Human action recognition with video data : research and evaluation challenges

Ramanathan, Manoj; Yau, Wei-Yun; Teoh, Eam Khwang

2014

Ramanathan, M., Yau, W.-Y., & Teoh, E. K. (2014). Human action recognition with video data : research and evaluation challenges, 44(5), 650 - 663.

https://hdl.handle.net/10356/102292

https://doi.org/10.1109/THMS.2014.2325871

Human action recognition with video data: Research and evaluation challenges

Manoj RAMANATHAN, *Student Member, IEEE,* Wei-Yun YAU, *Senior Member, IEEE,* and Eam Khwang TEOH, *Member, IEEE*

***Abstract*—Given a video sequence, the task of action recognition is to identify the most similar action among the action sequences learned by the system. Such human action recognition is based on evidence gathered from videos. It has wide application including surveillance, video indexing, biometrics, telehealth and human computer interaction. Vision-based human action recognition is affected by several challenges due to view changes, occlusion, variation in execution rate, anthropometry, camera motion and background clutter. In this survey, we provide an overview of the existing methods based on their ability to handle these challenges as well as how these methods can be generalized and their ability to detect abnormal actions. Such systematic classification will help researchers to identify the suitable methods available to address each of the challenges faced and their limitations. In addition, we also identify the publicly available datasets and the challenges posed by them. From this survey, we draw conclusions regarding how well a challenge has been solved and we identify potential research areas that require further work.**

***Index Terms*—Action recognition, view-invariance, execution rate, anthropometric variations, camera motion**

I. Introduction

Human action is not merely the pattern of motion of various body parts, but is the real world depiction of the person's intentions and thoughts. It is an important component in behaviour analysis and understanding which are essential for many applications, such as human computer interaction, surveillance, telehealth, biometrics, video indexing, training or virtual coaching etc.

'Action recognition' as the word suggests, means recognition of an action by using a system that typically analyses the video sequence to learn about the action and uses the learnt knowledge to identify similar actions. [1] broadly classifies human activities into four categories: gestures, actions, interactions (with objects and others) and group activities. Automatic recognition of such complex actions and behaviours has led to development of useful applications such as virtual coaches [2], understanding user environments and behaviours using wearable sensors [3], evaluation of robotic therapy as biofeedback device in dementia care [4] and development of home assistant robots for ageing society [5].

Actions recognition can be tackled using several strategies, namely, 3D markers [6] and wearable sensors [3], [7]. Apart from these strategies, video or images of a person's action provide ample clues. There have been several methods proposed in each of these categories. Since action recognition is a vast domain, in this survey we restrict ourselves to methods that gather evidence from action videos or images. Research in this domain has seen significant progress but is still impaired by several bottlenecks which include variation in viewpoint, occlusion, execution rate or speed, anthropometric variations, intra-class variations, camera motion and cluttered backgrounds. Other challenges include developing a generalized method to recognize any action, collection of adequate number of training samples and localizing the action spatially and temporally in video segments. The objective of this survey is to classify the methods according to their robustness to these challenges. Such taxonomy can help to identify shortcomings in the existing techniques.

The previous surveys classify the reported methods based on the features and classifiers used [8], [9] or framework adopted [1]. [8] described challenges in the domain that influence the choice of representation and classification algorithm. On the other hand [10] worked on a specific challenge, by reviewing recent advances in view-invariant action recognition and considering three issues namely, human detection, view-invariant pose detection and behaviour understanding. Unlike other surveys, our intention is not to classify the available methods according to any factors but to understand how well the challenges in action recognition domain have been solved. Therefore we classify the methods according to the challenges they can handle effectively.

In this survey, we mainly focus on vision based action recognition systems that use video camera as the primary sensor and incorporate video analysis component used to determine the action in the video. Excluded are methods that use wearable sensors or depth sensor such as Kinect as the primary sensor. We searched IEEE, ScienceDirect and Elsevier databases initially for previous reviews in the field and also collected other papers by searching these databases using keywords such as action recognition, activity recognition and vision based action recognition. We widened our search using publications cited in these collected papers. We identify the various challenges in vision-based action recognition and classify the selected papers according to their robustness to these challenges. Some non-vision methods are included in our discussion (in Section III) to compare how well such methods could be generalized for practical applications compared to the vision based approaches.

The rest of the paper is organized as follows: Section II deals with each of the challenges mentioned above and methods

Manuscript received on ,

M.Ramanathan and E.K. Teoh are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. E-mail: manoj005@e.ntu.edu.sg, eekteoh@ntu.edu.sg W.Y. Yau is with the Institute for Infocomm Research, A*STAR, Singapore. E-mail: wyyau@i2r.a-star.edu.sg.

used by the researchers to tackle them. Action classification methodologies and training strategies are discussed in section III. This is followed by section IV, which deals with datasets used for testing the challenges. Section V then concludes the paper and suggests potential research opportunities where more works need to be done to improve the usability of action recognition in day-to-day applications.

II. Challenges in Vision-based Action Recognition

In this section we list out some of the research challenges faced in action recognition and outline the different methods used by researchers to handle them. A quantitative performance comparison of the proposed techniques is difficult since datasets and testing strategy used vary significantly. Nevertheless, the amount of training data and ability to generalize the method to any type of action can be used as benchmarks to classify the methods as these are critical for successful real-world deployment.

A. Variation in View point

Most methods assume that action is performed from a fixed view point. Figure 1 shows why researchers tend to make this assumption. The figure shows four views of an actor performing walking action. In each camera angle the location and posture of the person varies considerably. Also motion patterns in each view would appear different, making recognition of the action not so trivial.

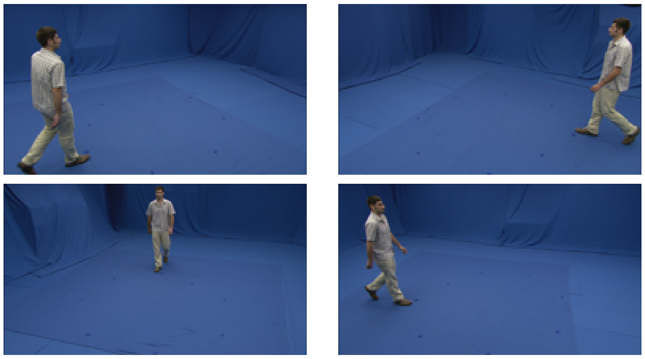

Fig. 1. Images depicting walking action from i3DPost [11] multi view dataset

The most common solution to tackle the change in camera view angle is to train the classifier using multiple camera views ([12]–[20]). In this approach, the viewing angle is discretized into evenly spaced segments or divisions each of which are captured by a camera. The features that are extracted from each view are combined to train a single classifier that handles changes in view point or used to train a set of classifiers, each one of them for a specific view point. Quantizing the space of viewpoints leads to several view-dependent representations of a single body pose, which causes many limitations during implementation [10].

This is a crude solution because the framework developed for one view is directly extended to a number of views. Once the extension is done, the performance of the method only depends on the features used to describe the action and the trained classifier. [12], [16] used motion history image (MHI) and motion energy image (MEI) to capture the underlying human motion in the images. It could be disrupted by the motion of background objects and also when more than one person is present in the camera's view. Methods such as [14], [15], [17] also require precise extraction of silhouettes for accurate results. These methods tend to ignore other challenges and assume that the background subtracted images or silhouettes are available. As a result the datasets used are mostly in a controlled environment setting (such as indoors, no illumination changes etc). Thus it is difficult to use such methods in a normal environment. Another setback is that multiple cameras have to be synchronized and processed at the same time. This tends to increase the complexity and computation time.

In order to obtain view invariant representation of the action, researchers also model the 3D or 2D body posture in an action ([14], [17], [21]–[33]). Human body is highly deformable and composed of several parts, each capable of undergoing separate deformations. But these deformations are constrained because of the underlying kinematic and skeletal structure of the body. The characterization of an action using 3D or 2D models depends on the representation used. For example, visual hulls ([28], [29]), envelope shape ([31]–[33]) and silhouettes ([14], [17]) have been used. Some researchers [24], [27] have also proposed detecting separate body parts and then combining the results to obtain the body model.

One main concern in these modelling approaches is how to track the changes in posture and reflect the same on the models created. Kalman filtering [14], object interactions [25], motion and appearance model [27] have been employed to track body parts. The use of object interactions for tracking limits the adoption of such methods to a very small class of actions where the type and number of objects are already known. [22], [26] try to track the dynamics involved in the action by selecting only the key frames that depict maximum pose variations. However, the main disadvantage is that dynamics of the whole action is not considered and is completely dependent on the key frames chosen.

Human body is highly articulated with many degrees of freedom to support complex movements. Creating a comprehensive model that can represent these movements using evidence collected from images is thus very challenging since it poses a problem in a very high dimensional space. The efforts to model in lower dimensions [21] (only 5 body joints used), [27], [24] (uses 10 and 5 body parts respectively) could only be an approximation and such reduced dimensional space might not be able to model the complex movements. Moreover, the representation obtained in the method [21] is not unique for each action; also testing has been done only on walking motion.

Lastly, researchers try to use view-invariant features or features cast in a space or model where view-invariant matching is possible ([30], [34]–[45]). Radon transform and Radon transform surface ([43], [45]), temporal self-similarities and dissimilarities that exist in an action [41], fundamental ratios for fixed cameras [30] and temporal order of action sequences [42] have been used successfully for recognition of certain limited actions. All these methods also have limitations based on the features. [43], [45] uses only shape information. Identifying subtle differences in actions and differentiating similar actions such as walking, running etc. would be difficult since motion features have been ignored. Fundamental ratios, which are the ratios among the elements in the upper left 2x2 sub matrix of the fundamental matrix of a camera, and is invariant to camera parameters are used in [30]. But this method is limited because it identifies actions by looking for similar planar motions from varying viewpoints. In [41], the descriptor is not strictly view invariant but experiments have shown that the method can handle very large changes in view.

Fourier transform and cylindrical co-ordinate systems ([40], [44]) are the most popular spaces where the features are cast for view-invariant matching. This is because a change in viewpoint is converted into a much simpler translation that can be easily computed and matched. Although wavelets filters in [44] helps in image denoising and background segmentation, a separate Maximum Average Correlation Height (MACH) filter must be synthesized for every action we want to recognize. Each of the methods mentioned design features in order to achieve view invariance but the other challenges are mostly ignored.

Among all these approaches, 3D modelling of the human body postures shows more robustness towards change in viewpoint. One reason is that all actions are represented using features that are extracted with human body as the reference frame. Using human body as reference helps to characterize actions in relative terms rather than absolute. But main concern is that inferring accurate 3D poses from the 2D images or video frames is difficult due to the large number of parameters that needs to be estimated and perspective projection involved may result in ambiguous poses to be recovered [10]. Another observation is that most multi-view datasets on which researchers ([17], [22], [28], [42], [45]) have reported around 80-90% recognition rates contain relatively uniform or fixed background. In order to truly gauge the performance of various methods it would be necessary to test those using actions recorded in real world settings.

B. Occlusion

The existing systems require that the action being performed be clearly visible in the video sequences. In a normal surveillance video, this is not possible because of the number of people in the field of view of the camera. Occlusions can be either self-occlusions or those created by other objects in the field of view of the camera during the video capture. This poses a big challenge to the research community because not all the body parts performing the action are visible in the video sequence.

Representing and analysing action as space time volumes can avoid limitations caused by occlusion [37]. Volumetric analysis ([29], [37], [46]–[51]) of actions are robust to self-occlusions because of two reasons. Firstly, the body parts are usually not occluded in the entire video. Secondly, by analysing spatio-temporal volumes as a whole, features extracted from other body parts in the entire interval when they are not occluded would be enough to match and classify the action. Employing local features such as local appearance model [52], edge maps [53], and space time interest points ([54]–[63]) in volumetric analysis would also solve the problem of cluttered backgrounds. Local representations create appearance models that can characterize the action in small patches. Using local features makes the method able to withstand pose, shape and illumination changes [64] compared to using the entire global features. The selection of appropriate patches or interest points that should be used to represent action is still a hurdle for researchers. A normal interest point detector [56] might wrongly identify local patches which are not in the foreground object. [46] used the Bhattacharyya coefficient to match the space time volumes which is robust to outliers (such as occlusions).

Probabilistic based methods such as Bayesian networks ([25], [65]–[67]) and Hidden Markov models (HMM) ([20], [25], [67], [68]) can also be used to create appearance models to model the limbs, heads and torso. These approaches consider the configuration of each body part as a state in their model and change of states is governed by a probability value. Since each body part can be considered as a separate entity, it also inherits the advantages of local representation. [67] segmented higher level activities into their component subactions using HMMs that are modified to handle missing data in the observation vector (occluded data such as limbs etc). A probable configuration of the body parts can be obtained by the available data. HMMs are good to represent simple actions but such a flat model cannot characterize the hierarchy and shared structure in all classes of actions [68]. Similar to probabilistic based methods, Kalman filters ([14], [27], [69]) and particle filters ([19]) also estimate the location of body parts based on the initially available data. These filters are simple methods that are used for tracking body parts.

Considering body parts separately is a plausible option to handle occlusions. Pose based constraints ([48], [70], [71]) can be imposed to locate parts that are occluded. Use of poselets ([72]–[76]) also follows the principle of body part detection. Poselets [72] were originally proposed for action recognition from static images. Therefore extension to video sequences is not so trivial. Poselets also require manually annotated training images to train the detectors for each body part. This limits the applicability of poselets for action recognition.

Researchers have also explored object interactions ([77], [78]) and multiple camera setup ([12], [16], [32]) to handle occlusions. [77], [78] used objects as image evidence to identify the location of occluded parts. Generalization of these methods to actions that do not involve objects is challenging. Similar to [77], [78], instead of using object interactions, contextual or scene related clues can be explored as a means to determine the occluded regions. However this approach can be biased towards the scene background and hence cannot be used separately to obtain reliable decisions on the location of the occluded parts. In the case of multiple camera set-up methods, [12], [16] are dependent on silhouettes which are not effective under self-occlusions.

Probabilistic methods and pose based approaches offer most acceptable performance when actions are being occluded.

They account for occluded parts by identifying most plausible location of these parts from the available image evidence and articulated human body constraints. One interesting research avenue will be to explore classifiers that can handle occlusions. Since feature extraction from occluded parts is not possible, it is important to come up with robust classifiers that can adapt according to the presence of occlusions.

C. Execution Rate

Each individual performs an action at his/her own pace. Also there is no guarantee that a person will repeat the action at the same speed every time. This variation in the rate of execution of an action has to be taken into account in an action recognition system.

Any method that provides a probabilistic framework such as Hidden Markov models ([20], [25], [43], [66]–[68], [79]–[84]), Bayesian networks ([24], [25], [65]–[67], [69], [85]–[90]), conditional random fields ([91]–[93]) and fuzzy based systems ([94]–[97]) are better suited to handle this challenge. All these modelling paradigms and probability values assigned to states govern when the state needs to be changed. Even though the action is performed at different speeds, the model will be updated only when the probability values indicate a state change. HMMs can model dynamic processes in nature [98]. Modelling complex interactions and simultaneous activities poses a challenge for researchers due to the rigid flat structure [68]. These models only provide architecture to represent action, but the performance of the system will depend on the effectiveness of the features extracted. As pointed out by experiments in [99], HMMs are more sensitive to the training examples and tend to require more training data for better performance.

One way to model and represent simultaneous activities is to employ Allen's predicate logic and past-now-future (PNF) networks ([100], [101]). These networks allow researchers to explicitly model temporal relationships, for example, *'during'*, *'before'*, *'after'* etc. PNF networks [101] represent the temporal structure of actions using an interval algebra constraint network, a constraint satisfaction network where the variables correspond to time intervals, and the arcs to binary temporal constraints between intervals. The aim of this network is to find the feasible values or minimal domain for the variables, which is a NP-hard problem. Calculations in PNF networks are easier since the number of states is reduced to three namely, *Past, Now, Future*. AND-OR graphs in conjunction with Allen's interval logic [59] provides a graphical model to encode variations in activities. This method can encode both causal and temporal relationships but requires weakly labelled data.

All of the above mentioned approaches use temporal domain to generate models. There has also been research in spatial domain by considering the whole video length. Time warping techniques ([13], [22], [23], [41], [44], [70], [102]–[105]) tend to convert the template and the given input video data into a common time scale that allows for easy comparisons. Histogram based methods, specifically bag of features, codebooks or dictionary based approaches ([50], [60], [71], [73], [75], [76], [106]–[114]) apply the same principle since they consider the image features from the complete video segments. The challenge in these methods is that they can only describe the complete actions and do not consider how the action is being performed. As such these methods cannot be directly used for action segmentation and localization. Codebooks or dictionaries developed suffer from quantization errors because the intra-class variations present are all generalized into clusters or codewords or visual words.

Considering temporal variations as intra-class variations, researchers ([12], [16], [21], [26], [30], [35], [38], [42], [115]–[120]) rely on the features used for action representation, classifier used and training data to handle the temporal variations. Changes in execution rate of an action are implicitly represented in datasets since more than one actor performs the action. Since the datasets encompass these variations, most of the methods tend to be robust to this challenge.

D. Anthropometric Variations

Each person has different body size, proportion and comfort zone while performing an action, For example, a waving gesture of a person might involve moving the hand above the head and then wave the hand but another person might not move his hand above his head and would just wave from a shoulder height. Thus researchers develop a generalized approach to capture and handle these variations.

Refer to Table I for the classification of the various methods. Since the major variations are in the pose and appearance of the actors, a common way is to avoid them and use other features such as motion, optical flow, frequency domain etc. These methods tend to ignore the shape/appearance of the subject. Motion and optical flow features are coarse and many scenes can exhibit similar flows over a short time period [118]. These features require a reliable and accurate background segmentation method to avoid the effect of background flow and motion [102], [53]. In order to reduce the effects of background variation, researchers use local feature representation [119], space time interest points [103], bag of features [106], motion trajectories ([38], [116]). These methods confine the features to smaller regions in the video frames, but they do not guarantee that the confined regions used will contain only the desired actor's motion.

Shape or appearance features are the simplest means of capturing anthropometric variations by determining the shape of the subject performing the actions. From the determined shapes, classifier is made invariant to these variations. Silhouettes, appearance models, pose based constraints and polygonal shapes are popular ways to capture the shape information. These methods are robust to variations in clothing and lighting [118]. Shape based techniques require a proper background model [118], [102], stationary cameras [118] and accurate tracking system [102]. Artefacts such as shadows, complex backgrounds, ghosts and moving objects in the videos can affect shape based methods since they affect the performance of background subtraction [121]. Silhouettes cannot identify internal motions such as motion of hands or legs within the body contour [118]. Other features such as power spectrum

TABLE I
CLASSIFICATION OF METHODS ROBUST TO ANTHROPOMETRIC VARIATIONS BASED ON FEATURES

Features	Robust Methods
Shape, Appearance	[14], [17], [21], [22], [27], [30], [36], [37], [43], [48], [52], [53], [57]–[59], [65], [69], [91], [100], [124]–[128]
Motion, frequency and others	[15], [16], [23], [24], [26], [28], [29], [34], [35], [38], [79], [81], [82], [85], [87], [103], [106], [116], [117], [119], [122], [129]–[134]
Hybrid	[25], [39], [40], [42], [47], [50], [54]–[56], [60], [67], [73], [88], [102], [107], [118], [135]

features [15] and discrete Fourier transform of image blocks [122] can handle these variations but are not robust to other challenges such as cluttered backgrounds, camera motion etc.

Using shape or motion feature alone has its own limitations and advantages. Hybrid features that combine both shape and motion information provide the right trade off. These features try to combine the advantages of both features while trying to avoid their limitations. Space time interest points ([39], [54]–[56], [60], [88]) and poselets [73] will encode the pose based constraints. Local patches thus generated are characterized by motion patterns [60], [107], optical flow [73], [123] etc. Poselets [72], [73] are also robust to internal motion. Experiments performed by the methods in Table I show they work better than considering them alone. Such methods can work very well in unconstrained, amateur videos that contain dynamic backgrounds and camera motion [107].

It can be noticed from Table II that only a few methods are not able to handle the changes introduced by this challenge. On the features side, it can be seen that hybrid features provide the best performance since they capture more information about the action than a single feature. Currently, datasets available capture anthropometric variations very easily since more than one actor is used. However, most of the datasets use controlled setting such as fixed background and stationary cameras. Therefore, it would be interesting to see the performance of these methods under unrestricted settings.

E. Camera Motion

In most action recognition cases, researchers assume static cameras, which might not be the case in unconstrained systems [62]. Camera motion severely affects motion features since erroneous and misleading motion patterns are induced in the videos. Shape features generally require a good tracking mechanism, background model and stationary cameras [102], [118]. Background subtraction required by these features is affected by moving cameras.

Researchers have focused on epipolar geometry and camera system directly to handle the variations. [36] handled non-stationary cameras by factorizing the tracking matrix into two matrices; one describing relative poses of the camera and the foreground object and the other describing the shape itself, thereby yielding invariance to camera motion. [34] extended the standard epipolar geometry to the dynamic scenes when cameras are moving using multi view geometry. They achieve this by deriving a temporal fundamental matrix and also analysing the rotation and translation motions. Even though these methods can easily handle camera motion, they require proper calibration and synchronization of cameras. Also, intrinsic and extrinsic camera parameters are required.

Another approach used to account for moving cameras is to include a motion compensation or subtraction component ([12], [82], [102]). In [102], dynamic cameras were accounted for in shape descriptors derived using appearance likelihood maps, which will be used to assign a probability of each pixel being part of a person in the bounding box. In the motion descriptor, they removed background and camera motion components by subtracting with the median of flow fields to obtain median compensated flow fields. To solve the camera motion problem, [12] proposed a method to use body centered motion field, where they subtract the motion caused by camera from the image. Videos used in [82] undergo a feature-based image alignment (homography from SIFT correspondences) to compensate for moving cameras before features are extracted. Motion compensation method is simple to implement but it might remove motion features that belong to the foreground object.

Researchers have tried to derive features that can be made invariant to camera motion. Optical flow [136], [137], and local space time features [54] account for the distortions in motion patterns caused by moving cameras by using velocity patterns observed in the video frames. [54] and [136] use velocity patterns in scale-space. This helps them to achieve invariance to scale changes as well. The difference in the partial derivatives of the local flow fields used in [137] cancels out most effects of camera motion in interest points generated. [54] (Space time interest points), [137] (motion boundary histogram) and [136] (optical flow based) all use local features to handle camera movements.

Codebook based methods ([57], [58], [88], [138]) rely on the generated dictionaries. If the newly observed local features contain patterns of scale changes and camera motion similar to those observed in the data used to form the codebook, they will be assigned to consistent memberships of the codebook [58]. Comprehensive training data is thus needed for these methods to work properly. Due to clustering, codebooks also tend to suffer from quantization errors.

On the whole, epipolar geometry based methods seem to provide the optimum solution when moving cameras are used. Since the camera system and its geometry are included while representing actions, they provide the best performance among all methods. One major drawback while comparing these methods is the non-availability of a standard dataset that contains camera motion [34]. Hence, most of the above mentioned methods ([54], [57], [58] etc) either report their performance on stationary camera datasets or create their own dataset [34], [138] for testing. Lack of benchmark dataset that portrays camera motion effectively is a major hurdle to progress in this area.

F. Cluttered Background

Dynamic or cluttered background is a form of distraction in the video sequence from the original action of interest as it introduces ambiguous information [58]. Flow based methods that calculate motion is affected as they detect unwanted background motion along with the actual required motion. Also color-based and region-based segmentation approaches require uniform non-varying background for reliable segmentation and tracking of the foreground object. To avoid the anomaly introduced, most applications assume a static background or a method to handle background segmentation from the videos prior to processing [15], [17].

Simple solutions proposed by researchers include pre-filtering or segmentation process ([53], [102], [129]) or normalize and threshold to separate foreground from background ([43], [56], [69], [79], [119], [139], [140]). Handling complex backgrounds is not so easy in these methods since they assume a uniform distribution in the background. Researchers have investigated spatio-temporal features based volumetric analysis ([54], [112], [115], [118], [130], [141]). Volumetric analysis does not rely on background subtraction or human body-part segmentation, and are relatively immune to noise, camera jitter, changing background, and variations in size and illumination [100].

Methods that model the background ([27], [52], [83], [142]) in order to separate the foreground object have been explored by researchers. Graphical models like latent semantic analysis etc ([57], [58], [88], [124]), Gaussian mixture models ([31], [82], [109], [122]) and Random Forests [143] are some of the paradigms explored. These methods are also robust to shadows and other artefacts in the videos [142]. These models provide really good background subtraction but modelling complex backgrounds over a long period of time will be difficult. Gaussian mixture models are intrinsically linear, which leads to relatively large fitting error to model complex and non-linear data [144]. Also, [27] can work only on gray images. Background subtraction results obtained using these methods are restricted by the modelling constraints imposed.

Features that are robust to background clutter have been explored such as space time saliency ([39], [46], [55], [100]), pruning of motion features ([51], [107], [137]), Bag-of-words approach ([62], [128]), PbHOG [87] and pose based representation ([48], [49], [71], [73], [74], [111], [145]). Pose-based representations can cope with background variations, occlusions and shifts in global representations if key poses are selected well [48]. Selecting key poses can also remove important information in the video, which will affect the recognition performance. Above all, [146] uses IR imagery and [138] uses depth information to remove clutter. This also limits the adoption of these methods.

In conclusion, color or region-based normalization and threshold methods work well for simple, uniform backgrounds. Reduction in computation time and complexity is an added advantage of these methods. But in day-to-day environments, methods based on modelling the background seem to be more robust than others. These models are dynamic and can adapt according to the background making them a good option for the real world. Even though they can provide good background segmentation, the final performance is dependent on the features extracted to represent the action. This is evident from the results reported by these methods. Another interesting observation from the reported performance of all methods is that average performance is only around 70-75% ([46], [49], [60]–[62], [88], [102], [107], [139]) on datasets containing challenging backgrounds, which suggests that more research efforts are still required to further improve the performance to a more acceptable level. Combining model based methods with local feature representations can help in reducing the effect of cluttered backgrounds.

Ambiguities in action recognition have been caused by at least one of the above issues. The presence of even one of the above mentioned issues can degrade the performance of the system drastically. Hence, researchers focus on solving more than one of these issues, thereby making it difficult to classify them according to the challenges they are robust to. To put this in perspective, we have summarized the various approaches according to all challenges it can handle in Table II. In Table II, ‘I’ represents an integrated approach and ‘S’ denotes a separate block is added to make the method robust to the challenge. For instance, using Hidden Markov Model that models temporal constraints inherently to tackle execution rate changes will be ‘I’, whereas including a time warping on the features or video sequences separately will be ‘S’. ‘$\surd$’ means that the method is generalizable and ‘$\times$’ represents that the method is not robust to the challenge.

III. Action Classification

Action classification and feature extraction are complementary to each other. They must mutually compensate for each other’s limitations so as to improve the overall system. To achieve efficient action classification, training the classifier becomes essential. The training scheme helps the classifier to learn about the intra and inter-class action variations. For reliable classification performance, classifier must be trained with adequate and diverse amount of training data to learn every action effectively. In this section, we try to determine the bottlenecks in the classifiers by discussing 3 aspects, namely, generalizability of an approach, abnormal action detection, and classifier approaches to reduce the amount of labelled training data needed.

A. Generalizability

An important concern for researchers when they develop an action recognition system is the capability of the method to learn actions other than what they have been trained for. For instance, can a recognition system originally created to study gait patterns of people, be used to recognize falling or sitting down action? The ability of a method to learn or to cope with actions other than what they were originally made for broadens the usage to encompass a variety of other applications of similar class.

Uniqueness in the representation is very important for action classification to be generalized to other actions. Canonical poses [21], person dependent features ([13], [80]) rigid shape formations [36] all suffer from this limitation. MACH filters ([44], [131]) extracted for each action, classifiers for gait patterns ([14], [21], [147] etc), makes extension to lots of activities with subtle variations difficult. Region based methods that rely on negative spaces ([70], [105]) are dependent on background subtractions and shapes of negative spaces making it difficult to generalize.

Moving Lights Display (MLD) [23], 3D Markers [6], solid coloured gloves [80] and RFIDs [7] have been used to capture 3D joint angles. However, extension of these methods to normal day-to-day activities which are unconstrained is not possible. Similar to the above mentioned examples, methods developed to identify objects from human interactions with them or vice versa, ([25], [66], [67], [77], [78], [86], [128]) also cannot be extended to actions without objects. Some classifiers cannot be employed in all circumstances, for instance, W^4 [27] and [146] are based only on monocular gray and infra-red imagery. Likewise, [138] requires depth information from depth cameras.

B. *Abnormality Detection*

One important task of a surveillance system is to identify an aberration or abnormality in user behaviour or action, such as detecting a bank robbery or suspicious persons in an airport. Generally this detection is done by a human observer, who sits in front of the monitor throughout the day. Abnormality can also be seen in normal day to-day life activities, for example, people trying to correct their postures, athletes trying to improve their performance in training sessions etc. Some of the approaches used for abnormal behaviour detection methods in intelligent video systems are discussed in [160].

One important concern is to collect training samples that contain these abnormalities. Movie clips [39], Youtube videos [107], and amateur videos [62] provide a good source for training data from which both normal and abnormal action examples can be collected. For learning actions from movie clips, [39] employs a text based classifier using scripts and subtitles. But these scripts and subtitles need to be accurately aligned with the scenes and coping with substantial variability of action expressions in the text.

Some of the methods ([104], [124], [126]) have been developed specifically for detecting abnormal behaviours. Abnormality depends upon the overall context where the action is performed. Detecting aberration requires these methods to define criteria which characterize the abnormality. Hybrid latent Dirichlet allocation (h-LDA) is applied to automatically learn the distribution of the spatio-temporal words and correspond to human action categories in [124]. h-LDA learns the probability distribution of motion text words in a scenario, which is used to define the criteria for abnormal action.

The exemplar based method used in [104] models the function space of an action using a set of time warping transformations on the computed action trajectory and can be extended to abnormal action detection since the method is independent of features chosen to represent action and requires lesser number of training examples. But to define an abnormality criterion would be difficult since it will vary with the type of feature used. To monitor and detect abnormal activities using a very low resolution image, [126] models

TABLE II
CLASSIFICATION OF APPROACHES BASED ON ROBUSTNESS TO CHALLENGES, GENERALIZABILITY, ABNORMALITY DETECTION

Method	VI[1]	Oc[2]	Er[3]	An[4]	Cam[5]	C B[6]	Gen[7]	Ab[8]
[79], [140]	×	×	I	I	×	S	×	×
[100], [112], [139], [142], [148]	×	×	I	I	×	I	√	×
[149]	×	×	I	I	×	I	√	√
[123], [141]	×	×	I	I	×	S	√	×
[145], [146], [150]	×	×	I	I	×	I	×	×
[151]	×	×	I	S	×	I	√	×
[152]	×	×	S	S	×	I	√	×
[13]	I	×	S	×	×	×	×	×
[103], [116]	×	×	S	I	×	×	×	×
[84], [109], [114], [120]	×	×	S	I	×	×	√	×
[106]	×	×	S	I	×	×	√	√
[21], [23]	I	×	S	I	×	×	×	×
[22], [33], [41]	I	×	S	I	×	×	√	×
[80]	×	×	I	×	×	×	×	×
[101]	×	×	I	×	×	×	√	×
[104]	×	×	S	×	×	×	×	√
[81], [85], [91]–[97], [108], [113], [127], [153], [154]	×	×	I	I	×	×	√	×
[89], [90], [117]	×	×	I	I	×	×	√	√
[86], [105], [134], [155]	×	×	I	I	×	×	×	×
[24]	S	×	I	S	×	×	×	×
[53], [75]	×	I	×	I	×	I	√	×
[52], [74], [78]	×	I	×	I	×	I	×	×
[66]	×	S	I	I	×	×	×	×
[69]	×	S	I	I	×	S	√	×
[129]	×	S	I	I	×	S	×	×
[27]	I	S	×	I	×	S	×	×
[14]	I	S	×	I	×	×	×	×
[12]	I	I	I	×	S	×	√	×
[26], [30], [38], [42]	I	×	I	I	×	×	√	×
[35], [44]	I	×	I	I	×	×	×	×
[34]	I	×	I	I	I	×	√	×
[83], [119]	×	×	S	I	×	S	√	×
[87], [111]	×	×	S	I	×	I	√	×
[17], [28], [40]	I	×	×	I	×	×	√	×
[36]	I	×	×	I	I	×	×	×
[39]	I	×	×	I	×	S	√	√
[132], [133], [135], [156]	×	×	×	I	×	×	√	×
[126], [131]	×	×	×	I	×	×	×	√
[125], [157], [158]	×	×	×	I	×	×	×	×
[55]	×	S	×	I	×	S	√	×
[56]	×	S	×	I	×	S	×	×
[54]	×	S	×	I	S	S	×	×
[47], [59], [65], [68]	×	I	I	I	×	×	√	×
[67], [70]	×	I	I	I	×	×	×	×
[130], [159]	×	×	×	I	×	I	√	×
[118]	×	×	×	I	×	I	×	×
[122]	×	×	×	I	×	S	√	×
[124]	×	×	×	I	×	S	√	√
[15], [29], [37]	I	I	×	I	×	×	√	×
[16]	I	S	×	I	×	×	√	×
[25], [32]	I	I	I	I	×	×	×	×
[20]	S	I	I	I	×	×	√	×
[19]	S	I	S	I	×	×	√	×
[57], [58]	×	I	×	I	I	I	√	×
[115]	×	×	I	×	S	I	×	×
[102], [107], [136], [138]	×	×	S	I	I	I	√	×
[88]	×	×	I	I	I	I	√	×
[82]	×	×	S	I	S	S	√	×
[46]	×	I	×	×	×	I	√	×
[43]	I	×	I	I	×	S	√	√
[31]	I	×	S	I	×	I	√	×
[48]–[51], [60]–[63], [71], [73]	×	I	I	I	×	I	√	×
[128]	×	×	×	I	I	I	×	×
[18]	I	×	×	×	×	×	√	×
[45]	S	×	S	S	×	×	√	×

[1] View-invariance [2] Occlusion [3] Execution rate [4] Anthropometric variations [5] Camera motion [6] Cluttered Background [7] Generalizability [8] Abnormality detection

an activity using the polygonal shape of the configuration of point masses and their deformation over time. Abnormality is defined by learning the mean shape and dynamics of the shape change using manually indexed location data. This approach is limited as it can be applied to static shape activities only.

C. Classifier Training

In this section we will look into the classifier approaches and training strategies used. Some of the classifiers have not been used in action recognition but have potential to be explored and employed in the domain. Researchers use classifiers such as nearest neighbour ([84], [100], [117], [118], [129], [136], [150], [153] etc) and SVM ([41], [54], [106], [118], [125]) due to their simplicity and good performance. Most of these methods use Euclidean distance between features which might not be efficient for distance calculation in high dimensional case such as action recognition features. Many other metrics such as Mahalanobis distance ([12], [16], [56]), Chamfer distance [103], Riemannian metric [136], weighted Euclidean distance [15] and normalized Levenshtein distance [150] have been employed to perform the distance calculations correctly. These methods make an assumption that the distribution of data used in training and test data are the same. However, this might not be true in many cases [110]. These classifiers also require lots of training data to capture all the possible variations.

Several researchers have devised classifiers that require less training data or techniques to collect these data automatically. One simple approach is to use web as a source of information for the classifier training [87]. A typical search engine is employed to retrieve the images and irrelevant ones are cleaned up using an incremental procedure. Transfer learning model ([110], [161], [162]) provides impressive properties to learn attributes and features in one dataset and apply them to another target dataset. These methods still use simple classifiers such as SVM or AdaBoost to learn conceptual clues from image databases and use them in the target action database. Active learning paradigm [163] requires a small amount of training data but is restricted in usage since wearable sensors are used to extract context information. Incremental learning methods [63] and slow feature analysis [151] do not require extensive training since they can learn and update the feature representation models based on new training samples. Even though these methods can work with insufficient training data, they do not make any distinctions between positive and negative training data.

A good classifier should uncover the most discriminative features and learn them with more weightage to make reliable decisions. Classifier [164] tries to learn from partially representative data in huge datasets that include negative data. Data mining has been used in ([51], [148]) to learn from compound features created from spatio temporal corners. Fuzzy rule based classifications ([94], [95]) have also been shown to discriminate between features. [95] introduced McFIS classifier that can be used in an incremental manner since it automatically decides what, when and how to learn based on its available knowledge and new training sample. These methods can handle most of the action recognition challenges. However, requiring the task adaptive classifiers such as Hough forests [19] and adaptive vocabulary forests [165] to handle all the challenges in action recognition might not be easy. For instance, a separate Hough forest [19] needs to be created for every view angle we want to handle. Also these methods are dependent on codebooks to learn the appearance, which are subject to quantization errors.

Fuzzy rule based classifiers combined with tree data structure [166] was used to create user profile based keyboard inputs. These classifiers are evolving and can adapt over time to behavioural changes. This structure can be extended to action recognition to create a hierarchical action recognition system. In addition, fuzzy empirical copula was used as classifier in [99] to identify human hand motion using finger joint angles captured using data gloves. Recently, Extreme Learning Machines (ELM) [50] is becoming popular because of the fast output and good multi-class performance to make distinctions between features. ELM can generate hidden nodes or parameters without seeing the training data [167]. Results obtained can fluctuate from ELM due to the random initialization. As a result, classification is done based on the average of a number of iterations.

IV. Datasets

Testing action recognition algorithm is essential as it provides qualitative and quantitative performance analysis. But for reliable analysis it is necessary that the datasets capture all the actions under various challenges and conditions that would prove the system is robust to them. Datasets that capture actions in all possible scenarios are very limited. Also, some datasets are not publicly available. This is the primary reason why many researchers create their own dataset for evaluation ([34], [35], [79], [103], [119], [126]). A detailed survey on the available datasets for performance evaluation is given in [168].

KTH & Weizmann human action datasets are the most popular for action recognition. KTH dataset [54] contains 6 action classes performed by 25 actors in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors (s4). KTH dataset is restricted because cameras are relatively stationary, and only zooming of camera is considered as camera motion. Weizmann dataset [37] contains 10 action classes by 9 actors using simple background and fixed camera [168]. They also provide background sequences so that silhouettes can be extracted easily.

IXMAS dataset [28], CASIA motion datasets [43], i3DPost Multi-view Dataset [11], MuHAVi [169], Virtual Human Action Silhouette (ViHASi) [170] [171] and West Virginia University (WVU) multi-view dataset [172] provide datasets for variations in viewpoint. In order to provide a good multi-camera video input, all actions are captured in an indoor controlled setting. Dynamic backgrounds are not considered in most of these datasets. UT-Tower dataset [173] provides videos of actions in the outdoor environment but mainly from aerial view in which people cannot be seen clearly. All of the above mentioned datasets provide only single person single action video data. This means spatial and temporal action localization cannot be directly tested using these datasets.

Recently, ChaLearn gesture dataset [174] and G3D, a gaming dataset [175], concentrated on continuous actions. These datasets do not consider the usual challenges in action recognition field since they provide multi-modal data such as skeleton data, depth and audio (only in ChaLearn) in addition to the video data. The main restriction is that ChaLearn considers only hand gestures and G3D only actions in gaming environment.

Datasets like Keck gesture set [176], UCF Sports dataset [131] and Hollywood2 human action (HOHA) datasets [106] provide good database for camera motion and background clutter. YouTube dataset [107] is collected from YouTube and broadcast videos that are captured under uncontrolled settings. UCF Sports dataset [131], Hollywood2 human action (HOHA) datasets [106] and YouTube dataset [107] are good datasets for dynamic background since they are extracted from sports, movies and web videos. They also can provide good benchmarks for multi-person actions and interactions.

V. Conclusion

Action recognition has received much interest due to the many research challenges that have not been satisfactorily addressed [8]. Another important reason is the vast range of potential applications such as surveillance, human computer interaction, telehealth, biometrics etc. In this survey, we have presented methods based on their robustness to the various challenges faced in action recognition including view invariance, occlusion etc.

Execution rate and anthropometric variations have been resolved as researchers have shown that they can be effectively handled as intra-class variations and by combining different features. View-invariance is the most motivating challenge in the field now for real-world applications in an uncontrolled setting. Most of the methods can only achieve moderate view-invariance.

Methods that tackle camera motion are even fewer. Methods employing motion compensation or the multi view geometry along with standard epipolar geometry to resolve the discrepancies have shown good performance but are still limited in nature. Efforts to solve occlusion and cluttered background have been high, but there is still room for improvement especially in real world scenarios. Other research potential includes new classifiers to handle these challenges robustly.

Development of datasets that can cater to all these challenges is also crucial to gauge and benchmark the performance of the proposed methods. Generalization of methods must be considered during the design stage of the system. As far as we are aware of, none of the methods proposed can handle all these challenges. Developing a unified approach that is robust to all these issues may require explorations into new fields and new ideas. This survey is the first step towards identifying challenges that have not yet been fully resolved. In turn, this will help researchers in this area focus their research effort on those issues identified as bottlenecks and to eventually develop a system robust to all major action recognition challenges.

References

[1] M. Ryoo and J. Aggarwal, "Human activity analysis: A Review," *ACM Computing Surveys, Article 16*, vol. 43, pp. 16:1 – 16:43, April 2011.

[2] D. Siewiorek, A. Smailagic, and A. Dey, "Architecture and applications of virtual coaches," *Proceedings of the IEEE, Invited Paper*, vol. 100, pp. 2472–2488, August 2012.

[3] T. Kanade and M. Hebert, "First-person vision," *Proc. of the IEEE, Invited Paper*, vol. 100, pp. 2442–2453, August 2012.

[4] T. Shibata, "Therapeutic Seal robot biofeedback medical device: qualitative and quantitative evaluations of robot therapy in dementia care," *Proceedings of the IEEE, Invited Paper*, vol. 100, pp. 2527–2538, August 2012.

[5] K. Yamazaki, R. Ueda, S. Nozawa, M. Kojima, K. Okada, K. Matsumoto, M. Ishikawa, I. Shimoyama, and M. Inaba, "Home-assistant robot for an aging society," *Proceedings of the IEEE, Invited Paper*, vol. 100, pp. 2429–2441, August 2012.

[6] P. Kelly, A. Healy, K. Moran, and N. E. O'Connor, "A virtual coaching environment for improving golf swing technique," in *ACM Multimedia Workshop on Surreal Media and Virtual Cloning*, pp. 51 – 56, October 2010.

[7] L. Palafox and H.Hashimoto, "Human action recognition using wavelet signal analysis as an input in 4W1H," in *IEEE Intl. Conf. on Industrial Informatics*, pp. 679 – 684, July 2010.

[8] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, pp. 976 – 990, June 2010.

[9] Y. Li and Y. Kuai, "Action recognition based on spatio-temporal interest points," in *Intl. Conf. on BioMedical Engineering and Informatics*, pp. 181 – 185, October 2012.

[10] X. Ji and H. Liu, "Advances in view-invariant human motion analysis - A Review," *IEEE Trans. on Systems, Man, and Cybernetic Part C: Applications and Reviews*, vol. 40, pp. 13 – 24, January 2012.

[11] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost multi-view and 3D human action/interaction database," in *Conf. for Visual Media Production*, pp. 159 – 168, November 2009.

[12] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257 – 267, March 2001.

[13] T. Darell and A. Pentland, "Space-time gestures," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 335 – 340, 1993.

[14] G. Rogez, J. Guerrero, and C. Orrite, "View-invariant human feature extraction for video-surveillance applications," in *IEEE Conf. on Advanced Video and signal based Surveillance*, pp. 324 – 329, 2007.

[15] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "Human action recognition using robust power spectrum features," in *IEEE Conf. of Image Processing*, pp. 753–756, October 2008.

[16] Y. Lu, Y. Li, Y. Chen, F. Ding, X. Wang, J. Hu, and S. Ding, "A Human action recognition method based on Tchebichef moment invariants and temporal templates," in *Intl. Conf. on Intelligent Human-Machine Systems and Cybernetics*, pp. 76–79, August 2012.

[17] A. Iosifidis, A. Tefas, and I. Pitas, "Neural representation and learning for multi-view human action recognition," in *IEEE World Congress on Computational Intelligence*, pp. 1–6, June 2012.

[18] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "View-independent behavior analysis," *IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics*, vol. 39, pp. 1028 – 1035, August 2009.

[19] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. Lempitsky, "Hough forests for object detection, tracking and action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2188 – 2202, November 2011.

[20] B. Chakraborty, O. Rudovic, and J. Gonzalez, "View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts," in *IEEE Intl. Conf. on Automatic Face & Gesture Recognition*, pp. 1 – 6, September 2008.

[21] V. Parameswaran and R. Chellappa, "Quasi-invariants for human action representation and recognition," in *Intl. Conf. on Pattern Recognition*, vol. 1, pp. 307–310, August 2002.

[22] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and Viterbi path searching," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, June 2007.

[23] D. Gavrilla and L. Davis, "Towards 3-D model-based tracking and recognition of human movement," *Intl. Workshop on Face and Gesture Recognition*, pp. 272 – 277, 1995.

[24] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," in *IEEE Intl. Conf. on Computer Vision*, pp. 120 – 127, 1998.

[25] P. Peursum, G. West, and S. Venkatesh, "Combining image regions and human activity for indirect object recognition in indoor wide-angle views," in *IEEE Intl. Conf. on Computer Vision*, vol. 1, pp. 82–89, October 2005.

[26] A. S. Ogale, A. Karapurkar, and Y. Aloimonos, "View-invariant modeling and recognition of human actions using grammars," in *IEEE Intl. Conf. on Computer Vision*, vol. 5, pp. 115 – 126, 2005.

[27] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22, pp. 809 – 830, August 2000.

[28] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision Image Understanding*, vol. 104, pp. 249 – 257, October 2006.

[29] D. Weinland, R. Ronfard, and E. Boyer, "Automatic discovery of action taxonomies from multiple views," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1639 – 1645, 2006.

[30] Y. Shen and H. Foroosh, "View-invariant action recognition using fundamental ratios," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1 – 6, 2008.

[31] F. Zhang, Y. Wang, and Z. Zhang, "View-invariant action recognition in surveillance videos," in *Asian Conf. on Pattern Recognition*, pp. 580–583, November 2011.

[32] F. Huang and G. Xu, "Action recognition unrestricted by location and viewpoint variation," in *IEEE Intl. Conf. on Computer and Information Technology Workshops*, pp. 433 – 438, July 2008.

[33] M. N. Kumar and D. Madhavi, "Improved discriminative model for view-invariant human action recognition," *Intl. Journal of Computer Science & Engineering Technology*, vol. 4, pp. 1263 – 1270, September 2013.

[34] A. Yilmaz and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *IEEE Intl. Conf. on Computer Vision*, vol. 1, pp. 150–157, 2005b.

[35] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of human action," in *IEEE Intl. Conf. on Computer Vision*, vol. 1, pp. 144–149, October 2005.

[36] S. M. Khan and M. Shah, "Detecting group activities using rigidity of formation," in *ACM Intl. Conf. on Multimedia*, pp. 403–406, 2005.

[37] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes," in *IEEE Intl. Conf. on Computer Vision*, vol. 2, pp. 1395–1402, October 2005.

[38] C. Rao and M. Shah, "View-invariance in action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. II–316 – II–322, 2001.

[39] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[40] T. Syeda-Mahmood and A. Vasilescu, "Recognizing action events from multiple viewpoints," in *IEEE workshop on detection and recognition of events in videos*, pp. 64 – 72, 2001.

[41] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 33, pp. 172 – 185, January 2011.

[42] Anwaar-Ul-Haq, I. Gondal, and M. Murshed, "On temporal order invariance for view-invariant action recognition," *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 23, pp. 203 – 211, February 2013.

[43] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on $\Re$ transform," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1 – 8, June 2007.

[44] T. Ang, W. Tan, C. Loo, and W. Wong, "Wavelet MACH Filter for omnidirectional human activity recognition," *Intl. Journal of Innovative Computing, Information and Control*, vol. 8, pp. 1 – 20, June 2012.

[45] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1 – 7, June 2008. DOI: 10.1109/CVPR.2008.4587552.

[46] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on spatiotemporal orientation analysis," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 35, pp. 527 – 540, March 2013.

[47] L. Liu, L. Shao, and P. Rockett, "Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition," *Pattern Recognition 46, Elsevier*, pp. 1810 – 1818, July 2013.

[48] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Trans. on Cybernetics, Issue: 99*, pp. 1– 11, January 2013.

[49] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," in *11th European Conf. on Computer Vision*, Springer, September 2010.

[50] R. Minhas, A. Baradarani, S. Seifzadeh, and Q. M. J. Wu, "Human action recognition using extreme learning machine based on visual vocabularies," *Neurocomputing 73, Elsevier*, pp. 1906 – 1917, March 2010.

[51] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 883 – 897, May 2011.

[52] O. Chomat and J. L. Crowley, "Probabilistic recognition of activity using local appearance," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 104 – 109, June 1999.

[53] H. Jiang, M. S. Drew, and Z.-N. Li, "Successive convex matching for action detection," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1646–1653, 2006.

[54] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Intl. Conf. on Pattern Recognition*, vol. 3, pp. 32–36, August 2004.

[55] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Joint IEEE Intl. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, October 2005.

[56] I. Laptev and T. Lindeberg, "Space-time interest points," in *IEEE Intl. Conf. on Computer Vision*, vol. 1, pp. 432–439, October 2003.

[57] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatio-temporal words," in *British Machine Vision Conf. - Preliminary version*, vol. 3, pp. 1249–1258, September 2006.

[58] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatio-temporal words," *Intl. Journal on Computer Vision*, vol. 79, pp. 299–318, September 2008.

[59] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots. Learning a visually grounded storyline model from annotated videos," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2012 – 2019, 2009.

[60] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1576 – 1588, August 2012.

[61] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural SVM," *IEEE Trans. on Circuits and Systems for video technology*, vol. 23, pp. 1422 – 1431, August 2013.

[62] J. Liu, J. Luo, and M. Shah, "Action recognition in unconstrained amateur videos," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 3549 – 3552, April 2009.

[63] K. K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," in *IEEE Intl. Conf. on Computer Vision*, pp. 1010 – 1017, September 2009.

[64] B. Jun, I. Choi, and D. Kim, "Local transform features and hybridization for accurate face and human detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1423 – 1436, June 2013.

[65] S. Park and J. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia Systems*, vol. 10, pp. 164 – 179, 2004.

[66] D. J. Moore, I. A. Essa, and M. H. Hayes, "Exploiting human actions and object context for recognition tasks," in *IEEE Intl. Conf. on Computer Vision*, vol. 1, pp. 80 – 86, 1999.

[67] P. Peursum, H. H. Bui, S. Venkatesh, and G. West, "Human action segmentation via controlled use of missing data in HMMs," in *IEEE Intl. Conf. on Pattern Recognition*, vol. 4, pp. 440–445, August 2004.

[68] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 955 – 960, 2005.

[69] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 831 – 843, August 2000.

[70] S. A. Rahman, S.Y.Cho, and M.K.H.Leung, "Recognising human actions by analysing negative spaces," *IET Computer Vision*, vol. 6, p. 197 213, May 2012.

[71] M. Raptis, K. Wnuk, and S. Soatto, "Flexible dictionaries for action classification," in *Intl. workshop on Machine Learning for Vision-based motion analysis*, October 2008.

[72] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *IEEE Intl. Conf. on Computer Vision*, pp. 1365 – 1372, September 2009.

[73] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2650 – 2657, October 2013.

[74] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *IEEE Intl. Conf. on Computer Vision*, pp. 1331 – 1338, November 2011.

[75] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3177 – 3184, June 2011.

[76] L. Bourdev, S. Maji, T. Brox, , and J. Malik, "Detecting people using mutually consistent poselet activations," in *European conf. on Computer vision: Part VI*, pp. 168–181, Springer-Verlag Berlin, Heidelberg 2010, September 2010.

[77] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 17 – 24, June 2010.

[78] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1691 – 1703, September 2012.

[79] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *IEEE Conf. on Computer Vision and Pattern Recognition*, p. 379 385, 1992.

[80] T. Starner and A. Pentland, "Real-time American sign language recognition from video using hidden Markov model," in *Intl. Sym. on Computer Vision*, pp. 265 – 270, 1995.

[81] P. Natarajan and R. Nevatia, "Coupled hidden semi-Markov models for activity recognition," *IEEE Workshop on Motion and Video Computing*, pp. 10 – 17, 2007. DOI: 10.1109/WMVC.2007.12.

[82] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1635 – 1648, July 2013.

[83] M. Ahmad and S.-W. Lee, "Human action recognition using multi-view image sequences features," in *Intl. Conf. on Automatic Face and Gesture Recognition*, pp. 523 – 528, April 2006.

[84] X. Ji, C. Wang, Y. Li, and Q. Wu, "Hidden Markov model-based human action recognition using mixed features," *Journal of Computational Information Systems*, vol. 9, pp. 3659–3666, May 2013.

[85] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 1325–1337, December 1997.

[86] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, June 2007.

[87] N. Ikizler-Cinbis and S. Scarloff, "Web-based classifiers for human action recognition," *IEEE Trans. On Multimedia*, vol. 14, pp. 1031 – 1045, August 2012.

[88] J. Zhang, B. Yao, and Y. Wang, "Auto learning temporal atomic actions for activity classification," *Pattern Recognition 46, Elsevier*, pp. 1789 – 1798, July 2013.

[89] Y. Zhang, Y. Zhang, E. Swears, N. Larios, Z. Wang, and Q. Ji, "Modeling temporal interaction with interval temporal Bayesian networks for complex activity recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2468–2483, October 2013.

[90] G. Lavee, M. Rudzsky, and E. Rivlin, "Propagating certainty in Petri nets for activity recognition," *IEEE Trans. on Circuits and Systems for video technology*, vol. 23, pp. 326 – 337, February 2013.

[91] G. Lin, Y. Fan, and E. Zhang, "Human action recognition using latent-dynamic condition random fields," in *Intl. Conf. on Artificial Intelligence and Computational Intelligence*, vol. 3, pp. 147 – 151, November 2009.

[92] Y. Wang and G. Mori, "Hidden part models for human action recognition: probabilistic versus max margin," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1310 – 1323, July 2011.

[93] Y. Song, L.-P. Morency, and R. Davis, "Action recognition by hierarchical sequence summarization," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3562 – 3569, June 2013.

[94] J.-Y. Chang, J.-J. Shyu, and C.-W. Cho, "Fuzzy rule inference based human activity recognition," in *Intl. Sym. on Part of IEEE Multi-conf. on Systems and Control Intelligent Control*, pp. 211 – 215, July 2009.

[95] K. Subramanian and S.Suresh, "Human action recognition using meta-cognitive neuro-fuzzy inference system," in *Intl. Joint Conf. on Neural Networks*, pp. 1 – 8, June 2012.

[96] W. Huang and Q. J. Wu, "Human action recognition based on self organizing map," in *IEEE Intl. Conf. on Acoustics Speech and Signal Processing*, pp. 2130 – 2133, March 2010.

[97] W. Huang and J. Wu, "Human action recognition using recursive self organizing map and longest common subsequence matching," in *Workshop on Applications of Computer Vision*, pp. 1 – 6, December 2009.

[98] H. Liu, "Exploring human hand capabilities into embedded multifingered object manipulation," *IEEE Trans. on Industrial Informatics*, vol. 7, pp. 389 – 398, August 2011.

[99] Z. Ju and H. Liu, "A Unified Fuzzy framework for human-hand motion recognition," *IEEE Trans. on Fuzzy Systems*, vol. 19, pp. 901 – 913, October 2011.

[100] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: video structure comparison for recognition of complex human activities," in *IEEE Intl. Conf. on Computer Vision*, pp. 1593–1600, October 2009b.

[101] C. S. Pinhanez and A. F. Bobick, "Human action detection using PNF propagation of temporal constraints," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 898 – 904, 1998.

[102] Z. Jiang, Z. Lin, and L. S. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 533 – 547, March 2012.

[103] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics*, vol. 36, pp. 710–719, June 2006.

[104] A. Veeraraghavan, R. Chellappa, and A. K. R. Chowdhury, "The function space of an activity," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 959 – 968, 2006.

[105] S. A. Rahman, L. Li, and M.K.H.Leung, "Human action recognition by negative space analysis," in *Intl. Conf. on Cyberworlds*, pp. 354 – 359, October 2010.

[106] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conf. on Computer Vision & Pattern Recognition*, pp. 2929 – 2936, June 2009.

[107] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild'," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1996 – 2003, June 2009.

[108] H. Wang, C. Yuan, G. Luo, W. Hu, and C. Sun, "Action recognition using linear dynamic systems," *Pattern Recognition 46, Elsevier*, pp. 1710 – 1718, June 2013.

[109] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Trans. on Circuits and Systems for video technology*, vol. 23, pp. 236 – 243, February 2013.

[110] W. Bian, D. Tao, and Y. Rui, "Cross-Domain human action recognition," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, pp. 298 – 307, April 2012.

[111] C. Thurau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1 – 8, June 2008. DOI: 10.1109/CVPR.2008.4587721.

[112] P. Bilinski and F. Bremond, "Contextual statistics of space-time ordered features for human action recognition," in *IEEE Intl. Conf. on Advanced Video and Signal-Based Surveillance*, pp. 228 – 233, September 2012.

[113] A.-P. Ta, C. Wolf, G. Lavoue, A. Baskurt, and J.-M. Jolion, "Pairwise features for human action recognition," in *Intl. Conf. on Pattern Recognition*, pp. 3224 – 3227, August 2010.

[114] Y.-H. Qin, H.-L. Li, G.-H. Liu, and Z.-N. Wang, "Human action recognition using PEM histogram," in *Intl. Conf. on Computational Problem-Solving*, pp. 323 – 325, December 2010.

[115] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 867 – 882, May 2011.

[116] L. W. Campbell and A. F. Bobick, "Recognition of human body motion using phase space constraints," in *IEEE Intl. Conf. on Computer Vision*, pp. 624–630, June 1995.

[117] Anwaar-Ul-Haq, I. Gondal, and M. Marshed, "Action recognition using spatio-temporal distance classifier correlation filter," in *Intl. Conf. on Digital Image Computing: Techniques and Applications*, pp. 474–479, 2011.

[118] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[119] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. II–123 – II–130, 2001.

[120] R. Filipovych and E. Ribeiro, "Determining the scale of interest regions in videos," in *IEEE Intl. Conf. on Image Processing*, pp. 985–988, November 2009.

[121] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1337 – 1342, October 2003.

[122] S. Kumari and S. K. Mitra, "Human action recognition using DFT," in *Third National Conf. on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pp. 239 – 242, 2011.

[123] K. Schindler and L. van Gool, "Action Snippets: how many frames does human action recognition require?," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1 – 8, June 2008.

[124] L. He, D. Wang, and H. Wang, "Human abnormal action identification method in different scenarios," in *Intl. Conf. on Digital Manufacturing and Automation*, pp. 594–597, August 2011.

[125] R. Lublinerman, N. Ozay, D. Zarpalas, and O. Camps, "Activity recognition from silhouettes using linear systems and model (in)validation techniques," in *Intl. Conf. on Pattern Recognition*, pp. 347–350, 2006.

[126] N. Vaswani, A. R. Chowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. II–633 – II–640, June 2003.

[127] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage, "Action recognition by learning discriminative key poses," in *IEEE Intl. conf. on computer vision workshops*, pp. 1302 – 1309, November 2011.

[128] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *IEEE Intl. Conf. on Computer vision*, pp. 1 – 8, October 2007.

[129] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," *IEEE Intl. Conf. on Computer Vision*, vol. 2, pp. 726 – 733, 2003.

[130] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 405–412, June 2005.

[131] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[132] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–6, June 2007.

[133] J. Rittscher, A. Blake, A. Hoogs, and G. Stein, "Mathematical modelling of animate and intentional motion," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 358, pp. 475 – 490, March 2003.

[134] S.-Y. Lin, C.-K. Shie, S.-C. Chen, M.-S. Lee, and Y.-P. Hung, "Human action recognition using action trait code," in *Intl. Conf. on Pattern Recognition*, November 2012.

[135] L. Wang, Y. Wang, T. Jiang, D. Zhao, and W. Gao, "Learning discriminative features for fast frame-based action recognition," *Pattern Recognition 46, Elsevier*, pp. 1832 – 1840, July 2013.

[136] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Trans. on Image Processing*, vol. 22, pp. 2479 – 2494, June 2013.

[137] Y. Chen, Z. Li, X. Guo, Y. Zhao, and A. Cai, "A spatio-temporal interest point detector based on vorticity for action recognition," in *IEEE Intl. Conf. on Multimedia and Expo Workshops*, pp. 1 – 6, July 2013.

[138] H. Zhang and L. E. Parker, "4-Dimensional local spatio-temporal features for human activity recognition," in *IEEE Intl. Conf. on Intelligent Robots and Systems*, pp. 2044 – 2049, September 2011.

[139] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3361 – 3368, June 2011.

[140] K. Ramirez-Amaro, E.-S. Kim, J. Kim, B.-T. Zhang, M. Beetz, and G. Cheng, "Enhancing human action recognition through spatio-temporal feature learning and semantic rules," in *IEEE-RAS Intl. Conf. Humanoid Robots*, October 2013.

[141] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1 – 8, June 2008. DOI:10.1109/CVPR.2008.4587628.

[142] S. Calderara, R. Cucchiara, and A. Prati, "Action signature: a novel holistic representation for action recognition," in *IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, pp. 121 – 128, September 2008.

[143] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking - learning - detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1409 – 1422, July 2012.

[144] Z. Ju and H. Liu, "Fuzzy Gaussian mixture models," *Pattern Recognition*, vol. 45, pp. 1146 – 1158, March 2012.

[145] L. Wang, L. Cheng, T. H. Thi, and J. Zhang, "Human action recognition from boosted pose estimation," in *Intl. Conf. on Digital Image Computing: Techniques and Applications*, pp. 308 – 313, December 2010.

[146] Y. Zhu and G. Guo, "A study on visible to infrared action recognition," *IEEE Signal Processing Letters*, vol. 20, pp. 897 – 900, September 2013.

[147] J.-H. Wang, J.-J. Ding, Y. Chen, and H.-H. Chen, "Real time accelerometer-based gait recognition using adaptive windowed wavelet transforms," in *IEEE Asia Pacific Conf. on Circuits and systems*, pp. 591 – 594, December 2012.

[148] A. Gilbert, J. Illingworth, and R. Bowden, "Scale invariant action recognition using compound features mined from dense spatiotemporal corners," in *10th European Conf. on Computer Vision: Part I*, pp. 222 – 233, 2008.

[149] H. Bouma, G. Burghouts, L. de Penning, P. Hanckmann, J.-M. ten Hove, S. Korzec, M. Kruithof, S. Landsmeer, C. van Leeuwen, S. van den Broek, A. Halma, R. den Hollander, and K. Schutte, "Recognition and localization of relevant human behavior in videos," *Proceedings of the SPIE*, vol. 8711, pp. 87110B.1 – 87110B.10, April 2013.

[150] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 8 – 13, June 2012.

[151] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 436 – 450, March 2012.

[152] B. Yousefi, C. K. Loo, and A. Memariani, "Biological inspired human action recognition," in *IEEE Workshop on Robotic Intelligence In Informationally Structured Space*, pp. 58 – 65, April 2013.

[153] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *IEEE Intl. Conf. on Computer Vision*, pp. 1 – 8, October 2007.

[154] C.-A. Lin, Y.-Y. Lin, H.-Y. M. Liao, and S.-K. Jeng, "Action recognition using instance-specific and class-consistent cues," in *IEEE Intl. Conf. on Image Processing*, pp. 1373 – 1376, September 2012.

[155] M. Shimosaka, T. Nishimura, Y. Nejigane, T. Mori, and T. Sato, "Fast online action recognition with boosted combinational motion features," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pp. 5851 – 5858, October 2006.

[156] V. Vo and N. Ly, "An effective approach for human actions recognition based on optical flow and edge features," in *Intl. Conf. on Control, Automation and Information Sciences*, pp. 24 – 29, November 2012.

[157] T. Dean, R. Washington, and G. Corrado, "Sparse spatiotemporal coding for activity recognition." Department of Computer Science, Brown University, CS 10-02, March 2010.

[158] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, "Realistic human action recognition with multimodal feature selection and fusion," *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, vol. 43, pp. 875 – 885, July 2013.

[159] Z. Gao, A. A. Liu, H. Zhang, G. P. Xu, and Y. B. Xue, "Human action recognition based on sparse representation induced by L1/L2 regulations," in *Intl. Conf. on Pattern Recognition*, pp. 1868 – 1871, November 2012.

[160] H. Liu, S. Chen, and N. Kubota, "Intelligent video systems and analytics: A Survey," *IEEE Trans. on Industrial Informatics*, vol. 9, pp. 1222 – 1233, August 2013.

[161] X. M. Lin and S. Z. Li, "Transfer AdaBoost learning for action recognition," in *IEEE Intl. Sym. on IT in Medicine & Education*, vol. 1, pp. 659 – 664, August 2009.

[162] A. P. B. Lopes, E. R. da S. Santos, E. A. do Valle Jr., J. M. de Almeida, and A. A. de Araujo, "Transfer learning for human action recognition," in *SIBGRAPI Conf. on Graphics, Patterns and Images*, pp. 352 – 359, August 2011.

[163] R. Liu, T. Chen, and L. Huang, “Research on human activity recognition based on active learning,” in *Ninth Intl. Conf. on Machine Learning and Cybernetics*, pp. 285 – 290, July 2010.

[164] S. Kaski and J. Peltonen, “Learning from relevant tasks only,” in *European conf. on Machine Learning*, pp. 608 – 615, Springer-Verlag Berlin, Heidelberg, 2007.

[165] T. Yeh, J. Lee, and T. Darrell, “Adaptive vocabulary forests by dynamic indexing and category learning,” in *IEEE Intl. Conf. on Computer Vision*, pp. 1 –8, October 2007.

[166] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, “Creating evolving user behavior profile automatically,” *IEEE Trans. on Knowledge and data engineering*, vol. 24, pp. 854–867, May 2012.

[167] G.-B. Huang, L. Chen, and C.-K. Siew, “Universal approximation using incremental constructive feedforward networks with random hidden nodes,” *IEEE Trans. on Neural Networks*, vol. 17, pp. 879 – 892, July 2006.

[168] M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa, “Action dataset A Survey,” in *SICE Annual Conf. ,Waseda University*, pp. 1650 – 1655, September 2011.

[169] S. Singh, S. A. Velastin, and H. Ragheb, “MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods,” in *IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, pp. 48–55, August 2010.

[170] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, “VIHASI: virtual human action silhouette data for the performance evaluation of silhouette based action recognition methods,” in *ACM/IEEE Intl. Conf. on Distributed Smart cameras*, pp. 1 – 10, September 2008.

[171] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, “VIHASI: virtual human action silhouette data for the performance evaluation of silhouette based action recognition methods,” in *ACM Intl. Workshop on Vision networks for behavior analysis (ACM Int’l Conf. on Multimedia), Vancouver, Canada*, pp. 77 – 84, October 2008.

[172] S.Ramagiri, R.Kavi, and V.Kulathumani, “Real-time multi-view human action recognition using a wireless camera network,” in *ACM/IEEE Intl. Conf. on Distributed Smart Cameras*, pp. 1–6, August 2011.

[173] C.-C. Chen, M.S.Ryoo, and J.K.Aggarwal, *UT-Tower dataset: aerial view activity classification challenge*, August 2010. Available online: http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html.

[174] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante, “Chalearn gesture challenge: design and first results,” in *CVPR Workshops*, pp. 1–6, June 2012.

[175] V. Bloom, D. Makris, and V. Argyriou, “G3D: A Gaming action dataset and real time action recognition evaluation framework,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 7 – 12, June 2012.

[176] Z. Lin, Z. Jiang, and L. S. Davis, “Recognizing actions by shape-motion prototype trees,” in *IEEE Intl. Conf. on Computer Vision*, pp. 444 – 451, September 2009.

Manoj Ramanathan (S’13)received his B.Tech degree in instrumentation and control engineering from the National Institute of Technology, Tiruchirapalli, India, in 2009. He was working as a software engineer in Toshiba Software India Pvt. Ltd till 2012. He is currently working towards his PhD degree in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision, action recognition, biometrics and biomedical imaging.

Wei-Yun Yau (SM’05)received his BEng from the National University of Singapore (1992), and MEng (1995) and PhD (1999) from Nanyang Technological University (NTU). Currently, he is a Programme Manager at the Institute for Infocomm Research, A*STAR and an Adjunct Associate Professor at NTU. He also serves as a member of the IAPRs Technical Committee on Biometrics and Chairman of the IPTV Working Group and the Biometrics Technical Committee, Singapore. He is the recipient of TEC Innovator Award 2002, Tan Kah Kee Young Inventors Award 2003 (Merit), IES Prestigious Engineering Achievement Awards 2006 and Standards Council Distinguished Award 2007. His research interest includes biometrics, active vision system, interactive TV, tele-care, and robotics and has published widely, with eight patents granted and over 100 publications. One of his papers received the Pattern Recognition Journal Honorable Mention 2010 and listed as top 25 most cited papers in Scopus as at April/May 2012.

Eam Khwang TEOH received his BE and ME degrees in Electrical Engineering from the University of Auckland, New Zealand in 1980 and 1982 respectively and the PhD degree in Electrical & Computer Engineering from the University of Newcastle, New South Wales, Australia in 1986. Since June 1985, he has been with the School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore as a Lecturer, Senior Lecturer and currently, an Associate Professor in the Division of Control & Instrumentation. Dr Teoh was a Vice Dean in the School of EEE, NTU from June 1996 till May 2005. In 1994 and 2012, he was voted by the students as NTU Teacher of the Year Award and Nanyang Award for Excellence in Teaching for the School of Electrical & Electronic Engineering respectively. Dr Teoh has published more than 250 journal and conference papers. In 2007, one of his papers titled ”Lane Detection and Tracking Using B-Snake” by Wang Y., Teoh E.K. and Shen D.G. in Image and Vision Computing, Vol 22, No 4, April 2004, pp 269-280 won the Most Cited Paper Award for the journal Image and Vision Computing. As at to-date, the total SCI citation count for this paper is 223. His research interests include Computer Vision, Pattern Recognition, New Media, Biometric Recognition Systems and Medical Image Processing. He is a member of the IEEE.

LIST OF FOOTNOTES

1) Page 7, Footnote explaining the names of the challenges in Table II ‘[1] View-invariance [2] Occlusion [3] Execution rate [4] Anthropometric variations [5] Camera motion [6] Cluttered Background [7] Generalizability [8] Abnormality detection’