Applied Video Processing in Surveillance and Monitoring Systems

Nilanjan Dey
Techno India College of Technology, Kolkata, India

Amira Ashour
Tanta University, Egypt

Suvojit Acharjee
National Institute of Technology Agartala, India

A volume in the Advances in Multimedia and Interactive Technologies (AMIT) Book Series

IGI GLOBAL
DISSEMINATOR OF KNOWLEDGE

www.igi-global.com

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Dey, Nilanjan, 1984- editor. | Ashour, Amira, 1975- editor. | Acharjee, Suvojit, 1989- editor.
Title: Applied video processing in surveillance and monitoring systems / Nilanjan Dey, Amira Ashour, and Suvojit Acharjee, editors.
Description: Hershey PA : Information Science Reference, [2017] | Series: Advances in multimedia and interactive technologies | Includes bibliographical references and index.
Identifiers: LCCN 2016033139| ISBN 9781522510222 (hardcover) | ISBN 9781522510239 (ebook)
Subjects: LCSH: Video surveillance. | Image processing--Digital techniques.
Classification: LCC TK6680.3 .A67 2017 | DDC 621.389/28--dc23 LC record available at https://lccn.loc.gov/2016033139

This book is published in the IGI Global book series Advances in Multimedia and Interactive Technologies (AMIT) (ISSN: 2327-929X; eISSN: 2327-9303)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 8
Encoding Human Motion for Automated Activity Recognition in Surveillance Applications

Ammar Ladjailia
University of Souk Ahras, Algeria

Nouzha Harrati
University of Souk Ahras, Algeria

Imed Bouchrika
University of Souk Ahras, Algeria

Zohra Mahfouf
University of Souk Ahras, Algeria

ABSTRACT

As computing becomes ubiquitous in our modern society, automated recognition of human activities emerges as a crucial topic where it can be applied to many real-life human-centric scenarios such as smart automated surveillance, human computer interaction and automated refereeing. Although the perception of activities is spontaneous for the human visual system, it has proven to be extraordinarily difficult to duplicate this capability into computer vision systems for automated understanding of human behavior. Motion pictures provide even richer and reliable information for the perception of the different biological, social and psychological characteristics of the person such as emotions, actions and personality traits of the subject. In spite of the fact that there is a considerable body of work devoted to human action recognition, most of the methods are evaluated on datasets recorded in simplified settings. More recent research has shifted focus to natural activity recognition in unconstrained scenes with more complex settings.

INTRODUCTION

Much research within the computer vision community is dedicated towards the analysis of and understanding of human motion. The perception of human motion is one of the most important skills people possess, and our visual system provides particularly rich information in support of this skill. Yet, attempts and efforts to understand the human visual system or to devise an artificial solution for visual perception have proven to be a difficult task. Human motion analysis has received much attention from

DOI: 10.4018/978-1-5225-1022-2.ch008

researchers in the last two decades due to its potential use in a plethora of applications. This field of research focuses on the perception and recognition of human activities. As computing becomes ubiquitous in our modern society, the recognition of human activities emerges as a crucial topic where it can be applied to many real-life human-centric scenarios (Aggarwal & Ryoo, 2011). Furthermore, given the immense expansion of video data being recorded in everyday life from security surveillance cameras, movies production and internet video uploads, it becomes an essential need to automatically analyse and understand video content semantically. This is to ease the process of video indexing and fast retrieval of data when dealing with large multimedia content and big data. Hence, the importance of automated systems for human activity recognition is central to the success of such applications (Turaga, Chellappa, Subrahmanian, & Udrea, 2008). Further, due to the proliferating number of crimes and terror attacks as well as the vital need to provide safer environment, it becomes a necessary requirement to improve current state of surveillance systems via the use of computer vision methods to automate procedures of detecting suspicious human activities.

Human activity recognition aims to automatically infer the action or activity being performed by a person or group of people. For instance, recognizing whether someone is walking, raising hands or performing other types of activities. This usually involves the analysis and recognition of different motion patterns in order to produce a high-level semantic description for the human activities or interaction between people. This is vital to apprehend the human behavior and to determine whether their behavior is abnormal or normal via the use of automated methods (Ko, 2008). There have been considerable amount of work by the computer vision community dedicated to activity recognition with numerous approaches and methods being proposed to address different aspects and contexts of this area of research (Aggarwal & Ryoo, 2011; Poppe, 2010). Many of the early approaches have considered the use of video sequences recorded using a single camera with people being asked to perform basic actions in simplified settings and conditions. Various low-features have been proposed for encoding the human activity either at a temporal or spatial level such as edges, curvatures or complex features such as interest point descriptors. The detection of human motion is considered as a rudimentary component for constructing the activity descriptor in the majority of approaches either explicitly or implicitly for recovering other high level features. In fact, it is infeasible to detect human action from a still frame as even though achieving pose recovery can be possible from a single image, the perception of human activity can be challenging. Vishwakarma and Agrawal (2013) grouped the methods object detection through motion estimation into six conventional methods: background subtraction, statistical methods, temporal differencing and optical flow. Various recent surveys can be found in the literature on the representation of different features for human activity recognition (Poppe, 2010; Turaga et al., 2008; Vishwakarma & Agrawal, 2013). Interestingly, a new trend of research has emerged on activities recognition through the use of wearable sensors mounted to the human body (Lara & Labrador, 2013).

Applications for Activity Recognition

Research into automated recognition of human activities is fueled by the wide range of applications where human motion analysis can be deployed such as smart automated surveillance, behavioral biometrics, human computer interaction, animation and synthesis in addition to sport refereeing and analysis.

Smart Automated Surveillance

Traditionally, it is impossible for human operators to work simultaneously on different video screens in order to track and identify people of interest as well as analyze their behaviors across different places. Thus, it has become a vital requirement for scientists from the computer vision community to investigate visual-based alternatives to automate the process for human activity recognition over different views. Recently, various approaches were published in the literature to accomplish this task based on using basic features such as shape or color information. However, their practical deployment in real applications is very limited due to the complex nature of such problem (Bouchrika, 2008; Bouchrika & Nixon, 2006; Ko, 2008; Vishwakarma & Agrawal, 2013). In fact, the inability of human operators to monitor the increasingly growing numbers of CCTVs (Closed-Circuit television) installed in highly sensitive and populated areas such as government buildings, airports or shopping malls, has rendered the usability of such systems to be useless. According to the British Security Industry Association, the number of surveillance cameras deployed in the United Kingdom was estimated to be more than 5 million in 2015; this figure is expected to increase rapidly particularly after the terrorist attacks that a number of cities in Europe have witnessed. Despite the huge increase of monitoring systems, the question whether current surveillance systems work as a deterrent to crime is still questionable (Bouchrika, 2008). Security systems should not only be able to predict when a crime is about to happen but more importantly, by early recognition of suspicious individuals who may pose security threats, the system would be able to deter future crimes as it is a significant requirement to identify the perpetrator of a crime as soon as possible in order to prevent further offences and to allow justice to be administered. Furthermore, the use of smart visual surveillance technology has a wide spectrum of potential applications in addition to behavior analysis such as access control, crowd flux and congestion analysis (Ko, 2008; Vishwakarma & Agrawal, 2013).

Human Computer Interaction

Gestural interaction is becoming an integral part for newly systems from smart televisions to gaming consoles. The visual cues are the most important mode of non-verbal communication and their effective employment holds promising and innovative ways for people to interact with computers. This can even help to improve the accessibility and usability level for people with special needs and requirements. As featured in numerous science fiction movies where the actor can interact with computer systems via moving their hands and tapping their fingers in the air, it is now becoming a reality with the introduction of Microsoft Kinect and the cheap prices of depth sensing devices that sparked the rapid and abrupt advancement of gestural interaction from the advent of commercial products to a myriad of research projects (Ren, Meng, Yuan, & Zhang, 2011). Game players instead of using pads or joysticks, they can use their full body, hands and legs as an input method to control the game without wearing any special sensors or markers. Furthermore, many consumer electronic devices such as smart televisions have been developed with the capability to let users interact using hand gestures to swap between different channels or control the volume level. There are various development framework and programming toolkits being proposed to ease the process of gestural interaction using Kinect and other sensors (Deshayes, Mens, & Palanque, 2013; Suma et al., 2013). Moreover, there is a stream of research for creating interactive environments such as smart rooms that can react to various human gestures (Kühnel et al., 2011).

Video Indexing and Retrieval

With video sharing websites as Youtube facing relentlessly growth with gigabytes of multimedia content being uploaded every day, it becomes necessary to develop efficient ways for indexing and retrieving video data beyond the use of simple textual information and tags. This can be achieved through semantic attributes that can be extracted from the actual content of the video data. Content-Based video summarization has been gaining interest with advances in content-based image retrieval (Rui, Huang, & Chang, 1999). Most of the early methods have used simple sematic traits as colors and basic shapes for searching videos. Recent research efforts were geared towards object detection using various approaches remarkably the use of visual bag of words. This is commonly implemented as a histogram of the number of occurrences of specific visual patterns in a given image. The visual patterns are called words which are pre-constructed in a codebook using clustering techniques. In spite of their simplicity, bag of visual words were successfully applied to various challenging computer vision cases including recent studies to explore their applicability in automated human activity recognition. However, indexing human activities is still in its infancy due to the cumbersome challenges and complexities involved. In (Niebles, Wang, & Fei-Fei, 2008), the authors presented an approach for the non-supervised classification of human actions into different categories from video sequences. The basis of their method is the extraction of a collection of spatio-temporal words via the use of latent topic models.

MOTION FOR HUMAN ACTIVITY PERCEPTION

In spite of the fact that people can discern the state of the subject from a single static image to infer that they are doing, motion pictures provide even richer and reliable information for the perception of the different biological, social and psychological characteristics of the person (Blake & Shiffrar, 2007) such as emotions, actions and personality traits of the subject. Furthermore, this notion was also observed by Darwin (1872) in his book "*The Expression of Emotions in Man and Animals*" where it was stated: "*Actions speak louder than pictures when it comes to understanding what others are doing*". The human visual system is very sensitive to motion as it tends to focus attention on moving objects. In contrast, static or motionless objects are not as straightforward to detect. Motion is a spatio-temporal event defined as the change of spatial location over time. Given some visual input, the visual perception of motion is regarded as the process by which the visual system acquires perceptual knowledge such as the speed and direction of the moving object (Derrington, Allen, & Delicato, 2004). Whilst this process is spontaneous for the human visual system, it has proven to be extraordinarily difficult to duplicate this capability into computer vision systems for automated understanding of human behavior.

Psychological studies carried out by the Swedish psychologist Johansson (1973), revealed that people are able to perceive human motion from Moving Lights Display (MLD). An MLD is a two-dimensional video of a collection of bright dots attached to the human body taken against a dark background where only the bright dots are visible in the scene. Different observers are asked to see the actors performing various activities. Based on these experiments observers can recognize different types of human motion such as walking, jumping, dancing and so on. Moreover, the observer can make a judgment about the gender of the performer (Kozlowski & Cutting, 1978), and even further identify the person if they are already familiar with their gait (Goddard, 1992). Cutting argued that the recognition is purely based on

dynamic gait features as opposed to previous studies which were confounded by familiarity cues, size, shape or other non-gait sources of information. Although the different parts of the human body are not seen in the points and no links exist between the bright dots to show the skeleton structure of the human body, the observer can recover the full structure of the moving object. Thereby, the motion of the joints contains sufficient information for the perception of human motion (Bingham, Schmidt, & Rosenblum, 1995; Dittrich, 1993). There is a wealth of research which strives to document the capability of the human visual system to perceive the human motion from a small number of moving points as argued by early medical studies by Johansson, Cutting and Murray. Nevertheless, the underlying perceptual process is poorly understood and there is still a lack of research which explains the underlying principles for representing and retrieving the biological motion (Troje, Westhoff, & Lavrov, 2005). Two main theories have been put forward for the perception of human motion from the MLD: *structure-based* and *motion-based* (Cedras & Shah, 1995). The former theory claims that the initial step is recovering the 3D structure from the motion information observed from the MLDs, and then uses the recovered structure for the purpose of recognition. In the motion-based approach, recognition is based directly on the motion information without recovering the skeleton structure of the human body from the MLD; instead the motion information is extracted from a sequence of frames.

ACTION VS. ACTIVITY

In the computer vision literature, both terms "*Activity*" and "*Action*" are used interchangeably and contentiously but every term has its rough and gray definition (Poppe, 2010). An *action* is considered as a simple activity referring to simple pattern performed by a person during a short period of time lasting a few seconds. Examples of actions may include raising hands, bending, sitting and even walking. Poppe (2010) described additionally the term *action primitive* which refers to an atomic movement at the limb level. Vishwakarma and Agrawal (2013) has described the word *gesture* to refer to an elementary movement made by a part of the human body occurring in a very short span of time with low complexity such as waving a hand or stretching an arm. The term *action* can be considered similar to an extent to the term *gesture*. On the other hand, an *activity* is considered as a composite sequence of actions executed by either a single person or several people interacting with each other. Examples of activities are like leaving an unattended bag, shaking hands or assaulting a pedestrian. There is the term *interaction* which defines an activity or activities performed by two or more people spanning over longer times (Vishwakarma & Agrawal, 2013).

VISION-BASED SYSTEM FOR ACTION RECOGNITION

An automated vision-based system for human activity recognition through the use of motion features is designed to extract kinematic-based features without the need to use markers or special sensors to aid the extraction process. In fact, all that is required is an ordinary video camera linked to special vision-based software. Marker-less motion capture systems are suited for applications where mounting sensors or markers on the subject is not an option as the case of visual surveillance. Typically, the system consists of two main components:

1. Hardware platform dedicated for data acquisition. This can be a single CCTV camera or distributed network of cameras.
2. Software platform for data processing and recognition.

The architecture of the software side for human activity analysis is composed broadly of three main components:

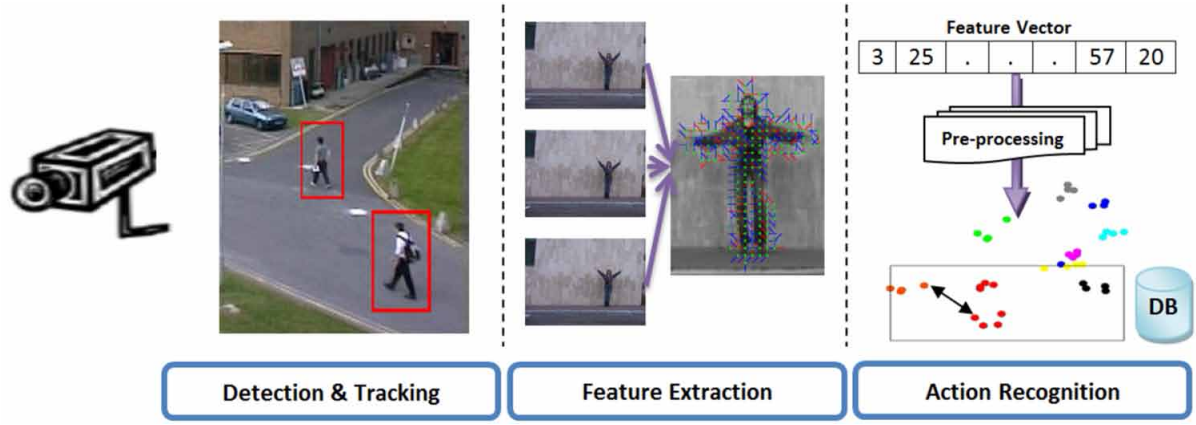
1. Detection and tracking of the subject,
2. Feature extraction and
3. Classification stage.

Figure 1 shows the flow diagram for human action recognition outlining the different subsystems.

Subject Detection and Tracking

People detection is the first major milestone for automated system of human activity recognition. A walking subject is initially detected within a sequence of frames using background subtraction techniques to detect moving objects or via the use of other methods such as the Histogram of Oriented Gradients (HoG) (Bouchrika, Carter, Nixon, Morzinger, & Thallinger, 2010; Dalal & Triggs, 2005) which is capable of detecting people from still images at real-time. The HoG method requires no background subtraction and therefore avoiding the need of maintaining and updating a model for the background. Subsequently, intra-camera tracking is performed to establish the correspondence of the same person across consecutive frames. Tracking methods are supported by simple low-level features such as blob size, aspect-ratio, speed and color in addition to the use of prediction algorithms to estimate the parameters of moving objects in the following frames. This is based on motion models which describe how parameters change over time. The most popular predictive methods used for tracking is the Kalman filter (Welch & Bishop, 2001), the Condensation algorithm (Isard & Blake, 1998), and the mean shift tracker (Comaniciu, Ramesh, & Meer, 2000).

Figure 1. Overview system for marker-less human activity recognition



Feature Extraction and Representation

This is the most important stage for automated marker-less capture systems whether for human identification, activity classification or other imaging applications. This is because the crucial data required for the classification phase are derived at this stage. Feature extraction is the process of estimating a set of measurements either related to the configuration of the whole body or the configuration of the different body parts in a given scene and tracking them over a sequence of frames. The features should bear certain degree of the discriminability between the different clusters of human activities. Various types of features are employed such as the trajectories of the joints positions estimated via pose recovery of the different parts of the human body. Contour-based features are used in a number of recent studies via analyzing silhouettes data. Textural features are proved to offer promising results on the detection of similar human actions even for the case of single frames. Irani *et al.* (Blank, Gorelick, Shechtman, Irani, & Basri, 2005; Shechtman & Irani, 2007) have proposed a descriptor based on analyzing adjacent patches based on their internal correlation for comparing images where they showed its potency for action detection. However, the majorities of studies consider the use of motion-based features for understanding human activities (Shah & Jain, 2013) . Depending on how kinematic features are represented based on the spatial properties, features estimated at this level can be categorized into two major types:

- **Global Features:** Where the whole image or body region is considered meanwhile.
- **Local Features:** Refer to the characteristics which are extracted from smaller portions of the image.

Classification Phase

This is mainly a pattern recognition process which involves matching a test sequence with an unknown label against a group of labeled references considered as the gallery dataset. At this stage, a high-level description is produced from the features extracted during the previous phases to infer or confirm the subject identity. The classification process is normally preceded by pre-processing stages such as data normalization, feature selection and dimensionality reduction of the feature space through the use of statistical methods. A variety of pattern recognition methods are employed in vision-based systems for human activity recognition including Neural Networks, Support Vector Machines (SVM) and K-Nearest Neighbor classifier (KNN). The latter is the most popular method for the classification due to its simplicity and fast computation and ease of comparison with most methods in the literature. The matching process during the classification phase is based on measuring the similarity between the test video against set of manually annotated actions to predict the class label for the unseen data. The similarity is computed using one of the distance metrics such the Euclidian or Mahalanobis distance.

MOTION FEATURES REPRESENTATION

The recognition of human activity is of prime importance for various applications as automated visual surveillance. The research area of human activity recognition is closely related to other fields of research that analyze human motion such as human computer interaction and biomechanical engineering. Although, there is a considerable body of work devoted to human action recognition, most of the

methods are evaluated on datasets recorded in simplified settings. More recent research has shifted focus to natural activity recognition in unconstrained scenes with more complex settings (Oshin, Gilbert, & Bowden, 2014). Poppe (2010) and Vishwakarma (2013) surveyed the recent methods, research studies and datasets devoted to this area of research. Existing methods can be broadly classified into two major categories in terms of image representation which are either global or local representation. There are recent studies that consider the use of a hybrid model by fusing both types of features as it was suggested to be more suitable for encoding human actions. In another study, Weinland *et al* have categorized three major classes of features for human action representation which are: body models, image models and sparse features (Weinland, Ronfard, & Boyer, 2011). The last two categories refer to the global and local features respectively discussed in most surveys. The body models aims to recover the spatial structure of the different parts of the human body via fitting a prior model. From another perspective, the temporal dimension is taken into account explicitly for most image representations in addition to the spatial information meanwhile other methods extract image features on a frame by frame basis. In this research work, three major categories for the various approaches devoted to markerless human activity recognition are considered and discussed in this section.

Pose-Based Approaches

For action recognition using pose-based representation, the parts of the human body are first recovered or reconstructed through the use of specific models. Although model-based approaches tend to be complex requiring high computational cost, these approaches are the most popular for human motion analysis due to their advantages (Yam & Nixon, 2009). The model can be either a 2 or 3-dimension structural model, motion model or a combined model. The structural model describes the topology of the human body parts as head, torso, hip, knee and ankle by measurements such as the length, width and positions. This model can be made up of primitive shapes based on matching against low-level features as edges. The stick and volumetric models are the most commonly used structural-based methods. Akita (1984) proposed a model consisting of six segments comprising of two arms, two legs, the torso and the head. Guo *et al* (Guo, Xu, & Tsuji, 1994) represented the human body structure by a stick figure model which had ten articulated sticks connected with six joints. Rohr (1994) proposed a volumetric model for the analysis of human motion using 14 elliptical cylinders to model the human body. Karaulova *et al*. (Karaulova, Hall, & Marshall, 2000) used the stick figure to build a hierarchical model of human dynamics represented using Hidden Markov Models (HMMs). Gavrila *et al.* (Gavrila & Davis, 1995) described a 3D model for pose recovery based on conducting a search of synthesized images against real images using the chamfer distance for different views. The main merit of using 3D models is the viewpoint invariance provided the pose estimation is done accurately (Weinland, Özuysal, & Fua, 2010).

Global-Based Approaches

For the global representations which are called occasionally holistic methods, the region of interest (ROI) of a person is encoded as a whole. In most cases, the labeling or detection of body parts are not required. Instead, the features are computed densely on a grid bounded by region of interest. The subject is usually derived from an image through applying background subtraction. The processing of global representations is based on low-level information taken from silhouettes, edges or optical flow (Poppe, 2010). However, these methods are susceptible to noise, occlusions and variations in camera viewpoint. Many research

studies argued that silhouette data provides strong cues for activity recognition with the benefit of being insensitive to texture, contrast and color changes (Weinland et al., 2011). However, silhouette-based methods depend on the accuracy of background segmentation which cannot be guaranteed in outdoor scenes. Recent studies argued that noisy silhouettes can be employed for activity recognition through the use of better matching techniques including the chamfer distance, phase correlation or shape context descriptor derived from silhouette data (Ogale, Karapurkar, & Aloimonos, 2007; Oikonomopoulos, Patras, & Pantic, 2005). Another important type of features used for global representation is optical flow extracted from consecutive frames to represent the motion whilst the subject performs an activity.

Wang (Wang, Huang, & Tan, 2007) applied the *R* transform on the extracted silhouettes reporting that the obtained representation is translation and scale invariant. The main benefit of the *R* transform is its low computational cost as well as it geometric invariance. A set of HMMs are employed for training the extracted features in order to recognize activities. Yamato *et al* quantized silhouette images into super pixels such that each pixel indicates the ratio of black to white pixels within the considered smaller region (Yamato, Ohya, & Ishii, 1992). Weinland (Weinland & Boyer, 2008) described a compact and efficient representation which is based on matching a set of discriminative static landmark pose models. The method does not depend on or take into account the temporal ordering of sequences. In their work, silhouette models are matched against edge data using the Chamfer distance and therefore eliminating the need for background segmentation. For the use of optical flow, Polana and Nelson computed the temporal texture to recognize events based on their motion. For human activity recognition, features are based on the optical flow magnitude contained within non-overlapping cells of a regular grid (Nelson & Polana, 1992; Polana & Nelson, 1994). In a different study, Ali and Shah derived a set of kinematic-based features from the optical flow such as divergence, velocity, symmetric and anti-symmetric flow fields. Multiple instance learning method is used together with Principal Component Analysis to determine the kinematic modes (Ali & Shah, 2010).

Local-Based Approaches

For activity recognition using local representations, a collection of independent patches within an image are analyzed to generate a discriminative feature vector for the observed activity. Local representations do not require accurate localization or background subtraction and enjoy the benefits of being to some extent invariant to appearance transformation, background clutter and partial occlusion (Poppe, 2010). Local patches are described by local grid-based descriptors that would summarize locally the observation within grid cells for the case of still frames. In contrast to the global representation, the local features are not linked or related to specific body parts or spatial positions of an image. Actions or activities are encoded based on the statistics of the sparse features. The main benefit of using local features is the un-necessity for people detection or the localization of the different body parts (Weinland et al., 2011). Space-time interest point descriptors which are analogous to classical 2D interest points as SURF and SIFT, have become the most popular type of local features being used for action recognition (Laptev, 2005).

For the use of motion-oriented features for human activity recognition, Yeffet (Yeffet & Wolf, 2009) proposed a local trinary pattern descriptor for encoding human motion from a sequence of frames. The trinary number is generated from a matching process of patches of a given frame against adjacent patches residing on both the previous and next frames respectively. The matching process is based on the self-similarity descriptor for textures (Shechtman & Irani, 2007). The encoding of action is done in the same way as the local binary operator to describe the displacement of patches between adjacent

frames. A histogram-based feature vector is constructed from the concatenation resulting from the image divided into a grid. As an extension of their work, Kliper-Gross (Kliper-Gross, Gurovich, Hassner, & Wolf, 2012) employed the same approach of the local trinary motion pattern renamed as Motion Interchange Pattern (MIP) for the automated recognition of human activities. Kliper-Gross presented a suppression mechanism in order to decouple static edges from edges related to motion. Further, in order to account for camera movement, motion compensation procedure is integrated within the actual local motion description based on affine transformation. For the classification stage, bag of visual features are used together with support vector machines. Oshin (Oshin et al., 2014) presented the Relative Motion descriptor for activity recognition in unconstrained scenarios using motion induced cues only. The descriptor is based on the relative distribution of spatio-temporal interest points my measuring the response strength of such points within localized regions.

The optical flow is used also for human action recognition via local representation of features. Chaudhry (Chaudhry, Ravichandran, Hager, & Vidal, 2009) argued that the recent use of complex histogram-based descriptors can fail at some point as they live on a non-Euclidean space. A Histogram of oriented optical flow (HOOF) is proposed with the merit of scale or motion direction invariance. The HOOF features are derived at every frame without the need for prior segmentation or background subtraction. The Binet-Chauchy kernels are extended to allow the matching of non-linear histograms of time series. The method was evaluated on the Weizmann human action dataset reporting a high classification rate of 95.66%. Ikizler (Ikizler, Cinbis, & Duygulu, 2008) combined the use of boundaries of a human figure fitted via small line segments together with motion information estimated via optical flow. The Hough transform is applied to detect line segments. The compact representation presented in their work was tested in different challenging conditions with high accuracy for action recognition. Feature selection is applied to compact the original feature space from 108 dimensions into a smaller space of 30 features. Martinez (Martínez, Manzanera, & Romero, 2012) computed optical flow to approximate the velocity for every pixel. The obtained flow vectors are accumulated into a per-frame histogram weighted by the norm whilst the motion orientations are quantized into 32 main directions. A histogram-based descriptor of 192 bins is obtained for every action. Results conducted on the Weizmann dataset shows the method can achieve an average accuracy of 95% using the support vector machine classifier.

In a different study by (Ladjailia, Bouchrika, Merouani, & Harrati, 2015b), the authors proposed an approach to encode a sequence of frames into a feature vector describing the performed action by a person. The method does not depend on background subtraction for the derivation of motion features. This is because it is computationally expensive and complex to deploy background subtraction for real-time surveillance applications due the process of updating the background model which is influenced by a number of factors such as background clutter, weather conditions and other outdoor environmental effects. Inspired by the work of Kliper-Gross et al., (2012) for proposing the Motion Interchange Pattern for action recognition together with the fact that local descriptors are known for their effectiveness and robustness for encoding texture for recognition purposes including biometrics, the local descriptor which captures the motion of the local structure based on estimating optical flow. Provided that there is a motion of a small patch at frame t to the next frame $t+1$, there is a high probability that a similar patch would be induced within the neighboring region of the original patch position at the previous frame. The proposed descriptor is based on constructing a feature that reflects the patch displacement from frame to frame based.

Because of the common increase of image self-similarity regions, the block matching using simple similarity operators can fail in distinguishing to between similarity caused by motion and similar static textures. In addition, the matching can be difficult as moving patches may have their appearances changed due to the non-rigid nature of the human motion. The optical flow is instead harnessed to better estimate the motion information from video sequences. Optical flow is one of the most active research areas in computer vision due to their central role in various fields of applications such as autonomous vehicle or robot navigation, visual surveillance and fluid flow analysis. The main basis of optical flow is to observe the displacement of intensity patterns (Fortun, Bouthemy, & Kervrann, 2015). This pattern is a result of the apparent motion of objects, surfaces, and edges in a visual scene caused by the relative movement between an observer and the scene (Burton & Radford, 1978). In other words, optical flow can arise either from the relative motion of the object or camera. For a given image *I*, the constraint for optical flow states that the gray intensity value of a moving pixel *I(x,y,t)* at time *t* stays constant over time as expressed as:

$$I\left(x,y,t\right)-I\left(x+V_{x},y+V_{y},t+1\right)=0 \tag{1}$$

To solve Equation 1 which has two unknowns, constraints are required to be added to ease finding a solution. There are several solutions proposed in the literature. Differential methods are the most used method. The method of Lucas, Kanade, et al., (1981) is considered for estimating the optical flow vector. The method is based on the principle that relative motion of brightness content between two successive images is small and approximately constant within a local neighborhood of a given point *p*. Therefore, the optical flow equation is assumed to hold for all points within the smaller neighborhood region centered at *p*. The lucas-kanade method solves the inherent ambiguity of the optical flow equation via combining information for several close pixels.

Based on a triplet of frames denoted as *previous*, *current* and *next*, a descriptor number *d* is constructed for every pixel for the current image through computing two optical flow images for v_{prev}: (*previous, current*) and v_{next}: (*current, next*). Thresholding is applied such that it is based on the magnitude of the velocity flow considering only values greater than *tau=0.5*. Based on the location of the angular values within the polar coordinate system which is equally divided into 8 numbered sections of 45 degrees from 1 to 8, the optical flow vector is converted into a number reflecting the order within the eighth circular portions. This is denoted using the function *AngIndex* as expressed in Equation 2. The zero indexes refer that there is no motion where the magnitude of the optical flow is less than the threshold *tau* . Both of the two digits resulting at every pixel from the next and previous frames are concatenated together to generate a number of base 9 which is converted to a decimal number.

$$d=\ AngIndex\left(v_{prev}\right)+AngIndex\left(v_{next}\right)*9 \tag{2}$$

The number *d* serves as a descriptor for the motion at a pixel level. Experimentally, it is observed that a simple action can be fully contained within only 15 frames based on video recorded at a frame rate of 25. Therefore, the encoding process is performed for every pixel for the seven different triplets

of consecutive frames taken from a video. The motion orientation histogram for a triplet is computed as shown in Equation 3. *b* is a Boolean function returning 1 for true cases and 0 for false conditions

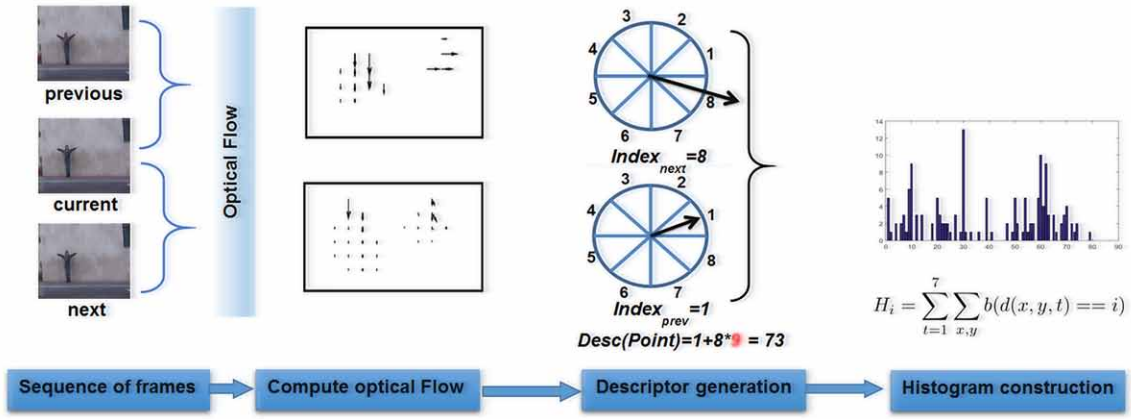
$$H_i = \sum_{x,y} b\left(d\left(x,y,t\right) == i\right) \tag{3}$$

Figure 2 outlines the procedure to estimate the histogram of motion-based features using optical flow. Various features that could potentially describe better the motion are generated based on simple fusion operations including summation and statistical operators being applied on the set of motion orientation histograms for the triplets of frames. Equation 4 shows the obtained feature vector by concatenation of different histograms. H^t refers the histogram obtained at t^{th} triplet of frames. STD is an abbreviation for the standard deviation. The resulting action vector consists of features describing purely local motion features of the human body without any information describing neither the global structure of the activity nor the anthropometric measurements of the human body.

$$F = \left[H^1 \ldots H^7 \quad Mean\left(H^{1\ldots7}\right) \quad STD\left(H^{1\ldots7}\right) \quad \sum_{t=1}^{7} H^t \right] \tag{4}$$

The feature selection process is considered in this research to derive the most discriminative features and suppress the redundant and irrelevant components which may degrade the classification rate. Because it is infeasible to conduct a brute force search procedure for all possible combinations of subsets to derive the optimal feature subset due to the high dimensionality of the raw feature vector. Alternatively, the Adaptive Sequential Forward Floating Selection (ASFFS) search algorithm (Somol, Pudil, Novovi\vcová, & Pacl\ik, 1999) is harnessed to reduce the number of features. The feature subset selection procedure is purely based on an evaluation function that assesses the discriminativeness of each component or set of features in order to derive the optimal subset of features for the classification process. Validation-based evaluation criterion is described to pick up the subset of features that would minimize the classification

Figure 2. Construction of histogram using motion-based features via optical flow



errors and ensure larger inter-class separability between the different classes. As opposed to the voting paradigm employed by the *KNN* classifier, the evaluation function utilizes coefficients *w* that signify the significance of the most nearest neighbors of the same class. The probability score for a candidate s_c to belong to a cluster *c* is expressed in the following Equation 5 as:

$$f\left(s_c\right) = \frac{\sum_{i=1}^{N_c-1} z_i w_i}{\sum_{i=1}^{N_c-1} w_i} \tag{5}$$

Where N_c is the number of instances within cluster *c*, and the coefficient w_i for the i^{th} nearest instance is inversely related to proximity as given:

$$w_i = \left(N_c - i\right)^2 \tag{6}$$

The value of z_i is defined as:

$$z_i = \begin{cases} 1 & if\ nearest\left(sc, i\right) \in c \\ 0 & otherwise \end{cases} \tag{7}$$

Such that the *nearest(s_c,i)* function gives the i^{th} nearest instance to the instance s_c. The Euclidean distance metric is used to deduce the nearest neighbours from the same class. The significance for a subset of features is based on the validation-based metric which is computed using the leave-one-out cross-validation rule. The human action signature is made as the subset of features *S* among the feature space *F* attaining the maximum value which is the average sum of *f* computed across the *N* instances *x* as expressed the following equation:

$$Signature = \arg\max\left(\frac{\sum_{x=1}^{N} f_S\left(x\right)}{N}\right) \tag{8}$$

After running the feature selection procedure on the obtained raw features, an optimal action signature is derived containing 648 features. The Correct Classification Rate is estimated using the K-nearest neighbour (KNN) classifier with *k=3* using the leave-one-out cross-validation rule. The KNN rule is applied at the classification phase due to its simplicity and therefore fast computation besides the ease of comparison to other existing methods. Using the Cumulative Match Score (CMS) evaluation method which was introduced by Phillips in the FERET protocol, we have correctly classified 95.02% of the 20 basic actions at rank *R=1* and 100% at rank *R=9*. Figure 3 shows the CMS curve for the classification process. The achieved results promising because the recognition is based purely on local motion information and this can be boosted through adding global features. The Receiver Operating Characteristics (ROC) curves are plotted in Figure 3 to express the verification results for estimating the similarity between two

instances across all pairs. In the verification process, the instances from database are verified to check if they belong to the claimed class labels based on computing the Euclidean distance. The thresholding function described in Feature Selection section is used to express whether the two pairs belong to the claimed class. In order to plot the False Acceptance Rate (FAR) versus the False Rejection Rate (FRR), different score thresholds are used. Using the human action signature derived from dynamics, the system achieved equal error rate of 1.89% is obtained (Ladjailia, Bouchrika, Merouani, & Harrati, 2015a)

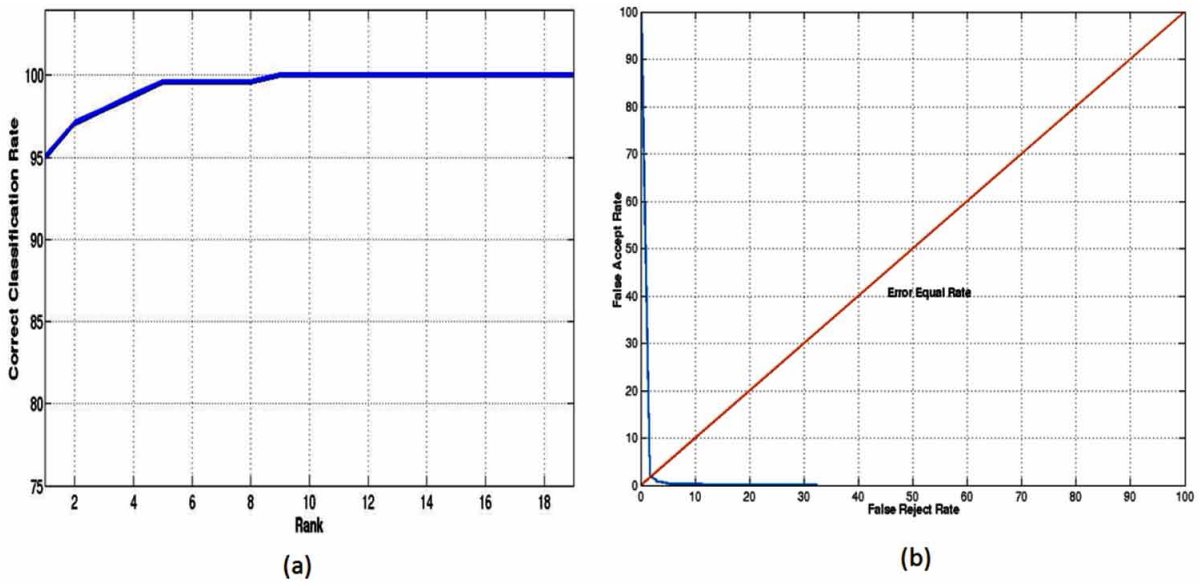
DATASETS FOR HUMAN ACTIVITIES

There are several datasets made publicly available to the research community to validate their methods for automated activity recognition and provide a common ground for researchers to compare their results on the same dataset. Most of the early datasets are constructed with a single camera containing a dozen of simple actions for a limited number of people. Recording is usually done in controlled environment with simple settings. There are recent emerging datasets based manual annotations of video clips being taken from movies and videos uploaded to online services.

KTH Dataset

The KTH dataset is constructed by the KTH Royal Institute of Technology in Stockholm, Sweden (Schüldt, Laptev, & Caputo, 2004). The dataset consists of 2,391 video sequences containing six types of human actions including: running, walking, jogging, boxing, hand clapping and hand waving. The

Figure 3. Classification results for human activity recognition: (a) cumulative match score plot. (b) receiver operating characteristic curve

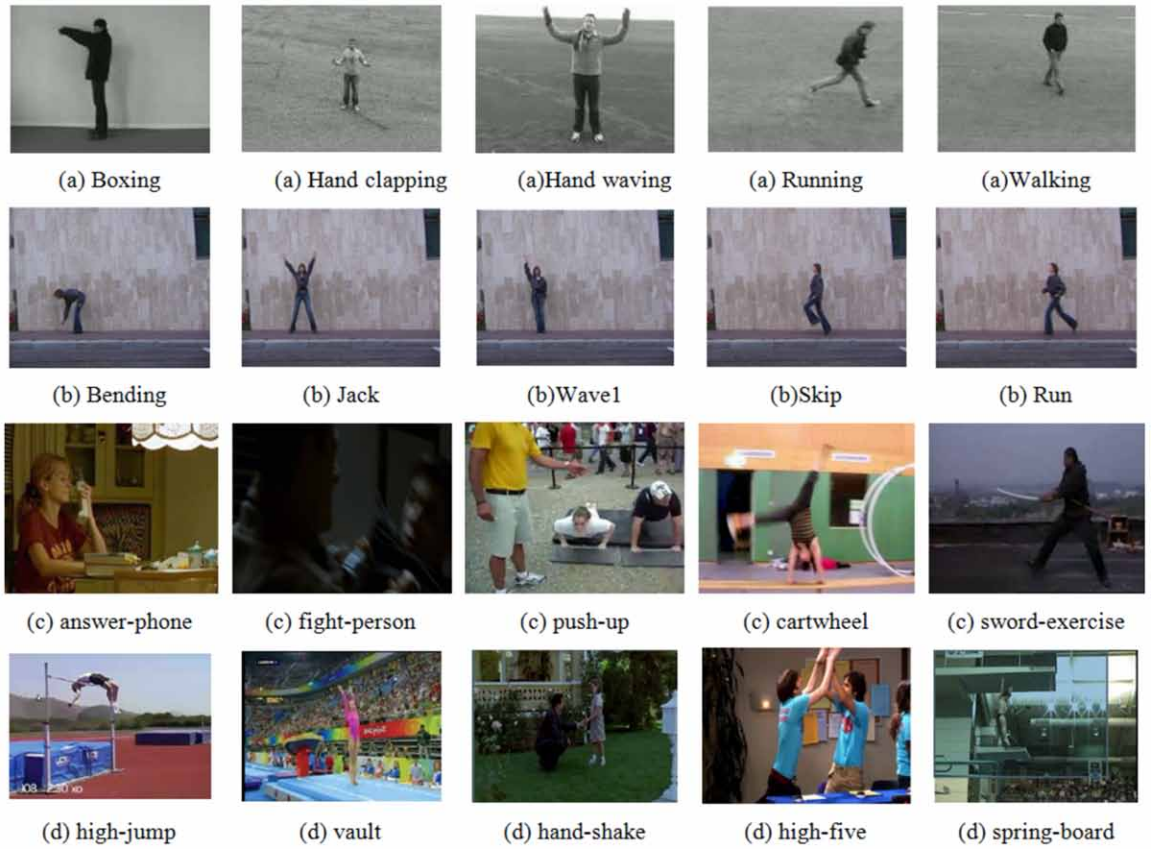


actions are performed by 25 people with three different outdoor scenarios and an indoor session. For the outdoor sessions, there are variations in terms of illumination, scale and clothing appearance. All video sequences were recorded over homogeneous backgrounds with a static camera with a frame rate of 25 frames per second. The videos are resized downward to the spatial resolution of 160 × 120 pixels with an average duration of four seconds. As for benchmarking, the authors suggested dividing the dataset with respect to the individuals into a training set containing 8 people, a validation set with subjects and a test dataset consisting of 9 persons. The dataset is made online publicly available for download as AVI video files Figure 4 shows examples taken from the KTH dataset.

Weizmann Dataset

The Weizmann dataset (Blank et al., 2005) contains 90 video sequences with low-resolution of *180x144* recorded at frame rate of 50 frames per second in de-interlaced mode. There are nine different people, each performing 10 natural activities. The performed actions include: walk, run, skip, jumping-jack (or shortly "jack"), jump forward on two legs (or "jump"), jump in place on two legs (or "pjump"), gallop

Figure 4. Human Activity Datasets: (a) KTH (Schüldt et al. 2004), (b) Weizmann (Blank et al. 2005), (c) HMDB51 (Kuehne, 2001), (d) Hollywood2 (Marszalek et al. 2009)



sideways (or "side"), wave two hands (or "wave2"), wave one hand (or "wave1") and bend. Silhouettes from the video sequences are provided with the dataset generated via subtracting the median background from each of the sequences. Examples from the Weizmann dataset are shown in Figure 4. In (Ladjailia et al., 2015b), the authors manually collected a dataset containing 241 video sequences for 19 different basic actions by decomposing an activity into primitive actions. Each video consists of 15 frames which are all checked to better describe the complete action.

Hollywood Dataset

There are two versions of the Hollywood dataset (Marszalek, Laptev, & Schmid, 2009). The first release covers only 8 basic actions with a limited number of video clips. The second version of the Hollywood dataset contains 12 different classes of human actions and 10 classes of scenes distributed over 3,669 clips with a total duration of approximately 20.1 hours of video footage. The dataset is setup with the aim to provide a comprehensive benchmark for human activity recognition in realistic and challenging environments. The dataset is constructed by taking video clips from 69 movies through the aid of automated movie script processing to retrieve scene descriptions. The list of actions contained in this dataset include: Answer Phone, Drive Car, Eat, Fight Person, Get Out Car, Hand Shake, Hug Person, Kiss, Run, Sit Down, Sit Up and Stand Up. The videos contained in the dataset are subjected to various factors as occlusions, camera movements and dynamic backgrounds which would make it more challenging. The database is split into a training and test subsets such that the two subsets do not share samples from the same film.

HMDB51 Dataset

The Human Motion DataBase (HMDB51) (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011) contains 51 different human action categories such that every activity class comes with at least 101 video clips with a total of 6,766 videos. The video are extracted from a wide variety of sources including Youtube.com. The authors claim that the HMDB51 dataset is the largest and most realistic database devoted for human activity recognition. Each video clip is manually annotated and validated by at least two people to ensure the consistency. Information meta tags are provided to allow better and precise selection of testing data and training for flexible evaluation of the performance of the proposed approach. The tags for each video describe the camera view-point, the presence of camera movement, the video quality, and the number of people in the scene. The original videos taken to extract the activity clips vary in size and frame rate. Therefore, In order to ensure consistency across the dataset, the heights of all clips are resized to 240 pixels. The width is rescaled accordingly to maintain the aspect ratio constant. The frame rate is resampled to 30 frames per second for all video clips.

CHALLENGES AND DIFFICULTIES

Despite the recent outstanding advancements in computer vision and pattern recognition technologies, the automated marker-less extraction and recognition of human activities are proven to be a challenging task. Although, the problem can be stated in simple terms, given a sequence of frames with one or more people performing a given activity, can an automated system recognize the activity being performed?

The solution is difficult to devise or implement. The difficulties stem from a large number of factors that can be related to one three following classes:

- **Person:** Most of the existing methods proposed for human activity recognition rely on sensors or special markers mounted on the subject (Lara & Labrador, 2013). For a marker-less approach, the articulated nature of human body which encompasses a wide range of possible motion transformations in addition to self-occlusion and appearance variability, exacerbate further complexity on the task of visual feature extraction for the process of human activity recognition (Moeslund, Hilton, & Krüger, 2006). Even through, there is ample research about pedestrian detection for real-time applications with reported higher accuracy, the localization of people is still hard to achieve in cluttered environments with the desired performance (Poppe, 2010; Tang, Andriluka, & Schiele, 2014). Furthermore, there is a substantial variation in terms of the appearance and the time needed for performing an action by different people. The variation is determined and influenced by various factors such as age, emotional state and fatigue which can severely change the way we perform actions.
- **Acquisition Environment:** Challenges related to the acquisition environment may include background clutter, illumination, camera movement and viewpoint as well as occlusion by other objects in the scene. Dynamic background adds further complexity for foreground segmentation and extracting motion or kinematic features related to people. Further, the challenges become even harder when using a moving camera. For the change of camera viewpoint or position, the same action can be represented and understood differently when changing the viewpoint or even the distance from the camera (Weinland et al., 2010). Low resolution and poor video quality due the temporal and spatial down-sampling are common in current surveillance technologies which exacerbate further obstacles (Rahman, See, & Ho, 2015). Even though recent research studies argued about the possibility of recognizing human actions from a number of limited frames (Schindler & Van Gool, 2008), it is still a difficult process to achieve an acceptable classification rate for cases of low-frame rates or frames being dropped.
- **Activity Understanding:** An activity can be performed at various ways by different people depending on the context (Zhu, Nayak, & Roy-Chowdhury, 2013) or even culture of the performer. For instance, human gestures or actions to express joy and happiness can take different ways and forms. Inversely, the same activity performed by different people can have different semantic meanings. Furthermore, activities can interleave within each other and performed in parallel rather than a sequential fashion. For instance a person can use their computer whilst eating at the same time or answering the phone. Hence, the system needs to infer between primary and secondary activities in the scene.

CONCLUSION

The perception of human motion is one of the most important skills people possess, and our visual system provides particularly rich information in support of this skill. Yet, attempts and efforts to understand the human visual system or to devise an artificial solution for visual perception have proven to be a difficult task. Human motion analysis has received much attention from researchers in the last two decades due to its potential use in a plethora of applications. This field of research focuses on the perception and

recognition of human activities. The recognition of human activity is of prime importance for various applications as automated visual surveillance. The research area of human activity recognition is closely related to other fields of research that analyze human motion such as human computer interaction and biomechanical engineering. Although, there is a considerable body of work devoted to human action recognition, most of the methods are evaluated on datasets recorded in simplified settings. More recent research has shifted focus to natural activity recognition in unconstrained scenes with more complex settings. Various types of features are considered for the representation of human actions that can be grouped in three major categories: Pose-based, global and local methods. There are several datasets made publicly available to the research community to validate their methods for automated activity recognition and provide a common ground for researchers to compare their results on the same dataset.

REFERENCES

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys*, *43*(3), 16. doi:10.1145/1922649.1922653

Akita, K. (1984). Image Sequence Analysis of Real World Human Motion. *Pattern Recognition*, *17*(1), 73–83. doi:10.1016/0031-3203(84)90036-0

Ali, S., & Shah, M. (2010). Human action recognition in videos using kinematic features and multiple instance learning. *Pattern Analysis and Machine Intelligence*. *IEEE Transactions on*, *32*(2), 288–303.

Bingham, G. P., Schmidt, R. C., & Rosenblum, L. D. (1995). Dynamics and the Orientation of Kinematic Forms in Visual Event Recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(6), 1473–1493. doi:10.1037/0096-1523.21.6.1473 PMID:7490589

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, *58*(1), 47–73. doi:10.1146/annurev.psych.57.102904.190152 PMID:16903802

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (Vol. 2, pp. 1395-1402).

Bouchrika, I. (2008). *Gait Analysis and Recognition for Automated Visual Surveillance*. University of Southampton.

Bouchrika, I., Carter, J. N., Nixon, M. S., Morzinger, R., & Thallinger, G. (2010). Using gait features for improving walking people detection.*20th International Conference on Pattern Recognition (ICPR)* (pp. 3097-3100). doi:10.1109/ICPR.2010.758

Bouchrika, I., & Nixon, M. S. (2006). Markerless Feature Extraction for Gait Analysis.*IEEE SMC Chapter Conference on Advanced in Cybernetic Systems*.

Burton, A., & Radford, J. (1978). *Thinking in perspective: critical essays in the study of thought processes*. Methuen.

Cedras, C., & Shah, M. (1995). Motion-based Recognition: A survey. *Image and Vision Computing*, *13*(2), 129–155. doi:10.1016/0262-8856(95)93154-K

Chaudhry, R., Ravichandran, A., Hager, G., & Vidal, R. (2009). *Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions.* Paper presented at the Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. doi:10.1109/CVPR.2009.5206821

Comaniciu, D., Ramesh, V., & Meer, P. (2000). Real-time Tracking of Non-Rigid Objects using Mean Shift.*Proceedings. IEEE Conference on Computer Vision and Pattern Recognition*, 2. doi:10.1109/CVPR.2000.854761

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. doi:10.1109/CVPR.2005.177

Derrington, A. M., Allen, H. A., & Delicato, L. S. (2004). Visual mechanisms of motion analysis and motion perception. *Annual Review of Psychology*, *55*(1), 181–205. doi:10.1146/annurev.psych.55.090902.141903 PMID:14744214

Deshayes, R., Mens, T., & Palanque, P. (2013). *A generic framework for executable gestural interaction models.* Paper presented at the Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on. doi:10.1109/VLHCC.2013.6645240

Dittrich, W. H. (1993). Action Categories and the Perception of Biological Motion. *Perception*, *22*(1), 15–22. doi:10.1068/p220015 PMID:8474831

Fortun, D., Bouthemy, P., & Kervrann, C. (2015). Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, *134*, 1–21. doi:10.1016/j.cviu.2015.02.008

Gavrila, D., & Davis, L. (1995). *Towards 3-d model-based tracking and recognition of human movement: a multi-view approach.* International workshop on automatic face-and gesture-recognition.

Goddard, N. H. (1992). *The Perception of Articulated Motion: Recognizing Moving Light Displays*. University of Rochester.

Guo, Y., Xu, G., & Tsuji, S. (1994). Understanding Human Motion Patterns. *Pattern Recognition, Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on, 2*.

Ikizler, N., Cinbis, R. G., & Duygulu, P. (2008). *Human action recognition with line and flow histograms.* Paper presented at the Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. doi:10.1109/ICPR.2008.4761434

Isard, M. C., & Blake, A. C. (1998). CONDENSATION: Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, *29*(1), 5–28. doi:10.1023/A:1008078328650

Johansson, G. (1973). Visual Perception of Biological Motion and a Model for its Analysis. *Perception & Psychophysics*, *14*(2), 201–211. doi:10.3758/BF03212378

Karaulova, I. A., Hall, P. M., & Marshall, A. D. (2000). A Hierarchical Model of Dynamics for Tracking People with a Single Video Camera. In *Proceedings of the 11th British Machine Vision Conference*, *1*, 352-361. doi:10.5244/C.14.36

Kliper-Gross, O., Gurovich, Y., Hassner, T., & Wolf, L. (2012). Motion interchange patterns for action recognition in unconstrained videos. *European Conference on Computer Vision*, (pp. 256-269). doi:10.1007/978-3-642-33783-3_19

Ko, T. (2008). *A survey on behavior analysis in video surveillance for homeland security applications.* Paper presented at the Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE. doi:10.1109/AIPR.2008.4906450

Kozlowski, L. T., & Cutting, J. E. (1978). Recognizing the Gender of Walkers from Point-Lights Mounted on Ankles: Some Second Thoughts. *Perception & Psychophysics*, *23*(5), 459. doi:10.3758/BF03204150

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). *HMDB: a large video database for human motion recognition.* Paper presented at the Computer Vision (ICCV), 2011 IEEE International Conference on. doi:10.1109/ICCV.2011.6126543

Kühnel, C., Westermann, T., Hemmert, F., Kratz, S., Müller, A., & Möller, S. (2011). I'm home: Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies*, *69*(11), 693–704. doi:10.1016/j.ijhcs.2011.04.005

Ladjailia, A., Bouchrika, I., Merouani, H. F., & Harrati, N. (2015a). Automated Detection of Similar Human Actions using Motion Descriptors. *16th international conference on Sciences and Techniques of Automatic control and computer engineering (STA).* IEEE.

Ladjailia, A., Bouchrika, I., Merouani, H. F., & Harrati, N. (2015b). On the Use of Local Motion Information for Human Action Recognition via Feature Selection.*4th IEEE International Conference on Electrical Engineering (ICEE)*. doi:10.1109/INTEE.2015.7416792

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2-3), 107–123. doi:10.1007/s11263-005-1838-7

Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, *15*(3), 1192–1209. doi:10.1109/SURV.2012.110112.00192

Lucas, B. D., & Kanade, T. et al. (1981). An iterative image registration technique with an application to stereo vision. *IJCAI*, *81*, 674–679.

Marszalek, M., Laptev, I., & Schmid, C. (2009). *Actions in context.* Paper presented at the Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. doi:10.1109/CVPR.2009.5206557

Martínez, F., Manzanera, A., & Romero, E. (2012). *A motion descriptor based on statistics of optical flow orientations for action classification in video-surveillance. In Multimedia and Signal Processing* (pp. 267–274). Springer.

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*(2), 90–126. doi:10.1016/j.cviu.2006.08.002

Nelson, R. C., & Polana, R. (1992). Qualitative recognition of motion using temporal texture. *CVGIP. Image Understanding*, *56*(1), 78–89. doi:10.1016/1049-9660(92)90087-J

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, *79*(3), 299–318. doi:10.1007/s11263-007-0122-4

Ogale, A. S., Karapurkar, A., & Aloimonos, Y. (2007). *View-invariant modeling and recognition of human actions using grammars. In Dynamical vision* (pp. 115–126). Springer.

Oikonomopoulos, A., Patras, I., & Pantic, M. (2005). Spatiotemporal salient points for visual recognition of human actions. *Systems, Man, and Cybernetics, Part B: Cybernetics. IEEE Transactions on*, *36*(3), 710–719.

Oshin, O., Gilbert, A., & Bowden, R. (2014). Capturing relative motion and finding modes for action recognition in the wild. *Computer Vision and Image Understanding*, *125*, 155–171. doi:10.1016/j.cviu.2014.04.005

Polana, R., & Nelson, R. (1994). *Low level recognition of human motion (or how to get your man without finding his body parts).* Paper presented at the Motion of Non-Rigid and Articulated Objects. doi:10.1109/MNRAO.1994.346251

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, *28*(6), 976–990. doi:10.1016/j.imavis.2009.11.014

Rahman, S., See, J., & Ho, C. C. (2015). *Action Recognition in Low Quality Videos by Jointly Using Shape, Motion and Texture Features.* Paper presented at the IEEE Int. Conf. on Signal and Image Processing Applications. doi:10.1109/ICSIPA.2015.7412168

Ren, Z., Meng, J., Yuan, J., & Zhang, Z. (2011). *Robust hand gesture recognition with kinect sensor.* Paper presented at the 19th ACM international conference on Multimedia. doi:10.1145/2072298.2072443

Rohr, K. (1994). Towards Model-Based Recognition of Human Movements in Image Sequences. *CVGIP. Image Understanding*, *59*(1), 94–115. doi:10.1006/ciun.1994.1006

Rui, Y., Huang, T. S., & Chang, S.-F. (1999). Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, *10*(1), 39–62. doi:10.1006/jvci.1999.0413

Schindler, K., & Van Gool, L. (2008). *Action snippets: How many frames does human action recognition require?* Paper presented at the Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. doi:10.1109/CVPR.2008.4587730

Schüldt, C., Laptev, I., & Caputo, B. (2004). *Recognizing human actions: a local SVM approach.* Paper presented at the Pattern Recognition. doi:10.1109/ICPR.2004.1334462

Shah, M., & Jain, R. (2013). *Motion-based recognition* (Vol. 9). Springer Science & Business Media.

Shechtman, E., & Irani, M. (2007). *Matching local self-similarities across images and videos.* Paper presented at the Computer Vision and Pattern Recognition. doi:10.1109/CVPR.2007.383198

Somol, P., Pudil, P., Novovičová, J., & Paclík, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, *20*(11), 1157–1163. doi:10.1016/S0167-8655(99)00083-5

Suma, E. A., Krum, D. M., Lange, B., Koenig, S., Rizzo, A., & Bolas, M. (2013). Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit. *Computers & Graphics*, *37*(3), 193–201. doi:10.1016/j.cag.2012.11.004

Tang, S., Andriluka, M., & Schiele, B. (2014). Detection and tracking of occluded people. *International Journal of Computer Vision*, *110*(1), 58–69. doi:10.1007/s11263-013-0664-6

Troje, N. F., Westhoff, C., & Lavrov, M. (2005). Person Identification from Biological Motion: Effects of Structural and Kinematic Cues. *Perception & Psychophysics*, *67*(4), 667–675. doi:10.3758/BF03193523 PMID:16134460

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology*. *IEEE Transactions on*, *18*(11), 1473–1488.

Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, *29*(10), 983–1009. doi:10.1007/s00371-012-0752-6

Wang, Y., Huang, K., & Tan, T. (2007). Human activity recognition based on r transform. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (pp. 1-8).

Weinland, D., & Boyer, E. (2008). Action recognition using exemplar-based embedding. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-7).

Weinland, D., Özuysal, M., & Fua, P. (2010). *Making action recognition robust to occlusions and viewpoint changes. In Computer Vision–ECCV 2010* (pp. 635–648). Springer.

Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, *115*(2), 224–241. doi:10.1016/j.cviu.2010.10.002

Welch, G., & Bishop, G. (2001). An Introduction to the Kalman Filter. *ACM SIGGRAPH 2001 Course Notes*.

Yam, C.-Y., & Nixon, M. (2009). Gait Recognition, Model-Based. In Encyclopedia of Biometrics, (pp. 633-639). Academic Press.

Yamato, J., Ohya, J., & Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. *Proceedings CVPR*, *92*, 1992.

Yeffet, L., & Wolf, L. (2009). Local trinary patterns for human action recognition. *Computer Vision, 2009 IEEE 12th International Conference on* (pp. 492-497).

Zhu, Y., Nayak, N. M., & Roy-Chowdhury, A. K. (2013). Context-aware activity recognition and anomaly detection in video. *Selected Topics in Signal Processing*. *IEEE Journal of*, *7*(1), 91–101.

KEY TERMS AND DEFINITIONS

Action: Is considered as a simple activity referring to simple pattern performed by a person during a short period of time lasting a few seconds. Examples of actions may include raising hands, bending, sitting and even walking.

Activity: Is defined as a composite sequence of actions executed by either a single person or several people interacting with each other. Examples of activities are like leaving an unattended bag, shaking hands or assaulting a pedestrian.

Feature Extraction: Is the process of estimating a set of measurements either related to the configuration of the whole body or the configuration of the different body parts in a given scene and tracking them over a sequence of frames.

Global Feature: This is the visual characteristics taken from an image in holistic fashion such that the region of interest of a person is encoded as a whole. In most cases, the features are computed densely on a grid bounded by region of interest.

Human Activity Recognition: Is the process to automatically infer the action or activity being performed by a person or group of people via the use of computer vision methods. This may involve the analysis and recognition of different motion patterns in order to produce a high-level semantic description for the human activities.

Local Feature: Is a type of low-level cues which are extracted from smaller portions of the image with no connection made their spatial locations within the human body.

Motion: Is a spatio-temporal event defined as the change of spatial location over time. Given some visual input, the visual perception of motion is regarded as the process by which the visual system acquires perceptual knowledge such as the speed and direction of the moving object.