ABSTRACT

Human activity recognition (HAR) aims to recognize activities from a series of observations on the actions of subjects and the environmental conditions. The vision-based HAR research is the basis of many applications including video surveillance, health care, human-computer interaction (HCI) and real-world applications. This review highlights the advances of state-of-the-art activity recognition approaches, especially for the activity representation and classification methods. For the representation methods, we sort out a chronological research trajectory from global representations to local representations, and recent depth-based representations. For the classification methods, we conform to the categorization of template-based methods, discriminative models, and generative models and review several prevalent methods. Next, representative and available datasets are introduced. Aiming to provide an overview of those methods and a convenient way of comparing them, we classify existing literatures with a detailed taxonomy including representation and classification methods, as well as the datasets they used. Finally, we investigate the directions for future research.

Keywords

Human activity recognition, real world applications

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1 Introduction:

Human activity recognition (HAR) is a widely studied computer vision problem. Applications of HAR include video surveillance, health care, and human-computer interaction. As the imaging technique advances and the camera device upgrades, novel approaches for HAR constantly emerge. This review aims to provide a comprehensive introduction to the video-based human activity recognition, giving an overview of various approaches as well as their evolutions by covering both the representative classical literatures and the state-of-the-art approaches.

Human activities have an inherent hierarchical structure that indicates the different levels of it, which can be considered as a three-level categorization. First, for the bottom level, there is an atomic element and these action primitives constitute more complex human activities. After the action primitive level, the action/activity comes as the second level. Finally, the complex interactions form the top level, which refers to the human activities that involve more than two persons and objects. In this paper, we follow this three-level categorization namely action primitives, actions/activities, and interactions. This three-level categorization varies a little from previous surveys and maintains a consistent theme. Action primitives are those atomic actions at the limb level, such as “stretching the left arm,” and “raising the right leg.” Atomic actions are performed by a specific part of the human body, such as the hands, arms, or upper body part. Actions and activities are used interchangeably in this review, referring to the whole-body movements composed of several action primitives in temporal sequential order and performed by a single person with no more person or additional objects. Specifically, we refer the terminology human activities as all movements of the three layers and the activities/actions as the middle level of human activities. Human activities like walking, running, and waving hands are categorized in the actions/activities level. Finally, similar to Aggarwal et al.’s review, interactions are human activities that involve two or more persons and objects. The additional person or object is an important characteristic of interaction. Typical examples

features need to be designed for the representation of activity images or videos. Thus, Sections 3 and 4, respectively, review the global and local representations in conventional RGB videos. Depth image-based representations are discussed as a separate part in Section 5. Next, Section 6 describes the classification approaches. To measure and compare different approaches, benchmark datasets act an important role on which various approaches are evaluated. Section 7 collects recent human tracking methods of two dominant categories. In Section 8 we present representative datasets in different levels. Before we conclude this review and the future of HAR in Section 8, we classify existing literatures with a detailed taxonomy including representation and classification methods, as well as the used datasets aiming at a comprehensive and convenient overview for HAR researchers.

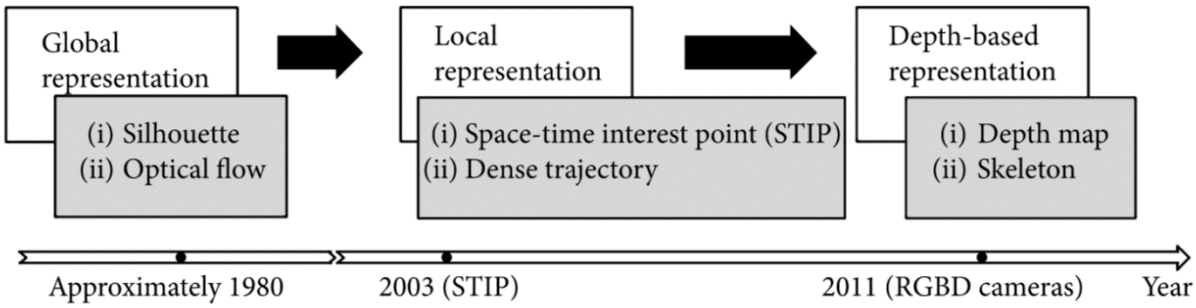

Figure 1.a Development of vision based HAR

1.2 Image Processing:

Image processing is any form of processing for which the input is an image, such as a photograph or video frame. The output of Image Processing can be either an image or a set of characteristics or parameters related to an image. The fundamental principle of Image processing operations carried out will assist us its greater perception and vision but doesn't add any information content the recent availability of sophisticated semi-conductor digital devices and compact powerful computers coupled with advances in Image processing algorithms has brought Digital Image processing to the fore front. Digital Image processing

has a broad spectrum. It has varied applications such as remote sensing via satellites and other space craft image transmission and automates inspection of industrial paths storage for business applications, medical processing, radars and acoustic image processing robotics. Image processing is necessary because human beings are adept at interpreting images of certain threshold beyond which we cannot detect just noticeable differences in the imagery. Human beings can detect only 8 to 16 shades of grey, even when data is recorded with 256 shades of grey. Therefore, one may not be able to interpret data in the remaining shades of grey. Also, it is necessary to continuously track large amounts of data and its storage is a problem. To avoid all these difficulties, one shall prefer processing of images by digital computers which processes at a much faster rate than human beings do. Major requirement of image enhancement is to restore a captured image from degradations arising from imperfect acquisition conditions. For example, to remove the noise imposed, to correct the colour cast and to sharpen the objects that appear in the image. To restore an image to improve 2 its contrast so that it is pleasant to a human viewer is one of the most demanding features. Image processing is a method to convert an image into digital form and accomplish some operations on it, in order to get and enhanced image or to extract some useful information's from it. It is a type of signal dispensation in which input is image like video frame or photograph and output may be an image or characteristics associated with that image. Usually, image processing system includes treating images as two-dimensional signals while applying already set signalling processing to them. Image processing is among rapidly technologies today, with its applications in various aspects of business. Image processing forms core research area within engineering and computer science disciplines too.

1.2.1 Image processing steps:

Image processing basically includes the following three steps.

- Importing the image with optical sensor or by digital photography.
- Analyzing and manipulating the image which includes data compression and image enhancement.
- Output is the last stage in which result can be altered image or report that is based on image analysis.

1.3.3 HSV color space:

The HSV stands for Hue, Saturation, Value, also known as HSB (Hue, Saturation, Brightness), defines a color space in terms of three constituent components.

- Hue represents a color which ranges from 0° to 360° (but normalized to 0-100% in some applications).
- Saturation indicates the range of grey in the color space. It ranges from 0 to 100%. Sometimes the value is calculated from 0 to 1.
- Value is the brightness of the color and varies with color saturation. It ranges from 0 to 100%. 0% black and 100% is white.

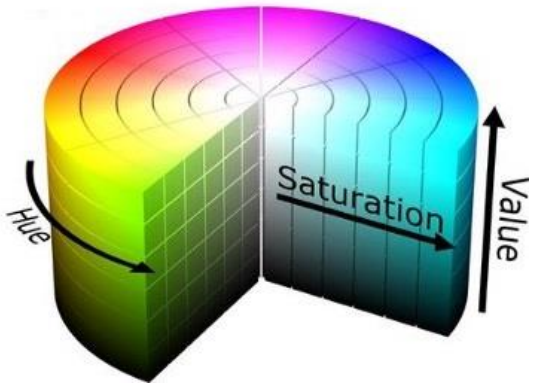

Figure 1. d HSV color model

1.4 Motivation of the work:

- Understanding people's actions and their interaction with the environment is a key element for the development of the aforementioned intelligent systems.
- Human activity recognition is the field that specifically deals with this issue through the integration of sensing and reasoning.

- In order to deliver context-aware data that can be employed to provide personalized support in many applications it requires sensing and reasoning.

1.5 Problem Statement:

The aim of the project is to train a neural network which recognizes the human's physical activity through video using smartphone camera.

2. LITERATURE SURVEY

2.1 Introduction:

The problem of action recognition in videos can vary widely and there's no single approach that suits all the problem statements. In this post, I will briefly touch upon a few approaches to get a sense of the existing research in this field. Traditional approaches to action recognition rely on object detection, pose detection, dense trajectories, or structural information. Here's a brief summary of the different approaches for action recognition:

Convolutional Neural Networks (CNN) extracts the features from each frame and pool the features from multiple frames to get a video-level prediction. The drawback of this approach is that it fails to capture sufficient motion information.

Motion information can be captured by combining optical flow containing short-term motion.

In addition to RGB and optical flow, information from other modalities such as audio, pose, and trajectory can also be used.

Wang et al. concatenated dense trajectory descriptors with appearance features

Choutas et al. encoded the movement of human joints, and the resulted heatmaps were aggregated temporally, obtaining PoTion

We can construct a spatiotemporal representation by fusion motion and appearance information in the way of two streams. It would work well for short duration clips, but would not be able to capture long-term temporal dynamics.

The two-stream network consists of two separate subnetworks, where one is for raw images and the other is for stacked optical flow, respectively, and captures spatiotemporal information by fusing the SoftMax scores of two streams.

Recurrent neural networks (RNNs), especially long short-term memory (LSTM), achieved impressive results in the sequence tasks due to the ability of long-term temporal modelling, so an alternative strategy is to adopt LSTM to model dynamics of frame-level features.