ACIT'2017
The International Arab Conference on Information Technology
Yassmine Hammamet, Tunisia
December 22-24, 2017

A General Structure for the Approaches Used in the Human Action and Activity Recognition from Video

Nawaf Alsrehin[1] ,Bilal Abul-Huda[2]
[1]Computer Information Systems Department, Faculty of Information Technology and Computer Science, Yarmouk University, Irbid, Jordan
[2]Management Information Systems Department, Faculty of Information Technology and Computer Science, Yarmouk University, Irbid, Jordan

Abstract: Recognizing human actions and activities from videos has become an important topic in computer vision, machine learning, and pattern recognition applications. Some of these applications include automatic video analysis, human behavior recognition, human machine interaction, robotics, and security aspects in real-time video surveillance. This paper provides a general review of most recent advances and approaches in human action and activity recognition during the past several years. It also presents a categorization of human action and activity recognition approaches and methods with their advantages and limitations. In particular, it divides the recognition process based on (a) the method used to extract the features from the input image/video, (b) learning classifier techniques. Moreover, it presents an overview of the existing and publically available human action and activity recognition datasets. This paper also examines the requirements for an ideal human recognition system and presents some directions for future research and report some exposed problems on human action and activity recognition.

I. Introduction

A human action is a sequence of human body movements generated by a human during the performance of an activity and might involve one or more human body parts simultaneously [2] [3]. Technically, action recognition is the process of labeling actions, which generally can be done by matching the observation with previously defined patterns and then assigns an action verb [2] [3]. Assigning a label for an action extracted from video containing human motion is one of the keys of several applications that are motivated from the real life requirements such as automatic surveillance for the elderly to support aging in smart homes.

Human activity recognition involves recognizing a sequence of human actions from a set of observations that are occurred at a given period of time. Automatically classify, recognize, analyze, and understand human actions and activities from video with low error is the ultimate goal for any human activity recognition system and it considered as a challenging task in most of the intelligent video systems and applications. This challenge generally comes from the challenges in: a) detecting and recognizing object/human body from video especially when partial occlusion occurs, b) classifying human actions and activities, c) changing the viewpoint, scale, lighting, frame resolution, and manifestation in the video, d) extracting the identity of a person, his personality and psychological state from a cluttered background, e) interpreting human activities, and f) determining action classes and distinguishing between actions of different classes [1].

Generally, action recognition has three major phases, which include: a) feature extraction, b) action learning and classification, c) action recognition and segmentation [4]. In the *feature extraction* phase, a set of features are extracted from a sequence of images ranging from complex body shapes (e.g., edges and corners) to motion information (e.g., changing in the object's position). These features must be rich enough to

allow for a robust classification of the action. There are two general representations of the image/video features: global and local representations based on the a) amount of information they describe, b) region of interest/localization in the image/video, c) background subtraction or tracking. These representations are more sensitive to viewpoint, noise and partial occlusions and variations in viewpoint [4]. In the *action learning and classification phase,* the extracted features are used to learn the classifier using some statistical models, and using those models to classify new feature observations. The classification algorithms are usually learned from training data. In the *action recognition and segmentation* phase, an action label is given for each frame or sequence of frames (video). Action segmentation is required to divide streams of motions into a set of activities. Action recognition might be done by comparing the observed sequence to labeled sequences or action class prototypes, or by using discriminative classifiers that discriminate between two or more classes by directly operating on the image features.

In this survey, we limit our focus to the application-based human action recognition approaches to address the characteristics and techniques that are typically used for specific domains and are not discussed in previous surveys. In addition, it provides a general categorization to the human activity recognition, shown in Figure 1. Also it compares different state-of-the-art researches and presents their advantages and limitations.

II. Classification based on Extracted Features

Here, we used the taxonomy proposed by [2] and [5] to classify the recognition techniques based on the method used to extract the features from the input image/video as follows:

- **Single Layered Approaches**: the activities in these approaches are recognized directly from the raw video data, which are usually simple videos or datasets, instead of primitive sub-actions or sub-activities. These approaches are basically focused on image/video representations, video feature extraction, and activity matching. Single layered approaches are used to detect simple primitive actions that can be employed to detect more complex actions. The single-layered approaches are divided into the following categories:
  - **Space-time approaches**: the space-time approaches use motion-related information to recognize the action or activities, these approaches are divided as follows**:**
    - **Space-time volumes**: these approaches used the entire 3D volume as a feature or template.
    - **Space-time trajectories**: these approaches are based on the information that are constructing from the tracking of joint or interest points in human body. This information should be sufficient to recognize an action.
    - **Space-time local features**: these approaches are based on the information extracted from local features in images such as: descriptions of interest points and their surrounding in the 3D volumetric data.
  - **Sequential approaches:** in the sequential approaches, the human action is defined as a sequence of observations. So these approaches are designed to capture temporal relationships of observations that are associated with either local or global features extracted from a frame or a set of frames.
    - **Exemplar-based approaches**: in these approaches, the focus is on defining how a new input video can be compared with a template or a sample sequence of action observations.
    - **State model-based approaches**: these approaches focus on learning a state model for each action and each action is represented in terms of a set of hidden states. So they associate each sequence of observations with an instance of the corresponding action.

- **Hierarchical Approaches:** these approaches try to recognize high level activities by composing a sequence of simple or low-level sub-activities. These sub-activities can be further decomposed into atomic ones and so on. The capability and flexibility in modeling the complex structure of human activities that include individual activities, human and/or object interaction, or group of activities is considered as an advantage of these approaches.
  - **Statistical-based approaches**: the hierarchal statistical approaches use multi-layer state-based statistical models to recognize human activities. These approaches recognize atomic actions from a sequence of feature vectors detected at the bottom-layer. The sequence of atomic actions is treated as observations generated by the second-level models. As a result and for each model, the probability is calculated to measure the likelihood between the activity and the input image sequence.
  - **Syntactic-based approaches**: human activities using syntactic-based approaches are represented as a set of production rules generating a string of atomic actions. Each atomic action is represented using a symbol of strings and recognized using some parsing techniques from the field of programming languages. Concurrent action recognition is one of the limitations of these syntactic approaches, in which these techniques are modeling the activities using a string of atomic actions, which assume the sequential order of these actions. Another limitation is that the user must provide a set of production rules for all possible events [3] [5].
  - **Description-based approaches**: these approaches recognize human activities using the description of the Spatio-temporal and logical structures. They are able to recognize both sequential and concurrent activities. In addition, they present a high-level human activity composed of a set of sub-activities.

Table 1 shows the comparison between single layered and hierarchical approaches. It provides the strength and limitations of each of them.

Table I. Comparison between single layered and hierarchical approaches

	Strength	Limitations
Single layered approaches	• Able to recognize common and simple primitive actions.	• Hard to detect concurrent actions or activities • A set of production rules should be provided or an algorithm that automatically extracts them should be given.
Hierarchical approaches	• More suitable in recognizing high-level activities, which are composed of sub-activities. • Require less training data. • Able to represent and recognize human activities with complex temporal structure. • Able to represent and recognize sequential and concurrent activities. • Capable of modeling the complex structure of human activities. • Flexible in recognizing individual	• Difficulties in representing and recognizing concurrent activities. • Inability to compensate for the failure of low-level components

	activities, interaction between humans and/or objects or group activities. • Able to integrate prior knowledge and understand the structure of activities.	

III. Classification based on Learning Techniques

Here, we used the taxonomy proposed by [6] to classify the existing human action recognition techniques, which is based on learning classifier techniques as follows:

- **Hand-crafted, knowledge based activity recognition**:

The general handcrafted approaches are based on the detected and extracted features from the image/video followed by a generic classifier for action representation and recognition that uses the trainable data.

Using general handcrafted approaches for handling the ambiguity and complexity involved in some real world applications is difficult and generally might generate inaccurate recognition results [7]. Learning based approaches are considered as better choices [7]. Learning-based approaches include knowledge- and logic-based approaches, which get rid of using the traditional handcrafted feature detectors and descriptors. These learning-based approaches use the capability of automatically learning features from the raw data. Thus the activity recognition is transformed from pixel level to action classification, which introduces the concept of end-to-end learning process [7].

- **Supervised Classifiers**

In the supervised classifier approaches, the

- **Human Body Model Based Methods**: these models use information extracted from the 2D or 3D representations of the human body parts, trajectories of joint positions, or landmark points to extract the pose of a human body.
- **Holistic Methods**: these methods use the features extracted from the shape and silhouette information to represent human body structure and its dynamics for action recognition in videos.
- **Local Feature Methods**: these methods use the local features extracted from the local feature detectors and then using the local feature descriptor to encode the spatio-temporal neighborhood around detected features. These methods can be classified as follows:
  - **Local Feature Detectors**
    - **Spatio-Temporal Interest Point Detector**: these methods detect interest points as features from the spatial and temporal dimensions extracted from videos.
    - **Trajectory Detectors**: these methods use information extracted from interest points that are tracked from consecutive frames.
  - **Local Feature Descriptors**: these methods use the description extracted from the captured shape and motion information that are surrounding to the interest points and trajectories.
  - **Collections of Local Features**: some of the limitations of the previous methods are they do not capture the relation among the discovered features because they are based only on the distinctive power or the global statistics of individual local features. Also, they do not consider either the position of the features of their local density. To overcome these limitations and

to improve the recognition accuracy, several approaches have been proposed to generate high-level feature representations and combine it with the bag-of-features approaches, which can be divided into:

- **Pair-wise Features**: these methods capture the relation between two features among the discovered features.
- **Contextual Features**: among the discovered features, these methods capture the relations between several neighboring features.

Encoding of Local Features: feature encoding is done by encoding global statistics of local features. First, the training video is used to extract a set of local feature. These local features are quantized using a special histogram and then used to create a visual vocabulary using unsupervised learning. Then, some normalization is applied to reduce the effect of varieties of detected local features and variable video size and content. Finally, some other techniques are then applied based on the visual vocabulary that is created using the bag-of-features model.

- **Other Classifiers**:

Once the video that contains a set of actions is represented using any of the above techniques, the goal is to recognize these actions automatically. There is large number of supervised learning approaches that are used to automatically recognize these actions, such as statistical approaches, graphical models, decision tree, artificial neural network, convolutional neural network, support vector machine, or different combination of them. These supervised learning approaches teach the classifier using labeled data, while the unsupervised learning approaches let the classifier learn by itself, for example, using clustering. There is a third approach that uses limited or less amount of training examples (i.e., labeled data) to define long-term activity models; this approach is called weakly-supervised classifier. The weakly-supervised classifiers might use some supervised approaches to get better recognition results.

Supervised learning approaches are the most common approaches for recognizing simple and short-term human actions and they achieve high recognition rates [6]. For complex and long-term activities, such as Activity of Daily Living (ADL) that are composed of a set of short-term actions, supervised learning approaches achieve low recognition rates because they require human interaction and calibration and most of them depend on the human skeleton that is hard to detect in the noisy and side-view situations.

The unsupervised approaches are a better choice when the learning is done from the whole video, such as traffic surveillance, when it contains abnormal activities. To allow high level analysis of abnormal activities, data should be clustering in different layers or trajectories, such as space, time, and speed. These clusters are used with other models to represent number of human motion states.

IV. Recognizing Human-Object Interactions

Reliable human activity recognition that includes objects and humans consists of: a) the identification of objects and motion, b) the integration between several components, and c) analysis of the interaction between these components. Human activity recognition might involve several objects in the recognition process; an example could be "a human is drinking from a cup", the cup, human arm, hand, and face or head could be involved in the activity process at different time slots. Here, the focus is on the activities that share the same object and exhibit interesting common properties and characteristics during the interaction and recognition processes.

To improve object and activity recognition, researchers have studied the relationship and dependences between objects, motion, and human during the activity and others have studies the relationship between object recognition and motion estimation. In the first approach, the identity of the object plays an important role in the way that human can interact with it meaning that they have high dependency between them. For example, the object "book" might have the same appearance as the object "tablet" and both might be involved in different types of human interactions "reading" and "playing game", respectively. These studies have designed different probabilistic approaches that result in that the human activity recognition can be enhanced by detecting and recognizing objects and the human activity recognition helps in classification of objects [5]. While in the second approach, objects are generally recognized first and then the object motion is analyzed, and finally the activities that involve these objects are recognized.

V. Recognizing Group of Human Activities

There are several forms of human activity recognition (HAR) such as: single-user activity recognition, (SAR), multi-user activity recognition (MAR) and group activity recognition (GAR). SAR is the process of recognizing what an actor is doing based on information taken from himself, things around him, or something happing in the environment. MAR is the recognition of separate activities of multiple actors in parallel, where two or more actors are involved [6]. GAR is the process of recognizing the behavior of a group of actors as a one entity rather than recognizing the activities of each separate actor, which has its own role that is different from the others. The activity that is characterized by the overall motion of entire group members represents the second type of group of activities, such as "marching", in which the motion of all the group members should be considered simultaneously.

To recognize group of human activities, it is important to analyze the activities of each individual actor and then analyze their structure and finally analyze the overall relation between all the recognized activities. There are different approaches that focus on the recognition of a single group of activities, the goal of these approaches is to recognize the activity with small number of humans who have non-uniform behavior. Example of this is a "school presentation" where the instructor is "talking" and the limited numbers of students are either "taking notes", "listening", or "asking questions".

VI. Datasets

In this section, we discuss public datasets available for the researcher to develop and evaluate their approaches. We will divide all these datasets based on the type of the media to either video-based or image-based datasets. More details for video-based datasets can be found in [9] and in [10] for image-based datasets.

A. The KTH video dataset

The KTH dataset is a large scale dataset used for human action recognition, which contains 2391 videos for six human actions performed several times by 25 subjects. These videos were taken indoor and outdoor and different clothes. The videos are down-sampled to the spatial resolution of 160x120 pixels.

B. The Weizmann video dataset

The Weizmann dataset includes 10 action categories of simple human actions performed by 9 subjects to generate 93 videos at 25 fps and 180x144 pixels as a spatial resolution. It also

contains 10 additional videos that were captured from different viewpoints.

C. The PETS video dataset

The PETS dataset includes realistic videos in uncontrolled environments that are collected from surveillance cameras. The PETS dataset has three variations PETS 2004, PETS 2006, and PETS 2007. The PETS dataset is provided to test the ability of recognition systems to identify and analyze human activities from realistic and specific activities. They include sets human category classes that are composed of a set of human actions; each class has number of videos with more than 30 videos as a total. Each video in the PETS 2004 is at 384x288 pixels as a spatial resolution and 768x576 pixels for PETS 2006 and PETS 2007 with different viewpoints.

D. The Hollywood video dataset

The Hollywood dataset contains two sub-datasets: Hollywood and Hollywood-2. Both of them provide natural actions that provide more challenges with only few action classes that were taken with controlled and simplified settings.

VII. Challenges in Human Activity Recognition Research Area

Analyzing and understanding human activities have become one of the most active topics in computer vision and machine learning. However, there are several challenges and limitations that face this research, some of them are:

- Lack of widely available and realistic dataset, such as true surveillance recording, sport and movies recording, and video data from internet and social media sites. This dataset should be: (1) rich enough to cover varieties of human actions and activities, (2) captured in high quality settings for still images or videos, (3) performed by different types of actors or subjects, and (4) annotated in a non-biased way. More details can be found in [3] [9] [10].
- Segmentation and tracking of multiple persons in video becomes harder due to poor lighting, crowded environment, and noisy images, and camera movements. In addition, controlled environments such as camera view, background subtraction models, and positions and appearances of persons should be considered carefully for the media collected.
- Detection, tracking, and segmentation of different body parts using temporal information is still a difficult issue, especially when there is a variety in motion quality. In addition, the motion speed of different part of human body in certain activities is different. For example the movement of head and face parts.
- Modeling the temporal and spatial structures of the person's movements is still not an easy task.

VIII. Future Directions

To achieve a high accuracy of recognizing human action and activity form different types of media, it is important to pay more attention to the recognition steps that automate the overall recognition process, namely:

1) Pre-processing step, which includes:

a) Feature extraction. Automatically detect the necessary information form the media type as a set of features that will help in the recognition step.

b) Background subtraction. Automatically subtract the background from the scene and focus only on the moving parts, or human body, will support the recognition step.

2) Detection and localization of human body or its parts. Automatically detect, localize, and track the moving parts in the scene and best utilize the temporal information will definitely benefit the overall

recognition step. Also, it is important to allow the recognition to be done in the presence of unknown actions or if the scenes contains multiple persons or contains some interaction between multiple persons. In addition, understanding psychological and physiological states of the human should be considered carefully in recognition of human activity. Moreover, adding a sequence of routine activities that human is doing every day and the relationship between these activities might help the recognition process.

3) Classification Step. Automatically and accurately and classify the detected activity into one of the predefined classes is the ultimate goal in any human recognition system. It is important here to achieve some scalability of action recognition systems with respect to vocabulary size that define the action classes

Conclusion

Automatically recognize human actions and activities from video is an important research in computer vision and machine learning areas and it has different applications in different fields. Some of these applications are tracking and monitoring elderly people or patients in hospitals, ambient assisted living (AAL) systems for smart homes, monitoring and surveillance systems for indoor and outdoor activities like train stations, underground subways or airports. A compete review of all the approaches is beyond reach, however, in this survey paper, we provided a general structure of the approaches used to automatically recognize human activates from video. Our goal is to summarize the recent survey papers [1] [2] [3] [5] [6] [7] [9] [10] and provide a general structure that is based on taxonomies that are reported in these survey papers. We provide an overview of these approaches and discuss the advantages and limitations of each one. This structure is based on extracted features, learning techniques, recognizing human-object interactions, and recognizing group of human activities. The video- and image-based general datasets are provided. After that, we present some challenges that face the human activity recognition and some future directions that are obviously encouraged to be done by researchers.

References

[1] D. Weinlanda, R. Ronfardb and E. Boyerc, "A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition," *Computer Vision and Image Understanding,* vol. 15, no. 2, p. 224–241, October 2010.

[2] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri and B. P. Buckles, "Advances in Human Action Recognition: A Survey," *eprint arXiv:1501.05964,* January 2015.

[3] V. Michalis, N. Christophoros and K. I. A., "A Review of Human Activity Recognition Methods," *Frontiers in Robotics and AI,* vol. 2, pp. 1-28, 2015.

[4] R. Poppe, "A Survey on Vision-Based Human Action Recognition," *Image and Vision Computing,* vol. 28, p. 976–990, 2010.

[5] J. K. A. a. M. S. RYOO, "Human Activity Analysis: A Review," *ACM*

Computing Surveys, vol. 43, no. 3, pp. 1-43, April 2011.

[6] F. N. a. F. Bremond, "Human Action Recognition in Videos: A Survey," INRIA Technical Report, Sophia Antipolis, France, 2016.

[7] A. B. Sargano, P. A. OrcID and Z. Habib, "A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition," *Applied Sciences*, vol. 7, no. 1, pp. 1-37, January 2017.

[8] T. Gu, Z. Wu, L. Wang, X. Tao and J. Lu, "Mining Emerging Patterns for Recognizing Activities of Multiple Users in Pervasive Computing," in *The 6th Annual International Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous*, Toronto, ON, Canada, 2009.

[9] J. M. Chaquet, E. J. Carmona and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633-659, 2013.

[10] G. G. a. A. Lai, "A survey on still image based human action recognition," *PatternRecognition*, vol. 47, no. 10, pp. 3343-3361, 2014.