Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **2**, March **2015**

The MOBISERV-AIIA Eating and Drinking multi-view database for vision-based assisted living

A. Iosifidis[†], E. Marami[†], A. Tefas[†], I. Pitas[†] and K. Lyroudia[*]

[†]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greeces
[*]School of Dentistry, Aristotle University of Thessaloniki, Thessaloniki, Greece
{tefas,pitas,lyroudia}@aiia.csd.auth.gr

Received June, 2014; revised November, 2014

Abstract. *Assisted living has a particular social importance in most developed societies, due to the increased life expectancy of the general population and the ensuing ageing problems. It has also importance for the provision of improved home care in cases of disabled persons or persons suffering from certain diseases that have high social impact. In this context, the development of computer vision systems capable to identify human eating and drinking activity can be really useful in order to prevent undernourishment/malnutrition and dehydration in a smart home environment targeting to extend independent living of older persons in the early stage of dementia. In this paper, we first describe the human centered interface specifications and implementations for such a system, which can be supported by ambient intelligence and robotic technologies. We, subsequently, describe a multi-view eating and drinking activity recognition database that has been created in order to facilitate research towards this direction. The database has been created by using four cameras in order to produce multi-view videos, each depicting one of twelve persons having a meal, resulting to a database size equal to* 59.68 *hours in total. Various types of meals have been recorded, i.e., breakfast, lunch and fast food. Moreover, the persons have different sizes, clothing and are of different sex. The database has been annotated in a frame base in terms of person ID and activity class. We hope that such a database will serve as a benchmark data set for computer vision researchers in order to devise methods targeting to this important application.*
Keywords: Nutrition assistance, Smart home environment, Activity recognition, Multi-view video database

1. **Introduction.** In the last years the need for developing efficient approaches for nutrition support and well-being has been increased. The objective of these methods is to help older persons that are in the last stages of their independent living period (e.g., people in the early stage of dementia [1]), trying to prolong their independent living period. To this end, human centered interfaces and methods following an anthropocentric approach should be developed, in order to monitor certain activities of older persons and their behavior in a smart home environment [2]. Such methods could recognize several Activities of Daily Living (ADLs) of the older persons [3, 4] and can be focused to eating and drinking activity recognition. The identification of eating and drinking abnormalities of persons of all ages is a necessary means for surviving and attaining good health at all times. Although, nutrition problems concern people of all ages, elderly persons, particularly those suffering from dementia, are more inclined to it [5]. The identification of eating/drinking activities can help in dietary menu planning, or replanning in case of deviations from the plan [6].

At first, studies on eating habits have been conducted mainly based on interviews [7]. On the other hand, attempts to monitor eating and drinking behavior of patients or elderly people have also been performed. A category of methods aiming at eating and drinking activity detection elaborates data acquired by ambient [8, 9] or body-worn (more invasive) sensors [10, 11]. The methods belonging to this category require various body-worn sensors in order to gather vital information for a person's activities such as arm movements, chewing or swallowing that reveal food or liquid intake activity. Ambient sensors mainly provide information concerning persons' location (kitchen, dining room etc.) and could be used in combination with other, more precise, techniques. A combination of sensors (ambient and wearable) has also been proposed to recognize ADLs of older persons [12, 13, 14].

The major disadvantage of the sensor-based ADLs recognition approach is that it sets restrictions on the operation scenario and requires person cooperation. That is, the person under consideration should agree to have physical contact with sensors. This generates several issues that should be taken into account [15]. It is evident that older people are not familiarized with such equipment. This fact results to non-cooperation due to fear. Even in the cases where an older person agrees to cooperate, he/she may forget to wear or make poor use of it. An alternative, less invasive, approach exploits visual information captured by cameras. However, in order to take into account the resistance to video observation by older persons that have privacy concerns, Privacy Preserving Technologies should be exploited in a certain extend. A system that monitors daily activities of elderly people at home is described in [16]. Considering that nutrition is essential to keep health, a system was installed at home in order to monitor daily behavior of the persons in the kitchen and the dining room. A concept of home health monitoring is also presented in [17]. Health parameters are automatically monitored at home without disturbing daily activities. Furthermore, a pilot project monitoring functional health status of the elderly at home by continuously recording parameters of daily living sensitive to changes in health is reported in [18]. In order to identify food consumption, a method that automatically detects chewing events in surveillance videos of a person is presented in [19]. Another work that deals with human events detection in a video sequence and can be extended in eating and drinking activity recognition is presented in [20]. A system exploiting information coming from cameras and microphones designed to be "tape on and forget" devices, is described in [21] for ADLs recognition. In the field of general activity recognition, there are studies that although do not directly address the eating and drinking activity recognition problem, they handle other related tasks, like meal preparation, interactions with cups, cutlery (e.g. 'grasp a cup', 'touch a fork'), recognition of objects and actions, etc [22, 23].

Another approach to eating/drinking activity recognition is based on a general human action recognition approach [24]. Human face and hands are detected [25, 26] and tracke-dover time [27, 28], while handling occlusions [29]. Then, the detected and tracked facial regions and possibly facial landmarks [30] and are used to perform eating and drinking activity recognition, thus achieving very good performance [24]. Besides its use in the home of elderly people, eating/drinking activity recognition plays an important role in other domains as well, notably in mastication and masseter myalgia studies [31, 32], as well as in the case of head and oral cavity abnormalities, e.g. clefts [33] or changes due to aging [34] that affect mastication, in correlation with other disorders [35].

In this paper, we describe the human centered interface specifications, research and implementations for a computer vision-based nutrition assistance system. We consider as target group of our study older persons that are in the early stage of dementia and suffer by mild memory loss. Two serious problems that the patients with early dementia face are underfeeding and dehydration. This is due to several reasons such as nerve deterioration,

loss of sense of smell, apraxia (loss of the ability or will to execute or carry out learned purposeful movements), etc. Specifically, we investigate the scenario of assisted living in a smart home environment providing several "intelligent functionalities", like computer vision-based automatic eating and drinking activity detection, automatic communication with relatives or physicians in case of abnormal nutrition activity, etc. Subsequently, we describe an eating and drinking activity recognition database that has been created in the context of the FP7 project MOBISERV[1] and can be used in order to facilitate research towards this direction. The database is publicly available for research purposes[2].

2. **System Description.** Our main objective is to develop and use up-to-date technology to support independent living of older persons as long as possible in their own homes. A system that automatically detects and recognizes eating and drinking activity using video processing techniques would greatly contribute to prolonging independent living of older persons in a non-invasive way. Such a system can be focused to monitor specific regions of the smart home and for pre-specified time intervals, corresponding to meal periods, in order to respect the privacy of the older persons. We consider monitoring of the dining table used by the older person for the daily meals. The nutrition support system should have the following functionalities/properties:

- Person appearance detection sitting on a chair in front of the dining table, in order to start operation.
- Activity detection and discrimination between eating/drinking activity from other activities (e.g., reading).
- Analysis of the eating/drinking activity during a session (pre-specified time interval).
- Ability to interact with the older person in case of abnormal nutrition activity (e.g., if he/she misses a meal).
- Ability to communicate with a relative/physician in case of emergency (e.g. report abnormal nutrition activity).
- Real-time operation.

In order to obtain the needed visual information, one or multiple cameras should be placed in such a way that they can capture the person's upper body during a meal. In our current implementation, we consider one camera placed at a distance of $1.5 - 2$ meters in front of the person and at a height of $20 - 40$cm above the dining table, so that the entire person's upper body is visible during a meal, as illustrated in Figure 1.

2.1. **Human appearance detection.** In order to detect the appearance of a person sitting on a chair in front of the dining table human body [36, 37], or face detection [38, 25, 26] techniques can be exploited. In the first case, the human body shape (specifically the shape of the person's head and shoulders), usually represented by the shape of its contour using shape descriptors, like the Histogram of Oriented Gradient (HOG) [39], is searched in each input video frame at multiple locations and scales in order to provide possible video frame locations belonging to the human body. In the second case, face detection techniques are applied at multiple video frame locations and scales in order to provide possible human face video frame locations. Haar-like features combined with a cascade of classifiers have been widely employed to this end [38]. Exploiting the fact that in our application scenario the person's face should be always visible during a meal, we have chosen face detection for person appearance detection. In order to discard background locations having shape similar to the human face (false positives), we also exploit the color properties of the human skin [40]. In order to fasten the operation of the face

[1] http://www.mobiserv.info/

[2] http://www.aiia.csd.auth.gr/MOBISERV-AIIA/index.html

FIGURE 1. *Example video frame depicting a person having a meal.*

localization process, we perform face detection periodically, e.g., every 15 frames, while we perform color-based face tracking between consecutive video frames. In addition, in order to further fasten the operation of the face localization process, we restrict face detection at a bounding box having a size twice the size of the person's face, which is centered to the previously determined facial region, as illustrated in Figure 2.

2.2. **Eating/Drinking activity representation.** After detecting the person's face for the first time, monitoring is started. The objective, at this stage, is to recognize the person's actions, i.e., his/her elementary movement patterns, in order to discriminate eating/drinking activity from other activities, e.g., reading at the dining table. Actions are usually described by using either features based on motion information and optical flow [41, 42], or features devised mainly for action representation [43, 23]. Although the use of such features leads to satisfactory action recognition results, their computation is expensive. Thus, in order to achieve (near-)real time operation, the use of simpler action representations is required. Neurobiological studies [44] have concluded that the human brain can perceive actions by observing only the human body configurations (poses or postures) during action execution. Thus, actions can be described as sequences of consecutive human body poses, in terms of human body silhouettes [45, 46]. In our case, the adopted human body poses have been chosen to denote the person's head and hands [24], as illustrated in Figure 3. In order to take into account the resistance to video observation by older persons that have privacy concerns, we keep only the the binary human body pose images for further processing, leading to a Privacy Preserving action representation.

Human body pose images are created by applying skin color segmentation techniques to the color video frames [47]. Experimental studies have concluded that the skin color of Caucasian people has several properties [40]. Specifically, it has been shown that the color values of the human skin lie in a known region of the HSV color-space, i.e., $0 < H < 0.1$, $0.23 < S < 0.68$ and $V > 0.27$, where:

$$h = cos^{-1} \frac{\frac{1}{2}\left((R-G)+(R-B)\right)}{\sqrt{(R-G)^2+(R-G)(G-B)}}, \tag{1}$$

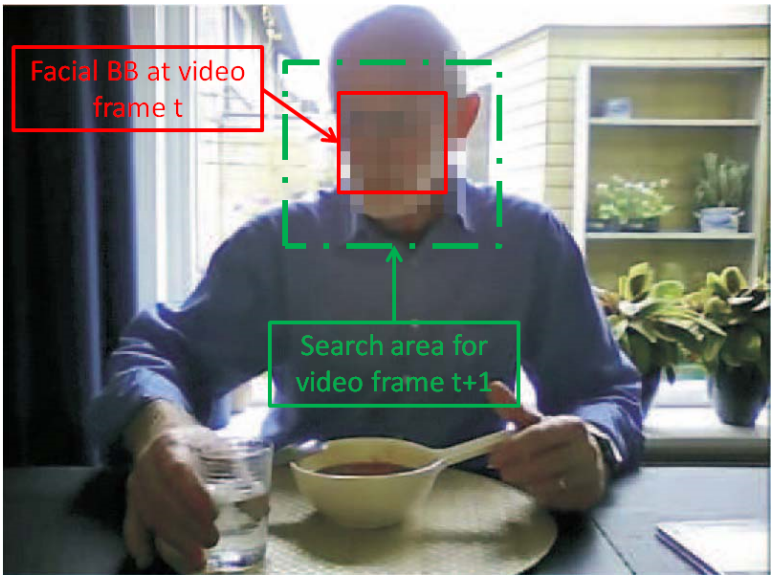

FIGURE 2. *Face detection on a restricted video frame area for faster operation.*

FIGURE 3. *Consecutive human body poses corresponding to an eating sequence.*

$$H = \begin{cases} h, & B \leq G \\ 2\pi - h, & B > G, \end{cases} \tag{2}$$

$$S = \frac{max(R,G,B) - min(R,G,B)}{max(R,G,B)}, \tag{3}$$

$$V = max(R,G,B). \tag{4}$$

R, G, B in (1) - (4) are the three channels of the color video frame. Despite the fact that the color of the human skin can be described by the above mentioned color boundaries, a skin color segmentation method should be able to cope with changes in lighting conditions of the scene. Specifically, in our case where eating/drinking activity recognition should be performed in time periods corresponding to several meals, i.e., breakfast, lunch and dinner, the different lighting conditions complicate the skin segmentation task. In order to address this issue, we calculate person-specific thresholds for skin segmentation by exploiting the information provided by the, previously detected, person's facial image. That is, at each video frame, the histogram of the person's facial image in the HSV color-space is calculated. Expecting that most of the pixels in the facial image are skin colored pixels, the person's skin color is approximated. To this end, we estimate the peak of the histogram calculated for each color channel. We use the width value of the peak to adjust properly the predefined thresholds, as illustrated in Figure 4, and use the modified thresholds in order to decide whether a pixel belongs to skin location or not. We check all the pixels in the video frame so as to obtain a binary bitmap denoting the skin-like video frame locations in white and the remaining video frame locations in black.

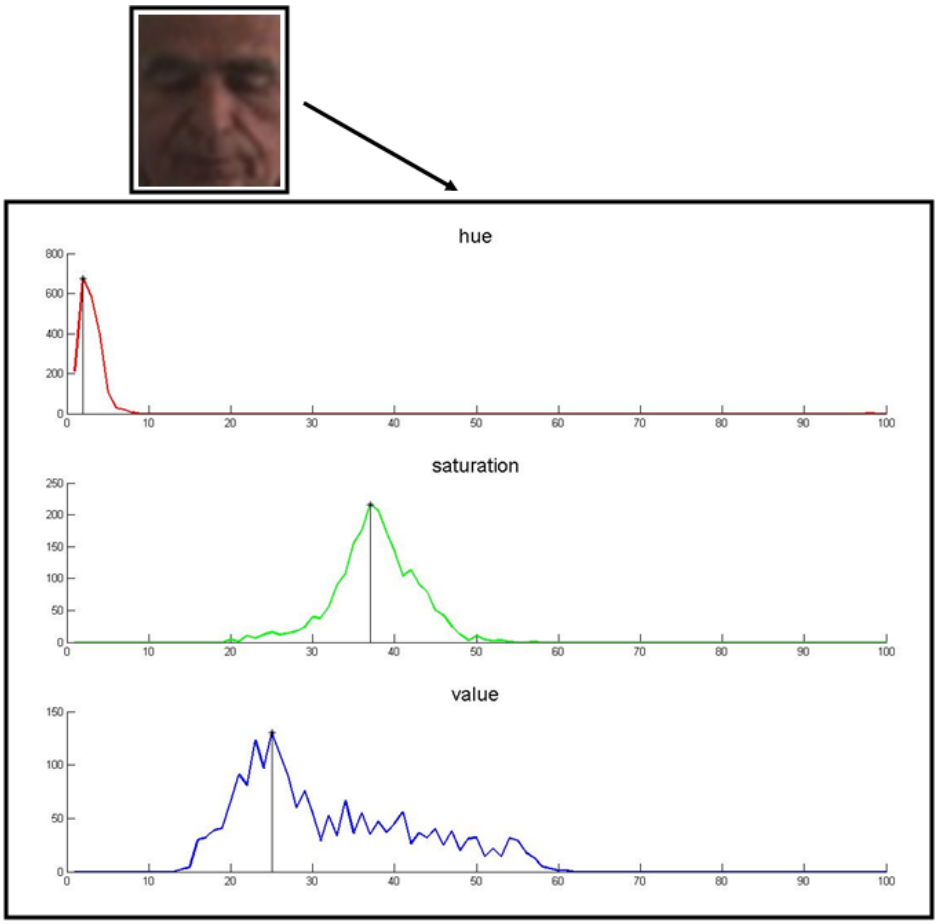

FIGURE 4. *Calculation of person-specific thresholds for skin color segmentation.*

In order to isolate individual elements, join disparate elements and remove noise in the obtained binary bitmap, we apply morphological operations (dilation and erosion) [48]. To discard background locations belonging to skin-colored objects we exploit the position of the, previously detected, human face and anthropometric ratios in order to determine a trapezoid enclosing the video frame locations that are probable to be occupied by the person's head and hands and keep the Regions Of Interest (ROIs) inside this trapezoid in order to form the so-called binary body image. To this end, we have used a trapezoid horizontally centered to the person's facial ROI and having bases equal to two and five times the width of the person's facial ROI. Finally, the resulted binary body image is centered to the person's ROIs center of mass, cropped to the ROIs region and resized to a fixed size ($H \times W$ pixels) image, the so-called posture image. We chose the size of binary posture images to be equal to 32×32 pixels, which has been found experimentally to be a good compromise between computational cost and activity recognition accuracy.

2.3. **Eating/Drinking activity recognition.** In order to perform eating/drinking activity recognition we employ a recently proposed action recognition scheme involving Fuzzy Vector Quantization (FVQ) and Artificial Neural Network (ANN)-based action video classification [49]. Let us assume that an action video, i.e., a video depicting an elementary movement pattern, is formed by N_t video frames. These video frames are preprocessed by applying the above described process in order to produce N_t binary posture images. Posture images are represented as matrices, which are vectorized column-wise in

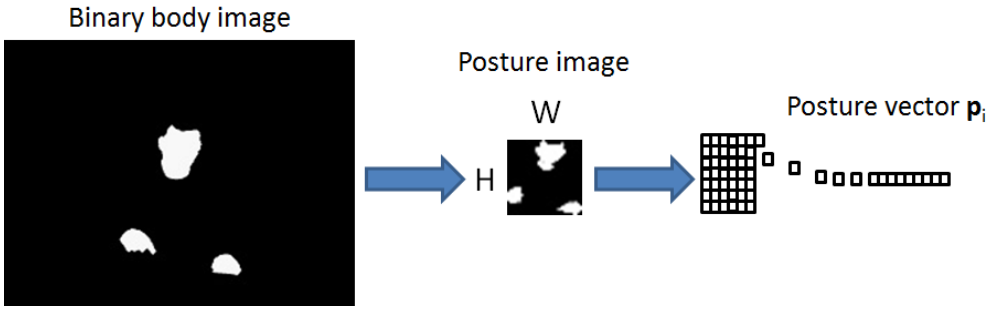

FIGURE 5. *Posture vectors creation.*

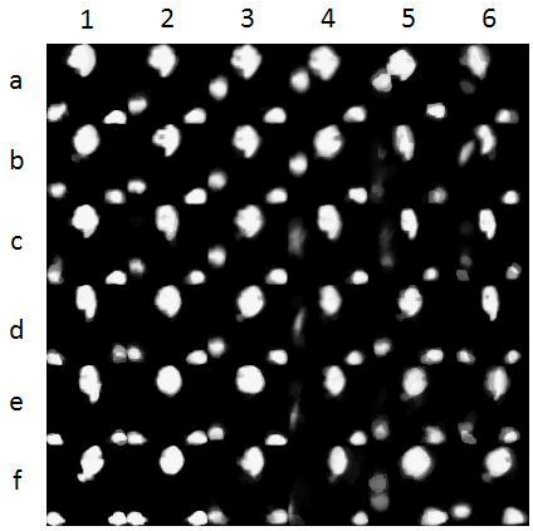

FIGURE 6. 36 *posture prototypes of actions 'eat', 'drink' and 'rest'.*

order to produce the so-called posture vectors $\mathbf{p}_i, \in \mathbb{R}^N$, $i = 1, \ldots, N_t$, where $N = HW$, as illustrated in Figure 5. By using the previously noted posture image resolution, i.e., 32×32 pixels, the dimensionality of the posture vectors $\mathbf{p}_i$ is equal to $N = 1024$.

After obtaining the posture vectors $\mathbf{p}_i$, the action video can be represented in an alternative way by exploiting the fuzzy similarity between $\mathbf{p}_i$ and the so-called dynemes $\mathbf{v}_d \in \mathbb{R}^{HW}$, $d = 1, \ldots, D$. Dynemes correspond to human body posture prototypes and are determined in a training phase by clustering the posture vectors of training action videos. Clustering can be performed by applying the D-Means algorithm or by training a self-organizing neural network (Self-Organizing Map - SOM). The number of dynemes D is specified by the problem at hand. In the case of eating/drinking activity recognition we have experimentally found that a small number of D is adequate for satisfactory performance, i.e., $D \leq 100$. A set of 36 dynemes are illustrated in Figure 6. Membership vectors $\mathbf{u}_i \in \mathbb{R}^D$ are calculated in order to encode the fuzzy similarity between posture vectors $\mathbf{p}_i$ and all the dynemes $\mathbf{v}_d$ according to a fuzzification parameter $m > 1.0$ by:

$$u_{id} = \| \mathbf{p}_i - \mathbf{v}_d \|_2^{-\frac{2}{m-1}}, \tag{5}$$

$$\mathbf{u}_i = [u_{i1} \ \ldots \ u_{iD}]^T. \tag{6}$$

TABLE 1. Indicative monitoring sessions and check points.

Meal	Session	Check Points
Breakfast	8:00 - 10:00	9:00, 9:30, 9:50
Lunch	12:00 - 14:00	13:00, 13:30, 13:50
Dinner	19:00 - 21:00	20:00, 20:30, 20:50

The resulted membership vectors $\mathbf{u}_i$ are employed in order to determine the so-called action vector $\mathbf{s} \in \mathbb{R}^D$ representing the action video:

$$\mathbf{s} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_1}. \tag{7}$$

Action vector $\mathbf{s}$ is scaled in order to have unit l_2 norm and normalized by exploiting the mean and standard deviation of the training action vectors. Finally, action vector $\mathbf{s}$ is introduced to a Single-hidden Layer Feedforward Neural (SLFN) network, which has been trained either by applying the Backpropagation algorithm [50], or the recently proposed Minimum Class Variance Extreme Learning Machine (MCVELM) algorithm [51] on the action vectors representing the training action videos, and classified to the action class corresponding to the network's highest output $\mathbf{o}$, i.e., $l = \underset{i}{argmax}\ o_i$. For a classification problem formed by N_A action classes, the adopted network consists of D input, L hidden and N_A output neurons, that is the network output vector consists of N_A neurons $\mathbf{o} = [o_1 \ \ldots \ o_{N_A}]^T$. The number of the network hidden layer neurons L should be appropriately set. We have experimentally found that a value of $L = 1000$ usually operates well in many classification problems.

2.4. **Online Eating/Drinking activity recognition.** The above described eating/drinking activity recognition method, as most activity recognition methods proposed in the literature, operate on pre-segmented action videos. However, in our case we would like to perform online eating/drinking activity recognition on the video frames captured by the camera placed at the dining table. To this end, we have employed the 'sliding window' technique in order to automatically segment action videos in an online manner and proceed with action video classification, as illustrated in Figure 7. According to this, the N_t most recent binary body images are stored in a buffer and are employed in order to form an action video that will be used in order to characterize the current video frame. We have experimentally found that a value of $N_t = 15$ video frames provides satisfactory action classification performance.

Regarding the computational complexity of the above-described online activity recognition scheme, it has been tested on a PC with a 2.4GHz Core 2 Quad processor and 4Gb RAM running Windows XP. By using a video frame resolution equal to 480×640 pixels, it operates in near real-time, i.e., 10-15 frames per second. We have experimentally found that such an operation speed is adequate for satisfactory action classification performance.

2.5. **Analysis of eating/drinking activity and decision making.** As it was previously mentioned, we consider a system monitoring the dining table used by the older person for the daily meals during pre-specified time intervals (sessions) corresponding to different meals, an example given in Table 1. Clearly, these sessions can be adjusted to the older person's schedule and habits.

When a monitoring session starts, a Control Unit (CU) placed in a control room inside the smart home initiates the nutrition support system operation. If the older person reaches the dining table and sits in order to have a meal, he/she is detected and monitoring

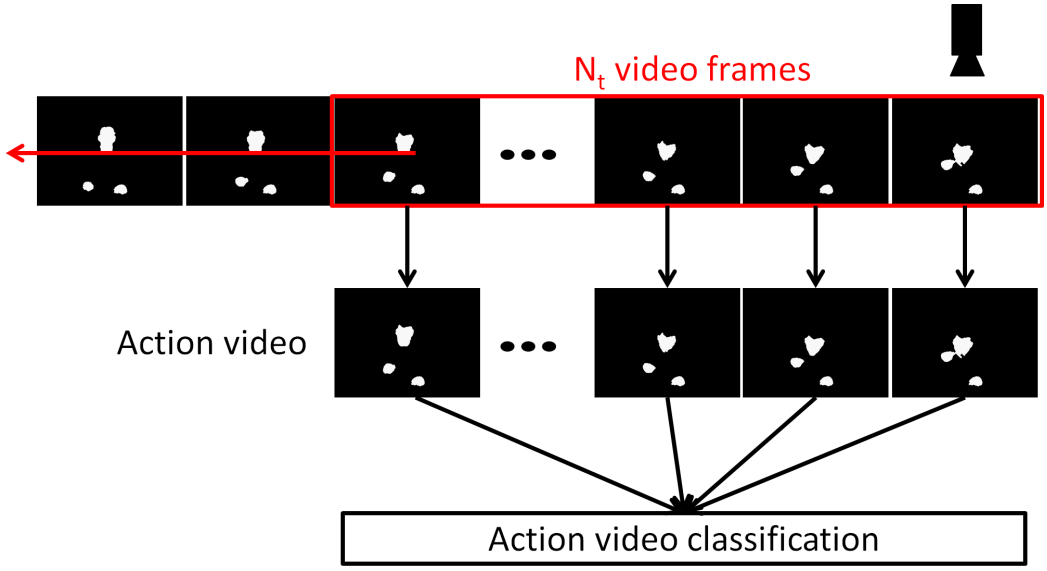

FIGURE 7. *Sliding window technique for online activity recognition.*

is started. By following the above described processing steps, the older person's activities are detected and recognized in a per video frame basis. The CU periodically analyses the detected activity by accumulating the corresponding per-frame activity recognition results. In the cases where the older person did not sit to the dining table, or he/she sits and has not eaten yet, a reminder message is triggered in order to remind him/her to take a meal. The same applies for drinking activity. The reminder message may be in the form of a suggestion, e.g."Would you like to drink some water?". In the case where the person has eaten his/her food and has consumed the necessary amount of liquid (such parameters should be personalized for each case), a successful session has been completed and the CU terminates the nutrition support system operation. In the case where the older person has missed a meal, this information can be stored to a data logger and, possibly, can be sent to a physician or a relative of the older person, in order to inform him/her and prevent possible future recrudescence. A block diagram of the above described eating/drinking activity analysis and decision making functionalities is illustrated in Figure 8.

3. **The MOBISERV-AIIA Multi-view Eating and Drinking Database.** In this Section, we describe an eating and drinking activity database that has been recorded in the context of the FP7 project MOBISERV. The database has been created in order to study eating and drinking behavior of individuals targeting to suggest and implement approaches to manage nutrition problems with non-invasive technological means. To the best of our knowledge, this is the first publicly available database that can be used for the evaluation of algorithms aiming at nutrition assistance exploiting multi-view information. In the following, we briefly describe existing publicly available databases aiming at both nutrition assistance and general activity recognition. Subsequently, we proceed with the description of the MOBISERV-AIIA database. Finally, we discuss research areas that can benefit from the use of the database and we focus on two application scenario usages, i.e., eating and drinking activity recognition and person identification based on nutrition activity.

3.1. **Existing Databases.** Nutrition assistance is a specific problem with limited publicly available datasets. The Pittsburgh fast-food image dataset (PFID database) [52] is

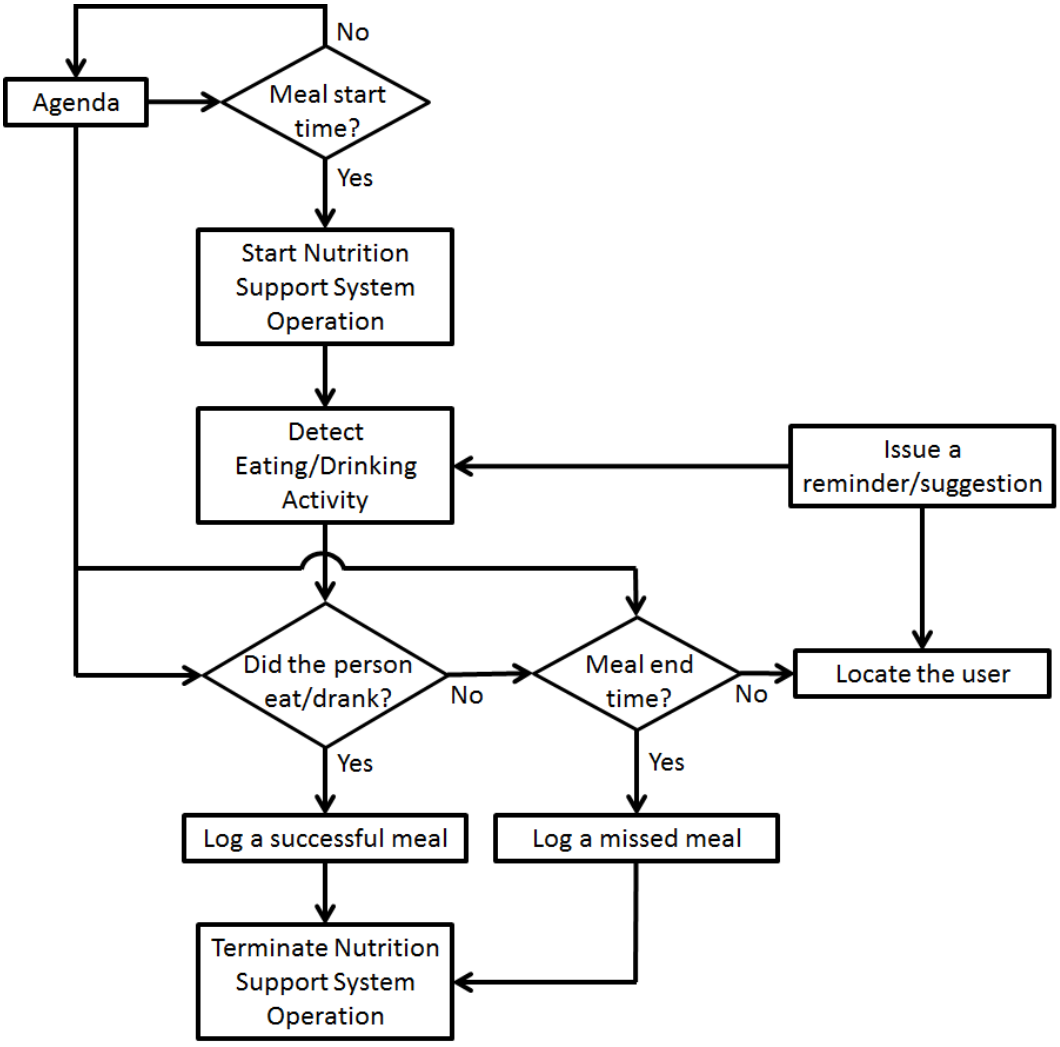

FIGURE 8. *Analysis of eating/drinking activity and decision making block diagram.*

offered to facilitate research in automated food recognition. It consists of 4.545 still images, 606 stereo pairs, 303 360^o videos for structure from motion and 27 privacy-preserving videos of eating events of volunteers. The data were collected by obtaining instances of foods from popular fast food chains and capturing images and videos in both restaurant conditions and a controlled lab setting. A series of experiments were conducted in [19] by using Multi Context Chewing Data Set. It is a dataset consisting of seven three-minute videos (captured at 24fps) of persons performing chewing and non-chewing actions (Data Set 2). Each video is comprised of five action segments: 1) closed-mouth chewing, 2) open-mouth chewing, 3) an assortment of facial expressions, 4) talking and 5) still face. There was no condition placed on the persons head movements, although the majority of the persons maintain a frontal head pose.

Hollywood2 is a Human Actions and Scenes Dataset [53] with 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips and approximately 20.1 hours of video in total. This dataset intends to provide comprehensive benchmark for human action recognition in realistic and challenging settings. The dataset is composed of video clips from 69 movies, including actions like “answer phone”, “drive car”, “eat”, “fight person”, “handshake”, “sit down”, “stand up”, and scenes took place in the house,

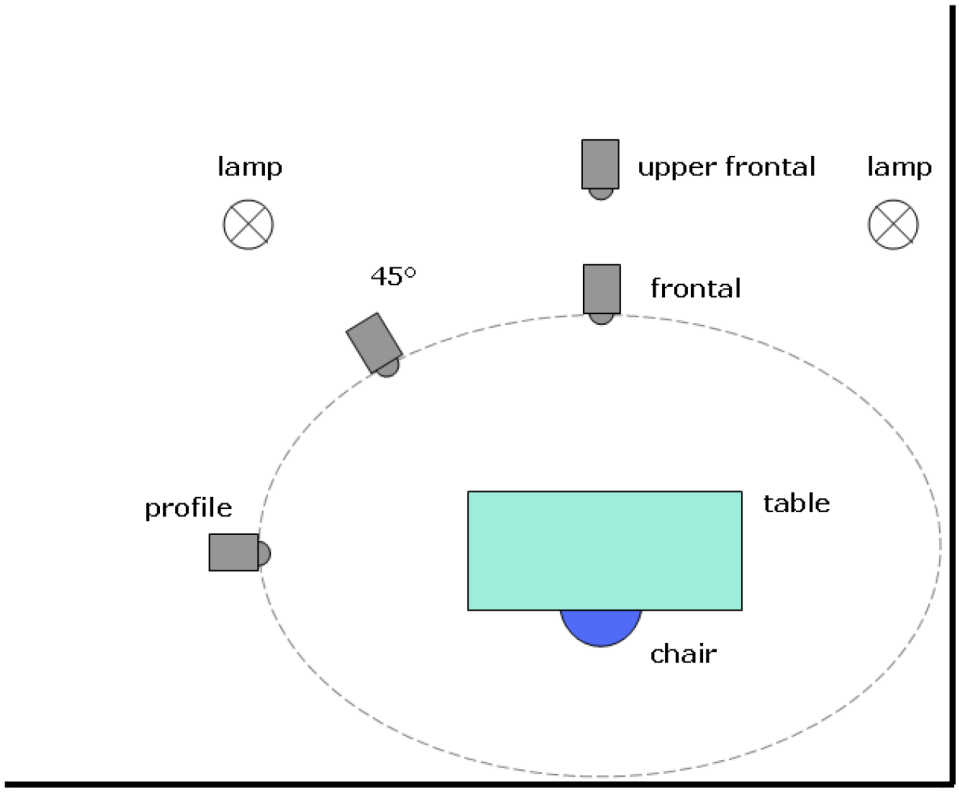

FIGURE 9. *The database camera setup.*

on the road, in a car, in the kitchen etc. Annotations for the action and scene samples are, also, available.

Regarding databases used in the context of general activity recognition, they, usually, contain actions that correspond to everyday human actions, like "walk", "run", "jump", "sit", etc. Tow publicly available single-view datasets that have been widely adopted for the evaluation of activity recognition methods are the Weizmann [54] and the KTH [55] datasets. Publicly available datasets containing sports and dance activities are the UCF sports [56], UCF 50, and the Ballet [57] datasets. However, due to the fact that the performance of action recognition methods is closely related to the camera observation angle, multi-view action datasets have been created. Two publicly available datasets, that have been widely adopted for the evaluation of multi-view activity recognition methods are the IXMAS [58] and the i3DPost [59] datasets.

3.2. **Studio Environment and Camera Setup.** The intake procedure was captured in a controlled laboratory environment including a camera and an illumination system. The details of the recording system, depicted in Figure 9, are described below. No audio was recorded.

3.2.1. *Camera Setup.* Four Sony XCD-V60CR digital video cameras (color model) were placed according to the setup illustrated in Figure 9. The frontal camera was mounted on a tripod at 1m height and at a distance of 1.4m from the chair where persons were sitting during recordings. The upper frontal camera was mounted on a tripod 1m higher from the frontal camera and at a distance of 1.77m from the chair. The profile camera was also mounted on a tripod at 1m height at the left hand of the person being recorded. The last camera was mounted on a tripod at 1m height 45^o to the person?s left. The two frontal cameras and the 45^o one can provide frontal or near frontal facial image poses

[60, 61], if needed. All sequences were saved in .avi format and they were captured at a resolution of 640 × 480 pixels and at a frame rate of 15 frames per second. The cameras were calibrated to extract their extrinsic (pose, orientation) parameters before shooting.

The 4 cameras were connected serially, using IEEE 1394 interface cabling system, which is a serial bus interface standard for high-speed communications and isochronous real-time data transfer, to a high performance PC. However, a delay between 1 to 32 frames per camera was observed. To overcome this issue, a light source was utilized; this served as signaling for the cameras synchronization. To begin the process, all cameras were turned on at the same time. Then, a lighter was used in front of all cameras to record an instant flash. This flash lasted a moment (one frame) and was captured from all the cameras. Before any further processing, during annotation phase, the first frames of the 4 videos were carefully compared to find the number of the frames containing evidence of the light. Appropriate time differences between each video and the video captured from the frontal camera were calculated, allowing the remaining parts of the videos to start simultaneously.

3.2.2. *Illumination Setup.* To improve and control the quality of the recorded video data, in addition to daylight lamps on the ceiling, two light sources, of 800W each, were used. The curtains were kept open during the recordings allowing illumination changes between different days. The position of the light sources (lamps) can be seen in Figure 9.

3.3. **Data Acquisition.** The recordings were conducted in a period of two months, in four sessions, and included 12 participants. Each of them had to accomplish four distinct recordings. Each recording consisted of two parts. In the first part the person had to act based on the scenario wearing a shirt or blouse with long sleeves, whilst in the second part he/she had to repeat the same actions wearing a T-shirt. Each participant was informed about the nature of the experiment and signed a consent form.

3.3.1. *Camera and Video Properties.* Video sequences were recorded and stored using Lagarith compression in order to reduce the required storage space. Lagarith is an open source lossless video codec which is not as good as MSU and FFV1 in terms of achieved compression rate and created video quality, however is significantly faster compared against both codecs. In our case, the time needed to record and compress 4 videos simultaneously is prohibited. The applied compression during recording mainly causes the delay between the connected cameras.

In order to export realistic video sequences regarding the color of the skin, clothes and background, various values for the camera software parameters were tried, in order to conclude to the following:

- Camera properties:
  - Aperture = 2.8,
  - Expo = 570,
  - Shutter = 91,
  - Gain = 0,
  - U/B = 2191,
  - V/R = 2082,
  - Hue = 2096,
  - Sat = 375,
  - Black level = 2047 and
  - Gamma = 0.
- Format Selection:
  - Pixel Format = Y_MONO16 bit,
  - Resolution = 640 × 480 and

TABLE 2. Participants' basic personal data

Person	Sex	Age
01	M	32
02	F	27
03	M	22
04	M	26
05	F	28
06	M	39
07	M	24
08	F	26
09	M	31
10	F	26
11	F	22
12	F	22

– Frame Rate = 15.0.

3.3.2. *Subjects.* The database includes 12 persons with Caucasian origin and ages between 22 and 39 years. There are 6 females and 6 males with or without beard. Three persons are not wearing glasses, while the rest of them were wearing according to their will. Apart from the expected occlusions caused during eating and drinking activity (hand, spoon, cup etc.), there were also some hair failing on the face.

Each person participated in four distinct sessions and recorded two videos for each one: one wearing a blouse or a shirt with long sleeves and one wearing a T-shirt or a blouse with short sleeves. Considering that there were used 4 cameras, 8 videos were created for each person at the end of each session. This results in a database consisted of 384 video sequences.

Each person recording wearing long sleeves or a T-shirt lasts from 7 to 11 minutes and its size ranges between 1.4 and 4.7GB. This results to a database of 59.68h and 1220GB initial size (in Lagarith compression). In order to be able to handle and use these videos, we compressed them using a more efficient compressor, lossy though. We used MPEG4 compressor which resulted in 107.5GB of data (91% compression rate).

3.3.3. *Recording Protocol.* Each participant of the database eats with spoon, cutlery, fork, hands and drinks from cup, glass and straw according the following scenario: He/she enters the room, sits down in front of the frontal camera and starts eating cereals with spoon (4 iterations) and he/she drinks water from a cup (4 iterations). The bowl is replaced by the experimenter with a plate and the cup with a glass of water. The participant continues eating with knife and fork (4 iterations) and after that he/she drinks ($1-2$ iterations). He/she eats the same food using only the fork (the hand holding the fork is changed, 4 iterations). He/she drinks again ($1-2$ iterations). The person puts a straw in a glass of water and continues eating with his/her hand (4 iterations) and drinks from straw (4 iterations in total). The person, also, eats with his/her hands a loaf of bread or croissant pretending to eat a sandwich (4 bites). Finally, the person spends some time doing something else besides eating, like thinking/wondering, paying attention, talking, reading, writing, touching his/her hair, ear, nose, glasses, chin or yawing with the assistance of the experimenter. He/she stands up and exits the room. The exact succession of activities that each person had to execute in each meal is illustrated in Table 3. Example video frames of the database are illustrated in Figure 10.

Table 3. The database recording protocol

MEAL	ITERATION	TYPE	ACTION
BREAKFAST	1^{st}	spoon	eat
BREAKFAST	1^{st}	cup	drink
BREAKFAST	2^{nd}	spoon	eat
BREAKFAST	2^{nd}	cup	drink
BREAKFAST	3^{rd}	spoon	eat
BREAKFAST	3^{rd}	cup	drink
BREAKFAST	4^{th}	spoon	eat
BREAKFAST	4^{th}	cup	drink
LUNCH	-	-	slicing
LUNCH	1^{st}	cutlery (knife+fork)	eat
LUNCH	-	-	slicing
LUNCH	2^{nd}	cutlery (knife+fork)	eat
LUNCH	1^{st}	glass of water	drink
LUNCH	-	-	slicing
LUNCH	3^{rd}	cutlery (knife+fork)	eat
LUNCH	-	-	slicing
LUNCH	4^{th}	cutlery (knife+fork)	eat
LUNCH	2^{nd}	glass of water	drink
LUNCH	-	-	slicing
LUNCH	1^{st}	fork	eat
LUNCH	2^{nd}	fork	eat
LUNCH	3^{rd}	glass of water	drink
LUNCH	3^{rd}	fork	eat
LUNCH	4^{th}	fork	eat
LUNCH	4^{th}	glass of water	drink
FAST FOOD	1st	hand	eat
FAST FOOD	2nd	hand	eat
FAST FOOD	1st	straw	drink
FAST FOOD	3rd	hand	eat
FAST FOOD	4th	hand	eat
FAST FOOD	2nd	straw	drink
FAST FOOD	1st	bite (both hands)	eat
FAST FOOD	2nd	bite (both hands)	eat
FAST FOOD	3rd	straw	drink
FAST FOOD	3rd	bite (both hands)	eat
FAST FOOD	4th	bite (both hands)	eat
FAST FOOD	4th	straw	drink
APRAXIA	-	-	thinking/wondering
APRAXIA	-	-	paying attention
APRAXIA	-	-	talking
APRAXIA	-	-	reading
APRAXIA	-	-	writing
APRAXIA	-	-	touching hair/beard/glasses or yawing using hands

3.4. **Database Annotation.** The database is annotated in a frame base according the structure of a template text file (.txt). For the videos recorded using the frontal camera, the starting and ending video frame numbers of each elementary movement are provided. The text files used to synchronize the video recordings from the four different camera

FIGURE 10. *Video frames depicting the twelve persons of the database from different viewing angles.*

views are included in the annotation file set. By using the information concerning the elementary movements, action videos depicting one elementary movement each, e.g. an eating instance, can easily be created. This is important for the comparison of methods evaluating their performance on the database.

3.5. **Database Application Examples.** In this subsection we discuss research areas that can benefit from the use of the MOBISERV-AIIA database and focus on two application scenarios, i.e., eating and drinking activity recognition and person identification from nutrition activity. The database can be used in order to facilitate research in:

- Single- and Multi-view action recognition: The database contains three actions, which consist of 13 sub-actions in total, as illustrated in Table 4. That is, action recognition can be performed on an action or on an sub-action base. The use of three action classes, i.e., 'eat', 'drink' and 'apraxia', may complicate the classification problem, due to the observed high intra-class and small inter-class variations. Intra-class variations in the case of the above described three-class classification problem are high, since due to different human body proportions and action execution style variations it is possible that an action video depicting a person eating with spoon is

more similar to an action video depicting another person drinking from a cup that from an action video depicting a sequence belonging to another eating subclass, e.g., 'eat with fork'. Two experimental settings can be employed in order to evaluate two application scenarios. The Leave-One-Person-Our (twelve fold) cross-validation scheme can be employed in order to evaluate the generalization ability of a method on persons that are not included in the training process. The Leave-One-Day-Out (four fold) cross-validation scheme can be employed in order to evaluate the performance of a method on a person-specific experimental setting.

- Person identification: The database can be used as a benchmark dataset for face recognition algorithms. The four cameras used for the recordings lead to facial images depicting the persons from different viewing angles, with different facial deformations (due to eating/drinking activity), different levels of occlusion and different facial appearance (beard, glasses). In addition, the database can be used to evaluate person identification methods exploiting motion information, i.e., person identification based on actions.
- Person de-identification: De-identification is the process which aims to remove all identification information of the person from an image or video, while maintaining as much information on the action and its context [62]. Identifying information captured on video can include face, silhouette, action, etc. Thus, the database can be used as a benchmark dataset for person de-identification methods.
- Visual tracking: The database can be used as a benchmark dataset for face and object tracking. Different facial poses, levels of occlusion and viewing angles set different levels of difficulty for face tracking algorithms. These facts apply also for tracking of hands and objects, e.g., cups, forks, etc.
- 3D reconstruction: By exploiting information coming from all the four cameras used in the recordings and the available camera setup calibration parameters, 3D reconstruction of the scene can be achieved. The reconstructed model can be used in order to facilitate all the above mentioned research areas.

3.5.1. *Eating and Drinking Activity Recognition.* We have used the videos captured by the frontal camera depicting the persons wearing T-shirt. The set of annotation files was used to segment the initial video recordings in smaller ones depicting elementary movements (action videos). We perform eating and drinking action recognition in videos based on the above described action recognition scheme. We have used the Leave-One-Person-Out (LOPO) cross-validation procedure in order to evaluate the impact of different human body sizes and action execution style variations among individuals on the action classification performance. That is, we used the action videos depicting all but one person in the database as labeled data and the action videos depicting the remaining one as test data, in order to perform one iteration (fold) of the cross-validation procedure. Twelve folds, equal to the number of persons appearing in the database, have been performed in order to complete an experiment. A parameter value $m = 1.1$ has been used, while the optimal number of posture prototypes D has been determined by performing multiple experiments for values $D = 5d$, $d = 10, \ldots, 20$. We have formed a three-class classification problem, i.e., each action video was followed by a label belonging to one of the three action classes: 'eat', 'drink' and 'apraxia'. Four experimental settings have been evaluated in order to simulate different experimental scenarios of incremental difficulty. The obtained action classification rates are illustrated in Table 4. As can be seen in this Table, action classification rates close to 80% have been obtained for all the four evaluated scenarios, ranging from 82.94% for the case where each action class is formed by two sub-actions, to

TABLE 4. Action classification rates for different experimental settings. The symbol 'x' denotes that the action videos depicting the corresponding sub-action are included in the experiment

eat					drink			apraxia					
cutlery	fork	spoon	hands	bite	cup	glass	straw	chew	slice	read	write	hands	Rate (%)
x	x	x			x	x		x	x				82.94
x	x	x	x	x	x	x	x	x	x				81.88
x	x	x	x	x	x	x	x	x	x	x	x		80.91
x	x	x	x	x	x	x	x	x	x	x	x	x	79.64

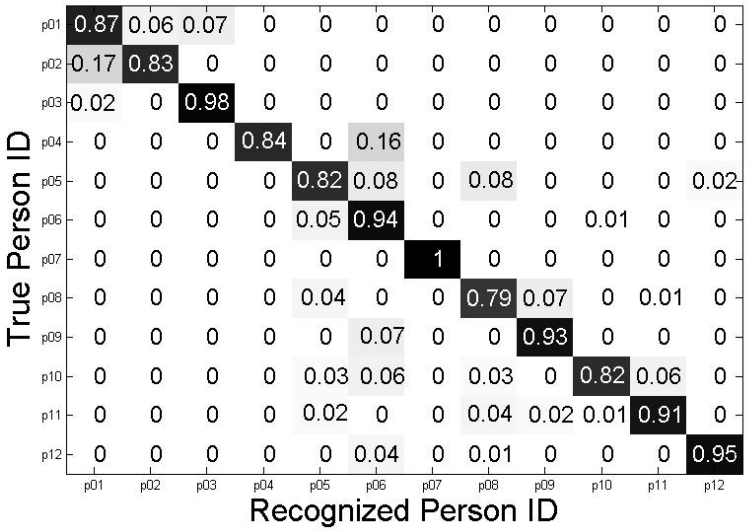

FIGURE 11. *Confusion matrix for person identification.*

79.64% for the case where all the action videos of the database are employed for evaluation.

3.5.2. *Person Identification from Activities.* By exploiting the person ID information of the action videos of the database, we have used the above described method in order to perform person identification based on actions appearing in meal intakes. We have used the videos captured by the frontal camera depicting the persons wearing T-shirt. We have used the Leave-One-Day-Out (LODO) cross-validation procedure in order to evaluate impact of variations in person appearance and action execution style between different days on the person identification performance. That is, we used the action videos of all persons in the database depicting the meals recorded for three days as labeled data and the action videos depicting the meals recorded for the remaining day as test data, in order to perform one iteration (fold) of the cross-validation procedure. Four folds, equal to the number of different recording days of the database, have been performed in order to complete an experiment. A parameter value $m = 1.1$ has been used, while the optimal number of posture prototypes D has been determined by performing multiple experiments for values $D = 10d$, $d = 10, \ldots, 30$. A person identification rate equal to 89.67% has been obtained. The confusion matrix of this experiment is illustrated in Figure 11. As can be seen in this Table, high identification rates have been obtained for most persons in the database. This fact confirms our assumption that action execution style variations among individuals are significant in actions belonging to meal intakes and can be used for action based person identification.

4. **Conclusions.** In this paper, we described the human centered interface specifications and implementations for a computer vision-based nutrition assistance system. Specifically, we investigated the scenario of assisted living in a smart home environment providing several "intelligent functionalities", like computer vision-based automatic eating and drinking activity detection, automatic communication with relatives or physicians in case of abnormal nutrition activity, etc. We have described a method exploiting a Privacy Preserving action representation requiring low computational cost. By using such an action representation and a classification scheme based on Fuzzy Vector Quantization and Neural Network-based classification, fast and effective operation can be achieved. In addition, we have described a publicly available multi-view eating and drinking activity recognition database that has been created in order to facilitate research towards non-invasive nutrition assistance. This database can be used for a number of other studies, notably to analyze mastication and the effects of aging on eating and drinking habits.

Acknowledgment. This work has been funded by the Collaborative European Project MOBISERV FP7-248434 (http://www.mobiserv.eu), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly.

REFERENCES

[1] J. Vogt, K. Luyten, J.V. den Bergh, K. Coninx and A. Meier, Putting dementia into context: a selective literature review of assistive applications for users with dementia and their caregivers, *International Conference on Human-Centered Software Engineering*, pp. 181-198, 2012.

[2] D. Vergados, A. Alevizos, A. Mariolis and M. Caragiozidis, Intelligent services for assisting independent living of elderly people at home, *International Conference on Pervasive Technologies Related to Assistive Environments*, vol. 97, pp. 1-4, 2008.

[3] Z. Lin, A. R. Hanson, L. J. Osterweil and A. Wise, Precise process definitions for activities of daily living: a basis for real-time monitoring and hazard detection, *Workshop on Software Engineering in Health Care*, pp. 13-16, 2011.

[4] C.A. Frantzidis and P.D. Bamidis, Description and future trends of ICT solutions offered towards independent living: the case of LLM project, *International Conference on Pervasive Technologies Related to Assistive Environments*, vol. 59, pp. 1-8, 2009.

[5] R. Watson, Measuring feeding difficulty in patients with dementia: perspectives and problems, *Journal of Advanced Nursing*, vol.18, no. 1, pp. 25-31, 1993.

[6] S.A. Noah, S.N. Abdullah, S. Shahar, H. Abdul-Hamid, N. Khairudin, M. Yusoff, S. Shahar, R. Ghazali, N. Mohd-Yusoff, N.S. Shafii and Z. Abdul-Manaf, DietPal: a Web-based dietary menu-generating and management system, *Journal of Medical Internet Research*, vol. 5, no. 3, 2003.

[7] C.H. Morris, R.A. Hope and C.G. Fairburn, Eating habits in dementia: a descriptive study, *The British Journal of Psychiatry*, vol. 154, pp. 801-806, 1989.

[8] K. Sim, G.E. Yap, C. Phua, J. Biswas, A.A.P. Wai, A. Tolstikov, W. Huang and P. Yap, Improving the accuracy of erroneous plan recognition system for activities of daily living, *International Conference on e-Health Networking Applications and Services*, pp. 28-25, 2010.

[9] N.M. Gil, N.A. Hine, J.L. Arnott, J. Hanson, R.G. Curry, T. Amaral and D. Osipovic, Data visualisation and data mining technology for supporting care for older people, *International ACM SIGACCESS conference on Computers and accessibility*, pp. 139-146, 2007.

[10] O. Amft and G. Troster, On-body sensing solutions for automatic dietary monitoring, *IEEE Pervasive Computing*, vol. 8, no. 2, pp.62-70, 2009.

[11] O. Amft and G. Troster, Recognition of dietary activity events using on-body sensors, *Artificial Intelligence in Medicine*, vol. 42, no. 2, pp. 121-136, 2008.

[12] A. Fleury, M. Vacher and N. Noury, SVM-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results, *IEEE Trans. on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 174-283, 2010.

[13] C. Phua, V.F. Foo, J. Biswas, A. Tolstikov, A.P.W. Aung, J. Maniyeri, W. Huang, M.H. That, D. Xu and A.W. Chu, 2-layer erroneous-plan recognition for dementia patients in smart homes, *IEEE International Conference on e-Health Networking, Applications and Services*, pp. 21-28, 2009.

[14] V. Di Lecce, C. Guaragnella, T. d'Orazio and R. Dario, Smart Postural Monitor for Elderly People, *19th IMEKO TC 4 Symposium and 17th IWADC Workshop Advances in Instrumentation and Sensors Interoperability*, 2013.

[15] V. Metsis, Z. Le, Y. Lei and F. Makedon, Towards an evaluation framework for assistive environments, *International Conference on Pervasive Technologies Related to Assistive Environments*, pp. 1-8, 2008.

[16] M. Ogawa and S. Ochiai and K. Shoji and M. Nishihara and T. Togawa, An attempt of monitoring daily activities at home, *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 786-788, 2000.

[17] T. Togawa, Home health monitoring, *Journal of Medical and Dental Sciences*, vol. 45, no. 3, pp. 151-160, 1998.

[18] B.G. Celler and E.D. Lisar and W. Earnshaw, Preliminary results of a pilot project on remote monitoring of functional health status in the home, *Bridging Disciplines for Biomedicine*, pp. 61-64, 1996.

[19] S. Cadavid and M. Abdel-Mottaleb, Exploiting visual quasi-periodicity for automated chewing event detection using active appearance models and support vector machines, *International Conference on Pattern Recognition*, pp. 1714-1717, 2010.

[20] P. Wu, J.W. Hsieh, J.C. Cheng, S.C. Cheng and S.Y. Tseng, Human smoking event detection using visual interaction clues, *International Conference on Pattern Recognition*, pp. 4344-4347, 2010.

[21] E. Tapia and S. Intille and K. Larson, Activity recognition in the home using simple and ubiquitous sensors, *PERVASIVE, 2nd International Conference*, pp. 158-175, 2004.

[22] Q. Fan, R. Bobbitt, Y. Zhai, A. Yanagawa, S. Pankanti and A. Hampapur, Recognition of repetitive sequential human activity, *Computer Vision and Pattern Recognition*, pp. 943-950, 2009.

[23] H.J. Seo and P. Milanfar, Action recognition from one example, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 867-882, 2011.

[24] A. Iosifidis, E. Marami, A. Tefas and I. Pitas, Eating and drinking activity recognition based on discriminant analysis of fuzzy distances and activity volumes, *International Conference on Acoustics, Speech and Signal Processing*, pp. 2201-2204, 2012.

[25] G. Stamou, M. Krinidis, N. Nikolaidis and I. Pitas, A monocular system for automatic face detection and tracking, *Visual Communications and Image Processing*, 2005.

[26] M.M Elmansori and K. Omar, An Enhanced Face Detection Method Using Skin Color and Back-Propagation Neural Network, *European Journal of Scientific Research*, vol. 55, no. 1, p. 80, 2011.

[27] H. Baltzakis, A. Argyros, M. Lourakis and P. Trahanias, Tracking of human hands and faces through probabilistic fusion of multiple visual cues, *Computer Vision Systems*, pp. 33-42, 2008.

[28] O. Zoidi, A. Tefas and I. Pitas, Visual Object Tracking based on Local Steering Kernels and Color Histograms, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 870-882, 2013.

[29] L. Goldmann, U.J. Monich and T. Sikora, Components and their topology for robust face detection in the presence of partial occlusions, *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 559-569, 2007.

[30] M. Pateraki, H. Baltzakis, P. Kondaxakis and P. Trahanias, Tracking of facial features to support human-robot interaction, *IEEE International Conference on Robotics and Automation*, 2009.

[31] C. Anagnostou, S. Psomiadis, A. Pavlides, G. Economou and L. Zouloumis, Protocols of conservative treatment of masseter myalgia, *29th Panhellenic Dental Congress*, 2009.

[32] E. Horjales-Araujo, N.B. Finnerup, T.S. Jensen and P. Svensson, Differential effect of visual and gustatory stimuli on experimental jaw muscle pain, *European Journal of Pain*, vol. 17, no. 6, pp. 811-819, 2013.

[33] S.A. Othman, R. Ahmad, S.M. Asi, N.H. Ismail and Z.A. Rahman, Three-dimensional quantitive evaluation of facial morphology in adults with unilateral cleft lip and palate, and patients without clefts, *British Journal of Oral and Maxillofacial Surgery*, vol. 52, no. 3, pp. 208-213, 2014.

[34] P. Koruga, M. Baca and M. Schatten, Analysis of craniofacial morphology changes during aging and their connection with facial age estimation, *International Conference on Information Technology Interfaces*, pp. 481-486, 2011.

[35] E. Gavopoulou, Investigation of correlation between dental and oral health and Alzheimer disease or other dementia forms, *32th Panhellenic Dental Congress*, 2012.

[36] Z. Lin and L.S. Davis, Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 604-618, 2010.

[37] N. Tsapanos, A. Tefas and N. Nikolaidis and I. Pitas, Shape matching using a binary search tree structure of weak classifiers, *Pattern Recognition*, vol. 45, no.6, pp. 2363-2376, 2012.
[38] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, *Computer Vision Pattern Recognition*, pp. 511-518, 2012.
[39] N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, *Computer Vision Pattern Recognition*, pp.886-893, 2005.
[40] I. Cherif, V. Solachidis and I. Pitas, A tracking framework for accurate face localization, *Artificial Intelligence in Theory and Practice*, vol. 1, pp. 385-393, 2006.
[41] S. Ali and M. Shah, Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288-303, 2010.
[42] J. Hoey and J.J. Little, Representation and recognition of complex human motion, *Computer Vision Pattern Recognition*, pp.752-759, 2000.
[43] W. Yang and G. Mori, Hidden Part Models for Human Action Recognition: Probabilistic versus Max Margin, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1310-1323, 2011.
[44] M.A. Giese and T. Poggio, Neural mechanisms for the recognition of biological movements, *Nature Reviews Neuroscience*, vol. 4, no. 3, pp. 179-192, 2003.
[45] A. Iosifidis, A. Tefas, N. Nikolaidis and I. Pitas, Multi-view human movement recognition based on fuzzy distances and Linear Discriminant Analysis, *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 347-360, 2012.
[46] A. Iosifidis, A. Tefas and I. Pitas, Multi-view action recognition based on action volumes, fuzzy distances and Cluster Discriminant Analysis, *Signal Processing*, vol. 93, no. 6, pp. 1445-1457, 2013.
[47] E. Marami, A. Tefas and I. Pitas, Nutrition Assistance based on Skin Color Segmentation and Support Vector Machines, *Man-Machine Interactions*, pp. 179-187, 2011.
[48] P. Soille, Morphological image analysis: principles and applications, *Springer-Verlag*, New York, 2nd, 2004.
[49] A. Iosifidis, A. Tefas and I. Pitas, View-Invariant Action Recognition Based on Artificial Neural Networks, *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no. 3, pp.412-424, 2012.
[50] P. J. Werbos, Beyond regression: New tools for prediction and analysis in the behavioral sciences, *Harvard University*, 1974.
[51] A. Iosifidis, A. Tefas and I. Pitas, Minimum Class Variance Extreme Learning Machine for Human Action Recognition, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp.1968-1979, 2013.
[52] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar and J. Yang, PFID: Pittsburgh fast-food image dataset, *IEEE International Conference on Image Processing*, pp. 289-292, 2009.
[53] M. Marszalek, I. Laptev and C. Schmid, Actions in Context, *Computer Vision Pattern Recognition*, pp. 1-8, 2009.
[54] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, Actions as space-time shapes, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247-2253, 2007.
[55] C. Schuldt, I. Laptev and B. Caputo, Recognizing human actions: A local SVM approach, *International Conference on Pattern Recognition*, pp.32-36, 2004.
[56] M.D. Rodriguez, J. Ahmed and M. Shah, Action mach: A spatio-temporal maximum average correlation height filter for action recognition, *Computer Vision Pattern Recognition*, pp.1-8, 2008.
[57] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, Behavior recognition via sparse spatio-temporal features, *Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp.65-72, 2005.
[58] D. Weinland, R. Ronfard and E. Boyer, Free viewpoint action recognition using motion history volumes, *Computer Vision and Image Understanding*, vol. 104, no.3, pp.249-257, 2006.
[59] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis and I. Pitas, The i3DPost multi-view and 3D human action/interaction database, *Conference on Visual Media Production*, pp. 159-168, 2009.
[60] I. Marras, N. Nikolaidis and I. Pitas, Frontal Facial Pose Recognition using a Discriminant Splitting Feature Extraction Procedure, *International Conference on Information Technology Interfaces*, 2011.
[61] F. Solina and R. Ravnik, Fixing missing eye-contact in videoconferencing systems, *International Conference on Information Technology Interfaces*, 2011.
[62] P. Agrawal and P.J. Narayanan, Person De-identification in Videos, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 299-310, 2011.