ITC 4/47

Journal of Information Technology
and Control
Vol. 47 / No. 4 / 2018
pp. 691-703
DOI 10.5755/j01.itc.47.4.21602

Human-Machine Interaction in Intelligent Technologies
Using the Augmented Reality

Received 2018/10/17 | Accepted after revision 2018/11/05

http://dx.doi.org/10.5755/j01.itc.47.4.21602

Human-Machine Interaction in Intelligent Technologies Using the Augmented Reality

Dawid Połap

Institute of Mathematics, Silesian University of Technology,
Kaszubska 23, 44-100 Gliwice, Poland; e-mail: Dawid.Polap@polsl.pl

Corresponding author: Dawid.Polap@polsl.pl

Intelligent homes are one of the most expanded technologies included in the definition of Internet of Things. These technology allow us to exercise control over our own apartment/house. As a control, we can understand energy consumption, monitoring, healthy eating as well health care. One of the elements of such a system are interface and interaction with it. In order to increase the possibilities and encourage people to invest in such technologies, it is important to increase the possibilities and forms of system visualization. One of the novel idea is to use augmented reality to increase the quality of Internet of Things applications. In this work, we propose a model to increase interaction between the user and the system installed at home. The proposed solution is based on the use of data obtained by the camera and microphone in smartphones to extract selected features and process them for the purpose of increasing interaction. Presented technique can be applied to image and voice samples using similar technique of processing (based on key-points searching algorithm). Such action can allow to increase the range of operation system, to simplify the life of the household members, in particular, the elderly and the disabled people. Model of the operation system has been described and tested to indicate the potential benefits of using augmented reality as an additional human-machine communication interface. Moreover, proposed technique of processing different type of input data like video, image and sound can be used for feature extraction and further classification purposes.

KEYWORDS: Augmented Reality, Intelligent System, Internet of Things, Artificial Intelligence, Neural Networks, Image Processing.

1. Introduction

Augmented reality is one of the most interesting achievements in the field of games or communication interfaces between the application and the user. Enabling manipulation of the reality image with the help of a camera available in every smartphone has opened new possibilities in other areas where phones, tablets or other equipment are used, enriched with certain sensors. Communication between these devices is particularly important, which contributes to the increase of quality as well the quantity and diversity of acquired data from the environment, which is the natural environment. Communication between various technological accessories such as smartwatch and smartphone allows to check the health of the user by examining, for example, the pulse and its analysis. The possibility of using the camera means that the photo can be quickly processed and analyzed. As the effect of its action, the obtained results can be presented to the user in a simple and comprehensible way. It is this use of new technologies that indicates the increase of our life's comfort as well its control.

Above all, enhancement of opportunities is visible in the areas of intelligent technologies as well games. Especially the second type which is the games causes, that players have more expectations from graphics and playability. It results in rapid technological development. The continuous increase of possibilities also means that not only hardware is improved, but also the software and technologies behind it. It is thanks to the games that the augmented reality gained such popularity. More precisely, thanks to the game built on the basis of classic creatures and the selection of a good age group of young people who, thanks to nostalgia, grabbed their smartphones and gave the game a chance. Talking about *Pokemon Go*, which has become one of the biggest phenomena of recent years [6, 12]. The use of augmented reality made it possible to locate creatures in reality thanks to the camera build in smartphones. An additional advantage was the possibility of catching these creatures, which is an interesting aspect of modifying reality with the use of additional sensors [15, 18].

However, it is worth noting that the game itself was just the beginning. Then, the augmented reality has been used for various purposes increasing the effects obtained using the camera itself. An example is the analysis of the sky by searching for different stars, placing the constellations or even observing sun path [23, 25]. Technology is being improved to now, and used primarily on the Internet of Things. It was the use of this technology in the application of image analysis and data measurement that turned out to be very useful. It is certainly particularly important to illustrate something to users, an example of which may be an additional interface or indication of certain things in the image which may not be visible or overlooked by the user. An example of a goal can be visually impaired people, for whom the augmented reality seen from up close can enlarge a given image and improve its quality or even modify it to increase their security.

It is worth noting that scientists are developing the technology of augmented reality as well as looking for practical applications to use the full potential of it. An important achievement is the construction of holographic near-eye displays that can affect the future of virtual modification of the reality [16]. The technology has also found its application in learning and navigation [10]. Difficult topics in the fields of biology and physics in schools were only possible by sketching general mechanisms. Now, by the use of phones or glasses, it is possible to create simulations of various phenomena. Such activities not only diversify learning in schools, but also provide the opportunity to better understand the issue and visualization of it. Increasing the presence of additional elements as part of the perception of reality is also analyzed through context awareness [7]. There is also the opportunity to raise people's awareness of existing threats, negative emotions to protect the environment and reduce the sales of nicotine [13]. Analysis of interfaces for the use of speech has been shown in [2, 9]. An interesting idea is the use of gamification through which some interfaces for such purposes can be created [5]. The use of augmented reality in different types of activities (like games, education) allows to assume, that it also can be useful in human machine interaction met in Internet of Things applications. In the presented applications, different algorithms are used, especially artificial intelligence techniques that are constantly being developed [19, 24].

In this work, we present the use of modified artificial intelligence methods to find selected objects and display the obtained information in an accessible form to the user. This type of technique can be used in various ways, including in the kitchen, where the camera will register objects in the refrigerator and display the possible combinations of recipes. Another application is to inform the user about a possible life threatening situation based on the found objects. In addition, we assume that in order to increase the capabilities of the offered interface to the user, the system should offer the possibility of voice control. The work presents a mathematical model containing processing voice and image sample in similar way. The process obtain some features of the samples that can be used to create input data for neural classifier. The theoretical part is supplemented with tests and their analysis depending on different configuration and due to wider practical application.

2. Image Processing

To describe the operation of the model and data processing, several assumptions need to be made. Suppose that a user uses a camera on a smartphone or tablet. In addition, suppose that the recorded video has a record of 24 frames per one second. Processing each frame would be too burdensome for the hardware, so we will take only one frame every two seconds. This action allows to minimize the number of calculations.

Having one frame, the next step is to locate objects. However, first you need to find key points, that is, those that have certain unique properties. An algorithm looking for these points is Speeded Up Robust Features (SURF) presented in [1]. The detector's operation is based on the approximate value of the Hessian matrix. This matrix defines blob detector, and the Haar's wavelet is used as a descriptor which is not used in this version. The Hessian matrix is defined as follows

$$H(x,\omega)=\begin{bmatrix} L_{xx}(x,\omega) & L_{xy}(x,\omega) \\ L_{xy}(x,\omega) & L_{yy}(x,\omega) \end{bmatrix}, \tag{1}$$

where the matrix coefficients are a convulsive operation on the integral picture I and the partial derivative using Gaussian kernels g(ω). The formula for this coefficients can be presented as

$$L_{xx}(x,\omega)=I(x)\frac{\partial^2}{\partial x^2}g(\omega), \tag{2}$$

$$L_{yy}(x,\omega)=I(x)\frac{\partial^2}{\partial y^2}g(\omega), \tag{3}$$

$$L_{xy}(x,\omega)=I(x)\frac{\partial^2}{\partial xy}g(\omega). \tag{4}$$

The algorithm assumes calculating the determinant of this matrix by the use of the following equation

$$\det\left(H_{approximate}\right)=D_{xx}D_{xy}-\left(wD_{xy}\right)^2, \tag{5}$$

where w is the coefficient identified with the weight of the input image *I*, and D_{xx} is associate with $L_{xx}(x, \omega)$ because of approximation of this value and discrete kernels. Having calculated this determinant, all local maximum values are considered to be extremes witch can be identified with key-points of a given image *I*.

Algorithm may return a very large set of key-points, that may causes that it will be distinguish to locate different objects from each other. For this purpose, it is necessary to minimalize this number of points. The solution is based on checking and evaluating the neighbors values. For each found key-point, Nearest neighbors are checks in relation to the current point in terms of quality. Quality is understood as the level of the edge of a given point relative to the entire neighborhood. To illustrate this process, let's assume that we have a given pixel at position (x, y), and the analyzed neighborhood is a size equal to n × n (hereinafter referred to as a grid), where a given pixel is located in the middle. The selected grid is binarized, i.e. for each pixel in position (i, j) in the grid, we replace a given color according to the formula below

$$\begin{cases} \text{if } \frac{R(I(i,j))+G(I(i,j))+B(I(i,j))}{3}<128, \\ \qquad \text{then } I(i,j)=\text{white}' \\ \text{else } I(i,j)=\text{black} \end{cases} \tag{6}$$

where I(i, j) is a specific pixel at selected position, and functions R(·), G (·), and B(·) refers to the components of RGB color system. Binarization has allowed to remove insignificant pixels. Depending on the color of the middle pixel, the average hue of the remaining pixels is counted and if the quotient of these two values meets the inequality below

$$\frac{H_{RGB}(I(x,y))}{\frac{1}{n^2-1}H_{RGB}\left(\Sigma_{i=0}^{i<n}\Sigma_{j=0}^{j<n,i!=\frac{n}{2},j!=\frac{n}{2}}I(i,j)\right)}\in\left\langle 0,\frac{2}{5}\right)\cup\left\langle \left\lceil\frac{n}{2}\right\rceil,\infty\right), \tag{7}$$

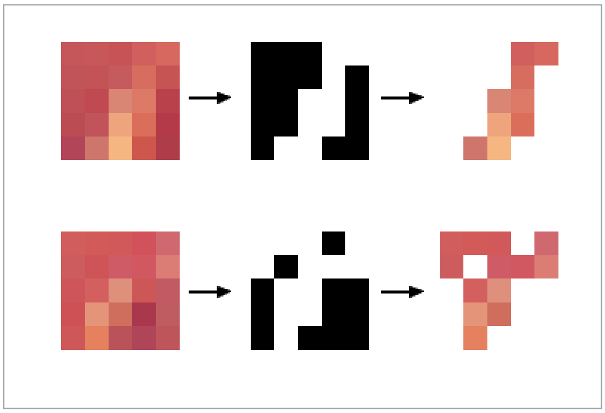
where $H_{RGB}(I(i,j))$ defines hue value which can be calculated as

$$H_{RGB}(I(i,j))= \\ \operatorname{atan2}\left(\sqrt{3}\left(G(I(i,j))-B(I(i,j))\right),2R(I(i,j))-G(I(i,j))-B(I(i,j))\right). \tag{8}$$

If a given pixel does not satisfy the Equation (7), it is removed. Illustrating this process is shown in Fig. 1.

Figure 1

Display of the neighborhood processing process for a given key point



The next step is to create sets representing the objects in the image. The problem is choosing key points that can reflect the object. To make this possible, the image is aligned with the edge detection filter [4]. In such a prepared image, it is checked whether there is a way to go through white pixels between these two points. If it exists, we assume that the key points belong to the same set. Otherwise, the second one automatically create a new set. We recognize the path between two pixels when the Euclidean distance between them is equal to 1.

Using these sets representing objects, for each of them, a new bitmap is created. Each of these points in a given set is placed on created bitmap, and all pixels around them are transferred from the original image. The size of the newly created bitmap is defined in advance for all samples. In the case when a given set goes beyond its area, a larger image is created and then reduced to a given one. This action is necessary that all objects could be correctly classified at a next stage.

The classification will take place via the Convolutional Neural Network (CNN). These type of classifiers are based on the action of primary cortex in human brain [11]. The biggest difference in relation to other types of neural structures is the type of entry. Classically, the input is a numeric vector, and in this case, the sample is saved as an image. Hence the earlier requirement to maintain one dimension for all samples representing objects. The network is composed of three types of layers. The first one is called convolutional which has one purpose – extract the features. It is done by the applying some filters on the image. These filters are marked as ω and they are written as a square matrix. There may be several layers of this type and each of them can have different filter. The most known filters used here are Gaussian blur or sharpening [8]. The filter works on the principle of modifying the image by moving this filter on the image by the given step *S*. The next type of layer is known as pooling. The layer is designed to reduce the size of the image by selecting the most important pixels. The choice is made using some the function describing the operation of minimizing, maximizing or averaging the pixel value in a given area. The idea of operation is similar to the use of a filter, with the difference that there is no filter, only the function that selects pixels to be transferred to a new, reduced image. The third type is named as fully connected and its structure and operation is almost the same as the classic neural structure. To describe it, these layer is composed of smaller parts, in classic reasoning knows as hidden layers and output layer. There may be many hidden ones, but only one output. As the entrance to first, hidden type, there is one image, where for each pixel, there is one neuron.

CNN is composed from three type of layer, but which are placed in a certain, specific arrangement. Layers named convolutional and polling are placed at the beginning, alternately and in the number selected by the user. Then there is exactly one layer names as fully connected, which is the end of this structure.

The structure itself is useless if it is not trained using sample datasets for a correct classification. The most well-known algorithm of training is the backpropagation algorithm. These algorithm operate on the principle of minimizing the error on the whole structure by modifying the values of weights, which are burdened by connections between layers. The error is calculated at the end of the network by a dedicated function f(·). To describe the idea of this algorithm, let us assume several signs. As the value returned by the neuron at position (i, j) on *l*-th layer will be a derivative of the form $\frac{\partial f}{\partial y_i^l}$, and the main equation we use is called chain rule can be described as

$$\frac{\partial f}{\partial \omega_{ab}} = \sum_{i=0}^{N-m} \sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^l} \frac{\partial x_{ij}^l}{\partial \omega_{ab}} = \sum_{i=0}^{N-m} \sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^l} y_{(i+1)(j+b)}^{l-1} . \quad (9)$$

Using the above rule, we are able to calculate the error that is achieved on the given layer as

$$\frac{\partial f}{\partial x^l_{ij}} = \frac{\partial f}{\partial y^l_{ij}} \frac{\partial y^l_{ij}}{\partial x^l_{ij}} = \frac{\partial f}{\partial y^l_{ij}} \frac{\partial\left(\sigma\left(x^l_{ij}\right)\right)}{\partial x^l_{ij}} = \frac{\partial f}{\partial y^l_{ij}} \sigma'\left(x^l_{ij}\right), \quad (10)$$

where $\sigma(x)$ is understood as a function that determines the activation of a given neuron in a layer. The algorithm operates from the last layer to the first one, so the error on the previous layer will be calculated using a different equation. However, it should be noted that the gradient for the convolutional layer will be defined as

$$\frac{\partial f}{\partial y^{l-1}_{ij}} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \frac{\partial f}{\partial x^l_{(i-a)(j-b)}} \frac{\partial x^l_{(i-a)(j-b)}}{\partial y^{l-1}_{ij}} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \frac{\partial f}{\partial x^l_{(i-a)(j-b)}} \omega_{ab}. \quad (11)$$

Having the formula on the gradient, the error on the convolutional layer can be calculated as

$$\frac{\partial x^l_{(i-a)(j-b)}}{\partial y^{l-1}_{ij}} = \omega_{ab} \, . \quad (12)$$

Pooling layer does not participate in the structure training.

In this way, the classifier is trained until the selected error or the given iteration is reached. Each of the objects found in the form of a picture is given to the network to automatically decide what the object is.

3. Audio Processing

In a similar way as image processing described in the previous section, such method can be used for sound processing. The recorded digital sound will be saved in the form of bits, so it is given in a discrete form. However, the sample itself is not possible to be analyzed, so it is necessary to calculate the given signal $s(n) = (s_0, s_1, ..., s_{N-1})$ by a discrete Fourier transform used as

$$S_k = \sum_{n=0}^{N-1} s_n \exp\left(-\frac{2\pi i n k}{N}\right), \quad 0 \leq k \leq N-1, \quad (13)$$

where all calculated values are complex number. The sound signal in this form can be written as an image, i.e. spectrogram. It is a time-frequency graph, where the values in the graph indicate intensity and are marked by the selected shade of color. The formula for calculating the spectrogram value is presented in the following way

$$\text{spectrogram}\{s(t)\}(t,f) = |S(t,f)|^2, \quad (14)$$

where function $S()$ is understood as short-time Fourier transform and define as

$$S(m,f) = \sum_{n=-\infty}^{\infty} S(n) w(n-m) \exp(-jfn), \quad (15)$$

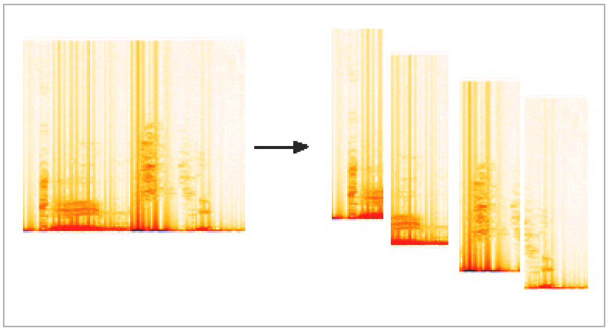
where $w()$ is a window function, an example of which can be a Welch function known as

$$w(n) = 1 - \left(\frac{n-0.5(N-1)}{0.5(N-1)}\right)^2. \quad (16)$$

Such a spectrogram may be cut in relation to a certain period of time. Assume that a given image will be cut every k seconds, then this samples can use be used as data to train previously described convolutional neural network (see Fig. 2).

Figure 2

An example of a spectrogram cut into training samples for a CNN



4. System Operation Model

The user using the smartphone has access to the camera and microphone. For example, suppose that user is in a smart home that is managed by intelligent technology. Additionally, let the system be connected to the user through the application on his smartphone. This solution is convenient because of communication and human-machine interaction.

All operations like recording video data through the camera and sound by the use of microphone are downloaded in real time and processed on smartphones. However, CNN training must be done on an external server. It is a long and burdensome process. Moreover, if the CNN should classify a large number of objects, the time needed for it is much longer. The larger is training database, the longer will be time to train classifier. For this reason, the user must have access to a trained network. So, the expanding database of new elements (which are added during whole living time of the system) should be on an external server or cloud, where the CNN would be trained all the time to increase its accuracy and extend its functionality. The user could use the update to download the new configuration of the classifier.

It is worth noting that if the user had to be able to control the voice, the application should contain two, separate classifiers. One designed for sound and the other for the image. Such a solution makes a lot of sense due to the hardware load, although the requirement for proper functioning is continuous access to the Internet network. Described idea is presented as visualization in Fig. 3.

5. Construction of the User Interface

The user interface should be simple and intuitive. When the technology of augmented reality is used, it should be borne in mind that the presented objects will overlap the visible image. It is important that the buttons or messages appearing are not on other objects. The basic elements will be the minimization and closing buttons, which should appear either at the top of the screen or on the left or right as an additional bar. The idea of an additional area, excluded from the view may be a good solution – in such a place all the necessary options can be placed depending on the needs of the system.

Image processing causes that every two seconds one video frame is downloaded, on which objects are extracted. Each of these objects can be marked on the

Figure 3

Visualization of the operation of the proposed system

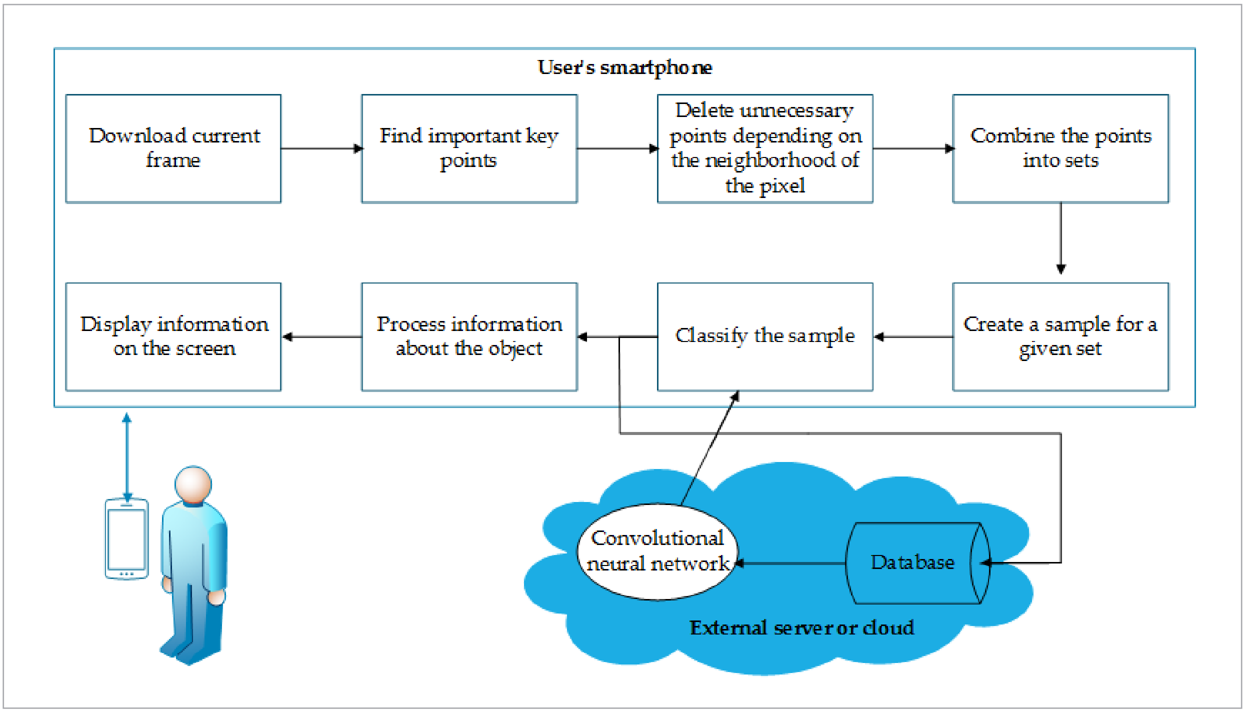
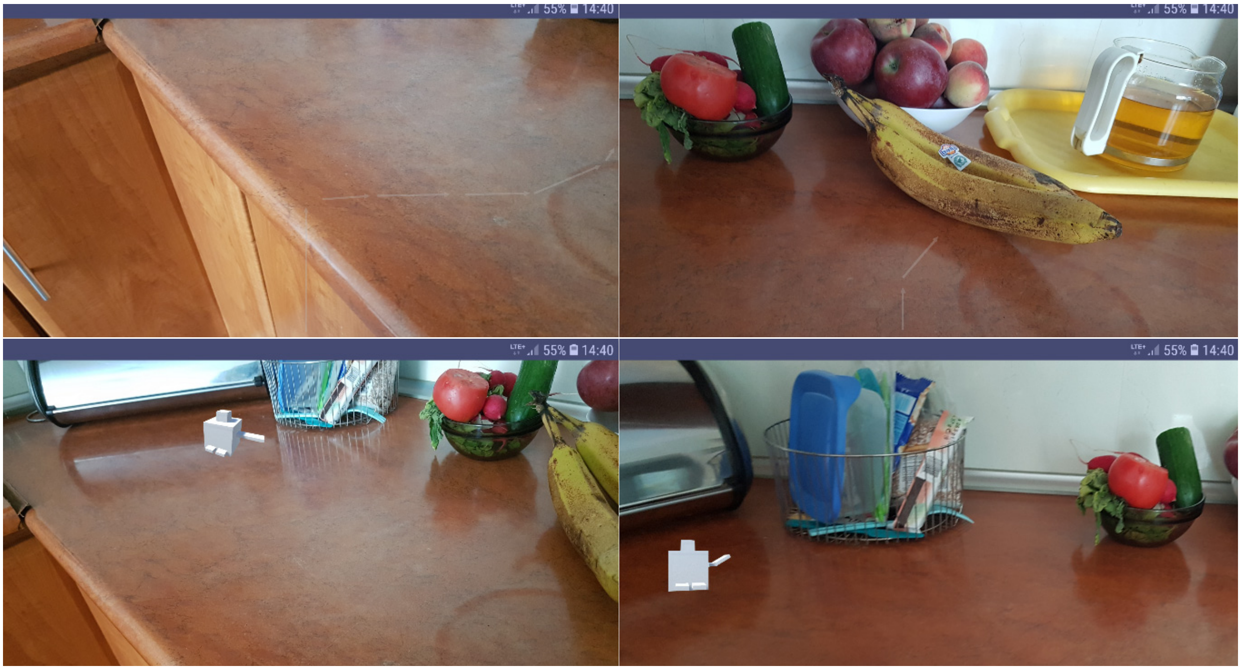

User's smartphone
Download current frame
Find important key points
Delete unnecessary points depending on the neighborhood of the pixel
Combine the points into sets
Create a sample for a given set
Classify the sample
Process information about the object
Display information on the screen
Convolutional neural network
Database
External server or cloud

Figure 4

Construction of exemplary interfaces based on augmented reality. In first two screenshots, there are small arrows indicating the location of bananas, and on the next ones, there is small figure indicating the same position



screen. If the database is extensive, the situation may arise that the registered image will be covered with information about found elements. To avoid this, objects only from a given category could be displayed. For example, when searching for a recipe for a dish, the camera should capture objects from the food category. Going further, the camera will register the objects that we have in the refrigerator, whether in the cabinet and based on these objects will find a recipe for the dish, which can be made using owned ingredients. Such a recipe must find its place on the screen, although it may be badly picked up by the elderly or visually impaired. Therefore, the name of the dish should be correctly displayed, and then the ingredients with their proportions should be indicated. Such targeting on the found object is possible in several ways. The first one is to display the arrows placed in the image, but in such a way that the user has the impression that they are in reality. The other one is to design an object (some kind of robot/bear) that will indicate and inform about the objects. On the basis of the described interface example, any other can be created. This construction of user interfaces with the use of augmented reality is presented in Figure 4.

6. Experiments

The work proposed three things - image and sound processing and interface construction. In the case of image processing, two databases were used. The first one is called Food-101 [3], which has 101 categories of prepared dishes, and exactly 101,000 photos. The second base consisted of a dozen or so categories of images of fruits and vegetables and other kitchen ingredients from the dataset ImageNet 2011 Fall (Marrow, vegetable marrow; Raw vegetable, rabbit food; Drupe, stone fruit; Fruit etc.) built with a total of 6348 images [11]. The algorithm for the detection and classification of objects has been tested on various configurations of bases consisting of three filters such as Gaussian blur, sharpen and emboss. CNN was trained in a ratio of 70:30 (trained to tested) depending on the archived error (to the error that was the stop criterion). In the case of fully-connected layer, the construction was chosen in empirical way. Several layer configurations were used with different numbers of neurons (in particular, the number of layers from the range 1-10, where the number of neurons ranged from

10-15) and the one with the higher accuracy was chosen. The final configuration was composed of 8 layers and 14 neurons in each of them.

Filters were selected in two configuration: I – {Gaussian blur, sharpen, emboss}, II – {sharpen, Gaussian blur, emboss}. The measurements of classification effectiveness are presented in Table 1. The highest efficiency was obtained for the error equal to 0.00001 where it value to nearly 90% when using the second configuration of convective layers. In the case of the first set, the effectiveness was slowly increased in contrast to the other. The reason for this is the poor selection of filters. In order to check the effectiveness, 2 000 other images were added to the database, which should be classified incorrectly. The correctness of the best classification is presented in the form confusion matrix in Fig. 5. Unfortunately, it took almost a month to train the CNN for such error. The smaller is the error, and the knowledge base is larger, the learning time will be relatively longer and it will counted in weeks.

Table 1

The obtained effectiveness for both configuration of CNN

Error	Configuration I	Configuration II
0.1	56%	60%
0.01	59%	66%
0.001	65%	73%
0.0001	66%	84%
0.00001	71%	89%

The proposed technique for voice processing, we tested for a dedicated core database for 5 people and composed of *5x40=200* samples, where the created samples had two simple sentences – *"Where are the bananas?", "Change the data display mode."*. For each of the samples, a spectrogram was created and trimmed with respect to *k* seconds (where k∈{1,2,3,4}). CNN was constructed with 3 convolutional layer with the previously described configuration.

The obtained results are presented in Tables 2 and 3. It should be noted that much better results are obtained using the second filter configuration (similar to image classification). Each classifier was checked against the error and time of the samples. The best results in both configurations were obtained for samples of length equal to 2 seconds. For the second configuration, efficiency was approximately 91%. For such a trained CNN, the correctness of classification for this database was tested, but with more samples –

Figure 5

Obtained confusion matrix for the classifier trained to error 0.00001 and second set of convolutional layer configuration (image)

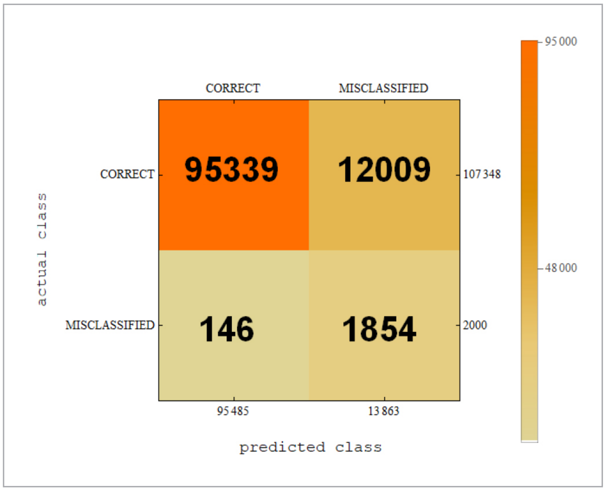


Table 2

Efficiency of the classifier relative to the first configuration

Error	Effectiveness			
	1	2	3	4
0.1	33%	34%	29%	23%
0.01	35%	34%	31%	25%
0.001	37%	37%	33%	30%
0.0001	40%	46%	38%	35%
0.00001	44%	51%	48%	43%

Table 3

Efficiency of the classifier relative to the second configuration

Error	Effectiveness			
	1	2	3	4
0.1	30%	34%	36%	32%
0.01	35%	52%	49%	38%
0.001	47%	67%	56%	42%
0.0001	61%	83%	70%	46%
0.00001	72%	91%	78%	51%

there were added an additional 100 samples with other sentences of similar length. The results are shown in Fig. 6. Obtained results for above best configuration was measured by some statistical coefficient like accuracy Γ, Dice's coefficient Λ, overlap Ψ, sensitive Υ and specificity Φ. Formulas for this coefficient are presented as

$$w(n)=1-\left(\frac{n-0.5(N-1)}{0.5(N-1)}\right)^2. \tag{17}$$

$$\Gamma=\frac{TP+TN}{TP+TN+FP+FN}, \tag{18}$$

$$\Lambda=\frac{2TP}{2TP+FP+FN}, \tag{19}$$

$$\Psi=\frac{TP}{TP+FP+FN}, \tag{20}$$

$$\Upsilon=\frac{TP}{TP+FN}, \tag{21}$$

$$\Phi=\frac{TN}{TN+FP}, \tag{22}$$

Figure 6

Obtained confusion matrix for the classifier trained to error 0.00001 and second set of convolutional layer configuration (voice)

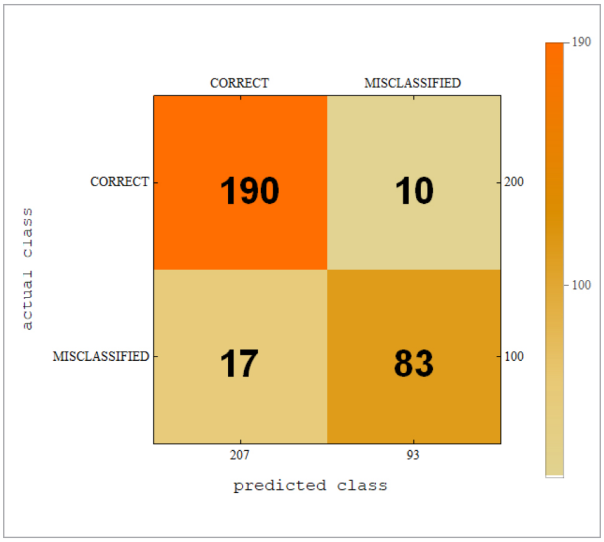


Table 4

The obtained grades on a scale of 0-10 for the use of a given interface

Age	7-15	15-18	18-25	25-40	40-65	Avg
No people	10	10	15	25	15	
Interface with trace	4.6	6.9	8.4	8.2	7.1	7.04
Interface with helper object	8	9.2	7.5	6.2	7.3	7.64

Table 5

The obtained statistical coefficients accuracy Γ, Dice's coefficient Λ, overlap Ψ, sensitive Υ and specificity Φ for CNNs, both trained to 0.00001 for image and voice

	Γ	Λ	Ψ	Υ	Φ
Image	0.889	0.994	0.873	0.998	0.134
Voice	0.91	0.934	0.671	0.918	0.892

where used parameters, in all formulas, were understood in the following way -- TP as true positive, FP as false positive, TN as true negative and FN as false negative. The values for these coefficient are shown in Table 5. Calculated coefficients for image and voice classification are particularly different in case of overlap and specificity. The specificity parameter determines the effectiveness in classifying and rejecting erroneous input data. The value for the image classifier is low due to the large difference between positive and negative samples in the test database. For classifying the sound samples, this problem did not appear because the database was smaller and similarly varied in terms of the samples number. This also confirms the value of overlap which is the measurement of the similarity between two sets. This results indicates that for such used database, the obtained results are good. This is particularly important in the case of sound, where the number of samples was not too large. The level of significance was at $p = 0.05$. Density function was calculated for both obtained results and described in Fig. 9, the formula was

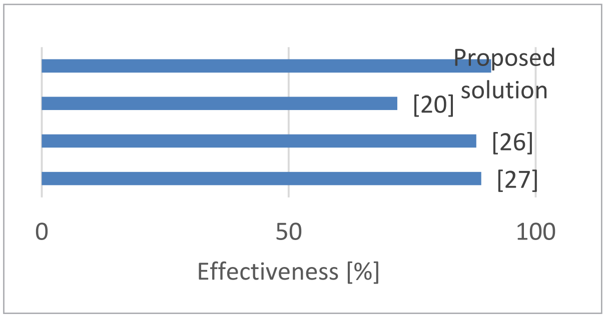
$$d(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{23}$$

where μ is the expected value and σ^2 is a variance.

The proposed solution was also compared with others, new solution presented in [20, 26, 27], with the same set of images. The results were presented in bar form in Fig. 7 where it can be seen that the effec-

Figure 7

Comparison of selected classification methods on the database used in the problem of sound classification in this paper



tiveness of the proposed solution for such a database. Proposed technique is better by approximately 2% than others, however, it is possible that in the case of more samples, the remaining classifiers would be better. Despite such a number of samples, the proposed technique for classifying sound files can be applied by the use of image processing approach. It is also worth noting that the process of training such a large number of samples in each method proved to be very time-consuming. For additional usability tests, the System Usability Scale (SUS) [14] was used and the average scores are presented in Table 6. The obtained scores were averaged and allow to calculate SUS score using the following equation

$$SUS=\left[\left(\sum_{i\in\{1,3,5,7,9\}} p_i -5\right)+\left(25-\sum_{i\in\{2,4,6,8,10\}} p_i\right)\right]\cdot 2, \quad (24)$$

where p_i means number of points for question number i. The obtained score was 77.5 which is considered a good results. The users indicated the simplicity and ease of use. The suggested implementation has not received the highest notes, although it is worth paying attention to the low scores for odd-numbered questions. Only in the groups of older people was it pointed out that the size of the added elements could be larger.

In the case of interface, we asked a group of 75 people to test the simplest application that indicates the location of bananas. Each of them was to evaluate both interfaces using a scale from 0-10. The obtained results are presented in Table 4. The average grades indicate a much better reception of a certain object, which indicates where something is located than the application of the idea of the trace. However, the highest scores received this system for younger users who appreciate more interaction and more objects. In the older age groups, the solution consisting in displaying a trace or arrows was considered more interesting. It is noting that in this proposed solution, human machine interaction is made constantly. It means that there are feedback loops. User can say some words, and camera can record some image and this information are processed. As the results some data are presented to user in augmented reality approach.

Table 6

The obtained scores for SUS

Question	Avg score
I think that I would like to use this system frequently.	3
I found the system unnecessarily complex.	1
I thought the system was easy to use.	4
I think that I would need the support of a technical person to be able to use this system.	2
I found the various functions in this system were well integrated.	4
I thought there was too much inconsistency in this system.	1
I would imagine that most people would learn to use this system very quickly.	4
I found the system very cumbersome to use.	1
I felt very confident using the system.	3
I needed to learn a lot of things before I could get going with this system.	2

The displayed information should be presented in quick and easy way. It is also important that the amount of information displayed on the screen may affect the speed of the application (the more information is presented to the user, the operation may be slower). For the presented interfaces, the number of arrows or the size of the creature has been selected due to the users' ratings on their own smartphones (the requirement was to have a built-in gyroscope). The obtained results have been depicted in Fig. 8. It is easy to notice that the size of displayed elements are quite small what is the cause of ratings. As the main argument of such action is a better view of reality, as well faster operation of the application.

Communication between the available system and the user can be done over the phone, but to allow manipulation of the reality seen by the user, the system must have access to it. Access can take place via built-

Figure 8

User rating for the size of displayed objects

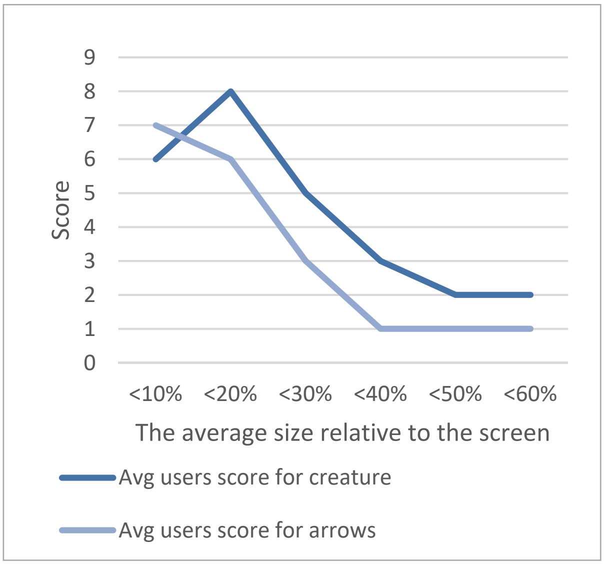
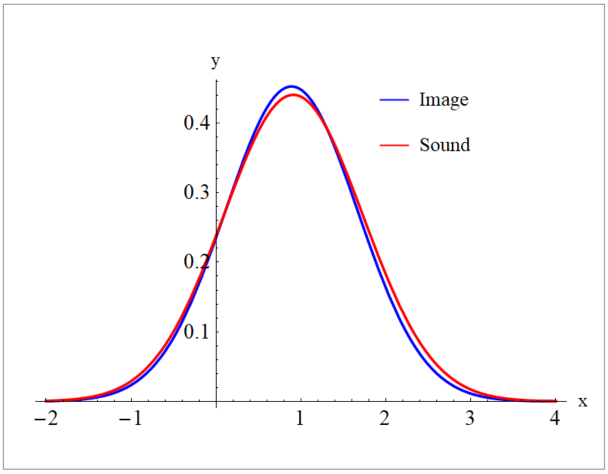


Figure 9

Graph of density function



in sensors on smartphone. Such activities using the augmented reality allow not only to manipulate what the user sees, but also to use it to improve it by displaying additional information or proposing a solution based on available objects that are previously found by the system.

The presented solution of the operating model, as well the construction of the human-machine interface, allows to increase the accessibility of smart technology and its improvement with additional advantages. As part of this work, the presented interfaces have proven to be intuitive and simple for each, tested age group. Interface should be simple and intuitive because of different target groups – from the youngest to the oldest. It is worth adding that except these two features, the interface should be characterized by clarity with the image being intercepted. In addition, the proposed techniques for extracting data from image and sound show great potential. The methods were tested only within selected knowledge databases and a few categories, but the achieved level of efficiency was almost 90%.

However, the presented solution is not perfect, because its biggest disadvantage is the time of training using databases consisting of a large amount of data. If the system requires detection and recognition of many object, the training can take even weeks and more. It is worth noting that once trained network can be attached to an application or system and no longer modified. However, such a solution brings with it the problem of different room lighting, as well the appearance of new objects unknown to the classifier. Despite this disadvantage, the operating model can be used with selected system modules, such as presented and tested at work, i.e. kitchen help, especially for older people.

Presented solution can be used in smart technologies like smart homes where can indicate some objects, or as an aid in the natural environment, and show the way. Moreover, it can be used as help for small children to help them to find parents (displayed on the TV). This can be created using smart technologies and build-in sensors to locate them and exchange data between apps. In addition, this can be a particularly interesting solution in the application as a help for drivers to improve the navigation system. It is worth noting that the technology used is slowly used in various life support applications [14, 17, 21, 22], which allows us to state that the proposed technique can be useful as a locator of certain objects in reality and used as an auxiliary module in other applications.

7. Conclusions

In this article, we presented the mechanism of analyzing data downloaded from smartphone by the use of built-in sensors. This technique uses image processing approach and can be used to analyze images as well sound samples after prior conversion to the spectrometer. It was used to extract some information from natural environment and presents the direction of the searched objects using augmented reality. Additionally, examples of interface structures for such use were presented.

It is worth noting that, despite good results, this is not an ideal tool. Especially in the case of increasing databases or possibilities of application. These types of problems can be the further subject of research to increase the efficiency of data processing methods. It is also worth paying attention to the growing use of gamification, which could be used to improve the interaction between human and machine.

Acknowledgement

Authors acknowledge contribution to this project to the Diamond Grant No. 0080/DIA/2016/45 funded by the Polish Ministry of Science and Higher Education.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. Speeded-Up Robust Features (SURF). Computer vision and image understanding, 2008, 110(3), 346-359. https://doi.org/10.1016/j.cviu.2007.09.014
2. Bier, A., Sroczyński, Z. Adaptive Math-to-Speech Interface. Proceedings of the Multimedia, Interaction, Design and Innnovation, 2015.
3. Bossard, L., Guillaumin, M., Van Gool, L. Food-101– Mining Discriminative Components with Random Forests. European Conference on Computer Vision, 2014, 446-461.
4. Canny, J. A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine In-telligence, 1986, 6, 679-698. https://doi.org/10.1109/TPAMI.1986.4767851
5. Darius A., Damasevicius R. Gamification of a Project MANAGEMENT system. Proceedings of International Con-ference on Advances in Computer-Human Interations (ACHI2014), 2014, 2000-2017.
6. Dorward, L. J., Mittermeier, J. C., Sandbrook, C., Spooner, F. Pokémon Go: Benefits, Costs, and Lessons for the Conservation Movement. Conservation Letters, 2017, 10(1), 160-165. https://doi.org/10.1111/conl.12326
7. Grubert, J., Langlotz, T., Zollmann, S., Regenbrecht, H. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(6), 1706-1724. https://doi.org/10.1109/TVCG.2016.2543720
8. Hammett, S. T., Georgeson, M. A., Gorea, A. Motion Blur and Motion Sharpening: Temporal Smear and Local Contrast Non-Linearity. Vision Research, 1998, 38(14), 2099-2108. https://doi.org/10.1016/S0042-6989(97)00430-6
9. Iniguez-Jarrin, C., Garcia, A., Roman, J. F. R., Lopez, O. P. Guidelines for Designing User Interfaces to Analyze Genetic Data. Case of Study: GenDomus. International Conference on Evaluation of Novel Approaches to Software Engineering, 2017, 3-22.
10. Joo-Nagata, J., Abad, F. M., Giner, J. G. B., García-Pe-alvo, F. J. Augmented Reality and Pedestrian Navigation Through Its Implementation in M-Learning and E-Learning: Evaluation of an Educational Program in Chile. Computers & Education, 2017, 111, 1-17. https://doi.org/10.1016/j.compedu.2017.04.003
11. Krizhevsky, A., Sutskever, I., Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 2012, 1097-1105.
12. LeBlanc, A. G., Chaput, J. P. Pokémon Go: A Game Changer for the Physical Inactivity Crisis? Preventive Medi-cine, 2017, 101, 235-237. https://doi.org/10.1016/j.ypmed.2016.11.012
13. Lee, J., Jung, S., Kim, J. W., Biocca, F. Applying Spatial Augmented Reality to Anti-Smoking Message: Focusing on Spatial Presence, Negative Emotions, and Threat Appraisal. International Journal of Human–Computer Interac-tion, 2018, 1-10. https://doi.org/10.1080/10447318.2018.1489581
14. Lewis, J. R., Sauro, J. The Factor Structure of the System Usability Scale. International Conference on Human Cen-tered Design, 2009, 94-103. https://doi.org/10.1007/978-3-642-02806-9_12
15. Lv, Z., Halawani, A., Feng, S., Ur Réhman, S., Li, H. Touchless Interactive Augmented Reality Game on Vision-Based Wearable Device. Personal and Ubiquitous Computing, 2015, 19(3-4), 551-567. https://doi.org/10.1007/s00779-015-0844-1

16. Maimone, A., Georgiou, A., Kollin, J. S. Holographic Near-Eye Displays for Virtual and Augmented Reali-ty. ACM Transactions on Graphics (TOG), 2017, 36(4), 85. https://doi.org/10.1145/3072959.3073624

17. Malūkas, U., Maskeliūnas, R., Damaševičius, R., Woźniak, M. Real Time Path Finding for Assisted Living Using Deep Learning. Journal of Universal Computer Science, 2018, 24(4), 475-487.

18. Morschheuser, B., Riar, M., Hamari, J., Maedche, A. How Games Induce Cooperation? A Study on the Relation-ship Between Game Features and We-Intentions in an Augmented Reality Game. Computers in Human Behav-ior, 2017, 77, 169-183. https://doi.org/10.1016/j.chb.2017.08.026

19. Nourani, V., Andalib, G., Dąbrowska, D. Conjunction of Wavelet Transform and SOM-Mutual Information Data Pre-Processing Approach for AI-Based Multi-Station Nitrate Modeling of Watersheds. Journal of Hydrology, 2017, 548, 170-183. https://doi.org/10.1016/j.jhydrol.2017.03.002

20. Price, M., Glass, J., Chandrakasan, A. P. A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Mod-els and Voice-Activated Power Gating. IEEE International Solid-State Circuits Conference (ISSCC), 2017, 244-245.

21. Shi, Z., Wang, H., Wei, W., Zheng, X., Zhao, M., Zhao, J., Wang, Y. Novel Individual Location Recommendation with Mobile Based on Augmented Reality. International Journal of Distributed Sensor Networks, 2016, 12(7), 1-13. https://doi.org/10.1177/1550147716657266

22. Spoladore, D., Arlati, S., Sacco, M. Semantic and Virtual Reality-Enhanced Configuration of Domestic Environ-ments: The Smart Home Simulator. Mobile Information Systems, 2017.

23. Tarng, W., Ou, K. L., Lu, Y. C., Shih, Y. S., Liou, H. H. A Sun Path Observation System Based on Augment Reality and Mobile Learning. Mobile Information Systems, 2018, 10, 1-10. https://doi.org/10.1155/2018/5950732

24. Wlodarczyk-Sielicka, M., Lubczonek, J., Stateczny, A. Comparison of Selected Clustering Algorithms of Raw Data Obtained by Interferometric Methods Using Artificial Neural Networks. 17th International Radar Symposium (IRS), 2016, 2016, 1-5.https://doi.org/10.1109/IRS.2016.7497290

25. Zhang, J., Sung, Y. T., Hou, H. T., Chang, K. E. The Development and Evaluation of an Augmented Reality-Based Armillary Sphere for Astronomical Observation Instruction. Computers & Education, 73, 2014, 178-188. https://doi.org/10.1016/j.compedu.2014.01.003

26. Zhang, Y., Chan, W., Jaitly, N. Very Deep Convolutional Networks for End-to-End Speech Recognition. IEEE In-ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, 4845-4849. https://doi.org/10.1109/ICASSP.2017.7953077

27. Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., Courville, A. Towards End-to-End Speech Recogni-tion with Deep Convolutional Neural Networks. Proceedings of Interspeech, 2016, 410-414.