HUMAN MOTION ANALYSIS, RECOGNITION AND UNDERSTANDING IN COMPUTER VISION: A REVIEW

Ahmed Nabil Mohamed [1, *], Mohamed Moanes Ali [2]

[1]Assistant Lecturer, Department of Computer and Information Systems, Sadat Academy
[2]Professor of Electrical Eng., Depart. of Computers and Systems, Eng. Faculty, Minia University

Received 21 February 2013; accepted 15 September 2013

ABSTRACT

The ultimate goal of computer vision is to understand the scene correctly through various steps of acquiring, processing, analyzing and understanding different kinds of information obtained by different kinds of sensors. Human motion analysis, recognition, and understanding is one of the very hottest topics within computer vision.

The purpose of this article is to shed some light on the very important subject of human motion understanding, so it can be a good insight for a novice computer vision researcher in this field. The article tries to spot many of the most cited and recent reviews in the field, indicating its wide demanding applications, different taxonomies used in structuring different surveys, various approaches used to solve different stages of the problem, different types of sensors used for data acquisition, some taxonomies used for classifying motions, and various processes involved in motion analysis, followed by a discussion and a conclusion.

Keywords: Human Motion Analysis, Action Recognition, Behavior Understanding, Computer Vision.

1. Introduction

Human motion understanding has been a desirable target for many researchers from different disciplines. Each discipline has a different aspect of the problem. With such different motivations for researchers to bear in mind, many contributions have been made in different disciplines. Integrating these contributions together provide us with a better understanding of human motion. However, this article concerns with the aspect of computer vision dealing with human motion understanding, with a focus on whole body movements. This article tries to recall some of the most cited surveys and recent ones in the past two decades trying besides stressing on the importance of the subject and its wide demanding applications, to reveal different approaches and taxonomies used in these reviews.

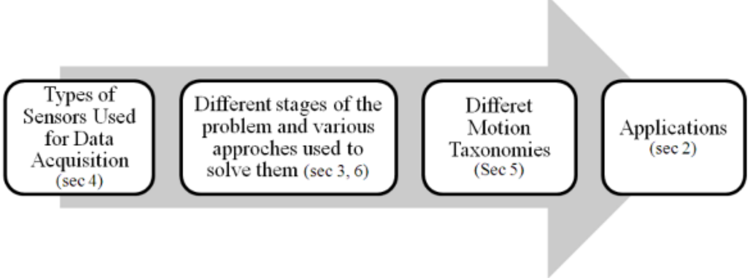
Types of Sensors Used for Data Acquisition (sec 4)
Different stages of the problem and various approches used to solve them (sec 3, 6)
Differet Motion Taxonomies (Sec 5)
Applications (sec 2)

Fig.1. a block diagram showing the proposed classification

* Corresponding author.
Email address: anabil.mohamed@live.com

The organization of the article will be as follows: Section 2 reviews some of the very potential and promising applications concerning human motion analysis and understating in computer vision. Section 3 discusses different taxonomies used by different surveys and briefly reviews many surveys ranging from the most cited to the most recent ones in the field covering many directions but mainly focus on vision-based techniques. Section 4 lists some kinds of the sensors that can be used for data acquisition. Section 5 introduces some motion taxonomies. Section 6 discusses different stages of the problem. Section 7 contains the discussion. Section 8 concludes the article.

2. Applications

Human motion analysis, recognition, and understanding has gained much interest and research in computer vision due to the wide range of demanding and promising applications. A review of several applications in different domains is listed as follows:

- Smart or Automated Surveillance: as the attention of a good, competent, and dedicated vigilant person decreases after 20 minutes into unacceptable levels due to boring and hypnotizing nature of monitoring video scenes [31], and with the growing numbers of cameras covering vast areas, the human operator becomes more costly and unreliable (e.g., a survey of CCTV (Closed Circuit Television) systems in one London borough revealed that over 75% of the institutions that apply the CCTV system had no dedicated monitoring staff [32]). Thus, the need for automated surveillance systems turns to be very urging. Some applications of smart surveillance are: suspicious behaviors and unlikely events [35] detections; understanding and describing human behaviors in dynamic scenarios (e.g., monitoring activities over a complex area using a distributed network of active video sensors [36]); access control in special areas such as military bases and important governmental units where the system should automatically obtain biometric features of the visitor and then decide whether the visitor can be cleared for entry; person-specific identification at a distance which can help the police in chasing and catching suspects by placing surveillance cameras in locations where suspects may appear such as subway stations and casinos; consumer demographics in shopping malls; crowd statistics [37, 38] and pedestrian congestion in public areas such as stores and travel sites, security applications in places such as banks [39], department stores, office buildings, parking lots, shopping centers, public transportation [40], borders, and homes.
- Behavioral Biometrics: recognizing humans based on their behavioral cues (e.g., human gait [41, 42], length, facial features, etc) does not require subject cooperation or intervention in their activities.
- Human-Computer Interaction: enables the user to control and command, e.g., gesture driven control, eye gaze tracking [43], speech recognition, sign language translation and understanding, signaling in high noise environments such as factories and airports, perceptual user interfaces [44] that allows a computer user to interact with the computer without having to use the normal keyboard and mouse by giving the computer the capability of interpreting the user's movements or voice commands.
- Virtual Reality: where the user is able to interact with a computer-simulated

environment, e.g., training of military soldiers, firefighters and rescue squads by learning in simulated environments. Virtual reality has also many applications in game and entertainment industries.

- Smart Environments: where extracting and maintaining awareness of a wide range of events and human activities take place, e.g., monitoring interactions of participants in a meeting room [45].
- Games Industry: several games use the gesture-based interactive technology where motion capture is employed to enable interaction between a player and a game through non-intrusive body movements. For example, the famous Microsoft Kinect Xbox [46, 47]
- Entertainment Industry: precise motion-capturing is used in Sci-Fi movies to replace actors with animation characters (digital avatars) [48].
- Video Annotation, Indexing, and Retrieval: as the number of videos increases rapidly due to the magnificent progress in capturing and recording technologies, accompanied by decreasing costs of cameras and storing media, the need to index and annotate various kinds of videos including personal videos, sports videos, news broadcasting, movies, surveillance videos, etc, becomes very insisting to save time and labor in retrieving them in a more easy, fast and convenient way, e.g., acquiring a certain highlight in a soccer game [49] or in news broadcasting. A review of recently developed information retrieval techniques can be found in [50].
- Physical Therapy: e.g., non-intrusive capturing of normal and pathological human movement [51], diagnosis of orthopedic patients, etc.
- Sports Motion Analysis: analyzing different sports such as the soccer sport [52] where verification of the following issues may be addressed such as referee decision, tactics analysis, automatic highlight identification, video annotation and browsing, content based video compression, automatic summarization of play, customized advertisement insertion, graphical object overlapping, player and team statistic evaluations, etc.
- Human motion analysis and synthesis: acquiring accurate movements of athletes, dancers, fighters, etc, for performance analysis, evaluation and enhancement and for training purposes.
- Robotics Learning for Imitation of Human Activities: e.g., using robots to set up or clean dinner tables, or using robots in dangerous situations or environments for experimental purposes such as car crash tests, skating when icy blast occurs, etc.
- Assisted Living or Proactive Services: assisting disabled people, elderly people, children as well as normal people, e.g., fall detection systems [53] that monitor the person's movements and call the corresponding emergency center if it detects a falling person. Chaaraoui et al. [54] provide a review on human behavior analysis for ambient-assisted Living.
- Intelligent Driver Assistance Systems: where the assisting process must be very efficient and in real time, e.g., monitoring driver awareness [55], sleep detection, airbag system control, predicting driver turn intent [56], pedestrian detection, etc.
- Safety Monitoring: e.g., detecting drowning in public swimming pools [57].

- Autonomous Mental Development [30]: this includes studying how the human brain develops its mental capabilities through examining autonomous real-time interactions with its environments using its own sensors and effectors, e.g., study the cognitive learning process of young children [58], examining the mental status of patients after traumatic accidents, etc.
- Video Compression: e.g., using model-based coding allows very low bit-rate compression.

These applications vary in their requirements (e.g., human modeling, real time processing, video resolution, controlled or uncotrolled environmets, active or passive sensing, types and number of sensors, performance robustness and accuracy, etc) to achieve human motion analysis and recognition.

3. Related Work

Many surveys have been written in the domain of human motion analysis and recognition, each with a specific focus and taxonomy to compare different publications. Factors that are used to classify previous work in human motion include: Model-based or non-model based, Explicit or implicit shape modeling, Model types (e.g., stick figures, volumetric models, surface models, etc), Human motion modeling, Human body parts involved in motion analysis, Full-body motion or body parts motion, Level of detail needed to understand human actions, space dimensionality (e.g., 2D approaches or 3D approaches), Sensor modality (e.g., visible light, infrared light, structured light, etc), Sensor multiplicity (monocular or stereo), Sensor placement (centralized or distributed), Sensor mobility (mobile or stationary), Active sensing or passive sensing, Marker-based or marker-free systems, Tracking one person or multiple persons, Various motion-types assumptions (e.g., rigid, articulated, elastic, etc), Functionality (initialization, tracking, pose estimation, and movement recognition), Image representation (global representations, local representations, application-specific representations), Object detection (e.g., shape-based, feature-based, depth-based, supervised learning, etc), Motion segmentation (e.g., background subtraction, statistical methods, temporal differencing, optical flow, etc), Object tracking (feature-based, shape-based, etc), View-invariant action representation and recognition, Spatial and temporal structure of actions, Human-object interactions and group activities, Pose representation and estimation (3D model-based, 3D model-free, example-based), Spatial action representations (body models, image models, spatial statistics), Action classification (direct classification, temporal state-space models), Action recognition (single-layered approaches or hierarchical approaches), Modeling and recognizing actions (nonparametric approaches, parametric methods, volumetric approaches), Human activity recognition (e.g., template matching approaches, state-space approaches, etc).

We will now briefly review many of these surveys to reveal various approaches used for implementing different stages of the problem. These surveys are published in the last two decades ranging from the most cited to the most recent. They cover most of the publications in this growing field of vision-based human motion analysis. The surveys are listed in a chronological order to reveal the progress of the field as follows:

Cedras and Shah [1] provide a review of motion-based recognition approaches prior to 1995, where they first discuss two theories about the interpretation of motion. After that, the authors identify two main steps involved in motion-based recognition. The first step deals with the extraction of motion information and its representation. The second step concerns with the matching of unknown inputs with constructed models. They also discuss tracking and recognition of human motion (e.g., walking, running). Moreover, they discuss several methods involved in the recognition of objects and motions such as cyclic motion detection and recognition, lipreading, and hand gestures interpretation.

Aggarwal et al. [2] provide an overview of articulated and elastic motion analysis prior to 1996. They discuss approaches used for recovering the 3D structure and motion of objects in a bottom-up strategy. These approaches are classified into two categories: model-based approaches and model-free (i.e., without a priori shape models) approaches.

Gavrila [3] discusses three approaches used in analyzing human gesture (hand motion) and whole-body motion. These approaches are: 2D approaches with explicit shapes, 2D approaches without explicit shapes, and 3D approaches.

Aggarwal and Cai [4] discuss three major areas related to human motion analysis: motion analysis of human body parts, tracking a moving human from a single view or multiple camera perspectives and recognizing human activities.

Moeslund and Granum [5] provide a comprehensive survey of human motion capture using 130 published papers from two decades prior to 2000 with much more concentration on the period (1994-2000). They use a taxonomy based on the system functionalities. Their taxonomy consists of four processes: initialization, tracking, pose estimation, and recognition.

Wang et al. [6] provide a comprehensive survey on three major issues involved in a general human motion analysis system, which are human detection, tracking, and activity understanding. The survey covers the research published from 1989 to 2001 with nearly 70% of the discussed papers were published after 1996.

Buxton [7] discusses generative models used in learning and understanding dynamic scene activity. She also discusses the use of these models in applications such as smart rooms and visual surveillance

Aggarwal and Park [8] discuss many aspects of high-level processing involved in understating human activities such as human body modeling, level of detail needed to understand human actions, approaches to human action recognition, and high-level recognition schemes with domain knowledge.

Hu et al. [9] discuss different stages involved in the visual surveillance system. These stages are environment modeling, motion detection, object classification, tracking, understanding and description of behaviors, human identification, and data fusion from multiple cameras.

Moeslund et al. [10] review over 350 publications reflecting the advances in human motion capture and analysis from 2000 to 2006. Following the taxonomy of Moeslund [5], they indicate how research has addressed novel methodologies for automatic initialization, reliable tracking and pose estimation in natural scenes, and automatic understanding of human actions and behaviors.

Yilmaz et al. [11] categorize the tracking methods based on object and motion representations. They provide detailed descriptions of representative methods in each category. They also examine their pros and cons. Moreover, they discuss the important issues related to tracking.

Poppe [12] discusses the characteristics of human motion analysis by studying pose estimation process in model-based (or generative) approaches and model-free (or discriminative) approaches. For model-based approaches, the pose estimation process consists of two phases: a modeling phase and an estimation phase. The modeling phase is the construction of a likelihood function. The estimation phase searches for the most likely pose given the likelihood surface. For model-free approaches, the pose estimation process can be accomplished through learning-based approaches or example-based approaches.

Krüger et al. [13] analyze different approaches for the representation, recognition, synthesis and understanding of action within the computer vision, robotics and artificial intelligence communities. They, first, discuss approaches that recognize human actions with and without body parts. They also discuss the grammars approach. Then, they discuss action learning and imitations for robots. In the end, they discuss plan and intention recognition.

Pantic et al. [14] discuss how far are we from enabling computers to understand human behavior. They discuss the design of human-like interactive functions including understanding and imitating certain kinds of human behaviors where requirements (such as what is communicated, how the information is passed, and in which context the information is passed on) need to be explored. They also discuss some tasks involved in modeling human behavior and understanding displayed patterns of behavioral signals (e.g., human sensing, context sensing).

Turaga et al. [15] present a comprehensive survey on representing, recognizing, and learning human actions and activities from video and related applications. They discuss the problems at different levels of complexity starting from atomic or primitive actions, where they briefly discuss low-level feature extraction. Then, they deal with actions with more complex dynamics, where they categorize the used approaches into three major classes: nonparametric, volumetric and parametric approaches. Finally, the authors consider complex activities, where they categorize various approaches into three categories: graphical models, syntactic approaches, and knowledge and logic-based approaches.

Lavee et al. [16] discuss two main components of the event understanding process: abstraction and event modeling. Abstraction is the process of molding the data into informative units to be used as input to the event model. Event modeling is devoted to describing events of interest formally and enabling recognition of these events as they occur in the video sequence.

Ji and Liu [17] provide a survey on view-invariant human motion analysis with the emphasis on view-invariant pose representation and estimation, and view-invariant action representation and recognition. They categorize view-invariant pose representation and estimation into three categories, which are: 3D model-based, model-free, and example-based. They also categorize view-invariant action representation and recognition into template-based approaches and state space approaches.

Poppe [18] discusses global and local image representations. Then he discusses human action recognition as a classification problem where he addresses issues like direct classification, temporal state-space models and action detection.

Weinland et al. [19] give an overview of the approaches used in action representation, segmentation and recognition concentrating on full-body motions, such as kicking, punching, and waving. They categorize these approaches according to how they represent the spatial and temporal structure of actions; how they segment actions; and how they learn a view-invariant representation of actions.

Aggarwal and Ryoo [20] provide a recent review of human activity recognition. They discuss the methodologies of recognizing simple actions and high-level activities. They use an approach-based taxonomy to compare the advantages and limitations of each approach. They classify all activity recognition methodologies into two categories: single-layered approaches and hierarchical approaches. They also discuss the recognition of human-object interactions and group activities.

Chen and Khalil [21] briefly discuss vision-based and sensor-based activity recognition approaches indicating a new emergent object-based approach that deals with a sensorized environment where activities are characterized by objects being used during their operation. They also discuss activity recognition algorithms. However, the main interest of their article is given to describe the general framework and the lifecycle of the ontology-based activity recognition approach. The authors also provide an exemplar case study "MakeDrink" to demonstrate the ontology-based approach indicating its support to progressive activity recognition at both coarse-grained and fined-grained levels.

Yang et al. [22] discuss different feature descriptors of an object (gradient features, color features, texture features, spatio-temporal features). Then, they categorize the tracking progresses into three groups: online learning methods, context information, and Monte Carlo Sampling. They provide detailed descriptions of representative methods in each group, and examine their positive and negative aspects.

Holte et al. [23] provide a review and a comparative study of multi-view approaches for human 3D pose estimation and activity recognition. They first deal with model-based approaches aimed to extract 3D postures indicating common steps involved in these approaches. Next, the authors deal with 2D and 3D approaches used for human action recognition. In the end, they discuss some shortcomings of the multi-view camera systems and the pros and cons of using sensors like ToF range cameras to acquire 3D data.

Cristani et al. [24] analyze a new perspective of human behavior analysis that brings in concepts and principles from the social, affective, and psychological literature. This is called Social Signal Processing (SSP). The authors introduce, first, a short review about classical activity analysis. Then, they give three examples of problems that encode social events indicating that employing SSP in these problems would be apparently fruitful. Then, they discuss various behavioral cues that represent heterogeneous and multimodal aspects of a social interplay. These cues are categorized into five categories: physical appearance (e.g., attraction, height, somatotype), body postures and gestures, facial expression and gazing, vocal characteristics (prosody, linguistic and non linguistic vocalization, silence, turn taking patterns), and space and environment (interpersonal distances, spatial arrangements of interactants). They also discuss crowd behavior analysis.

In the end, they indicate that combining sociological notions with computer vision algorithms may lead to novel applications such as design of public spaces and learning spaces.

4. Types of Sensors Used for Data Acquisition

As the main goal of computer vision is to derive information from the observed scene, several types of sensors can be used for data acquisition such as: still cameras, video cameras, night-vision cameras, markers on the human body, special body suits and gloves, laser rangefinder (used to determine the distance to an object by applying a laser beam, it may operate on technologies such as time of flight), light detection and ranging “LiDAR” (is a remote sensing technology that measures distance by sending pulses of laser light that strike and reflect from the object surface, it could be used in robotics in order to percieve the surrounding environment), structured light (calculates the depth and surface information of the objects in the scene by projecting a known pattern of pixel,e.g., grids, and measure the deformation), sound navigation and ranging “Sonar” (which uses sound propagation to detect objects), radio-frequency identification “RFID” (used to transfer data, for the purposes of automatically identifying and tracking tags attached to objects), radiometers, millimeter wave radar, microwave radar, synthetic aperture radar, tomographic motion detection, x-ray sensors (can give us a complete image of a whole human body without any occlusions), inertial measurement units (electronic devices that measure object's velocity, orientation, and gravitational forces, using a combination of accelerometers and gyroscopes, sometimes also magnetometers), fiber optic sensors (used to measure strain which can, in turn, be used to recognize body postures), pressure-sensitive foam sensors (to measure respiration rate), etc.

4.1. Active versus passive sensing

Sensors can be classified into two categories based on power supply requirement: active sensors and passive sensors. Active Sensors require power supply (i.e. they provide their own energy source) and are placed on the human subject or in his surroundings. These sensors transmit and receive generated signals. They are suitable for applications in well controlled environments. Laser rangefinder is an example of active sensors. Passive Sensors do not require power supply. They deal with natural signal sources such as visual light, require no wearable devices, and only detect the transmitted energy. They are useful for surveillance applications but they can be used for all applications. Video cameras are examples of passive sensing.

5. Motion Taxonomies

Understanding human motion requires its classification into various levels of abstractions or details. Different taxonomies that categorize motions into different levels already exist in the literature, however, some terms, such as action, activity, simple action, complex action, etc, have different meanings in different taxonomies. Here, we will review some of these taxonomies as follows:

Nagel [27] classified motion into five levels: change, event, verb, episode, and history, where a change refers to a discernable motion in a sequence, an event is a change that is considered as a primitive of a more complex description, a verb describes some activity, an episode describes a complex motion that may be consisted of several actions, and a history which is an extended sequence of related activities. Nagel's goal was to generate conceptual descriptions of image sequences. He used this taxonomy to reflect different dimensions of the motion understanding problem.

Bobick [28] used another taxonomy: movement, activity, and action where movements are the most atomic primitives requiring no contextual or sequence knowledge to be recognized, activity refers to a sequence of movements where the only required knowledge is the statistic of the sequence, and actions are larger scale events that typically include interactions with the environment and causal relationships.

Moeslund et al. [10] used the following action hierarchy: action/motor primitives, actions and activities. Action/motor primitive is an atomic movement that can be described at the limb level such as moving a leg. Action is a sequence of action primitives that may describe a possibly cyclic whole-body movement such as running. Activity consists of a set of actions that gives an interpretation of what is being performed (may be described as an understanding of the situation) such as playing football. Thus, activities are larger scale events that typically depend on the context of the environment, objects, or interacting humans. Other surveys that use the same taxonomy are [13,18] (Note: Kruger is a coauthor in both [10] and [13]). Turaga et al. [15] present a very similar taxonomy to that used by Moeslund et al. [10]. They used the taxonomy of atomic or primitive actions, actions, and activities. Atomic or primitive action is the simplest of action classes. Action refers to simple motion patterns usually performed by a single person and last for a short period of time (e.g., bending, walking, etc). Activities refer to complex sequences of actions performed by several humans who could be interacting with each other in a constrained manner. Activities last for much longer durations (e.g., a gang of robbers attacking a bank). The authors added, in this taxonomy, that the boundary between action and activities is not hard and that there may be some motions that lie in this grey area, where they can neither be described as simple as an "action" nor as complex as an "activity" such as of a music conductor conducting an orchestra using his gestures. Chaquet et al. [25] followed Moeslund et al. [10] and Turaga et al. [15] in structuring their taxonomy into primitive actions, actions, and activities where action is used to fulfill a simple purpose such as walking, or kicking a ball, and activity is defined as a sequence of actions over space and time such as playing football. They also related interactions as an additional feature of activities and indicated that sometimes there is no clear distinction between action and activities.

Aggarwal and Ryoo [20] categorize human motion, depending on the complexity of the motion itself, into four categories: gestures, actions, interactions, and group activities. Gestures are elementary or atomic movements performed by a part of a human body, e.g., moving a leg. Actions are activities performed by a single person and may be composed of multiple gestures, e.g., walking. Interactions are activities that involve two or more persons and/or objects, e.g., two persons fighting each other, a man shoot another one with a gun. Group activities are activities performed by conceptual groups composed of multiple persons and/or objects, e.g., a group of persons marching, two teams playing football.

Lavee et al. [16] present a very different taxonomy that is called “event terminology”. They defined an event as “an occurrence of interest in a video sequence”. Inspired by some other researchers, they used prefixes to the term “event” to describe different types of events with varying properties. They used atomic and composite prefixes to describe the composition property, pixel-based and object-based prefixes to reflect the content properties, single-threaded and multi-threaded prefixes to reflect temporal properties, and sub and super prefixes to reflect the relation to the event of interest. They also introduced another term “event domain” to address the context issue by providing a description of the type of the target events, e.g., gestures in an interactive environment.

Cedras and Shah [1] considered that the recognition of higher level movements, like walking or running, should take into account that those movements consist of a complex and coordinated series of events. They defined motion events as significant changes or discontinuities in motion (e.g., a stop, a pause, a sudden change in direction or in speed, etc).

Motion can also be classified according to its type. For example, Kambhamettu et al. [33] developed a taxonomy for various types of objects motions based on the degree of nonrigidity of the object, see fig. 2. A brief review of this taxonomy is described as follows:

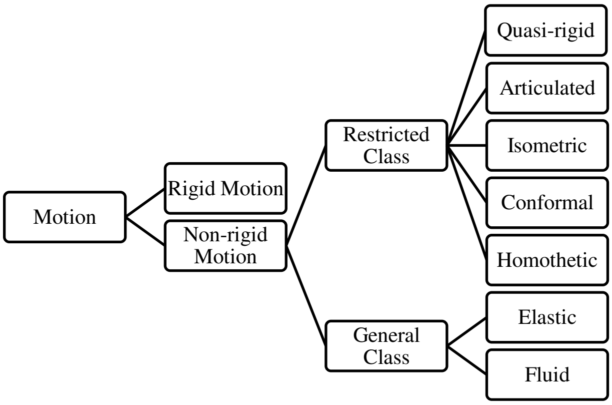

Fig. 2. the classification tree for various types of objects motions as defined by Kambhamettu [33]

- Rigid motion has all distances and angles unchanged.
- Quasi-rigid motion has a small deformation; a general motion is quasi-rigid if viewed in a sufficiently short interval of time.
- Articulated motion is a piecewise rigid motion. The overall motion of the object is not rigid but its constituent parts conform to the constraints of the rigid motion.
- Isometric motion is a nonrigid motion that preserves the angles between the curves on the surface and the distances along the surface.
- Conformal motion is a nonrigid motion that preserves the angles between the curves on the surface, but not the distances.

- Homothetic motion is a nonrigid motion with a uniform expansion or contraction of the surface.
- Elastic motion is a nonrigid motion that preserves some degree of continuity or smoothness.
- Fluid motion is a nonrigid motion that violates even the continuity assumption. It may involve topological variations and turbulent deformations.

6. Different Stages of the Problem

In this section, we will, very briefly, review some classifications of various processes involved in human motion analysis and behavior understanding that are used throughout the literature.

Wang et al. [6] categorize processes involved in human motion analysis in three levels. The low level deals with human detection and contains motion segmentation and object classification processes. The intermediate level deals with human tracking. The high level deals with behavior understanding and contains action recognition and semantic description.

Aggarwal and Park [8] categorize processes involved in human activities understanding into two levels: low–level vision processes such as segmentation, tracking, pose recovery, and trajectory estimation; high-level vision processes such as body modeling and action representation.

Hu et al. [9] use a three-level hierarchy for classification of different processes according to the general framework of visual surveillance: low-level vision, intermediate-level vision, and high-level vision. The hierarchy starts with environment modeling, motion segmentation, and object classification, then continues with object tracking and ends with behavior understanding and person identification.

Turaga et al. [15] categorize processes involved in real-life activity recognition systems into three levels. At the lower levels, there are the modules of background–foreground segmentation, tracking and object detection; the main challenge of this level is to achieve robustness against errors. At the midlevel, there are action–recognition modules; the main challenge of this level is to achieve view and rate-invariant representations. At the high level, there are the reasoning engines that encode the activity semantics based on the lower level action primitives; the main challenge of this level is to achieve effective semantic representations of human activities.

Marco et al. [24] categorize background subtraction/object segmentation and object detection into the low-level stages; while grouping object tracking and activity analysis into the high-level stages.

Shah [26] uses a three-level hierarchy for automatically understanding human behavior from motion imagery. The first level deals with the extraction of relevant visual information from a video sequence. The second level represents that information in a suitable form. The third level concerns with the interpretation of the represented information for the purpose of understanding and recognizing human behavior.

Chellappa and Chowdhury [29] categorize the processes into three levels with vaguely defined boundaries: low-level processes such as extracting features, segmenting regions and tracking feature over a sequence of frames, intermediate-level processes such as grouping features, depth estimation, and motion and structure estimation, high-level processes such as description of objects and scenes.

A simple general frame work for different stages of the problem may be depicted as in fig. 3.

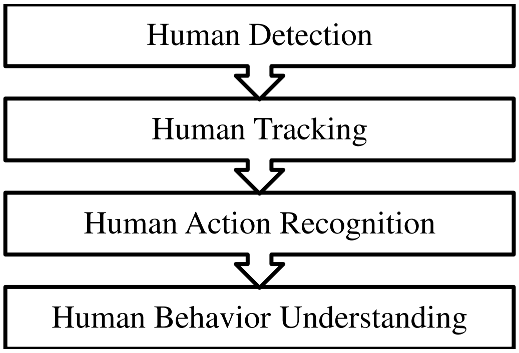

Fig. 3. a simple general framework for Different stages of the problem

7. Discussion

The research in computer vision has started since 1960s, and its interest in recognizing and understanding human motion has started since 1980s. Since then, computer vision researchers, in their quest to solve this problem, have used various techniques and algorithms found in many disciplines such as applied mathematics, statistics, geometry, signal processing, image processing, physics, artificial intelligence, neural networks, neuroscience, psychophysics, biological vision, etc. However, they have faced several problems in detecting, segmenting, analyzing and recognizing different human motions from video sequences. Some of these problems are: data loss resulted from projection of a 3d scene to a 2d image, image noise (e.g., salt and pepper noise, fixed pattern noise, banding noise), dynamic or cluttered background, lighting conditions (e.g., indoor, outdoor, morning, night, etc), weather conditions (e.g., rainy, foggy, sunny, windy, etc), illumination change, light reflections, shadows, temporal textures motion (e.g., tree leaves motion, waving clothes in the air, flying birds, etc), object appearance change, object shape change, wearing excessively sloppy clothes, full or partial occlusion (whether it is an object-to-object occlusion or a scene-to-object occlusion), number of humans in the scene, articulated type of the human motion, complex motion, camera motion, distance of the object from the camera and if zooming is required or not, different viewpoints of the performed motion, data fusion from multiple cameras or different types of sensors. For actions and activities recognition, the researchers encounter extra problems such as the interclass variations between different performers (e.g., speed, pace, anthropometric variation), different execution rates of the same action by the same performer, performing the same action with slight variations (e.g., walking while carrying objects, walking on crutches or with a walking stick, walking while holding on to a walker, nordic walking, walking with a dog, etc), distinguishing between similar classes that are close in performance (e.g., walking, line walking, marching, etc), performance similarity of different classes.

In order to overcome many of these difficulties, researchers imposed some constraints on background appearance, and/or object appearance, and/or object motion, and/or number of objects, etc. The suitability of a certain algorithm depends on the proposed constraints. Some of these constraints are: using controlled environment (e.g., indoors, markers placed on the performer), constant lighting, static background, uniform or simple background, tight clothes, disagreement between the colors of the background and the object clothes, no occlusion, no camera motion, object motion is parallel to the motion plane, moving on a flat ground plane, motion periodicity, constant speed, manual initialization of the human body pose, reducing dimensions of human body modeling, one person in the scene, known camera parameters, etc. Another way to deal with these difficulties is to introduce more information about the scene such as using multiple cameras, audio sensors, radar, ultrasonic, etc. Although, for example, using multiple cameras proves effectiveness in handling the occlusion problem, building a 3D model of the object, finding more suitable features, etc, but this, of course, adds more costs, computations, and complexities to the problem such as increasing installation costs of using more cameras, increasing processing time, calibrating cameras, searching for features in each camera image separately and then combining the information or combining the information early in order to reconstruct a 3D model, or selecting an active viewpoint to determine which camera is the most suitable for more clearer information. Fusing different types of features (e.g., color, shape, position) from different viewpoints of an object is not also an easy task. Fortunately, recent technological advances in real-time image capture, transfer, and processing have been an encouraging factor to lessen the burden of these computations and costs, and also further the research on human motion analysis through using more powerful and complex algorithms.

In designing algorithms to solve the problem of human motion recognition and understanding, several factors should be considered such as: the generality of the proposed algorithm or its applicability to a specific domain or context, efficiency or real time processing requirements (e.g., it is very critical to some applications such as intelligent driver assistance systems, and it is required in applications such as assisted living, human-computer interaction, gesture-based interactive games, smart environments, visual surveillance, but it is not essential in applications like entertainment industry or sports motion analysis), robustness (which is important for continuity and can be tested through using a large amount of data, using different performers, employing dynamic environments, changing conditions, etc), accuracy or precision (high accuracy is required in sports motion analysis, movies industry, etc, but for applications such as human-computer interaction, gesture-based interactive games it may vary between medium and high).

The selection of the most discriminative features of an object is a very important issue in designing algorithms. For example, the features should not change significantly over time, be robust to transformations (e.g., translation, rotations, scaling), be robust to illumination change conditions (e.g., edges and textures are less sensitive to illumination change compared to color feature), etc. Employing multiple features and combining them in a weighted manner is also of great importance to ensure the continuity of the algorithm. It is also better to have an algorithm that can select the appropriate features online automatically. In the same vein, incorporating prior and contextual information is very helpful to adjust the algorithm to a particular scenario in which it is used. Behavior recognition may be the hardest to implement because the same behavior may have several different interpretations depending on many factors such as scene, task, object, and cultural contexts, intention, attention, etc.

It is important to indicate that the different approaches used to solve the problem may be suitable for different situations, which means that there is no single approach to be claimed superior to another. Moreover, many approaches or algorithms may be combined together to yield a better recognition results. For example, Model-based approaches (which use a prior shape model) can deal more efficiently with complex motions because of its ability to integrate shape knowledge and visual input; however, they require more processing to match the model with the input image. Appearance-based approaches (no shape model is used) can be applied in more applications, but they are sensitive to noise [8]. Many researchers also indicate that the use of spatio-temporal features instead of trajectories is gaining more popularity due to their finer analysis of individual human behavior and their robustness to noise, small camera motion, and lighting changes [24]. Again, if we take Dynamic Time Warping “DTW” as an example for measuring similarity between two sequences where the two sequences are warped non-linearly in time to determine a measure of their similarity independent of certain non-linear time variations, we find that this sequence matching method deals successfully when variation in time or speed occur between the two sequences because it handles the differences by operations of deletion-insertion, compression expansion, and substitution, of subsequences. However, DTW lacks the consideration of interactions between nearby subsequences occurring in time [8]. We have said earlier that action recognition algorithms may be classified into discriminative and generative algorithms. Discriminative algorithms have lower error rates when dealing with larger training sets. On the other hand, generative algorithms tend to converge to their optimal performance much quicker even with lesser training examples. Moreover, they have been found to show more flexibility when dealing with incomplete training sets and they are better suited for learning complex patterns [59]. If we take Hidden Markov Model “HMM” as an example of generative algorithms, we find that it combines together the benefits of a temporal evolution model (such as a finite state machine “FSM”) and a probabilistic model (such as a Bayesian network “BN”) [16]. However, Conditional Random Field “CRF”, an example of discriminative algorithms, outperforms HMM for similar action recognition tasks because of their ability to choose an arbitrary dependent abstraction scheme (which may be defined as a categorization of low-level inputs into pixel- based, object-based and logic-based abstractions), and that the abstraction feature selection can consider any combinations of past and future observations [16]. Moreover, many researchers advise to use CRFs instead of HMMs to deal with view-invariant human motion analysis because of their ability of modeling dependencies between features and observations [17].

8. Conclusions

We have shown the importance of analyzing, recognizing, and understanding human motion in computer vision by surveying many promising and demanding applications. We have also discussed different types of sensors used for data acquisition, various processes involved in motion analysis, some taxonomies used for classifying motions, different taxonomies used in structuring different surveys, and brief summaries of selected surveys. In addition, we have discussed various problems faced by computer vision researchers and how they tried to confront them. We have also discussed several factors and important issues that should be considered in designing algorithms.

To sum up the current progress in the field, we can say that the low levels of data processing such as object detection and tracking have reached a reasonable degree of maturity. Also, many algorithms can deal with multiple persons in the scene, and can even deal with occlusion. Recently, the trend has been escalating from dealing with semi-realistic actions performed with simple and static backgrounds to deal with more realistic actions, interactions and activities in more complex situations and conditions, but still there is much more to be done. In literature, we find that different algorithms have been developed for dealing with specific-domain applications, however, the ultimate goal of computer vision researchers in the field of human motion analysis and recognition is to enable computer systems to have human-level recognition of any types of motion, but we still far from reaching this end. So, calls for a unified framework that benefits all motion recognition tasks are growing stronger [34].

9. References

[1] Cedras, Claudette, and Mubarak Shah. "Motion-based recognition a survey." *Image and Vision Computing* 13.2 (1995): 129-155.

[2] Aggarwal, J. K., Cai, Q., Liao, W., & Sabata, B. "Nonrigid motion analysis: Articulated and elastic motion." *Computer Vision and Image Understanding* 70.2 (1998): 142-156.

[3] Gavrila, Dariu M. "The visual analysis of human movement: A survey." *Computer vision and image understanding* 73.1 (1999): 82-98.

[4] Aggarwal, Jake K., and Qin Cai. "Human motion analysis: A review." *Computer Vision and Image Understanding,* Vol. 73.3, March, pp. 428–440, 1999.

[5] Moeslund, Thomas B., and Erik Granum. "A survey of computer vision-based human motion capture." *Computer Vision and Image Understanding* 81.3 (2001): 231-268.

[6] Wang, Liang, Weiming Hu, and Tieniu Tan. "Recent developments in human motion analysis." *Pattern recognition* 36.3 (2003): 585-601.

[7] Buxton, Hilary. "Learning and understanding dynamic scene activity: a review." *Image and vision computing* 21.1 (2003): 125-136.

[8] Aggarwal, J. K., and Sangho Park. "Human motion: Modeling and recognition of actions and interactions." *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*. IEEE, 2004.

[9] Hu, Weiming, Tieniu Tan, Liang Wang, and Steve Maybank. "A survey on visual surveillance of object motion and behaviors." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 34.3 (2004): 334-352.

[10] Moeslund, Thomas B., Adrian Hilton, and Volker Krüger. "A survey of advances in vision-based human motion capture and analysis." *Computer vision and image understanding* 104.2 (2006): 90-126.

[11] Yilmaz, Alper, Omar Javed, and Mubarak Shah. "Object tracking: A survey." *Acm Computing Surveys (CSUR)* 38.4 (2006): 13.

[12] Poppe, Ronald. "Vision-based human motion analysis: An overview." *Computer vision and image understanding* 108.1 (2007): 4-18.

[13] Krüger, V., Kragic, D., Ude, A., & Geib, C. "The meaning of action: a review on action recognition and mapping." *Advanced Robotics* 21.13 (2007): 1473-1501.

[14] Pantic, Maja, Alex Pentland, Anton Nijholt, and Thomas S. Huang. "Human computing and machine understanding of human behavior: A survey." In *Artifical Intelligence for Human Computing*, pp. 47-71. Springer Berlin Heidelberg, 2007.

[15] Turaga, Pavan, Rama Chellappa, Venkatramana S. Subrahmanian, and Octavian Udrea. "Machine recognition of human activities: A survey." *Circuits and Systems for Video Technology, IEEE Transactions on* 18.11 (2008): 1473-1488.

[16] Lavee, Gal, Ehud Rivlin, and Michael Rudzsky. "Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 39.5 (2009): 489-504.

[17] Ji, Xiaofei, and Honghai Liu. "Advances in view-invariant human motion analysis: a review." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 40.1 (2010): 13-24.

[18] Poppe, Ronald. "A survey on vision-based human action recognition." *Image and vision computing* 28.6 (2010): 976-990.

[19] Weinland, Daniel, Remi Ronfard, and Edmond Boyer. "A survey of vision-based methods for action representation, segmentation and recognition." *Computer Vision and Image Understanding* 115.2 (2011): 224-241.

[20] Aggarwal, J. K., and Michael S. Ryoo. "Human activity analysis: A review." *ACM Computing Surveys (CSUR)* 43.3 (2011): 16.

[21] Chen, Liming, and Ismail Khalil. "Activity recognition: approaches, practices and trends." *Activity Recognition in Pervasive Intelligent Environments*. Atlantis Press, 2011. 1-31.

[22] Yang, Hanxuan, Shao, L., Zheng, F., Wang, L., & Song, Z. "Recent advances and trends in visual tracking: A review." *Neurocomputing* 74.18 (2011): 3823-3831.

[23] Holte, Michael B., Cuong Tran, Mohan M. Trivedi, and Thomas B. Moeslund. "Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments." *Selected Topics in Signal Processing, IEEE Journal of* 6.5 (2012): 538-552.

[24] Cristani, Marco, Raghavendra, R., Del Bue, A., & Murino, V. "Human behavior analysis in video surveillance: a social signal processing perspective." *Neurocomputing* (2012).

[25] J.M. Chaquet, E.J. Carmona, A. Fernández-Caballero, A Survey of Video Datasets for Human Action and Activity Recognition, *Computer Vision and Image Understanding* (2013)

[26] Shah, Mubarak. "Understanding human behavior from motion imagery." *Machine Vision and Applications* 14.4 (2003): 210-214.

[27] Nagel, H-H. "From image sequences towards conceptual descriptions." *Image and vision computing* 6.2 (1988): 59-74.

[28] Bobick, Aaron F. "Movement, activity and action: the role of knowledge in the perception of motion." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 352.1358 (1997): 1257-1265.

[29] Chellappa, Rama, and Amit K. Roy Chowdhury. "Computer Vision, Statistics in." *Encyclopedia of Statistical Sciences*.(2006).

[30] Weng, Juyang, James McClelland, Alex Pentland, Olaf Sporns, Ida Stockman, Mriganka Sur, and Esther Thelen. "Autonomous mental development by robots and animals." *Science* 291.5504 (2001): 599-600.

[31] Hampapur, Arun, Lisa Brown, Jonathan Connell, Sharat Pankanti, Andrew Senior, and Yingli Tian. "Smart surveillance: applications, technologies and implications." In Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on, vol. 2, pp. 1133-1138. IEEE, 2003.

[32] Norris, Clive, Mike McCahill, and David Wood. "The Growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space." Surveillance & Society 2.2/3 (2004).

[33] C. Kambhamettu, D. B. Goldgof, D. Terzopoulos, and T. S. Huang, Nonrigid motion analysis, Handbook of PRIP: Computer Vision, Vol. 2, 1994.

[34] Aggarwal, J. K., and M. S. Ryoo. "Toward a unified framework of motion understanding." *Image and Vision Computing* 30.8 (2012): 465-466.

[35] Zhong, Hua, Jianbo Shi, and Mirkó Visontai. "Detecting unusual activity in video." *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. Vol. 2. IEEE, 2004.

[36] Collins, Robert, Alan Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, and Lambert Wixson. "*A system for video surveillance and monitoring*", Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep., CMU-RI-TR-00-12, 2000.

[37] Zhan, Beibei, Dorothy N. Monekosso, Paolo Remagnino, Sergio A. Velastin, and Li-Qun Xu. "Crowd analysis: a survey." Machine Vision and Applications 19, no. 5-6 (2008): 345-357.

[38] Jacques Junior, Julio Cezar Silveira, Soraia Raupp Musse, and Claudio Rosito Jung. "Crowd analysis using computer vision techniques." Signal Processing Magazine, IEEE 27.5 (2010): 66-77.

[39] B. Georis, M. Maziere, F. Bremond, and M. Thonnat, "A video interpretation platform applied to bank agency monitoring," in Proc. 2nd Workshop Intell. Distributed Surveillance Syst., 2004, pp. 46–50.

[40] Siebel, Nils T., and S. Maybank. "The advisor visual surveillance system." *ECCV 2004 workshop Applications of Computer Vision (ACV)*. Vol. 1. 2004.

[41] Drosou, Anastasios, et al. "Spatiotemporal analysis of human activities for biometric authentication." Computer Vision and Image Understanding 116.3 (2012): 411-421.

[42] Gafurov, Davrondzhon. "A survey of biometric gait recognition: Approaches, security and challenges." *Annual Norwegian Computer Science Conference*. 2007.

[43] Morimoto, Carlos H., and Marcio RM Mimica. "Eye gaze tracking techniques for interactive applications." *Computer Vision and Image Understanding* 98.1 (2005): 4-24.

[44] Turk, Matthew, and George Robertson. "Perceptual user interfaces." *Communications of the ACM* 43.3 (2000).

[45] Mikic, Ivana, Kohsia Huang, and Mohan Trivedi. "Activity monitoring and summarization for an intelligent meeting room." Human Motion, 2000. Proceedings. Workshop on. IEEE, 2000.

[46] Shotton, Jamie, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. "Real-time human pose recognition in parts from single depth images." *Communications of the ACM* 56.1 (2013): 116-124.

[47] Zhang, Zhengyou. "Microsoft kinect sensor and its effect." *Multimedia, IEEE* 19.2 (2012): 4-10.

[48] Tamura, Hideyuki, Takashi Matsuyama, Naokazu Yokoya, Ryosuke Ichikari, Shohei Nobuhara, and Tomokazu Sato. "Computer vision technology applied to MR-based pre-visualization in filmmaking." In Computer Vision–ACCV 2010 Workshops, pp. 1-10. Springer Berlin Heidelberg, 2011.

[49] Assfalg, Jürgen, et al. "Semantic annotation of soccer videos: automatic highlights identification." Computer Vision and Image Understanding 92.2 (2003): 285-305.

[50] Jones, Simon, and Ling Shao. "Content-based retrieval of human actions from realistic video databases." Information Sciences (2013).

[51] Mündermann, Lars, Stefano Corazza, Ajit M. Chaudhari, Thomas P. Andriacchi, Aravind Sundaresan, and Rama Chellappa. "Measuring human movement for biomechanical applications using markerless motion capture." In Electronic Imaging 2006, pp. 60560R-60560R. International Society for Optics and Photonics, 2006.

[52] D’Orazio, Tiziana, and Marco Leo. "A review of vision-based systems for soccer video analysis." Pattern recognition 43.8 (2010): 2911-2926.

[53] Rougier, Caroline, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. "Robust video surveillance for fall detection based on human shape deformation." *Circuits and Systems for Video Technology, IEEE Transactions on* 21.5 (2011): 611-622.

[54] Chaaraoui, Alexandros André, Pau Climent-Pérez, and Francisco Flórez-Revuelta. "A review on vision techniques applied to human behaviour analysis for ambient-assisted living." Expert Systems with Applications 39.12 (2012): 10873-10888.

[55] Murphy-Chutorian, Erik, and Mohan M. Trivedi. "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness." Intelligent Transportation Systems, IEEE Transactions on 11.2 (2010): 300-311.

[56] Cheng, Shinko Y., and Mohan M. Trivedi. "Turn-intent analysis using body pose for intelligent driver assistance." Pervasive Computing, IEEE 5.4 (2006): 28-37.

[57] Eng, H-L., K-A. Toh, Alvin Harvey Kam, Junxian Wang, and W-Y. Yau. "An automatic drowning detection surveillance system for challenging outdoor pool environments." In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 532-539. IEEE, 2003.

[58] Yu, Chen, Linda B. Smith, Hongwei Shen, Alfredo F. Pereira, and Thomas Smith. "Active information selection: Visual attention through the hands." Autonomous Mental Development, IEEE Transactions on 1, no. 2 (2009): 141-151.

[59] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes." in Neural Information Processing Systems, 2001, pp.841-848.

تحليل حركة الإنسان والتعرف عليها وفهمها من خلال الرؤية الحاسوبية

الملخص العربى

إن غاية الرؤية الحاسوبية هو فهم ما يدور حولها في البيئة المحيطة بصورة صحيحة من خلال عدة خطوات تشمل عملية جلب المعلومات ومعالجتها وتحليلها ثم فهمها. يعد تحليل حركة الإنسان والتعرف عليها وفهم الأنشطة التي يقوم بها من أكثر الموضوعات بحثا في مجال الرؤية الحاسوبية.

الغرض من هذا المقال هو تسليط بعض الضوء على هذا الموضوع بحيث يكون مفيدا للباحثين في مجال الرؤية الحاسوبية. هذا المقال يستعرض الكثير من المقالات السردية (المراجعات) الأكثر شهرة و الأكثر حداثة في هذا المجال، مشيرا إلى العديد من التطبيقات الهامة و الواعدة، وكذلك أنواع الحساسات أو المجسات المختلفة المستخدمة للحصول على المعلومات، كما أنه يستعرض المستويات المختلفة اللازمة لتحليل حركة الإنسان والتعرف عليها وفهم الأنشطة المختلفة التي يقوم بها. هذا المقال يستعرض أيضا بعض التصنيفات الخاصة بالحركة طبقا لمدى بساطتها أو تعقدها وكذلك التصنيفات التي تنتهجها بعض المراجعات في عرض العديد من الدراسات المختلفة. وفي النهاية يقدم المقال مناقشة عامة موضحا بعض المشاكل المختلفة وطرق حلها.