IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2019, accepted June 5, 2019, date of publication July 1, 2019, date of current version July 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926206

A Fusion-Based Framework for Wireless Multimedia Sensor Networks in Surveillance Applications

ADNAN YAZICI[1], (Senior Member, IEEE), MURAT KOYUNCU[2], SEYYIT ALPER SERT[3], AND TURGAY YILMAZ[4]

[1]Department of Computer Science, Nazarbayev University, 010000 Astana, Kazakhstan
[2]Department of Information Systems Engineering, Atilim University, 06830 Ankara, Turkey
[3]Strategic Development and Preparation Directorate, NATO Supreme Headquarters Allied Powers Europe (S.H.A.P.E.), 7010 Mons, Belgium
[4]Elsevier B.V., 1043 NX Amsterdam, The Netherlands

Corresponding author: Murat Koyuncu (mkoyuncu@atilim.edu.tr)

This work was supported in part by the Nazarbayev University Faculty Development Competitive Research Grants Program under Grant FY2019-FGP-1-STEMM, and a research grant from The Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 114R082.

ABSTRACT Multimedia sensors enable monitoring applications to obtain more accurate and detailed information. However, the development of efficient and lightweight solutions for managing data traffic over wireless multimedia sensor networks (WMSNs) has become vital because of the excessive volume of data produced by multimedia sensors. As part of this motivation, this paper proposes a fusion-based WMSN framework that reduces the amount of data to be transmitted over the network by intra-node processing. This framework explores three main issues: 1) the design of a wireless multimedia sensor (WMS) node to detect objects using machine learning techniques; 2) a method for increasing the accuracy while reducing the amount of information transmitted by the WMS nodes to the base station, and; 3) a new cluster-based routing algorithm for the WMSNs that consumes less power than the currently used algorithms. In this context, a WMS node is designed and implemented using commercially available components. In order to reduce the amount of information to be transmitted to the base station and thereby extend the lifetime of a WMSN, a method for detecting and classifying objects on three different layers has been developed. A new energy-efficient cluster-based routing algorithm is developed to transfer the collected information/data to the sink. The proposed framework and the cluster-based routing algorithm are applied to our WMS nodes and tested experimentally. The results of the experiments clearly demonstrate the feasibility of the proposed WMSN architecture in the real-world surveillance applications.

INDEX TERMS Data fusion, fuzzy clustering, fuzzy routing, multimedia, object detection, surveillance applications, wireless communication, wireless multimedia sensor networks.

I. INTRODUCTION

Over the past two decades, wireless sensor networks (WSN) have become one of the most promising technologies with integrated microprocessors, low-power analog and digital electronics, and advances in wireless communication. These technological developments have made possible for producers to implement inexpensive sensor nodes supporting standard and efficient communication protocols [1], [2]. In addition, recent developments in the field of information technology are introducing low cost and small cameras and microphones. New enhanced video and audio sensors, combined with some conventional scalar sensors, may be more useful for more accurate identification in real-time settings. In parallel with these significant technological developments, researchers and users have begun to explore the possibility of obtaining more accurate and realistic information from real-world applications in rapidly changing environments. As a result, distributed systems with more powerful sensor nodes are introduced as Wireless Multimedia Sensor Networks (WMSN) [3], [4]. As such, it has become possible to capture, store, process and resolve multimedia content

The associate editor coordinating the review of this manuscript and approving it for publication was Kaigui Bian.

in WMSNs. With the use of multimedia sensors, it is easier to obtain more precise and detailed information in applications such as smart cities and smart homes, environmental monitoring, smart patient care and personalized medical control, intrusion detection, command control and fire detection. In addition, the demand for the collection and use of multimedia data in WMSNs has increased in recent years.

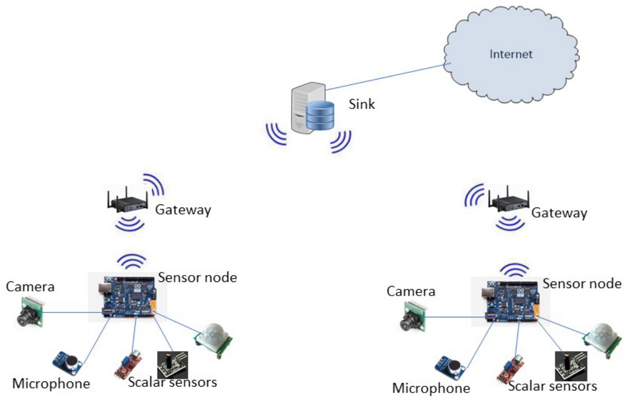

FIGURE 1. A typical WMSN with scalar and multimedia sensors.

Video and audio recordings are effectively used as complementary mechanisms in existing surveillance systems to counter potential threats that may require the involvement of law enforcement officials. With the participation of multimedia sensors, it is possible to obtain more precise and detailed information in the fields of an application. Current video and audio sensors, as well as some conventional scalar sensors, can be used for more accurate identification in real-time situations. Fig. 1 illustrates a typical WMSN in which nodes with scalar and multimedia sensors communicate with the gateway to transmit bulk data to the sink. However, the simple transmission of raw multimedia data possesses many problems to be solved. In smart environments and urban applications, WMSNs are considered one of the major sources of high-volume data traffic because they provide big data/information from multimedia devices such as cameras and microphones [5]. Intelligent transport systems [6], intelligent farming systems [7] and intelligent health systems [8] are among the most common WMSN applications. There is an increasing demand for high volume WMSNs, which make it difficult to manually transfer and process data. As a result, the development of efficient and lightweight solutions for managing data traffic on WMSNs has become vital for the fields of Internet of Things and big data analytics.

Wireless sensor nodes are typically battery-powered devices that need replacement of battery at regular intervals. However, they can be deployed where access can be difficult. They can even be in an enemy zone where it is impossible to replace the battery. Therefore, the efficient use of energy is one of the main research areas for WSNs [9]. For communication, the use of energy efficient components alone cannot produce satisfactory results and requires the incorporation of more complex methods into different layers of a sensor node architecture. It is important to consider energy efficiency techniques not only in the sensor node but also in the network, as prolonging the life of a network is very important for many real-life scenarios [10].

Extension of multimedia sensors aggravates energy consumption problem of WMSNs and requires implementation of energy efficient solutions. One of the most common ways to reduce the energy consumption of WMSNs is to decrease the size of image/video/audio data to be transmitted, thereby minimizing power consumption by reducing transmission requirements, which consume more energy than other processes. Various studies aim to reduce the size of the multimedia data to be transmitted using different techniques. For example, Chen *et al.* [11] add an energy-efficient image processing strategy through motion detection. Instead of sending the image, the camera sends a simple message indicating that it cannot detect moving objects. Although this approach reduces the amount of video transmitted, misperceptions are still a significant disadvantage. In [12], an architecture is provided for sensor nodes in which a passive infrared (PIR) sensor is used in addition to the camera. The presence of people (or other objects) is first detected by the PIR sensor, and then the video processing module is triggered. The camera captures a series of video sequences and uses the background subtraction technique to detect any intrusion by processing video frames on a Raspberry Pi-based node. Background Subtraction (BS) is a commonly used method for extracting foreground objects in the videos. For effective use of BS in WMSNs, Sukumaran *et al.* [13] propose a base station based on compression detection to reduce the energy consumed during the object extraction and data transmission steps. The compression is applied to the difference frame calculated by subtracting the reference background frame from the current frame. The differences are compared to a threshold to determine which ones to use for reconfiguring the foreground. Thus, instead of sending the entire video, only the most useful data for the reconfiguration of the foreground is transmitted. Although there are numerous studies in the literature to develop methods to reduce the amount of data to be transmitted, as explained in Section II, this issue remains one of the most popular research topics due to the need to reduce energy in WMSNs.

A WMSN is defined as a set of peripherals equipped with cameras and/or microphones. However, many studies in the literature ignore auditory data and focus only on visual data. This is mainly because visual data almost always contains more valuable information than auditory data. In addition, current audio studies treat audio data separately from visual data [14], [15]. Nevertheless, it is possible to develop a multimodal learning solution by combining several data modalities collected from sensors. More specifically, for some surveillance applications, particularly for environmental and industrial surveillance applications, auditory data can be used to improve the accuracy of information obtained from visual data and other sensor data.

In the context of this introduction, the present study aims at developing a new approach to WMSNs proposing an effective solution to the mentioned problems. The main goal is to develop an effective framework for using WMSNs in surveillance applications. It shows that multimodal information collected using multimedia sensors capable of capturing, storing and communicating multimedia information is useful for the early detection of objects and activities. We introduce effective algorithmic procedures on three major issues, namely:

•**Designing a wireless multimedia sensor (WMS) node to detect objects using machine learning methods:** The transmission of big multimedia data to the sink for processing is one of the problems of WMSNs as mentioned above. Processing multimedia data in the node level to convert it in a more informative format is an essential requirement. To achieve this, a WMS node architecture is developed in conjunction with a camera and a microphone, in addition to three scalar sensors, including passive infrared (PIR), acoustic and vibratory sensors. Multimedia sensors are normally kept in sleep mode to save energy and are awakened by scalar sensors, which are always active. When the camera and microphone are turned on, both types of audio and visual data are processed using automatic learning methods and fused at the sensor node. One of the main contributions of this study is to capture and process visual and audio data in addition to scalar sensor data, and fuse all of them to accurately detect and recognize objects in the monitored area of the WMS nodes. This is one of the distinctive features of our study compared to other studies reported in the literature.

•**Develop a method to improve accuracy while reducing the amount of information to be transmitted to the base station:** Although we propose to process the data collected in the node, it is not possible to collect and process all data at a single level for different reasons such as node processing capacity, limited power and network-wide analysis requirement. As a result, an object extraction method using data fusion at three different layers is developed by collecting pre-fused data from the sensor nodes to the base station, which extends the wireless sensor network lifetime by reducing the amount of data to be transmitted over the network. In this context, the data obtained from the PIR, vibratory and acoustic sensors are used at the first layer. The data of these scalar sensors are fused and the first decision is made as to the presence of an object such as a human being or a vehicle in the scene controlled by the sensor node. According to this decision, the next layer, including multimedia sensors (camera and microphone), is activated. More precise information about the objects in the monitored area can be obtained by processing the frame (image) captured by the camera and the sound recorded from the microphone. In the context of second-layer fusion, image and audio data undergoes a fusion process to increase the accuracy of object classification. After these operations on the sensor node, the generated synthesis information is transmitted to the base station via the Zigbee wireless protocol. The third-layer fusion and classification processes are performed at the base station (in the sink). Here, a more sophisticated recognition process is performed using correlations between intra-modal and inter-modal data of different modalities. This process is performed at the base station because it requires high energy and heavy CPU usage. The use of such a three-layer fusion method is another important contribution of this study that distinguishes it from existing studies.

•**Development of a cluster-based routing algorithm that consumes less power than currently used algorithms:** The efficient transfer of information/data by the WMS nodes to the sink is another important problem for the WMSNs because of the limited energy of the WMS nodes. To support the contributions given above with a complementary and efficient routing algorithm, a cluster-based routing algorithm, which consumes less energy than the state-of-art algorithms, is designed and tested. The proposed algorithm presents an unequal clustering approach designed in a distributed and lightweight structure that is easy to use on real sensor nodes. With the proposed algorithm, a clustered sensor network capable of efficiently collecting data can be obtained from a non-clustered wireless sensor network comprising nodes deployed using various methods. The clustering algorithm is a fuzzy logic-based algorithm that uses the distance to the base station, the remaining energy, and the relative connectivity parameters of a node to determine the competitive radius of the tentative cluster heads. The routing algorithm is also based on fuzzy logic and uses the average residual energy of the link and the relative distance parameters in order to make better routing decisions. The cluster-based fuzzy routing algorithm presented in this paper is another contribution of this study.

The initial architecture of the WSN node with only video capability was already presented in [16] and its extension with audio functionality was introduced in [17]. Sert et al. introduce the cluster-based routing algorithm of the proposed framework [18]. However, the main purpose of this article is to provide an overview of the entire structure, including the WMSN three-layer architecture and the cluster-based routing algorithm. At the first layer of the architecture, data collected from different scalar sensors is fused to determine if there is an object of interest in the monitored area. In the context of second layer fusion, information obtained from visual and auditory data is fused to increase the accuracy of object classification. In the third layer, additional recognition is obtained by fusing the results of the classification processes, as well as the intra-mode and inter-mode correlations between the data obtained from the different channels of the base station. Although our earlier papers include some specific parts of the study, this paper presents the general framework with the latest developments and extensions, including our third-layer fusion and classification approaches.

The rest of the paper is organized as follows: Section II summarizes the relevant studies in the literature. Section III provides an overview of the system architecture, followed by the details of the three-layer data processing and fusion steps in Section IV. The proposed routing algorithm based on fuzzy clusters is presented in Section V. Finally, Section VI presents the conclusions of this study.

II. RELATED WORK

Wireless sensor nodes are typically battery-powered devices with limited energy, which can result in the unexpected death of a sensor node before the end of the assigned tasks. Therefore, conservation of energy to extend the life of WSNs is one of the critical issues that is attracting the attention of researchers. Several solutions are proposed to solve the problem of energy consumption of battery-powered sensor nodes. Anastasi *et al.* [9] study the power consumption of components of a typical sensor node and discuss key aspects of energy conservation in WSNs. They present a systematic and comprehensive taxonomy of energy conservation techniques, including topology control, energy management, data reduction, and energy-efficient data collection. Rault *et al.* [19] divide existing approaches to energy conservation into five classes: radio optimization, data reduction, standby/wake-up systems, energy-efficient routing, and recharging solutions. For each of these categories, various studies on specific methods of energy saving can be found in the literature.

On the other hand, multimedia content in WMSNs exacerbates the problem because of its complex structure and size. Although multimedia content can be processed on such networks, WMSNs have certain limitations. Unlike traditional network systems, the WMSNs have limited computing power, small storage capacity, short-term power sources, and limited transmission bandwidth. Since sensor nodes with video and audio capture capabilities consume more power than traditional scalar sensor nodes, the development of more energy efficient methods is a necessity for WMSNs [20]. Akyildiz *et al.* [3] explain that special equipment and algorithms should be developed to process the multimedia data in the network to prevent the passage of large amounts of raw streams into the sink. Some studies show that data processing on wireless sensor nodes consumes much less energy than that used for network transmissions [10], [19]. As such, in-network processing is one of the recommended methods for saving energy by reducing the amount of data to be transmitted and thus extending the life of WMSNs.

One of the approaches used to reduce the amount of video data to be transported in WMSNs is motion detection [11], [21], [22]. A background image is captured and stored as a frame of reference when the camera is initialized. Then, the new images are compared to the reference image to understand if there is a moving object in the scene using techniques such as background subtraction [13]. At the end, only images with motion or moving object regions are transmitted to reduce the amount of video data. Some studies enable cameras, which result in more power consumption in a WMS node, based on scalar sensors such as PIR, when motion is detected [12]. In this way, unnecessary power consumption is avoided. Compression or encoding of the video data to be transmitted is another technique widely used to minimize communication needs [23], [24]. Some studies combine motion detection and compression techniques to further reduce the size of video data [13], [21], [22]. Video summarization [25] and wireless line sensor usage [26] may be mentioned as other techniques proposed to reduce the size of the video data to be transmitted on WMSNs. Depending on technological developments, it has become possible to develop sensor nodes with higher processing capabilities in recent years. Dependently, there are studies performing object classification at the sensor node and sending only information about the detected objects in text format and, thus, reducing drastically the size of data to be transmitted [27], [28]. While there are studies on video data processing at sensor nodes to minimize the size of the data to be transmitted, as explained above, additional research is needed to find energy efficient solutions for WMSNs.

Another interesting point related to WMSNs is that there are very few studies that use acoustic sensor data for object/event detection and classification. Although a multimedia sensor node is defined as a sensor node with audio and video capture and processing capability, studies typically focus only on video data, as shown in the studies above. In [14], the authors propose a two-level WMSN architecture consisting of low-cost audio nodes deployed in a dense manner at the first level and high-cost video nodes placed at the second level for surveillance applications. Audio events are first detected by the audio nodes and then the video nodes are activated by the base station on demand. In the study, auditory and visual data are treated separately. In another study, a multimedia sensor node based on Mica2 motes is developed to capture and process audio data for event detection [15]. The developed system is trained and tested to detect tree cutting events in a forest area. The authors claim that the analytical and experimental results prove the effectiveness of the proposed event detection system. In this study, only audio data is used while ignoring video data. Existing studies in the literature show a great lack of studies on the use of audio data in WMSNs, which is one of the modalities discussed in the current study.

Data fusion combines data collected from different sources to improve the overall performance of a system. It is also used widely in WSN applications to improve reliability of networks [29]. Multiple sensors send their decisions about the observed phenomenon to a central fusion center where a final decision is declared using a decision fusion rule. Different fusion rules for WSNs with space diversity are studied in [30]. Channel and jamming aware decision fusion in multiple-input multiple-output (MIMO) WSNs is examined under Rician fading channels where the sensors transmit their decisions simultaneously and the fusion center which is equipped with multiple antennas is tested for various decision rules. Kailkhura et al. proposed a robust distributed weighted average consensus algorithm and devised a learning technique to estimate the weights of the nodes [31]. This enables an adaptive design of the local fusion or update rules to mitigate the effect of data falsification attacks. Salvo Rossi et al. investigated a system for MIMO decision fusion in underwater sensor networks based on energy detection and concluded that it achieves a good performance even with low-quality sensors [32]. Wimalajeewa and Varshney investigated multimodal data fusion for detection purposes with heterogeneous dependent data in a compressed domain [33]. They tested the proposed model with a dataset consisting of raw observations from several acoustic, seismic and PIR sensors that were deployed in an outdoor space to record human and animal activities. They concluded that their approach with a small number of compressed measurements per node leads to enhanced performance compared to detection with uncompressed data under certain conditions. The two sub-optimal decision fusion algorithms are presented in the context of distributed classification of multiple moving targets in [34]. It was shown that the complexity of the proposed approaches is lower both in time and space dimensions with respect to the joint Optimal Decision Fusion (ODF). These studies show the usage of data fusion in WSN studies. However, there is no study using multimode data fusion for visual and auditory data in WMSNs, which is one of the contributions of the current study.

On the other hand, choosing the best route to transmit the data collected by the WMS nodes to the sink is another important problem to solve for WMSNs [35]–[38]. Therefore, an energy-efficient routing protocol should be integrated with data reduction techniques to extend the life of WMSNs. Existing routing protocols can be categorized as flat routing, cluster-based routing, and location-based routing protocols. Cluster-based routing algorithms have gained more attention from researchers because of its advantages [39], [40]. Clustering is one of the effective ways to achieve an energy-efficient network, in which some nodes are assigned as cluster heads used to relay data from sensor nodes. There are many studies in the literature about clustering algorithms and their properties. A detailed review of well-known clustering algorithms such as Low Energy Compliant Clustering Hierarchy (LEACH), Hybrid Energy Efficient Distributed Clustering (HEED), Concentric Clustering Scheme (CCS), Energy Efficient and Distance-Based Clustering (EEDC), and Energy Efficiency Clustered Diagram (EEHC) is discussed in [38] and [41]. In addition, in [42], a general approach is provided to adjust clustering algorithm parameters to optimize WSN performance criteria. With the use of simulated annealing algorithms and K-means to adjust the parameters of grouping and routing protocols, a systematic and efficient method is presented in the Castalia framework with OMNET++, a discrete event system simulator (DESS). The use of the actual parameters is necessary to obtain the most efficient configuration. In addition to the data provided, various studies on WMSN routing and clustering can be found. However, both clustering and routing are broad areas where many unresolved issues require more comprehensive research efforts.

Based on the summarized literature above, a comparison of different features existing in the current study and similar studies are given in Table 1. The table presents an overview of the contributions of our study.

III. MULTIMEDIA SENSOR NODE AND NETWORK ARCHITECTURE

The architecture proposed in this study is a multilayer automatic surveillance system consisting of wireless multimedia and scalar sensors for outdoor applications. The architecture of the system is given in Fig. 2. At the WMS node level, the system consists of two layers. The first layer contains scalar sensors with acoustic, vibratory and motion detection capabilities. This layer activates the second layer comprising multimedia sensors having audio and video capture capabilities.

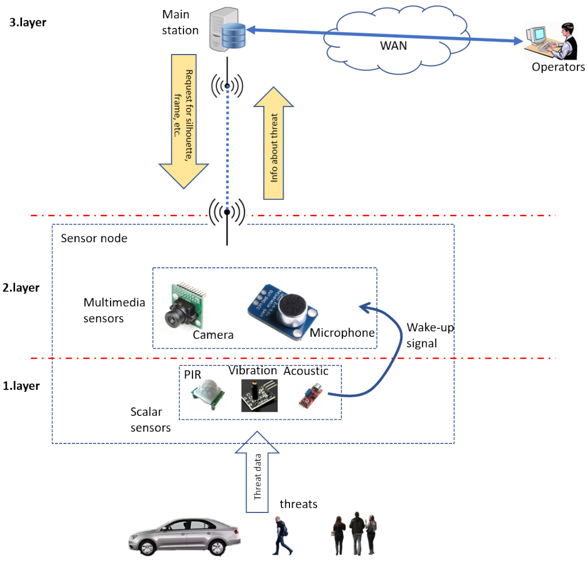

FIGURE 2. System architecture.

The first operation on the WMS node is to determine the object entering the monitoring area by scalar measurements. Scalar sensors provide digital signals to the sensor node. These signals are combined (fused) in rule-based decision making and it is then decided whether the microphone and/or camera should be activated. The WMS nodes work by checking whether the signals from scalar sensors are stable for a predefined time interval, thus avoiding unnecessary operations.

The scalar sensors in the first layer perform the initial detection of the potentially dangerous objects. Once the objects are detected by the first layer, the multimedia sensors become active. In addition to the scalar data, the image obtained by the video camera and the sound obtained by the microphone are used to improve the accuracy of semantic information relating to the object entering the controlled area.

The sensor nodes are developed on Raspberry Pi (RPi) Model B+ after a detailed examination of single-card computers available on the market. The camera and microphone required for the node as multimedia sensors, PIR, acoustic sensor, vibration sensor, Xbee communication module,

TABLE 1. Comparison of features supported by existing work in the literature.

Features	Chen et al.[11]	Pham and Aziz [21]	Ur Rahman et al. [22]	Magno et al.[12]	Alhilal et al. [27]	Bhatt and Datta [14]	Singh et al.[15]	**Our study**
Exploiting scalar sensor data?	No	No	No	Yes	No	No	No	Yes
Exploiting visual data?	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Exploiting audio data?	No	No	No	No	No	Yes	Yes	Yes
Fusion on multiple modalities (scalar, audio and video data)?	No	No	No	No	No	No	No	Yes
In-node processing to reduce the size of data to be transmitted?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Object detection in wireless multimedia sensor node?	No	No	No	No	Yes	Event detection	Yes	Yes
Test and deploy AI models at the edge, locally?	No	No	No	No	No	No	Yes	Yes
Layered node architecture (use multimedia sensors when needed)?	No	No	No	Yes	No	Yes	No	Yes
Efficient cluster-based routing algorithm included in the framework?	No	No	No	No	No	No	No	Yes

SD card and battery are connected as shown in Fig. 3. A sensor node performs two main tasks before the operation. The first is to train the system to learn the features of each object category under consideration in order to perform an automatic classification later. The training dataset is derived from pre-recorded video images and audio data captured by the sensor node itself. When the sensor node is activated, it learns the specified features of the object classes using the training dataset. The second main task is to generate a statistically updated background template for the image to be used for background subtraction. An initial background is detected at the beginning of the operation and stored in memory for later use.

Once the camera is enabled, the WMS node begins image processing operations to identify objects in the frame. When a new object is detected, the low-level features of the new objects are matched to the features of the learned objects and evaluated according to the most relevant category. The same goes for the processing of audio data. The previously trained node classifies the voice according to the training data. The accuracy of the extracted semantic information is increased by transmitting the results of the classification of the audio and visual data through a fusion process. Instead of sending the raw multimedia data, the WMS node sends the semantic information extracted from the multimedia data.

A two-layer fuzzy logic-based protocol, namely Two-Tier Distributed Fuzzy Logic Based Protocol for Efficient Data Aggregation in Multi-Hop Wireless Sensor Networks Protocol (TTDFP), was developed for efficient transfer of data from sensor nodes to the base station. WMS nodes create clusters to send data. They are connected to a cluster head and transmit their data to the cluster head using a wireless interface. The cluster head node sends its own data to the base station along with the cluster data. Radio frequency

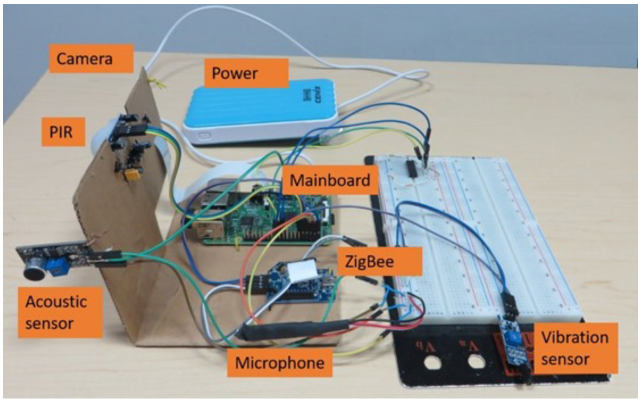

FIGURE 3. Wireless multimedia sensor node.

serial communication is used between sensor nodes to limit power consumption. The sensor nodes send the results of the classification in a text format. Thus, the amount of data transmitted and received by the nodes is minimized and power is saved in significant amounts. Nodes are also designed and developed for the generation and transmission of multimedia data such as low-level features of objects, silhouette or foreground image at the request of users of the system.

The data from different sensor nodes are combined in the base station. The base station stores the data in its own database and makes it accessible to operators. By processing this aggregated data, it becomes possible to draw conclusions about the surveillance area and decide which action is appropriate. The base station has data collected not from a single sensor node, but from several sensor nodes. It contains semantic data extracted by the sensor nodes, as well as silhouette or foreground images. Even low-level features extracted by sensor nodes can be transmitted to the main station instead of the raw data. In addition, the main station may have data from different modalities such as visual and auditory data. Therefore, complicated detection algorithms requiring more powerful resources can be used at the main station. In this study, we use a fusing technique that benefits from intra-modal and inter-modal correlations in addition to the classification performed for each category separately.

IV. DATA PROCESSING AND FUSION

The data collected by scalar and multimedia sensors are processed and fused at three different layers. In this section, we discuss the details of our fusion and classification approaches at each layer.

A. FIRST LAYER FUSION AND CLASSIFICATION

The purpose of the first-layer fusion and classification process is to determine whether there is a requirement to reactivate the camera and/or microphone, resulting in high power consumption at the node. In case of an object detection, the camera and/or the microphone are activated and the collected data are used for the second-layer fusion and classification. We have tried to find a simple method for the first-layer fusion and classification, which does not overload

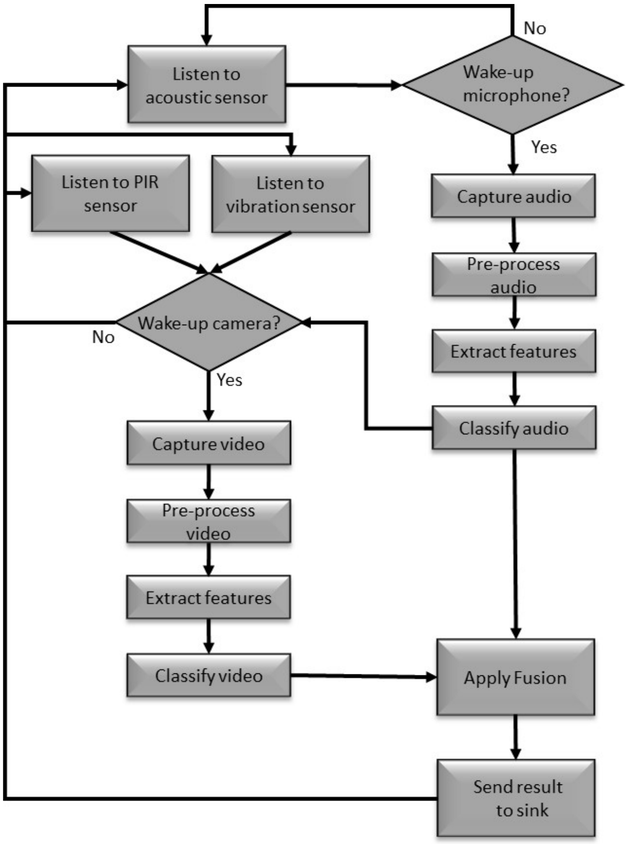

FIGURE 4. Flow diagram of a WMS node processing.

the node and causes only minimal power consumption for continuous operation. Therefore, it is decided to apply a rule-based data fusion method to activate the multimedia sensor nodes. In other words, the outputs of the scalar sensors are evaluated by a set of IF-THEN rules and the decision to activate the multimedia sensors is produced when the conditions of a rule are met. Fig. 4 shows the flowchart of the fusion and classification processes of the first and second layers in a WMS node application.

In terms of real-life scenarios, when there is no activity in the controlled area, the microphone and camera remain in sleep mode for energy efficiency. When a high-level sound is detected by acoustic sensor, the microphone is activated. For example, when there is a sound from a vehicle passes through the controlled area, the microphone is activated and the collected sound is processed. However, the camera is activated taking into account the output of three sensors (PIR, vibration sensor and microphone) as shown in Algorithm 1. These signals are expected until they become stable. One of them or a combination of them can activate the camera. For example, when a truck passes into the controlled area of the sensor node, the PIR and vibration sensors can easily detect it and are used to activate the camera. The high signal of PIR is used for activation. On the other hand, the vibration sensor produces a signal whose value is between 0 and 1024. We have experimentally determined a threshold value and if

Algorithm 1: Camera Wake-Up

```
    INPUT: cameraState, PIR, VIB, MIC
    OUTPUT: New status
    getValues (PIR, VIB, MIC, cameraState)
     if (PIR or VIB)//Alarm by PIR or VIB
       if (cameraState is OFF)
           wait() // Signal stability check for PIR and VIB
          if (PIR or VIB or MIC) // PIR/VIB still active
    or alarm by MIC
             return ON // Turn camera on
          else // wrong alarm
            return OFF // Keep camera off
           endif
          else if (cameraState is ON)
           return
    ON // Continue detection by keeping camera on
          endif
        elseif (!PIR and !VIB and !MIC)
          return OFF // No alarm, turn camera off.
        endif
```

the signal level is higher than the predefined threshold value, it is used for camera activation. We have used a two-level activation procedure for camera activation depending on the audio. If an audio signal exists in the environment, the acoustic sensor activates the microphone, the audio data captured by the microphone is processed, and if an object is detected, the node's camera is activated. The reason for this implementation is to prevent the unnecessary activation of the camera, which consumes energy and computing power, with meaningless audio signals (noise). Although our first-layer algorithm is simple, it avoids to some extent unnecessary use of multimedia sensors to maximize the lifetime of a sensor node.

The purpose of the first-layer fusion and classification process is to activate the camera and/or microphone if there is a potential risk. There is a light-weight decision system which fuses the output of three scalar sensors and the state of the camera. In our implementation, there are 16 rules with 4 conditions in the decision system. The computational complexity of layer 1 algorithm is O(1), which means that its time complexity is constant. According to the experiments, it takes about 20 ms in the implemented WMS node.

B. SECOND LAYER FUSION AND CLASSIFICATION

At the second layer of the architecture, multimedia sensors consisting of a camera for video capture and a microphone for audio capture are used. The activation of the camera has been implemented as described above. Once the camera is turned on, a snapshot is taken and the existing objects in the front view are extracted using the connected component analysis. If the area covered by the extracted objects is below a certain threshold, these regions are considered noise and are ignored. If the area covered by an object is greater than a certain threshold, object classification on the image of the object begins by using a machine learning algorithm, namely Support Vector Machine (SVM)-based classification.

A C ++ application has been developed for extracting and classifying objects on the sensor node. The application uses OpenCV Library for image processing functions [43]. The Gloox XMPP Client Library is used to send messages to the main station and to receive operator commands [44].

A two-level cascading classification approach was chosen to extract image objects after testing different models, as explained at the end of this section. When a BLOB is detected by background subtraction, its features are first extracted. In this scope, 3 shape-based features, which are the width-to-height ratio, the compactness and the ratio of the blob area, are calculated and the SVM-based classification is performed using calculated values. The best matching label is subtracted from the training set. When testing sensor nodes, three categories of objects are considered: vehicle, human and group of people. According to our experiments, vehicle-type objects can be well ranked. However, this method of classification can pose some problems when classifying objects of human type and group of people. The second step of the cascaded classification is carried out when the result of the first step is not the type of vehicle. At this point, the Speeded Up Robust Features (SURF) descriptors of the detected BLOBs are extracted and placed in a bag of word (BoW) structure and mapped to the SURF descriptors of the training set using the SVM-based classification. The first reason for choosing such a cascading classification approach is that shape-based functions are suitable for lightweight systems, especially wireless sensor nodes. However, they are insufficient in the classification of humans and groups of people. The second reason is that the video quality of the WMS node is low and requires more detailed processing for some types of objects.

The microphone has been added to the WMS node as the second additional multimedia device. In this way, it becomes possible to collect peripheral voices and to perform an additional classification based on these collected voices. Two types of sound have been studied: human (single or group) and vehicle. The classification of the sound is used for two purposes: 1) Activate the camera according to the result of the sound classification, as shown in Fig.4 (the camera is activated if the result of the sound classification is human or vehicle); 2) The result of the audio classification is fused with the video classification result to improve the overall performance of the node classification.

An additional application that captures and classifies audio data coming from the RPi's SPI pins is integrated into the sensor node. Because the classification is done using SVM, the application starts with the training process when it is first started. The training dataset is composed of different voices recorded in classes of humans and vehicles collected by the node itself. The recorded sounds are stored as a training set in a comma-separated value (CSV) file, taking the 13 Mel Frequencies Cepstral Coefficients (MFCC). A CSV file has been created for each sound class (human and vehicle).

The application trains with these CSV files and learns the MFCC properties of the classes during initialization.

The sound application continuously monitors the output of the acoustic sensor when the microphone is in a passive state. When the acoustic sensor receives a strong signal, the application starts acquiring raw audio data from the SPI of the RPi that the MCP800 ADC and the microphone connected to it, and continues to acquire audio data for a certain period of time (test time is 1 second). Raw audio data at a frequency of 10 KHz and a resolution of 16 bits are buffered. The MFCC properties of the buffered data are subtracted and sent to SVM to detect the class. Instead of giving the exact matrix of MFCC properties, the average value of each column of the matrix is calculated and a vector of 13 MFCC coefficients is used in the classification. The result of the audio classification is sent to the main application (video) via a UDP socket.

As noted above, the result of the audio classification is fused with the video classification results to improve the overall performance of the node classification. In this context, the classification result obtained from image data and the classification results obtained from audio data are used. As a fusion method, a function-based high-level fusion combining the audio classification results and the video classification results is applied based on the decision function given in Equation (1). Algorithm 2 describes the fusion implemented at the second layer to fuse the classification results of video and audio classifications.

$$d\left(\overrightarrow{v_n}, a_n\right) = \begin{cases} \overrightarrow{v_n} + a_n, & \textit{when human or vehicle are detected} \\ & \qquad\qquad\textit{from voice} \\ \vec{v}_n, & \textit{when there is no object detected from voice} \end{cases} \tag{1}$$

Here, $\overrightarrow{v_n}$ is a list of the object classes detected after processing a video frame, and a_n is the object class detected after the sound processing.

Algorithm 2: Second Layer Fusion

```
INPUT: Frame, Audio
OUTPUT: listofObjectsinFrame
$listofObjectsinFrame = classifyFrame(Frame)$
$(audioCategory, membershipRatio) = getClass(Audio)$
if $(membershipRatio > membershipThreshold)$ and
   audioCategory not exists in listofObjectsinFrame
      add audioCategory to listofObjectsinFrame
endif
```

At the second layer, the captured video and audio data are processed. Video data processing is the main source of complexity and it is done as a set of activities including background subtraction, pre-processing, segmentation (cutting out bounding box of objects), feature extraction, classification and fusion with audio. The complexity of each one for a frame processing can be defined as follows:

- Background subtraction: O(w*h), where w is the width, h is the height of a frame.
- Pre-processing: O(2*w*h), where w is the width, h is the height of a frame, 2 is for cleaning and sharpening operations.
- Segmentation (cutting out bounding box (BB) of objects): O(k), where k is the number of BBs covering objects in a frame.
- Feature extraction: O(n*m*k), where n is the width, m is the height of a BB, k is the number of BBs (for SURF).
- Classification: O(e*f), where e is the number of features, f is the size of training data
- Fusion: When there is a corresponding audio with the processed video frame, the class of this audio is compared with the classes of objects extracted from video. Since there is only one class obtained from audio for a frame and maximum several classes obtained from video, its time complexity is defined as O(1).

Depending on the complexities given, background subtraction and pre-processing are the dominant activities for computational complexity with O(3*w*h) for the second layer. Time measurements show that these two activities consume most of the time. In other words, approximately 88% of the time used to analyze a video frame is consumed by these two activities and remaining 12% is consumed by other activities.

A series of tests were performed on the sensor node developed to measure the performance of different visual data algorithms in the first step. For training and test data, the video captured by the node itself is annotated and used. As classes of objects, vehicle, human and group of people are tested. The k-NN and SVM classifiers are tested using shape-based features, namely width/height ratio, compactness and BLOB ratio, in order to obtain a light classification. Their results appear in the first and second rows of Table 2. Although we observe high performances for the vehicle class, the results obtained were not satisfactory for the others. Therefore, the SVM classifier with the SURF descriptor is tested as a third test to determine a better method and found that it better distinguishes human and group of people classes than the previous ones. However, its performance for the vehicle was lower and the treatment takes longer. As such, we decided to also test cascaded two classifiers (k-NN+SVM or SVM+SVM), as shown in the last two rows of Table 2. The first classifier (k-NN or SVM) with the shape-based features is used to understand whether there is a vehicle or not. If there is a vehicle, there is no need to continue processing. However, if the object is not a vehicle, the second classifier, which uses SVM as a classifier and BoW of SURF as a feature, is activated to check if there is a human object or a group of people in the scene. The fifth model, which uses two cascading SVM classifications, gives the best performance based on the average f-measure, as shown in the last column of Table 2.

After determining an appropriate classification method for the visual data, the sensor node is expanded with the audio capability as explained above and tested in the second step to

TABLE 2. Classification performance of video-only algorithms.

	Human			Group of people			Vehicle			
Model	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Average f-measure**
k-NN (shape-based features)	0.51	0.35	0.42	0.47	0.71	0.57	0.85	0.95	0.90	0.63
SVM (shape-based features)	0.55	0.32	0.40	0.5	0.84	0.63	0.86	0.93	0.89	0.64
SVM (Bag of SURF)	0.49	0.46	0.47	0.5	0.61	0.55	0.87	0.68	0.76	0.60
k-NN(shape-based features) + SVM(bag of SURF)	0.55	0.39	0.46	0.5	0.71	0.59	0.84	0.96	0.90	0.65
SVM(shape-based features) + SVM(bag of SURF)	0.57	0.48	0.52	0.5	0.77	0.61	0.86	0.93	0.89	**0.67**

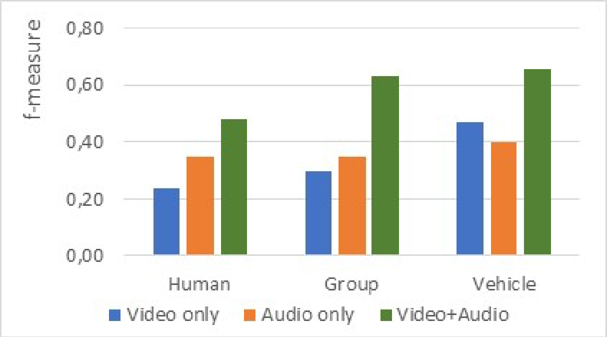

FIGURE 5. Comparison of performance results of three test cases.

observe the effect of the audio data on the object classification. As with the previous test, the visual data and auditory data captured by the sensor node itself are used for training and testing purposes. The recorded video has a frame rate of 25 fps, a resolution of 640 × 480 pixels and H.264 video coding. The corresponding audio is recorded at a frequency of 10 KHz and a resolution of 16 bits. Video and audio are annotated using different tools. The only video algorithm giving the best result of Table 2 is used in this case. In this perspective, three cases are tested to see the contribution of different modalities (video and audio separately) and the fusion of these two modalities. The results of the experimental tests obtained are given in Fig. 5. As the figure shows, the fusion of two modalities gives better performance for all classes. Another important point is that it is possible to increase the performance of the video-only case by decreasing the area threshold, which corresponds to the ratio between the area of the region of interest and the area of the frame. The regions detected below the area threshold are considered noise and eliminated without processing. In the given results, 0.03 is used as the threshold value. When a lower value, for example 0.01, is used, the performance of the video data improves. However, this requires more energy and processing resources because the visual data has a complex structure. On the other hand, audio data, which has a simpler structure and requires less processing power, can be used to compensate for such visual data loss, while providing better results when audio data is available. Interested readers may refer to [16] and [17] for more detailed information on the architecture of our WMS nodes and the other performance results obtained.

C. THIRD LAYER FUSION AND CLASSIFICATION

In the third-layer fusion and classification process, a more sophisticated recognition process was investigated using intra-mode and inter-mode correlation with data obtained from different channels. This process takes place in the base station because it requires more energy and resource usage costs.

1) BOW-BASED FUSION STRUCTURE

In the recent studies of multimedia and multimodal information access literature, Bag of Words (BoW) is frequently used [45] and good results are obtained with this model. In this direction, a unique method has been developed for the third layer data fusion to work according to the BoW model. This is an intermediate level of knowledge between low-level features and high-level concepts. Thus, it can still use valuable information found in low-level features, but it also includes features such as high-level concepts; because words represent parts or regions in concepts.

Fig. 6 illustrates a general fusion framework. The proposed approach assumes that all inputs are in the form of BoWs. However, if it is not in the form of BoW, converting information into a BoW form is not a complicated process. Fusion entries can be classified into two types, as shown below:

- Type-1 (work with low-level features): Low-level features of multimedia data can be used as input for the fusion. However, there are key points or local part requirements. After getting the key points, the words are clustered to form a vocabulary (dictionary). Then, the training data is converted into BoW format using the dictionary. Any other type of information that cannot be represented by local parts is treated as the second type indicated below.

- Type-2 (work with high-level concepts): High-level (semantic) concepts are the second type of entry for the fusion. All of the high-level concepts that occur in the training set of multimedia data are assumed to be the vocabulary and each high-level concept is a word. After that, the data can be easily converted to BoW format. In this way, depending on the type of classifier or target class of the classifier, it can produce many word bags such as bag of objects, bag of activity, bag of process, bag of silhouette, etc.

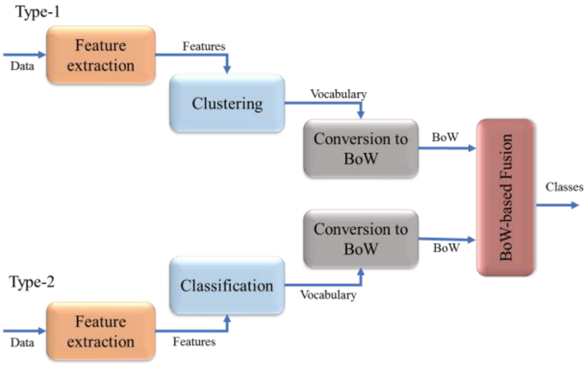

FIGURE 6. General fusion frame.

Using the general fusion frame described in Fig. 6 for the third level data fusion, both the first and second level classification results (classification results obtained at the sensor node), and low-level features obtained from camera such as SIFT, SURF and STIP (visual features) or low-level features obtained from microphone such as MFCC and ZCR (auditory features), which are frequently used in current multimedia information extraction studies, can be combined in the same integration process.

In addition to the BoW model applied in the general fusion frame, the objective is to use both complementary and correlated information from different modalities in the fusion process, and to maximize the contribution to be obtained at the result. In this direction, in addition to each modality classification result, intra-modal and inter-modal correlations are included in the fusion process. The use of intra-modal correlations allows the detection of co-occurring words for different types of objects in a particular subset of BoW. In the same way, inter-modal correlation analysis reveals the coexistence of words in different modalities. To give a simple example; in the case where our target class is a vehicle, the words to be obtained from visual mode (each word is actually a local feature) define various points of interest (headlight, wheels, mirror, etc.) of the vehicle. In auditory mode, different words are produced in relation to the audio signals of the vehicle. Making an intra-modal correlation analysis on the words that describe the points of interest on the vehicle (visual modality) and on the words that express the voice of the vehicle (audio modality), and ultimately making an inter-modal correlation analysis on the words of images and audio, which can be related to each other, can produce useful information for a more effective recognition process. The method developed for this purpose schematically is illustrated in Fig. 7 and comprises the following four steps:

1. Classification of information in each modality
2. Intra-modal correlation analysis for each modality and classification of information obtained
3. Inter-modal correlation analysis between modalities and classification of information obtained
4. Late fusion of all classification results.

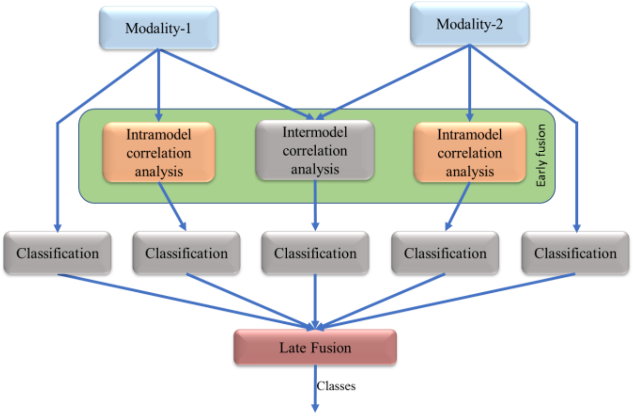

FIGURE 7. Flow of BoW based fusion.

2) CLASSIFICATION OF INFORMATION IN EACH MODALITY

It is important to note that recognition performance prior to the fusion is very important and that the fusion process enhances this recognition performance to a certain extent. For this reason, the learning process for each mode (modality) should be as effective as possible. In this study, the approach proposed by Jiang *et al.* [45] is used as summarized below:

a: CLASSIFIER AND KERNEL SELECTION

Support Vector Machine (SVM) is one of the most popular classifiers for BoW-based classification. Choosing an appropriate kernel function for SVMs is a critical issue for classification performance. Jiang *et al.* [45] experimentally showed that the RBF kernel gives better results when used for classifying BoW based feature vectors. For this reason, the use of SVM with the RBF kernel has been preferred for classification. Another issue with the classification procedure is how to deal with multi-class classification. Given that multimedia data is often multi-labelled, the classification procedure must be multi-class and multi-labelled accordingly. For such a purpose, a one-against-all approach, in which k classifiers are trained for k different class labels, is preferred.

b: WEIGHTING SCHEME

The weighting in BoW is statistical information on the repetition of words in multimedia document. The most basic schema is the binary weighting that shows the presence/absence of a word in each document. More complex schemes include term frequency (TF) and/or inverse document frequency (TF-IDF), which work better than binary weighting. TF weighting is preferred in this study.

c: VOCABULARY SIZE

For BoW modeling, vocabulary is a set of key points in the clustering process. Having a small vocabulary can lead to difficulties in distinguishing differences because two key points can be assigned to the same set, although they are not similar. On the contrary, a large vocabulary is less generalizable, less tolerant of noise and causes more processing time. Studies in the literature work with a vocabulary dimension of between 100 and 10,000. Jiang *et al.* [45] have shown that the effect of vocabulary size is less important when complex weighting systems are used. For this reason, a medium level dimension, 4096, is preferred.

3) INTRAMODAL CORRELATION ANALYSIS AND CLASSIFICATION

For intra-modal correlation, each mode is treated separately. It is a good idea to group words that occur on the assumption that parts of a particular object or scene are often together in different instances of that object/scene. For this reason, we recommend to use phrases as groups of words that are frequently repeated together. In order to find the phrases, a graph-based data mining algorithm is used in the training dataset. Thanks to the mining algorithm, we find significant phrases for each modality. Thus, intra-modal relationships within each modality are revealed. Then, in order to extract phrases and find the words included in each phrase, a graphical representation is constructed. After creating the graph, phrases are drawn by processing the graph.

It should be noted here that samples belonging to each class must be treated separately. This means that the algorithm runs separately for each class. During frequent repetitive word mining, the support thresholds for frequent patterns are very different for various classes. Therefore, class-specific support thresholds must be applied to each class.

After obtaining phrases with the procedure outlined above, phrase-based feature vectors must be extracted from the training and test data. Since each phrase contains several words, we need an aggregation method to assign numeric values to each phrase. To this end, we prefer a simple averaging approach. In this approach, the average of the TF values of the words of each phrase is calculated as a phrase value. After performing the aggregation task, phrase-based feature vectors are obtained for each training and test document. The learning procedure using feature vectors extracted from extracted phrases is similar to the procedure described in the previous section. The data is classified with an SVM classifier based on an RBF kernel.

4) INTERMODAL CORRELATION ANALYSIS AND CLASSIFICATION

For the discovery part of the inter-modal correlation problem, all the modalities are processed together and the correlations between words and phrases of different modalities are extracted. The idea of intra-modal correlation analysis also applies to inter-modal correlation. The data of a particular scene, collected by different channels (modalities), usually have parts that exist together in different instances of that scene. For example, if different samples of a vehicle video are processed, it is very likely that some visual words belonging to the vehicle appear with certain vehicle audio signals.

In order to find multimodal phrases, a correlation and a graph-based grouping algorithm are applied to the training data set. Thanks to the algorithm, the correlations between pairs of phrases of different modalities are first calculated. The correlation is calculated based on the Pearson correlation coefficient. Then, by selecting a single phrase of each modality, the other phrases associated with the selected phrase are determined and groups of multimodal phrases are formed in this manner. Similar to intra-modal analysis, the given algorithm is executed separately for each class. Thus, inter-modal relationships are revealed for each modality. As in the intra-modal correlation analysis, a graph is generated by including the phrases of all modalities as nodes of the graph in order to facilitate the extraction of the phrases.

After deriving the multimodal phrases, the feature vectors for the multimodal phrase must be extracted from the training and test data. Similar to intra-modal analysis, a simple averaging approach is used to aggregate multiple phrases into a single multimodal phrase and to assign a numerical value to each phrase. After averaging the TF values of the phrases and assigning these values as multimodal phrase values, a feature vector based on multimodal phrases is obtained for each training and test document.

The learning procedure for inter-modal analysis is similar to the intra-model analysis procedure. For learning and querying, extracted phrase-based multimodal feature vectors are used. The data is classified with an SVM classifier based on an RBF kernel.

5) LATE FUSION

Although the fusion process helps improve information retrieval performance, the success rate of each modality contributes the most to final performance. Methods of learning provided by different modalities and intra-modal/inter-modal analyzes make it possible to produce abstract videos in different ways. Each of these learning methods is likely to complement each other. If an object is misclassified by one of the learning procedures, it is still possible to be correctly classified by others. For this reason, all of these methods must be combined (fused) to improve their recognition capability.

After performing the classification procedures for each modality, as well as the intra-modal/inter-modal analyzes, the results of the classification are combined with the late fusion scheme, as shown in Fig. 7. The results are integrated using the linear weighted averaging approach. Because of its simplicity and reasonable performance, it is the most commonly used approach in the fusion literature [46]–[49]. The approach requires a good selection of weights in order to obtain positive results; it is therefore supported by the RELIEF-MM algorithm for the modality/feature weighting [50].

The time complexity of the third layer fusion algorithm is bounded by the intermodal correlation calculation operation given in Algorithm 3, which is the most complex operation. The number of modalities (such as video and audio) is very small depending on the number of multimedia documents. Thus, the complexity of the given algorithm is in linear in terms of the number of multimedia documents, because the algorithm requires two passes over the entire dataset. If the number of modalities is concerned, the complexity of the algorithm is in quadratic (or sub-quadratic time, depending on the implementation) in terms of the number of modalities, since all pairs of phrases from different modalities are calculated.

The fusion method described above as the third layer of the WMSN was tested on the TRECVID 2011 dataset. Since we could not collect enough video data with various objects using our WMS nodes, the TRECVID dataset was chosen for these tests. In Fig. 8, the obtained MAP values are illustrated for different test cases which are summarized in Table 3.

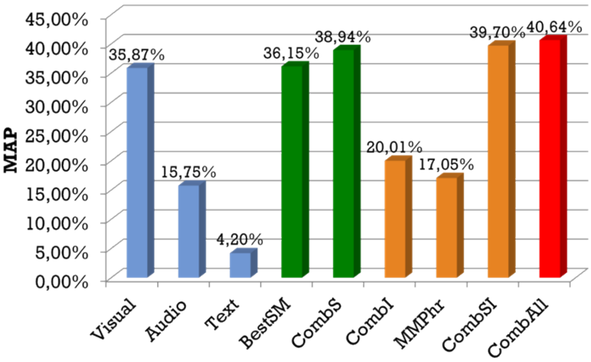

FIGURE 8. MAP comparison for different test cases.

Here we want to share the fusion gains since our goal is the fusion. Table 4 shows the fusion gains of each fusion configuration relative to the others in percentage. As shown in the table, the proposed fusion algorithm (Comb$_{\text{all}}$) greatly improves the object recognition performance of the system.

Algorithm 3: Intermodal Correlation Analysis

```
Input: Modalities Ad $\mathcal{M} = \{m_i\}_{i=1}^{n}$, multimedia
documents $\mathcal{D} = \{d_i\}_{i=1}^{t}$ phrase vocabulary list of all
modalities $\mathcal{PW} = \left\langle PW^i \right\rangle_{i=1}^{n}$ s.t. each phrase vocabulary
$PW^i = \left\{phr_j\right\}_{j=1}^{r}$
Output: Multimodal phrases list $MMP$
begin
    // Correlation calculation
    for $d_k \in \mathcal{D}$ do
        for $m_i \in \mathcal{M}$ do
            $P^i \leftarrow getPhraseV\ ector\ (d_k, m_i)$;
            for $p_a \in P^i$ do
                $mean\,[m_i]\,[p_a] \leftarrow$
                $mean\,[m_i]\,[p_a] + value\,(p_a)\,/size(\mathcal{D})$
            end
        end
    end
    for $d_k \in \mathcal{D}$ do
        foreach $\left\{m_i, m_j\right\} \in \mathcal{M} \times \mathcal{M}, i\ \neq j$ do
            $P^i \leftarrow getPhraseVector\,(d_k, m_i)$;
            $P^j \leftarrow getPhraseVector\,\left(d_k, m_j\right)$;
            for $\{p_a, p_b\} \in P^i \times P^j, p_a \in P^i \wedge p_b \in P^j$ do
                // Pearson's corr.coeff. calculation
                $partX \leftarrow value\,(p_a) - mean\,[m_i]\,[p_a]$;
                $partY \leftarrow value\,(p_b) - mean\,[m_i]\,[p_b]$;
                $partCov \leftarrow partX \times partY$;
                $cov\,[m_i]\left[m_j\right][p_a]\,[p_b] \leftarrow$
                $cov\,[m_i]\left[m_j\right][p_a]\,[p_b] + partCov$;
                $sdtDev\,[m_i]\,[p_a] \leftarrow$
                $sdtDev\,[m_i]\,[p_a] + partX^2$;
                $stdDev\left[m_j\right][p_b] \leftarrow$
                $sdtDev\left[m_j\right][p_b] + partY^2$;
            end
        end
    end
    foreach $\left\{m_i, m_j\right\} \in \mathcal{M} \times \mathcal{M}, i\ \neq j$ do
        foreach $\langle phr_k, p\ hr_l\rangle \in PW^i \times PW^j, p\ hr_k \in$
        $PW^i \wedge phr_l \in PW^j$ do
            $r\,[m_i]\left[m_j\right][phr_k]\,[phr_l] \leftarrow$
            $cov\,[m_i]\left[m_j\right][phr_k]\,[phr_l]\,/\,(stdDev\,[m_i]$
            $[phr_k]\ \cdot stdDev\left[m_j\right][phr_l]\big)^{1/2}$
        end
    end
    // Phrase extraction
    $MMP \leftarrow \langle\rangle$;
    // Inittialize multimodal phrases list
    for $m_i \in \mathcal{M}$ do
        for $phr_k \in PW^i$ do
            $mmPhr_i \leftarrow \left\{phr_j\right\}$;
            for $m_j \in \mathcal{M} - m_i$ do
                $phr_l \leftarrow argMax\left(r\,[m_i]\left[m_j\right][phr_k]\right)$;
                //Get max correlated phrase
                $mmPhr_i \leftarrow mmPhr_i + \{phr_l\}$;
            end
            $add\,(MMP, m\ mPhr_i)$
        end
    end
end
```

V. CLUSTERING AND ROUTING ALGORITHM

In this study, we present an approach for the processing of scalar and multimedia sensor data in WMS nodes and also in the base station. The object and activities extracted by the

TABLE 3. Test configurations.

Configuration	Description
Visual, Audio, Text	Each single modality
BestSM	Best single modality
Comb_S	Combination of all modalities
Phr_V, Phr_A, Phr_T	Intramodal analysis of each single modality
Comb_I	Combination of all intramodal analyses outputs
MMPhr	Intermodal analysis
Comb_SI	Combination of modalities and intramodal analyses outputs
Comb_All	Combination of all inputs

TABLE 4. Fusion gains for different cases.

	BestSM	Comb_S	Comb_SI
Comb_S	8.92%		
Comb_SI	13.12%	3.31%	
Comb_All	17.12%	6.89%	3.24%

WMS nodes is transferred to the sink to reduce the amount of data to be transferred and the power consumption. However, information such as computed low-level features, silhouettes or foreground object images can also be transferred at the request of the system users, when needed. Therefore, an efficient data collection and transfer protocol is important for such a framework [9], [19]. As such, an efficient cluster-based routing algorithm, called Two-Tier Distributed Fuzzy Logic Based Protocol (TTDFP) for efficient aggregation of data in multi-hop wireless sensor networks, is developed in the proposed framework.

TTDFP is a two-level, fuzzy logic protocol that improves the efficiency of data collection in multi-hop wireless sensor networks. In a cluster network, the member nodes transmit the resulting or obtained data to the base station via cluster heads (CH). In multi-hop wireless networks, transmission from one CH passes through the other CHs. Due to the adoption of a multi-hop topology, hot spots and/or energy-hole problems may occur. In order to avoid these problems and extend the lifespan of these networks, TTDFP, as a new protocol, has been proposed and developed. It is a distributed and scalable protocol that works effectively for sensor network applications. In addition, an optimization framework is used to adjust the value of the parameters used during the fuzzy clustering phase in order to optimize the performance of a particular wireless sensor network, in combination with the two-tier approach.

In the first tier (distributed fuzzy clustering phase), the TTDFP decides final CHs through an energy-based competition of tentative leaders, which are primarily chosen using a probabilistic model. The TTDFP protocol is a fully distributed and optimized competitive protocol that takes into account the lifetime requirements of WSNs. The TTDFP does not require the inclusion of a central decision point during its phases. This distributed operation architecture protects the protocol from single point of failure situations. The fuzzy clustering phase manages the uncertainty in the clustering phenomenon more efficiently than its crisp counterparts and other fuzzy counterparts. This tier is designed taking into account three crucial elements. The first is energy efficiency, the second is the distributed operating requirements that provide scalability and the last is the optimized configuration of execution.

It can be noted that many studies in the literature are primarily concerned with energy-efficient clustering and none of them take into account the efficiency of the clustering and routing phases together. In TTDFP, the optimization framework described in [42] is used to tune the two first-tier parameters, which are the radius and maximum competition threshold, rather than using an empirical approach to find the right mixture of these parameters. The optimization framework uses the Simulated Annealing algorithm to tune the aforementioned parameter pair in order to optimize the WSN performance metrics. In addition, the fuzziness in the second tier (fuzzy routing phase) is a novelty that also improves the performance of routing compared to its crisp counterparts. Two essential factors are taken into account when designing the second tier of fuzzy routing. The first factor is energy efficiency in the tier, which is essential for the overall efficiency of the TTDFP, and the second is the simplicity of the computational aspect. Like the previous tier, this tier also uses a distributed approach because the sink is not included in the routing route selection procedure.

The proposed protocol was compared with selected state-of-the-art algorithms. In the experiments, we use three metrics for the evaluation of the energy efficiency of the described protocols. These metrics are First Node Dies (FND), Half of the Nodes Die (HND), and Total Remaining Energy (TRE).

In order to evaluate the performance of the TTDFP, we can say that it has been applied to two different scenarios. In Scenario 1, the performance of the fuzzy clustering tier is evaluated independently regardless of the fuzzy routing level. In this way, it is possible to deduce or highlight the gain resulting solely from the use of the fuzzy clustering approach. Table 5 shows the average of the results obtained for the fuzzy clustering tier, including different cases in Scenario 1, which is designed to measure the performance gain of the new fuzzy clustering algorithm.

In Scenario 2, the performance of the fuzzy routing tier is tested to be evaluated independently. For this reason, LEACH is not included in the scenario because it is specifically designed for direct transmission. Table 6 shows the average results of different test cases obtained from the fuzzy routing tier.

Since there may be heterogeneous nodes whose initial deployed energy differs in the network, the performance of the proposed approach is also tested in this type of configuration. The reduction of the TRE of each protocol in a heterogeneous network scenario is averaged and presented

TABLE 5. Average results of fuzzy clustering (Tier-I).

Algorithm	*FND*	*HND*	*TRE*
	(# of rounds)	*(# of rounds)*	*(joule)*
LEACH	8	45	76.45
CHEF	6	76	200.41
EEUC	7	83	246.12
MOFCA-Optimized	9	96	269.45
TTDFP	10	104	300.87

TABLE 6. Average results of fuzzy routing (Tier-II).

Algorithm	*FND*	*HND*	*TRE*
	(# of rounds)	*(# of rounds)*	*(joule)*
CHEF	10	139	159.05
EEUC	12	151	220.07
MOFCA-Original	16	155	232.99
MOFCA-Optimized	16	167	242.58
TTDFP	18	179	258.47

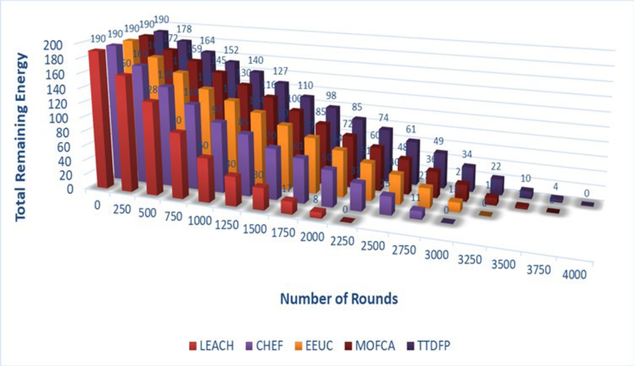

FIGURE 9. Average performance results for compared algorithms in heterogeneous networks.

in Fig. 9. In this heterogeneous environment, our proposed TTDFP approach maintains its efficient operation and consumes TRE much more slowly than the compared algorithms.

In the overall, the test results show that the TTDFP performs better than other protocols based on the metrics used to compare the energy efficiency of the protocols and the lifetime of the network. Much more detailed and comprehensive information on the proposed approaches to fuzzy routing and clustering, and their performance results are provided in [18].

VI. CONCLUSION

In this paper, we studied: 1) developing a WMS node to detect and recognize objects using machine learning; 2) developing methods that increase the accuracy rate while reducing the amount of information to be transferred to the base station; 3) developing a cluster-based routing algorithm that consumes less power than currently used approaches. The conclusions that can be drawn from the study can be summarized as follows:

- As a result of technological developments, new platforms capable of processing image and audio data in the nodes have become available on the market. With these platforms, in addition to scalar sensor data, the multimedia data can also be processed in the sensor nodes to a certain extent.
- Due to the large size of the multimedia data, their processing in the node is a crucial requirement. As indicated in the corresponding works, different methods related to the subject have been proposed. However, as suggested in this paper, the object extraction process in the node significantly reduces the amount of data to be transmitted to the base station and thus contributes significantly to extending the lifetime of the WMSNs.
- Visual data for object extraction on multiple sensor nodes is the main data. However, fusing image data with audio data greatly increases the success rate of object detection.
- The contribution of data obtained using scalar sensors to the retrieval/classification of objects seems rather limited. For this reason, it is considered that such sensors should be used to activate multimedia sensors, which normally wait in sleep mode to save energy.
- Although intra-sensor node image processing is possible on existing platforms on the market, we have encountered some limitations. For example, if the images are taken in HD quality and processed in real time or if the number of processed images in one second increases, the power of their processors is not sufficient and some images are lost.
- The combination of different information search modalities (fusion) provides more accurate results than a single modality. For example, the fusion of three modes (image, voice and text) results in an improvement of 8.92% compared to the best simple modality. Here we must not forget that mode selection and weight determination are critical issues. Otherwise, a bad choice can lead to worse results than those of the best single mode.
- The ability to process multimedia data, including wireless video and audio data provided in wireless sensor nodes, has increased the importance of energy efficient clustering and routing algorithms.
- It has been found that when the connectivity is used in place of the density parameter for the distributed operating architecture, the result is unchanged in the worst case, but provides resistance to single point-of-failure situations in the clusters.
- In clustering, among the parameters tested and implemented, the highest value-added parameter is the remaining energy. However, it is corroborated that the distance parameter, which is not as efficient as the remaining energy parameter, also contributes significantly to the lifetime of the network. Although the connectivity parameter has an effect on the improvement

of the network lifetime, it was found that this effect was not as important as the other two parameters (the remaining energy and the distance to the station) and acted as a fine-tuning parameter.

- The fuzzy approach used in the two steps of TTDFP (Tiers I and II) proved superior to the existing methods. This superiority stems from the gradual membership rather than crisp membership to a value and this situation can be explained by the concept of relaxation in the literature.

This study introduces a unique perspective on the clustering and routing algorithms of wireless multimedia sensor networks that consume less energy and on methods that increase the accuracy rate while reducing the amount of information routed to the base station. However, since this is a very comprehensive subject, there are more problems to be studied in this area. Based on the experience and knowledge gained in this study, the possible research topics that should be researched in this area are as follows:

- In addition to the scalar sensors used in this study, the contributions of other scalar sensors to the object detection and recognition process can be examined.
- Different features and classification algorithms can be tried to improve the performance of object retrieval from visual and audio data.
- In this study, different data were fused at three different layers to increase the success rate of object extraction. More efficient fusion methods can be studied for each layer to improve system performance.
- Third-level fusion studies were conducted on sensor information collected in the base station. It is considered that data collected in the base station will gain big data quality in case of continuous flow from a large number of sensors. As a result, big data analytics can be applied to these sensor data to extract the knowledge to be used for optimization of the WMSN system and to make predictions.

ACKNOWLEDGMENT

The authors would like to thank the METU Multimedia Database Research Group members, Enver Ever, Ahmet Cosar, Mustafa Sert, Muhsin Civelek, Saeid Sattari, Shakiba Rahimi, Gizem Karagoz, Ezgi Selver, Ali Nail Ozdin, and Cihan Kucukkececi for their very valuable supports and contributions for this study.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[2] P. Rawat, K. D. Singh, H. Chaouchi, and J. M. Bonnin, "Wireless sensor networks: A survey on recent developments and potential synergies," *J. Supercomput.*, vol. 68, no. 1, pp. 1–48, 2014.

[3] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Comput. Netw.*, vol. 51, no. 4, pp. 921–960, Mar. 2007.

[4] I. T. Almalkawi, M. G. Zapata, J. N. Al-Karaki, and J. M. Pozo, "Wireless multimedia sensor networks: Current trends and future directions," *Sensors*, vol. 10, no. 7, pp. 6662–6717, Jul. 2010.

[5] N.-S. Vo, T. Q. Duong, M. Guizani, and A. Kortun, "5G optimized caching and downlink resource sharing for smart cities," *IEEE Access*, vol. 6, pp. 31457–31468, May 2018.

[6] F. Al-Turjman, A. Radwan, S. Mumtaz, and J. Rodriguez, "Mobile traffic modelling for wireless multimedia sensor networks in IoT," *Comput. Commun.*, vol. 112, pp. 109–115, Nov. 2018.

[7] T. Ojha, S. Misra, and N. S. Raghuwanshi, "Wireless sensor networks for agriculture: The state-of-the-art in practice and future challenges," *Comput. Electron. Agricult.*, vol. 118, pp. 66–84, Oct. 2015.

[8] A. B. Noel, A. Abdaoui, T. Elfouly, M. H. Ahmed, A. Badawy, and M. S. Shehata, "Structural health monitoring using wireless sensor networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1403–1423, 3rd Quart., 2017.

[9] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad Hoc Netw.*, vol. 7, no. 3, pp. 537–568, May 2009.

[10] V. Raghunathan, C. Schurgers, S. Park, and M. B. Srivastava, "Energy-aware wireless microsensor networks," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 40–50, Mar. 2002.

[11] M. Chen, S. Gonzalez, H. Cao, Y. Zhang, and S. T. Vuong, "Enabling low bit-rate and reliable video surveillance over practical wireless sensor network," *J. Supercomput.*, vol. 65, no. 1, pp. 287–300, 2013.

[12] M. Magno, F. Tombari, D. Brunelli, L. Di Stefano, and L. Benini, "Multimodal video analysis on self-powered resource-limited wireless smart camera," *IEEE J. Emerging Sel. Topics Circuits Syst.*, vol. 3, no. 2, pp. 223–235, Jun. 2013.

[13] A. N. Sukumaran, R. Sankararajan, and M. Swaminathan, "Compressed sensing based foreground detection vector for object detection in wireless visual sensor networks," *AEU-Int. J. Electron. Commun.*, vol. 72, pp. 216–224, Feb. 2017.

[14] R. Bhatt and R. Datta, "A two-tier strategy for priority based critical event surveillance with wireless multimedia sensors," *Wireless Netw.*, vol. 22, no. 1, pp. 267–284, 2016.

[15] V. K. Singh, G. Sharma, and M. Kumar, "Compressed sensing based acoustic event detection in protected area networks with wireless multimedia sensors," *Multimedia Tools Appl.*, vol. 76, no. 18, pp. 18531–18555, 2017.

[16] M. Civelek and A. Yazici, "Automated moving object classification in wireless multimedia sensor networks," *IEEE Sensors J.*, vol. 17, no. 4, pp. 1116–1131, Feb. 2017.

[17] M. Koyuncu, A. Yazici, M. Civelek, A. Cosar, and M. Sert, "Visual and auditory data fusion for energy-efficient and improved object recognition in wireless multimedia sensor networks," *IEEE Sensors J.*, vol. 19, no. 5, pp. 1839–1849, Dec. 2019.

[18] S. A. Sert, A. Alchihabi, and A. Yazici, "A two-tier distributed fuzzy logic based protocol for efficient data aggregation in multihop wireless sensor networks," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 6, pp. 3615–3629, Dec. 2018.

[19] T. Rault, A. Bouabdallah, and Y. Challal, "Energy efficiency in wireless sensor networks: A top-down survey," *Comput. Netw.*, vol. 67, pp. 104–122, Jul. 2014.

[20] E. Harjula, T. Mekonnen, M. Komu, P. Porambage, T. Kauppinen, J. Kjällman, and M. Ylianttila, "Energy efficiency in wireless multimedia sensor networking: Architecture, management and security," in *Greening Video Distribution Networks: Energy-Efficient Internet Video Delivery* Cham, Switzerland: Springer, 2018, pp. 133–157.

[21] D. M. Pham and S. M. Aziz, "Object extraction scheme and protocol for energy efficient image communication over wireless sensor networks," *Comput. Netw.*, vol. 57, no. 15, pp. 2949–2960, 2013.

[22] Y. A. U. Rehman, M. Tariq, and T. Sato, "A novel energy efficient object detection and image transmission approach for wireless multimedia sensor networks," *IEEE Sensors J.*, vol. 16, no. 15, pp. 5942–5949, Aug. 2016.

[23] Q. Lu, W. Luo, J. Wang, and B. Chen, "Low-complexity and energy efficient image compression scheme for wireless sensor networks," *Comput. Netw.*, vol. 52, no. 13, pp. 2594–2603, 2008.

[24] T. Sheltami, M. Musaddiq, and E. Shakshuki, "Data compression techniques in wireless sensor networks," *Future Generat. Comput. Syst.*, vol. 64, pp. 151–162, Nov. 2016.

[25] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 165–179, Feb. 2015.

[26] M. Chitnis, Y. Liang, J. Y. Zheng, P. Pagano, and G. Lipar, "Wireless line sensor network for distributed visual surveillance," in *Proc. 6th ACM Symp. Perform. Eval. Wireless Ad Hoc, Sensor, Ubiquitous Netw.*, 2009, pp. 71–78.

[27] M. S. Alhilal, A. Soudani, and A. Ai-Dhelaan, "Low power scheme for image based object identification in wireless multimedia sensor networks," in *Proc. Int. Conf. Multimedia Comput. Syst. (ICMCS)*, Apr. 2014, pp. 927–932.

[28] H. Oztarak, T. Yilmaz, K. Akkaya, and A. Yazici, "Efficient and accurate object classification in wireless multimedia sensor networks," in *Proc. 21st Int. Conf. Comput. Commun. Netw.*, Aug. 2012, pp. 1–7.

[29] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[30] D. Ciuonzo, A. Aubry, and V. Carotenuto, "Rician MIMO channel- and jamming-aware decision fusion," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 3866–3880, Aug. 2017.

[31] B. Kailkhura, S. Brahma, and P. K. Varshney, "Data falsification attacks on consensus-based detection systems," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 1, pp. 145–158, Mar. 2017.

[32] P. Salvo Rossi, D. Ciuonzo, T. Ekman, and H. Dong, "Energy detection for MIMO decision fusion in underwater sensor networks," *IEEE Sensors J.*, vol. 15, no. 3, pp. 1630–1640, Mar. 2015.

[33] T. Wimalajeewa and P. K. Varshneyet, "Compressive sensing-based detection with multimodal dependent data," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 627–640, Feb. 2018.

[34] D. Ciuonzo, A. Buonannoy, M. D'Ursoy, and F. A. N. Palmieri, "Distributed classification of multiple moving targets with binary wireless sensor networks," in *Proc. 14th IEEE Int. Conf. Inf. Fusion*, Jul. 2011, pp. 1–8.

[35] A. Sarkar and T. S. Murugan, "Routing protocols for wireless sensor networks: What the literature says?" *Alexandria Eng. J.*, vol. 55, pp. 3173–3183, Dec. 2016.

[36] S. Ehsan and B. Hamdaoui, "A survey on energy-efficient routing techniques with QoS assurances for wireless multimedia sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 265–278, May 2012.

[37] H. D. Al-Ariki and M. N. Swamy, "A survey and analysis of multipath routing protocols in wireless multimedia sensor networks," *Wireless Netw.*, vol. 23, no. 6, pp. 1823–1835, 2017.

[38] M. M. Afsar and M.-H. Tayarani-N, "Clustering in sensor networks: A literature survey," *J. Netw. Comput. Appl.*, vol. 46, pp. 198–226, Nov. 2014.

[39] S. P. Singh and S. C. Sharma, "Survey on cluster based routing protocols in wireless sensor networks," *Procedia Comput. Sci.*, vol. 45, pp. 687–695, Jan. 2015.

[40] P. X. Britto and S. Selvan, "A hybrid soft computing: SGP clustering methodology for enhancing network lifetime in wireless multimedia sensor networks," *Soft Comput.*, vol. 23, pp. 2597–2609, Apr. 2019.

[41] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Comput. Commun.*, vol. 30, nos. 14–15, pp. 2826–2841, Oct. 2007.

[42] A. Alchihabi, A. Dervis, E. Ever, and F. Al-Turjman, "A generic framework for optimizing performance metrics by tuning parameters of clustering protocols in WSNs," *Wireless Netw.*, vol. 25, no. 3, pp. 1031–1046, 2018.

[43] *OpenCV Open Source Computer Vision Library*. Accessed: Jul. 2018. [Online]. Available: http://opencv.org

[44] *Gloox*. Accessed: Nov. 2017. [Online]. Available: https://camaya.net/api/gloox-1.0

[45] Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 42–53, Jan. 2010.

[46] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 942–956, Jun. 2005.

[47] K. Turner and J. Ghosh, "Linear and order statistics combiners for pattern classification," 1999, *arXiv:9905012*. [Online]. Available: https://arxiv.org/abs/cs/9905012

[48] R. Yan and A. G. Hauptmann, "The combination limit in multimedia retrieval," in *Proc. 11th ACM Int. Conf. Multimedia*, New York, NY, USA, vol. 3, 2003, pp. 339–342.

[49] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Multi-classification approaches for classifying mobile app traffic," *J. Netw. Comput. Appl.*, vol. 103, no. 1, pp. 131–145, Feb. 2018.

[50] T. Yilmaz, A. Yazici, and M. Kitsuregawa, "RELIEF-MM: Effective modality weighting for multimedia information retrieval," *Multimedia Syst.*, vol. 20, no. 4, pp. 389–413, 2014.

ADNAN YAZICI received the Ph.D. degree in computer science from the Department of EECS, Tulane University, LA, USA, in 1991. He is currently a Full Professor with the Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, and the Chair of the Department of Computer Science, Nazarbayev University, Astana, Kazakhstan. He has published over 200 international technical papers and coauthored/edited three books entitled *Fuzzy Database Modeling* (Springer), *Fuzzy Logic in its 50th Year: New Developments, Directions and Challenges* (Springer), and *Uncertainty Approaches for Spatial Data Modeling and Processing: A Decision Support Perspective* (Springer). His current research interests include intelligent database systems, multimedia and video databases and information retrieval, wireless multimedia sensor networks, data science, and fuzzy data modeling. He is also a member of ACM, the IEEE Computational Intelligence Society, and the Fuzzy Systems Technical Committee. He was a recipient of the IBM Faculty Award, in 2011, and the Parlar Foundations Young Investigator Award, in 2001. He was the Conference Co-Chair of the 23rd IEEE International Conference on Data Engineering, in 2007, the 38th Very Large Data Bases, in 2012, and the 23rd IEEE International Conference on Fuzzy Systems, in 2015. He is also an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS.

MURAT KOYUNCU received the Ph.D. degree in computer engineering from Middle East Technical University, Ankara, Turkey, in 2001. He is currently an Associate Professor with the Department of Information Systems Engineering, Atilim University, Ankara. His research interests include fuzzy logic, object-oriented databases, knowledge-based systems, multimedia databases, computer networks, and multimedia wireless sensor networks.

SEYYIT ALPER SERT received the B.S.E.E. degree from the Turkish Military Academy, Ankara Turkey, in 2004, and the M.Sc. and Ph.D. degrees from the Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, in 2014 and 2018, respectively. He is currently a Staff Officer with the Strategic Development and Preparation Directorate, NATO Supreme Headquarters Allied Powers Europe (S.H.A.P.E.). His current research interests include wireless sensor networks, uncertainty modeling, algorithms, optimization, and computational intelligence.

TURGAY YILMAZ received the B.Sc. degree from the Department of Computer Engineering, Bilkent University, in 2004, and the M.Sc. and Ph.D. degrees from the Computer Engineering Department, Middle East Technical University, in 2008 and 2014, respectively. During his Ph.D. studies, he was a Visiting Research Associate with the Toyoda-Kitsuregawa Lab, The University of Tokyo. He worked in various roles, including a Software Developer, a Software Architect, and the Team Leader at EES, the Rakuten Institute of Technology, TURKSAT, HAVELSAN, and Powerhouse, respectively. He has 15 years of professional software development experience in backend, desktop, and web environments. In addition, he has over five years of experience for managing/leading agile software teams. He is currently the Software Team Leader with Elsevier, Amsterdam. He is an experienced Software Engineer/Software Architect.