity using both spatial & temporal distribution of foreground pixels.	▪ It fails to extract a foreground object that has a color closer to that of a background. ▪ A shadow may be incorrectly determined as a foreground object.
Optical flow	▪ It avoids explicit object segmentation & tracking ▪ It can be used in highly crowded scenes with broad clutter & dynamic occlusions. ▪ It provides rich information about motion direction and speed.	▪ Computationally complex ▪ It works well only for small displacements
Appearance-based features	▪ It shows promising results.	▪ Computationally expensive ▪ It fails with: o Low-frame rate surveillance videos o Severe occlusion o Large number of objects with complex activities

3. ACTION RECOGNITION

The problem of analyzing behaviors in video has been the focus of several researchers' efforts and several systems have been described in the literature. Action recognition is the process of labeling image sequences with action labels. The task is challenging due to variations in motion performance, recording settings and inter-object differences. Generally, the action recognition process can be divided into two steps [25]: (1) feature extraction and representation and (2) action class prediction. The first step deals with the extraction and encoding of features to describe motions of interest. Multiple features might be extracted for motion modelling prediction. Many techniques have been used in this step including parametric models, appearance descriptors (such as silhouettes or skeletons), and local motion descriptors (such as optical flow) [25]. Parametric models suffer from the difficulty of recovering model parameters to fit the target. Appearance descriptors describe how the target looks but it fails towards partial occlusions of the target. Local motion descriptors describe the apparent motion of the pixels, providing a very strong cue for action recognition. The second step aims to transform features into semantic descriptions in order to predict action class [25]. It is possible to apply exemplar-based models with different distances measures to select action labels. However, most of the systems employ probabilistic graphical models such as Hidden Markov Model (HMM). Although, HMM is effective, discriminative graphical models have shown a better performance in action class prediction but there are many open problems concerning their use [25].

The existing methods for action recognition in realistic, uncontrolled video data can be categorized into three categories: human model based methods, holistic methods, and local feature methods (Fig. 5) [26]. Human model based methods employ a full 3D or 2D model of human body parts, then action recognition is carried out using information of body part positioning as well as movements. Holistic methods use knowledge about the localization of humans in video and consequently learn an action model that captures characteristic, global body movements without any information of body parts. Local feature methods are entirely based on descriptors of local regions in a video, no prior knowledge about human positioning nor of any of its limbs is given. In the following subsections, these categories are discussed in more detail.

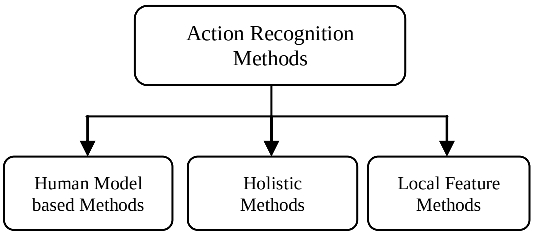


Fig. 5. Action recognition methods

3.1 Human Model Based Methods

Human model based methods recognize actions by employing information such as body part positions and movements. A significant amount of research is dedicated to action recognition using trajectories of joint positions, body parts, or landmark points on the human body with or without a prior model of human kinematics [27, 38]. Approaches in this field depend on a previous psychophysical work on visual interpretation of biological motion. This work shows that humans are able to recognize actions from the motion of a few moving light displays attached to the human body.

The localization of body parts in movies has been investigated in the past and some works have shown impressive results

[28]. However in general, the detection of body parts is a difficult problem in itself, and results are still limited especially for the case of realistic and less constrained video. Some recent approaches try to improve their results by assuming particular motion patterns, hence improving body parts tracking. However, this also limits their application to action recognition [26].

3.2 Holistic Methods

Holistic methods do not require the localization of body parts. Instead, global body structure and dynamics are used to represent human actions [26]. The main idea is that, global dynamics are discriminative enough to characterize human actions, given a region of interest centered on the human body. Moreover, holistic representations are much simpler than other approaches that explicitly use a kinematic model or information about body parts, since they only model global motion and appearance information. Therefore their computation is in general more efficient and robust. This characteristic is especially important for realistic videos in which background clutter, camera ego-motion, and occlusion make the localization of body parts difficult. In general, holistic approaches can be roughly divided into two categories. The first category employs shape masks or silhouette information, stemming from background subtraction or difference images, to represent actions [2, 29]. The second category is mainly based on shape and optical flow information [30, 37].

3.3 Local Feature Methods

Local space-time features keep characteristic shape and motion information for a local region in video. They provide a relatively independent representation of events with respect to their spatio-temporal shifts and scales as well as background clutter and multiple motions in the scene. These features are usually extracted directly from video and hence avoid possible failures of other pre-processing methods such as motion segmentation or human detection [31]. In the literature, various approaches are proposed under this category [34, 36, 39].

4. ACTION DATASETS

The methods of evaluating the performance of object detection, object tracking, object classification, and behavior detection and identification in a visual surveillance system are more complex than some of the well-established biometrics identification applications, (such as fingerprint or face), due to uncontrolled environments and the complexity of variations found in the same scene [13]. Due to the increasing research in video surveillance systems over the last years, there are several public datasets that try to evaluate the performance of such systems. These datasets are necessary to fairly evaluate algorithms under different conditions and to compare new algorithms with existing ones. The datasets used to evaluate the behavior analysis work can be classified into three categories: surveillance datasets (such as PETS and ViSOR), video retrieval datasets (such as TRECVid), and action recognition datasets (such as Weizmann, KTH, UCF, YouTube, and Hollywood) [1, 9, 26]. Also, there are some other datasets designed only for specific surveillance applications: driver assistance systems, people detection walking through a busy pedestrian zone, very specific scenarios or even very general video security systems. It should be mentioned that among all these categories of datasets, action recognition datasets are used widely to evaluate the behavior analysis work.

PETS dataset (Performance Evaluation for Tracking and Surveillance) (http://www.cvg.rdg.ac.uk/slides/pets.html) is a good starting place when looking into performance evaluation (see Fig 6.a). PETS has several good datasets for both indoor and outdoor tracking evaluation and event/behavior detection. PETS datasets include outdoor people and vehicle tracking using single or multiple cameras, indoor people tracking (and counting) and hand posture classification, annotation of a smart meeting, including facial expression, gaze and gesture/action, multiple sensor (camera) sequences for unattended luggage, multiple sensor (camera) sequences for attended luggage removal (theft), and multiple sensor (camera) sequences for loitering.

In addition to PETS datasets, ViSOR dataset (Video Surveillance Online Repository) (http://imagelab.ing.unimore.it/visor/) is a video repository, designed with the aim of establishing an open platform for collecting, annotating, retrieving, and sharing surveillance videos, as well as evaluating the performance of automatic surveillance systems (see Fig 6.b). The repository is free and researchers can collaborate sharing their own videos or datasets. Most of the included videos are annotated. It is not just one dataset for one specific topic but it includes a lot of videos for different video surveillance applications.

CAVIAR dataset (Context Aware Vision using Image-based Active Recognition) (http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/) includes a number of recorded video clips acting out the different scenarios of interest (see Fig 6.c). These include people walking alone, meeting with others, window shopping, fighting and passing out, and leaving a package in a public place. All video clips were filmed with a wide angle camera lens. CAVIAR consists of two sets of video clips. The first set was filmed in the entrance lobby of the INRIA Labs at Grenoble, France. The second set of data was filmed along and across the hallway in a shopping center in Lisbon, Portugal.

Also, there are efforts, like TRECVid evaluation datasets (http://trecvid.nist.gov/), with the goal to support the development of technologies to detect visual events through standard test datasets and evaluation protocols (see Fig 6.d). In fact, various types of data have been involved in the TRECVid workshops and the availability of these datasets varies by type and year. Originally, the TREC (Text REtrieval Conference) conference series aim to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. In 2001 and 2002 the TREC series sponsored a video "track" devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, this track became an independent evaluation (TRECVID - TREC Video Retrieval Evaluation) with a workshop taking place just before TREC.

Table 2 shows different datasets used for action recognition. The Weizmann actions dataset (http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html) includes 10 different types of action classes: bending downwards, running, walking, skipping, jumping-jack, jumping forward, jumping in place, galloping sideways, waving with two hands, and waving with one hand (see Fig. 6.e). Each action class is performed once (sometimes twice) by 9 subjects. In total, the dataset consists of 93 video sequences. The background in the videos is homogeneous and static. In the original experimental setup by the authors, testing is carried out using leave-one-out cross-fold validation approach, i.e., testing is performed for one sequence at a time while training is executed on all remaining sequences. Performance is given in terms of average accuracy (error rate).

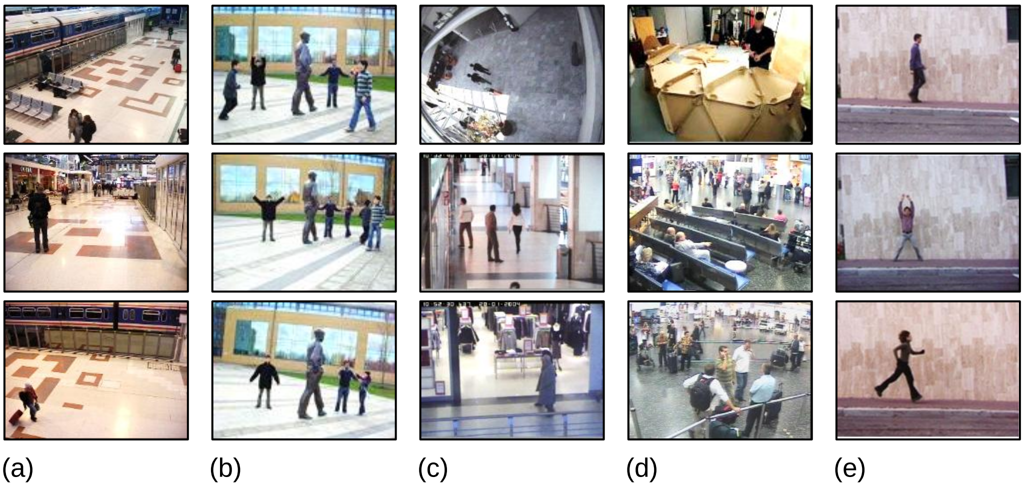


Fig. 6. Example frames of (a) PETS dataset, (b) ViSOR dataset, (c) CAVIAR dataset, (d) TRECvid dataset and (e) Weizmann dataset

Table 2. Different Datasets Used for Action Recognition

Action dataset	**Dataset name comes from**	**No of action classes**	**No of video samples**
Weizmann	The name of the institute in which the dataset is prepared (Weizmann Institute of Science)	10	93
KTH	The name of the university in which the dataset is prepared (Swedish: Kungliga Tekniska högskolan, abbreviated KTH). It means “The Royal Institute of Technology”	6	2391
UCF	The name of the university in which the dataset is prepared (University of Center Florida)	10	150
YouTube	Its video samples are collected from YouTube	11	1600
Hollywood1	Its video samples are collected from Hollywood movies	8	663
Hollywood2	Its video samples are collected from Hollywood movies	12	2517

The KTH actions dataset (http://www.nada.kth.se/cvap/actions/) consists of 6 different human action classes: walking, jogging, running, boxing, waving, and clapping (see Fig. 7.a). Each action class is performed several times by 25 subjects resulting in 2391 video samples in total. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. The background is homogeneous and static in most sequences. In the original experimental setup by the authors, samples are divided into a test set (9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22) and training set (the remaining 16 subjects). Evaluation on this dataset is done via multi-class classification. Classification performance is evaluated as average accuracy over all classes.

The UCF sport actions dataset (http://crcv.ucf.edu/data/UCF_Sports_Action.php) contains 10 different types of human actions: swinging (on the pommel horse and on the floor), diving, kicking (a ball), weight-lifting (see Fig. 7.b), horse-riding, running, skateboarding, swinging (at the high bar), golf swinging and walking. The dataset consists of 150 video samples that show a large intra-class variability. The performance criterion for the multi-class task is the average accuracy over all classes. The original setup employs leave-one-out approach for testing.

The YouTube dataset (http://crcv.ucf.edu/data/UCF_YouTube_Action.php) contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog (see Fig. 7.c). This dataset is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions etc. It contains a total of 1600 video sequences. In the original setup, the evaluation is performed using cross validation for a set of 25 folds that is defined by the authors. Average accuracy over all classes is used as performance measure.

The Hollywood1 dataset (http://www.di.ens.fr/~laptev/actions/) contains 8 different action classes: answering the phone, getting out of the car, hand shaking, hugging, kissing, sitting down, sitting up, and standing up (see Fig. 7.d). Action samples have been collected from about 32 different Hollywood movies. In total, the full dataset contains 663 video samples, divided into a clean training set (219 sequences) and a clean test set (211 sequences), where training and test sequences were obtained from different movies. The additional noisy training set consists of 233 sequences.

Hollywood2(http://www.di.ens.fr/~laptev/actions/hollywood2/) is the extended version of Hollywood1 dataset (see Fig. 7.e). It consists of video samples collected from 69 different Hollywood movies. The initial eight action classes were extended by adding four additional ones: driving car, eating, fighting, and running. In total, there are 2517 action samples split into a manually cleaned training set (823 sequences) and a test set (884 sequences). The noisy training set contains 810 sequences. Train and test sequences are obtained from different movies. The performance for both, Hollywood1 and Hollywood2, is evaluated by computing the average precision for each of the action classes and reporting the mean AP over all classes..

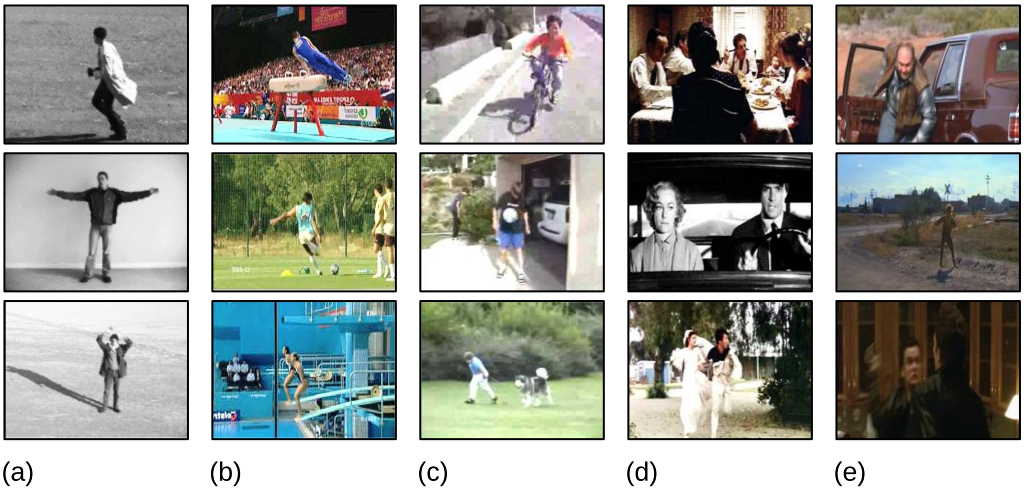


Fig. 7. Example frames of (a) KTH dataset, (b) UCF sports action dataset, (c) YouTube dataset, (d) Hollywood1 dataset and (e) Hollywood2 dataset

Table 3. State-of-the-art results on different datasets reported as average accuracy achieved

Year	Method	Dataset used	Accuracy achieved	Year	Method	Dataset used	Accuracy achieved
2008	Fathi et al. [32]	UCF	71.00%	2012	Nowozin et al. [45]	Weizmann	90.00%
2009	Chen et al. [33]	UCF	66.00%	2012	Nagendar et al. [46]	UCF	93.50%
2009	Gilbert et al. [34]	Hollywood1	53.50%			KTH	98.90%
		KTH	94.50%			YouTube	86.60%
2009	Han et al. [35]	Hollywood1	47.50%	2012	Acar et al. [47]	Weizmann	96.03%
		KTH	94.10%			KTH	87.84%
2009	Wang et al. [36]	YouTube	86.60%	2013	Sadek et al. [48]	Weizmann	98.00%
2010	Guo et al. [37]	Weizmann	97.40%			KTH	93.50%
2010	Ali et al. [38]	Weizmann	87.70%	2013	Hernández et al. [49]	Weizmann	96.66%
2010	Kovashka et al. [39]	Weizmann	94.50%	2013	Sun et al. [50]	KTH	78.60%
		YouTube	87.27%	2013	Vrigkas et al. [51]	Weizmann	98.80%
2010	Lui et al. [40]	Weizmann	97.00%			KTH	96.71%
2010	Bregonzio et al. [41]	YouTube	86.90%	2013	Goudelis et al. [52]	Weizmann	95.42%
						KTH	93.14%
2011	Seo et al. [42]	Weizmann	95.10%	2013	Liu et al. [53]	KTH	95.5%
2011	Oshin et al. [43]	Weizmann	89.30%	2013	Wang et al. [54]	KTH	90.7%
2011	Lui et al. [44]	Weizmann	97.00%	2013	Derpanis et al. [55]	UCF	81.50%
		YouTube	88.00%			Hollywood2	48.40%

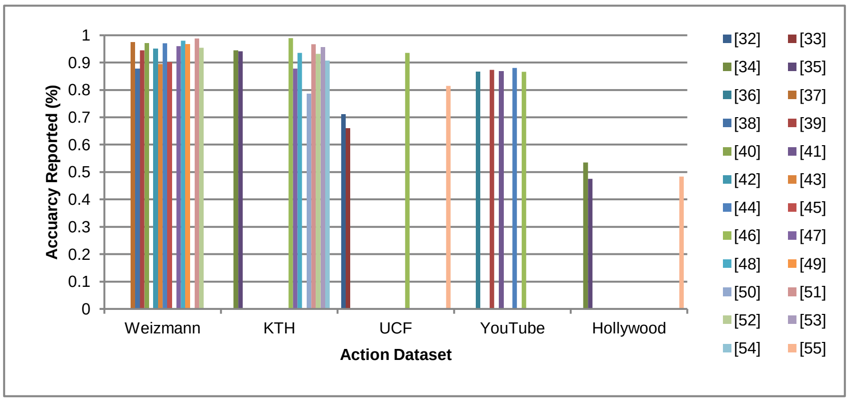


Fig. 8. Accuracy reported for each action dataset

Table 3 shows comparative results of different methods on different datasets. The two popular datasets, which are currently used by most of the approaches are: Weizmann and KTH datasets, however both are all not very realistic and share strong simplifying assumptions, such as static background, no occlusions, given temporal segmentation, and only a single actor [13]. Note that several authors report high performance exceeding 90% for both Weizmann and KTH datasets while the performance is degrading for UCF, YouTube, and Hollywood datasets (Fig. 8). That is, the UCF sports dataset is a collection of TV sport events. It offers a large variety of action classes while being limited in its size. Also, the YouTube and Hollywood datasets are considered the most challenging and extensive datasets published in the literature.

Although behavior analysis techniques perform rather strongly in selected datasets, a real-world behavior analysis is still extremely challenging, due to complicated environments, cluttered backgrounds, occlusions, illumination changes, multiple activities, and numerous deformations of an activity. Simplifying the problem by adding more assumptions may significantly improve the results but it will limit its applicability in real world. The algorithms developed therefore often have specific strengths and limitations, and are designed for particular domain. A particular algorithm may be optimal for a specific application and may perform effectively without modification. However, due to the complex nature of many environments, adaptive and/or hybrid forms of existing behavior representation and action recognition approaches may best be able to meet the needs of dynamically changing conditions. This survey has identified the need for further research in this direction, which will require a comprehensive analysis of the specific environment, and its dynamic nature, prior to the determination of optimal combinations taking into account the real-time challenge.

5. CONCLUSION

Recently, the automatic surveillance system has extremely progressed due to the high applicability in public institutions, private firms, and houses. Hence, the area of behavior analysis in video surveillance is a hot issue of extensive research. In this paper, we present a survey of some of the important studies in the area by grouping them in consistent contexts. We have classified approaches with respect to how they represent the actions, and how they recognize actions from a video stream. Although many proposed behavior analysis techniques perform strongly in selected datasets, a real- world surveillance video archive is still extremely challenging, due to complicated environments, cluttered backgrounds, occlusions, illumination changes, multiple activities, and numerous deformations of an activity. Behavior analysis still faces great difficulties, including variance in the appearance of particular events, similarity in the appearance of different events, lack of specific background information, which may contain large amount of prior knowledge, etc.

Despite clear advances in the field of action recognition, evaluation of these methods remains mostly heuristic and qualitative. Most of the datasets do not include ground-truth and detailed pose information for the researchers. There is a need to find some meaningful datasets and areas to work, rather than keeping efforts in trivial action recognition datasets.

To sum up, with investigating all of dominant algorithms which are widely used in behavior analysis, the survey reveals important progress made in the last five years. However, many issues are still open and deserve further research. Future work needs to come up with more efficient ways to detect complex actions where there is some interaction between different blobs in an environment. Also, more research should be directed to solve difficulties in behavior detection such as the strong appearance variation in semantically similar events (e.g., people performing actions with different clothing), the viewpoint variation, and the duration of the action. Finally, robust and realistic surveillance datasets are needed to effectively evaluate different proposed methods.

This survey can be considered as a starting point for those interested in pursuing further work in this area and it suggests that further exploration is still required. Behavior analysis will continue to remain an active research area since the computer understanding of behavior is exceedingly important for many military and civil applications.

6. REFERENCES

[1] NemanjaSpasic, "Anomaly Detection and Prediction of Human Actions in a Video Surveillance Environment," Master Thesis, Computer Science Dept., Cape Town University, South Africa, December 2007.

[2] MahfuzulHaque, and ManzurMurshed, "Robust Background Subtraction Based on Perceptual Mixture-of-Gaussians with Dynamic Adaptation Speed," In

Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops (ICMEW '12), Melbourne, Australia, pp.396-401, July 2012.

[3] H. Wang, and Paul Miller, "Regularized Online Mixture of Gaussians for Background Subtraction," In Proceedings of the 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '11), Klagenfurt, Austria, pp. 249-254, September 2011.

[4] Vikas Reddy, Conrad Sanderson, Andres Sanin, and Brian C. Lovell, "MRF-Based Background Initialisation for Improved Foreground Detection in Cluttered Surveillance Videos," In Proceedings of the 10th Asian Conference on Computer Vision (ACCV'10), Queenstown, New Zealand, Volume Part III, pp. 547-559, November 2010.

[5] PawelForczmanski, and MarcinSeweryn, "Surveillance Video Stream Analysis Using Adaptive Background Model and Object Recognition," In Proceedings of the 2010 International Conference on Computer Vision and Graphics (ICCVG'10), Warsaw, Poland, Part I, pp. 114–121, September 2010.

[6] Chris Poppe, Gaetan Martens, Peter Lambert, and Rik Van de Walle, "Improved Background Mixture Models for Video Surveillance Applications," In Proceedings of the 8th Asian Conference on Computer Vision (ACCV'07), Volume Part I, pp. 251–260, November 2007.

[7] Liman Liu, Wenbing Tao, Jin Liu, and JinwenTian, "A Variational Model and Graph Cuts Optimization For Interactive Foreground Extraction," In the Signal Processing Journal, Volume 91, Issue 5, May 2011.

[8] Xavier Suau, Josep R. Casas, and Javier Ruiz-Hidalgo, "Multi-Resolution Illumination Compensation for Foreground Extraction," In Proceedings of the 16th IEEE International Conference on Image Processing (ICIP'09), pp. 3189-3192, November 2009.

[9] Chen Change Loy, "Activity Understanding and Unusual Event Detection in Surveillance Videos," PhD dissertation, Queen Mary University of London, 2010.

[10] YanivGurwicz, RaananYehezkel, Boaz Lachover, "Multiclass Object Classification for Real-time Video Surveillance Systems," In Pattern Recognition Letters Journal, Volume 32, issue 6, pp.805–815, April 2011.

[11] Bahadır KARASULU, "Review and Evaluation of Well-Known Methods for Moving Object Detection and Tracking in Videos," In the Journal of Aeronautics and Space Technologies, Volume 4, Number 4, pp.11-22, July 2010.

[12] Wenming Yang, Fei Zhou1, and Qingmin Liao, "Object Tracking and Local Appearance Capturing in a Remote Scene Video Surveillance System with Two Cameras," In Proceedings of the 16th International Multimedia Modeling Conference (MMM 2010), Chongqing, China, pp. 489–499, January 2010.

[13] Daniel Weinland, RemiRonfard, and Edmond Boyer, "A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition," In the Journal of Computer Vision and Image Understanding, Volume 115, Issue 2, pp. 224-241, February, 2011.

[14] Frederick Tung, John S. Zelek, and David A. Clausi, "Goal-Based Trajectory Analysis for Unusual Behaviour Detection in Intelligent Surveillance," In the Journal of Image and Vision Computing, Volume 29, Issue 4, March 2011.

[15] A. F. Bobick, and J. W. Davis, “The Recognition of Human Movement Using Temporal Templates,” In IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 23, Number 3, pp.257–267, 2001.

[16] S. Gong, and T. Xiang, “Scene Event Recognition Without Tracking,” In ActaAutomaticaSinica Journal, Volume 29, Number 3, pp. 321–331, 2003.

[17] E. L. Andrade, S. Blunsden, and R. B. Fisher, “Hidden Markov Models for Optical Flow Analysis in Crowds,” In Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06), pp. 460-463, 2006.

[18] N. Dalal, and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” In Proceedings of the International IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893, June 2005.

[19] L. Zelnik-Manor, and M. Irani, “Statistical Analysis of Dynamic Actions,” In IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 28, Number 9, pp. 1530–1535, 2006.

[20] L. Kratz and, K. Nishino, “Anomaly Detection in Extremely Crowded Scenes Using Spatiotemporal Motion Pattern Models,” In IEEE Conference on Computer Vision and Pattern Recognition, pp. 1446–1453, 2009.

[21] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly Detection in Crowded Scenes,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1975 - 1981, June 2010.

[22] J. Kim, and K. Grauman, “Observe Locally, Infer Globally: A Space-Time MRF for Detecting Abnormal Activities with Incremental Updates,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2928, 2009.

[23] T. Xiang, and S. Gong, “Video Behaviour Profiling for Anomaly Detection,” In IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 30, Number 5, pp. 893–908, 2008.

[24] S. Park, and M. M. Trivedi, “A Two-Stage Multi-View Analysis Framework for Human Activity and Interactions,” In Proceeding of IEEE Workshop on Motion and Video Computing, University of California, San Diego, pp. 29-34, February 2007.

[25] Rodrigo Cilla, Miguel A. Patricio, Antonio Berlanga, and José M. Molina, "Human Action Recognition with Sparse Classification and Multiple-View Learning," In Expert Systems Journal, Wiley Publishing Ltd, August 2013.

[26] Ronald Poppe, “A Survey on Vision-Based Human Action Recognition,” In Image and Vision Computing Journal, Volume 28, Issue 6, pp. 976–990, June 2010.

[27] V. Parameswaran, and R. Chellappa, “View Invariance For Human Action Recognition,” In the International Journal of Computer Vision, Volume 66, Issue 1, pp. 83-101, January 2006.

[28] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, “Progressive Search Space Reduction for Human Pose Estimation,” In Proceedings of the 2008 IEEE

Conference on Computer Vision and Pattern Recognition (CVPR 2008), June 2008.

[29] Ziming Zhang, Yiqun Hu, Syin Chan and Liang-Tien Chia, “Motion Context: A New Representation for Human Action Recognition” In the European Conference on Computer Vision (ECCV), Marseille, France, October 2008.

[30] M. Rodriguez, J. Ahmed, and M. Shah, “Action MACH a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition,” In the IEEE Conference on Computer Vision and Pattern Recognition Action (CVPR 2008), pp. 1 - 8, June 2008.

[31] Shu-Fai Wong, "Extracting Spatiotemporal Interest Points Using Global Information," In Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV 2007), pp. 1-8, October 2007.

[32] Alireza Fathi, and Greg Mori, "Action Recognition by Learning Mid-Level Motion Features," In Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, Alaska, USA, pp. 1-8, June 2008.

[33] Chia-Chih Chen, J. K. Aggarwal, "Recognizing Human Action From a Far Field of View," In Proceeding of the 2009 International Conference on Motion and Video Computing (WMVC'09), Snowbird, Utah, pp. 119-125, December 2009.

[34] Andrew Gilbert, John Illingworth, and Richard Bowden, "Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features," In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, Japan, pp. 925-931, October 2009.

[35] Dong Han, Liefeng Bo, and Cristian Sminchisescu, “Selection and Context for Action Recognition,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV), University of Bonn, Bonn, Germany, pp. 1933 - 1940, September 2009.

[36] Heng Wang, Muhammad MuneebUllah, Alexander Kläser, Ivan Laptev, CordeliaSchmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," In Proceedings of the British Machine Vision Conference (BMVC 2009), London, UK, pp. 124-134, September 2009.

[37] Kai Guo, PrakashIshwar, and JanuszKonrad, "Action Recognition Using Sparse Representation on Covariance Manifolds of Optical Flow," In Proceedings of the Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2010), pp. 188 - 195, September 2010.

[38] Saad Ali, and Mubarak Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," In the IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 32 , Issue 2, pp. 288 - 303, February 2010.

[39] Adriana Kovashka, and Kristen Grauman, "Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, California, USA, pp. 2046-2053, June 2010.

[40] Yui Man Lui, J. Ross Beveridge, and Michael Kirby, “Action Classification on Product Manifolds,” In Proceedings of the Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, CA, USA, pp. 833-839, June 2010.

[41] Matteo Bregonzio, Jian Li, Shaogang Gong, and Tao Xiang, "Discriminative Topics Modeling for Action Feature Selection and Recognition", In Proceedings of the British Machine Vision Conference (BMVC 2010), Aberystwyth, UK, pp.1-11, September 2010.

[42] Hae Jong Seo, "Action Recognition from One Example," In IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 33, Issue 5, pp. 867 - 882, May 2011.

[43] Olusegun Oshin, Andrew Gilbert, and Richard Bowden "Capturing the Relative Distribution of Features for Action Recognition," In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG 2011), Santa Barbara, California, USA, pp. 111-116, March 2011.

[44] Yui Man Lui, and J. Ross Beveridge, "Tangent Bundle for Human Action Recognition," In Proceedings of the Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, California, USA, pp. 97-102, March 2011.

[45] Sebastian Nowozin, Jamie Shotton, "Action Points: A Representation for Low-latency Online Human Action Recognition", Technical Report, Microsoft Research Cambridge, July 2012

[46] G Nagendar, Sai Ganesh, Mahesh Goud and C.V Jawahar "Action Recognition using Canonical Correlation Kernels," In Proceedings of the 11th Asian Conference on Computer Vision (ACCV 2012), Daejeon, Korea, November 2012.

[47] Esra Acar, Tobias Senst, Alexander Kuhn, Ivo Keller, Holger Theisel, Sahin Albayrak, and Thomas Sikora, "Human Action Recognition using Lagrangian Descriptors," In the 2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP), Banff, Canada, pp. 360-365, September 2012.

[48] Samy Sadek, Ayoub Al-Hamadi, Gerald Krell, and Bernd Michaelis, "Affine-Invariant Feature Extraction for Activity Recognitio," ISRN Machine Vision Journal, volume 2013, Article ID 215195, 7 pages, July 2013.

[49] J. Hernández, R. Cabido, A. S. Montemayor and J.J. Pantrigo "Human Activity Recognition Based on Kinematic Features", Expert Systems Journal, doi: 10.1111/exsy.12013, Volume 2013, February 2013.

[50] Tanfeng Sun, Xinghao Jiang, Chengming Jiang, Yaqing Li, "A Video Content Classification Algorithm Applying to Human Action Recognition," In the journal of Electronics and Electrical Engineering (Elektronika ir Elektrotechnika), ISSN 1392-1215, DOI: http://dx.doi.org/10.5755/j01.eee.19.4.4056, Volume 19, Number 4, 2013

[51] Michalis Vrigkas, Vasileios Karavasilis, Christophoros Nikou and Ioannis Kakadiaris, "Action Recognition by Matching Clustered Trajectories of Motion Vectors," In Proceedings of the 8th International Conference on Computer Vision Theory and Applications, Barcelona, Spain, February 2013.

[52] Georgios Goudelis, Konstantinos Karpouzis, and Stefanos Kollias, "Exploring Trace Transform for Robust Human Action Recognition," In Pattern Recognition

Journal, Volume 46, Issue 12, pp. 3238–3248, December 2013.

[53] Li Liu, Ling Shao, and Peter Rockett, "Boosted Key-Frame Selection and Correlated Pyramidal Motion-Feature Representation for Human Action Recognition," In Pattern Recognition Journal, Volume 46, Issue 7, pp. 1810–1818, July 2013

[54] Li Wang, Ting Yun, and Haifeng Lin, "Boost Action Recognition through Computed Volume," In the Electrical Engineering Journal (TELKOMNIKA), Volume 11, Number 4, pp. 1871-1876, April 2013.

[55] Konstantinos G. Derpanis, Mikhail Sizintsev, Kevin J. Cannons, and Richard P. Wildes, "Action Spotting and Recognition Based on a Spatiotemporal Orientation Analysis," In the IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume $_{35}$, Issue 3, pp. 527 - 540, March 2013