Applications of human action analysis and recognition on wireless network infrastructures: State of the art and real world challenges

Georgios Tsatiris*, Kostas Karpouzis*, Stefanos Kollias*†
*School of Electrical and Computer Engineering
National Technical University of Athens
9, Heroon Politechniou str., 15780, Athens, Greece
gtsatiris@image.ntua.gr, kkarpou@cs.ntua.gr
†School of Computer Science
University of Lincoln
Brayford Pool, Lincoln, Lincolnshire, UK
skollias@lincoln.ac.uk

Abstract—Human action recognition and analysis has given life to a wide variety of real-world applications, ranging from surveillance and human-computer interaction to patient monitoring and rehabilitation. Most action recognition systems, especially smart-home or assistive living applications, depend on network infrastructures for easy data fusion and integration of different sensing modalities. However, despite the fact that action recognition methods have extensively been evaluated for their accuracy and there is a consensus on the ways to provide quality of service in various network infrastructures, there is poor coverage of the inherent challenges of performing human action in real world network-based applications. In this work, we attempt to document these challenges based on representative, state of the art techniques and venture to report on the open issues that need to be resolved by new techniques aiming to provide viable real world applications.

I. INTRODUCTION

Human activity analysis is currently one of the hottest areas in pattern recognition and machine intelligence, due to its applications in gaming [1], robotics [2], assistive environments [3][4][5], qualitative assessment [6], automated surveillance [7], as well as its intrinsic scientific challenges.

The problem of human action recognition refers to the automatic detection, classification and analysis of human activities, using information acquired from a variety of sensing modalities. Although the idea is simple, the specific task is notably challenging as any relevant system has to overcome a large number of restrictive parameters. In the paradigm of vision-based techniques, illumination variations, camera view angle, complicated backgrounds and occlusions are only a fraction of the existing set of problems. In addition to these, individuality is another very important factor that cannot be neglected, as every person performs the same set of movements (action) in a unique and different to every other person’s way.

It could be said that analysis tasks are generally carried out either on segmented poses (videos that only contain the action in question) [8] [9], or “in the wild” [10]. Research in the field of human activity recognition has now shifted towards the latter domain due to the abundance of camera feeds and better networking infrastructures used to transmit captured video feeds. Trending topics include generic activity recognition (online or in real-time), emotion classification by modeling actions, as well as human-human and human-computer interaction recognition [11]. In the meantime, researchers seem to abandon conventional activity recognition techniques that depend only on a single modality (e.g only audio or visual features), or controlled environments, such as segmented or acted actions. In this trend, network-based techniques have arisen, as network infrastructures appear to provide a basis for more intuitive fusions of multiple modalities and sensing strategies.

A. Human action recognition

In a recent and comprehensive survey by Chen et al. [12], action recognition methods are classified into two main categories: the vision based methods and the wearable sensor based ones. Vision based action recognition original utilized conventional RGB cameras. Methods based on video sequences can be further categorized into template-based approaches, focused on lower level features, and model-based approaches, performing higher-level feature extraction [13]. Notably, modeling human action in video has been mainly dealt by using a variety of methods, such as spatio-temporal interest points [14], to capture the motion essence in the space and the time domain. Other methods include modeling using trajectories [15] and image/frame-based encoding [16]. Naturally, usage of common RGB cameras has some limitations, mainly associated with the lack of 3D data in conventional imaging. 3D information is a significant factor to incorporate when analyzing human movement, and the fact that most modern action recognition systems leverage depth information showcases that importance [2].
3D information can be acquired using one of three main approaches. The first utilizes marker-based motion capture systems, such as MoCap\(^1\). Motion capture systems perform optical sensing of markers placed on the human body, and use triangulation from multiple cameras to estimate the 3D real world coordinates of skeleton joints. This approach is relatively expensive. The second approach is based on stereo imaging. Here, partially overlapping views of two or more RGB cameras are combined and 3D depth data can be obtained via stereo matching. However, stereoscopic 3D reconstruction algorithms are computationally expensive and exhibit sensitivity to noise and clutter issues that plague any conventional RGB image based technique.

The third approach is based on depth sensing. In recent years, depth sensors (such as the, now discontinued, Microsoft Kinect\(^2\) and the Intel RealSense series\(^3\)) have become publicly accessible and are able to provide real-time depth information. Based mainly on infrared technology, depth sensing shows insensitivity to lighting changes. Another key factor that has led to its extensive use in action recognition applications is the emergence of algorithms that can extract the human skeleton from depth images in real time \([17][18]\).

Vision based action recognition continues to be an active and thriving research area, albeit being subject to challenges stemming from intrinsic issues such as occlusion, camera placement, background clutter, etc. Camera placement also sets a predefined field of view, thus limiting the application of the method in a constrained space. To address these issues, a common alternative is the utilization of wearable inertial sensors with accelerometer and gyroscope functionalities. When no cameras are involved, occlusion, lighting conditions and camera placements are not an issue. Furthermore, the evolution of sensor hardware with enhanced computational capabilities, as well as lower power consumption, has further enabled heavier computations, longer recording and usability. Wearable inertial sensors are able to provide 3D information using acceleration and angular data. However, as with the previous modalities, sensor based systems come with their own shortcomings. The most important one is the intrusiveness of any device that needs to be constantly carried on the subject’s body, limiting the subject’s movement. In addition, measurements are sensitive to sensor location on the body.

Table I summarizes the advantages and disadvantages of the three different modalities (RGB, depth and inertial sensors), as presented in [12].

B. Wireless ad hoc and sensor networks

The rise of pervasive computing systems and the evolution of mobile wireless computing into a viable framework for computationally demanding applications, triggered the emergence of human-centered computing techniques, based on wireless sensing and learning. Focusing in the field of human activity recognition and analysis, this study attempts to document some of the most typical works and recognize issues and challenges that may arise from the fusion of typical action recognition pipelines and wireless computing.

A wireless ad hoc network is a decentralized wireless network which does not rely on any existing infrastructure, such as routers or access points. Instead, each node participates in routing by forwarding data. Data forwarding and routing choices are made dynamically, based on network connectivity and the routing algorithm in use [19]. Wireless sensor networks, on the other hand, are a special type of ad hoc networks, composed by scattered, low-cost and low-power devices (sensors)[20]. Each sensor has its own processing, storing and transceiver capabilities and it measures environmental conditions like temperature, pressure, sound, etc. Each node of the network should be able to process as much information as possible in a local level. However, in many cases, there is need for higher processing power, as well as incorporation of information from a large number of nodes.

Quality of Service: Research on networks of sensors and other “computationally-enabled” devices has long left the “proof-of-concept” status and has already shifted towards real-world applications. The quality of the provided services (QoS) to the applications based on wireless sensor networks is the focal point [21]. Consequently, in this study, we will ponder on the challenges and the demands a real-time action recognition system may present on a wireless framework.

As shown in [21], the most important QoS challenges in wireless sensor networks are:

- Limited resources and capabilities: The first and foremost thing to keep in mind is that we are dealing with machines of limited energy, bandwidth, memory, and processing and communication capabilities. Authors point out that there should be a balance between QoS and energy consumption.
- Network topology and node deployment: Due to their dynamic nature, wireless sensor networks, especially mobile networks, have a non deterministic behavior in terms of connectivity, path discovery, etc. They do have a lower deployment cost, though.
- Scalability: Changes of the number of nodes should not affect QoS.
- Multi-source multi-sink systems: The provision of QoS in critical infrastructures, taking platform heterogeneity, service-oriented design and security requirements into account.
- Application heterogeneity: It is expected that different applications have different QoS demands from a wireless sensor framework. According to [21], the most demanding applications can be found in:

  - Real-time systems: in applications such as surveillance and healthcare, there are specific requirements on bandwidth and delay. There are applications completely intolerant to any delay and others that can sacrifice some of their real-time capabilities for better results and interactivity.
### Table I

<table>
<thead>
<tr>
<th></th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RGB cameras</strong></td>
<td>• Cost effective&lt;br&gt;• Widely available&lt;br&gt;• Easy to use&lt;br&gt;• Provide texture information</td>
<td>• The subject needs to be visible&lt;br&gt;• Sensitive to lighting conditions and illumination changes&lt;br&gt;• Sensitive to occlusion and background clutter&lt;br&gt;• Sensitive to camera calibration&lt;br&gt;• Algorithms can be computationally expensive</td>
</tr>
<tr>
<td><strong>Depth cameras</strong></td>
<td>• Cost effective&lt;br&gt;• Widely available&lt;br&gt;• Insensitive to lighting conditions and illumination changes&lt;br&gt;• Can work in total darkness&lt;br&gt;• Provide 3D information&lt;br&gt;• Easy to use&lt;br&gt;• Not sensitive to color and texture change</td>
<td>• The subject needs to be visible&lt;br&gt;• Noise generally present in the images&lt;br&gt;• Sensitive to materials with different reflection properties (e.g., transparent materials, light absorbing materials, etc.)&lt;br&gt;• No color information</td>
</tr>
<tr>
<td><strong>Inertial sensors</strong></td>
<td>• Cost effective&lt;br&gt;• Widely available&lt;br&gt;• High sampling rate&lt;br&gt;• Can work in total darkness&lt;br&gt;• Can work in unconfined environments</td>
<td>• Sensitive to sensor location on the body&lt;br&gt;• Sensor drift&lt;br&gt;• Power consumption for sensor battery&lt;br&gt;• Require multiple sensors for capturing full body motion&lt;br&gt;• Intrusiveness of constantly carrying one or more devices</td>
</tr>
</tbody>
</table>

- Multimedia systems: Audio, video and images are demanding data to circulate in a network. In cases when some computations need to be executed on the sensor device, there is even need for considerable processing power (with respect to the limitations of these devices). The way these demands are dealt with affects other aspects, such as the interactivity of the application.
- Variations in traffic: Generated data may generate different traffic patterns.
- Reliability of Wireless connections: Environmental factors can affect the performance of a wireless network. Shorter links are preferable in many cases.
- Data redundancy: In real-time applications and multimedia systems, having multiple copies of data is very important, especially when coping with data loss. In many cases, though, it is discouraged or even impossible to provide abundance of computationally demanding or "heavy" data.

#### C. QoS in action recognition applications

In the previous paragraphs we attempted a concise presentation of the fields of human action recognition and quality of service on wireless networks. Modern human action recognition is performed online (with respect to the flow of the signal that represents the human motion) or in real-time. Intrinsically, such applications, when based on network infrastructures, are expected to demand quality of service relative to real-time availability, multimedia traffic and processing. In the following section, a number of current human action and affect modeling techniques, which leverage underlying wireless network frameworks, will be presented. We will document on how these techniques perform in realistic scenarios and investigate the ways they cope with the related QoS challenges.

### II. Wireless Network Enabled Action Recognition: State of the Art and Challenges

One can find a variety of interpretations as to what constitutes a wireless network based application of an action recognition pipeline. Techniques based on simple sensor networks or body area networks are mainly dominant in the relevant literature, with early representative works such as the one by Jovanov et al. [22]. Beginning with this study, we will present a number of illustrative examples of the main trends in the area and ascertain how they dealt with the aforementioned QoS challenges.

#### A. Action analysis with body area networks

Considering general human monitoring scenarios, early works such as [22] presented the use of simple wearable sensor networks, called body area networks. In the aforementioned work, the authors present a telemedicine application that utilizes commercial wireless sensors, such as an electrocardiogram sensor, motion sensors (gyroscopes and accelerometers), temperature sensors, etc. Thus, a wireless body area network can be comprised of a set of wearable sensors on the human
body, as well as a centralized personal server aggregating the data and possibly performing all computationally demanding tasks. The system can be further connected to the Internet or a cloud infrastructure, in order to access external information and services. A simple data flow and topology diagram of such a system can be seen in Figure 1.

The proposed system performs real-time data analysis, providing medical feedback to rehabilitating patients. A QoS guarantee can be detected in the fact that no intensive processing is handled by the sensors themselves. In a modern setup, the personal server in Figure 1 (which can be running on a smartphone or other mobile device), may even be equipped with the appropriate hardware to perform expensive computations, considering the power of modern handheld devices. Moreover, a body area network cannot be considered a typical wireless sensor network. The structure of a body area network is more or less predefined and, in some cases, such as the one in [22], without any particular routing challenges and scalability needs, as few or no additional nodes are usually added without reconfiguring the setup.

On the other hand, such a dedicated framework may be unable to handle malfunctions of even a single node. It is natural that if a node, such as an accelerometer or a microphone, stops transmitting live data to the application, a valuable input to the pipeline becomes silent, considering that it is not certain that the other nodes will adequately fill its gap. This is an issue in similar sensor-based systems (e.g. accelerometers and gyroscopes on each of the subject’s joints) as well, because the failure of a particular sensor means that motion of a limb may go undetected. Especially in real-time or online applications, where processing time needs to be constrained and based on data from a concise time window, this can end up being a fatal issue. For instance, considering one of the case studies presented in [23], there are time-critical signals, such as sounds (in the event of an emergency) which cannot be appropriately complemented by other signals, consequently leaving an emergency undetected. In offline applications, as the survey by Lara and Labrador [24] implies, more information is available (as processing and action classification are not conducted as the action sequence progresses) and the failure of a single node (or perhaps even a number of nodes) may not lead to severe malfunctions. However, offline processing is not desirable in real world action recognition systems, as it cannot support interactive applications.

B. Vision based wireless systems

Networks of wearable sensors, while appearing to not pose particular QoS challenges when it comes to routing, structure and data overloading, they could be vulnerable to single node failures as there is no guarantee that a specific node signal can be complemented by others. This can be considered a limitation of a wireless sensor based system (in the traditional use of the term). In contrast, vision based techniques, such as the one presented in [25], may not be prone to such limitations. Here, a setting of a number of cameras monitors a room and a subject’s movement inside it (Figure 2). The idea is that multi-view action recognition (i.e. action monitored from different angles) can produce better results, because it combines features not limited to a preset viewing angle and, consequently, to occlusion or perspective issues. This demands the incorporation of multiple RGB cameras, as well as, in

![Fig. 1. Topology and data flow in a body area network. The personal server can also be a portable device, such as a smartphone.](image-url)
In many cases, depth cameras, to further enhance the calculated 3D information with accurate depth readings.

In [25], an interesting requirement for a viable real-time application is defined. Notably, local processing and extraction of features is important, in order to avoid overloading the network with data. It is proposed that a selective, compressed representation of the original data should be sent to a centralized center for the final classification task. There is also a need to avoid computationally expensive feature extraction techniques, in order to keep the processing rate high and not deviate from the interactive nature of the application. By adhering to these proposed guidelines, the goal is to keep a balance between the data transmitted through the network and the workload imposed on the individual nodes. This is not straightforward and there is no universally accepted way to determine if this balance is achieved. For instance, a feature extraction pipeline such as the one presented in [9] cannot be delegated to the nodes, due its high computational cost. If we were to design a system based on that technique, it would be preferable to send direct video streams from the cameras to a powerful centralized server.

Concerning the robustness of a camera-based wireless system, we can assert that it would not show the limitations of a sensor based system. Either by performing some sort of feature extraction at node level, or by simply sending raw video data to a computation server, each node does nothing more than observe and analyze an action from a specific viewing angle. Even simpler techniques, based on the evolution of the human silhouette, such as the one in [25], can produce meaningful results, even using a single camera [26]. This way, camera-based systems could theoretically operate adequately, even when parts of the network (a number of cameras) go offline or malfunction. Adding information from other view perspectives enhances the results. More advanced multi-view techniques, such as the ones presented in [27] and [28] are more tailored to the task, although it cannot be measured if they adhere to the aforementioned guideline of transmitted-to-locally-processed data balance and to what extent.

III. OPEN ISSUES

In the previous sections, we presented the state of the art on real-time human action analysis and recognition using wireless network infrastructures. The challenges and limitations that these techniques face have also been documented. This documentation can lead to a concise and well-defined list of open issues that need to be addressed, in an attempt to equip the field of wireless-powered action recognition with a set of universal evaluation rules and criteria, not limited to classification accuracy, but also assessing network-related aspects that are critical for any relevant system’s applicability to the real world.

A. Sensor-based methods

To the best of the authors’ knowledge, there is still no unified practice to supplement the failure of a number of nodes on sensor-based or body area networks. We can agree that no other serious QoS issues arise from the use of such networks, due to the small amount of data and the centralized computational manner, as well as the predefined and simplified structure of the network. However, the fact that losing a node may result to the loss of information that is otherwise hard to replace is a major flaw that needs to be dealt with. There exist fusion techniques that blend vision (color or depth) based imagery with sensors to create more robust recognition pipelines [29], but this does not solve the intrinsic limitation of sensor-based techniques.

B. Vision-based methods

Although more robust in single node failures and presenting a similar virtue as the sensor-based techniques (i.e. simplified predefined network structure), there is no globally endorsed way to evaluate their viability in real world scenarios, mainly because the focus of the existing literature up until now was to report on the classification accuracy. The balance between transmitted data and locally processed data is a fine intuitive measure, but it cannot be evaluated in a structured manner. Video distribution related QoS criteria [30] can be used for this task, although sharing raw video streams is only a subset of the available methodologies and generally not recommended.

IV. CONCLUSION

In this study, we have presented the contemporary definition of the fields of human action recognition, wireless sensor networks as well as how they can be both utilized to create real world human action monitoring applications. The main focus of this work was to report on the challenges that real world applications of this fusion face and how (and if) they are dealt in state of the art methodologies. We documented on the limitations of the existing literature and pointed out a set of open issues, resolving which will empower research in this almost multidisciplinary field and provide more concrete evaluation for new methods, concerning their applicability in the real world.
ACKNOWLEDGMENT
This work was supported by the EC-funded project ECoWeB - Emotional Competence and Well-Being (grant agreement No 754657).

REFERENCES