Journal Pre-proof

Human activity recognition from UAV-captured video sequences

Hazar Mliki, Fatma Bouhlel, Mohamed Hammami

PII: S0031-3203(19)30441-8
DOI: https://doi.org/10.1016/j.patcog.2019.107140
Reference: PR 107140

To appear in: *Pattern Recognition*

Received date: 27 May 2019
Revised date: 5 November 2019
Accepted date: 27 November 2019

Please cite this article as: Hazar Mliki, Fatma Bouhlel, Mohamed Hammami, Human activity recognition from UAV-captured video sequences, *Pattern Recognition* (2019), doi: https://doi.org/10.1016/j.patcog.2019.107140

highlights

- Human activity recognition from UAV-captured video sequences.
- The adaptation of the classic CNNs to detection.
- Handling the detection task in non-annotated datasets.
- The classification of both video frames, and entire video sequences for human activity categorization.

Human activity recognition from UAV-captured video sequences

Hazar Mliki[a], Fatma Bouhlel[b,*], Mohamed Hammami[c]

[a] *University of Sfax, MIRACL-ENETCOM, Sfax, Tunisia*
[b] *University of Sfax, MIRACL-FSEG, Sfax, Tunisia*
[c] *University of Sfax, MIRACL-FS, Sfax, Tunisia*

Abstract

This research paper introduces a new approach for human activity recognition from UAV-captured video sequences. The proposed approach involves two phases: an offline phase and an inference phase. A scene stabilization step is performed together with these two phases. The offline phase aims to generate the human/non-human model as well as a human activity model using a convolutional neural network. The inference phase makes use of the already generated models in order to detect humans and recognize their activities. Our main contribution lies in adapting the convolutional neural networks, normally dedicated to the classification task, to detect humans. In addition, the classification of human activities is carried out according to two scenarios: an instant classification of video frames and an entire classification of the video sequences. Relying on an experimental evaluation of the proposed methods for human detection and human activity classification on the UCF-ARG dataset, we validated not only these contributions but also the performance of our methods compared to the existing ones.

Keywords: Scene stabilization, Human detection, Human activity recognition, Deep learning, Convolutional neural networks, UAV.

*Corresponding author
Email address: fatmabouhlel@fsegs.u-sfax.tn (Fatma Bouhlel)

1. Introduction

Human behavior analysis is considered as an active area of research that continues to evolve owing to its potential application in a variety of fields. This interesting research area involves detecting, tracking, and analyzing humans physical behavior [1]. Within this framework, human activity recognition is regarded as a crucial task in the analysis of human behavior. It relies upon identifying the label of the performed activity. Thus, a great interest has been allocated to human activity recognition by the research community [2]. In particular, human activity recognition from UAV-captured video sequences has attracted the attention of several researchers thanks to the ability of the UAV to overcome the problem of fixed coverage and reach difficult access areas [3]. Moreover, the UAV exhibits an increased flexibility allowing it to overcome the occlusion problem and therefore to ensure the surveillance of open spaces (forests, festivals, open-air concerts).

Methods of human activity recognition have prospered and evolved over the last decades. However, the human activity recognition from UAV-captured video sequences still remains a thorny problem that was not fully deciphered due to multiple constraints related to the acquisition platform such as: the dynamic and complex background as well as the variation in point of view, in the camera altitude, and in human appearance. Furthermore, the intra and inter human activity classes variability is considered as a major deficiency in the human activity recognition field.

In this paper, we displayed a new approach for human activity recognition from UAV-captured video sequences for a video surveillance application. The suggested approach involves two phases. An offline phase to generate the human/non-human and the human activity models using a convolutional neural network as well as an inference phase allowing a human detection and an activity recognition through the already generated model.

The main contributions of the proposed approach can be summerised as follows:

- Although the convolutional neural networks like the R-CNN [4], Fast R-CNN[5], Faster R-CNN [6], have been used for objects detection, they require an annotated dataset of human objects position in the video sequence frames. In order to overcome such deficiency, we proposed a new approach allowing the classic CNNs dedicated to the classification task to address the problem of detection, through a module that generates and selects regions of interest (RoI). Our contribution lies in the possibility to way the detection task is handled in non-annotated datasets.

- The proposed approach is based on two scenarios of human activity classification; an instant video frames classification, and an entire video sequence classification. The instant video frames classification is more suitable for video surveillance systems as an alert can be triggered in case of a suspicious activity. However, the entire video sequence classification consists in assigning a single activity label to the video sequence. Although the existing works perform only an entire classification, which is not adequate in the context of video surveillance, we suggest classifying the human activity according to the two already mentioned scenarios.

The remaining sections of the paper are organized as follows. Section 2, identifies certain related works in scene stabilization, human detection and human activity classification methods. In section 3, the different steps of the proposed approach are reported. The experimental results and discussion are provided in Section 4. The last section provides the conclusion and offers some perspectives for the future works.

2. Related Works

The general process of a human activity recognition approach from UAV-captured video sequences is illustrated in figure 1.

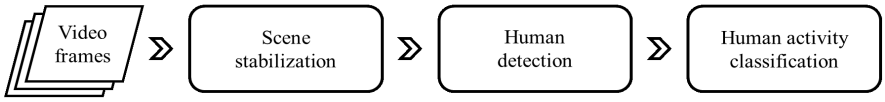

Figure 1: General process of human activity recognition approach

- Scene stabilization is a preprocessing step performed in order to eliminate the ego-motion of the acquisition sensor. This motion, mainly resulting from the UAV motion, affects the scene appearance.
- Human detection is a special case of object detection, which involves the detection of the human location in each frame.
- Human activity classification is the final step in this process. It consists in identifying the label of the performed activity.

2.1. Scene stabilization

The scene stabilization is an important preprocessing step in the video surveillance domain that uses the UAV as a sensor. It is performed in order to reduce the ego-motion of the acquisition platform.

The scene stabilization generally consists of three steps [7]: features matching, motion estimation, and motion compensation. The first step allows identifying the correlations between the points or the regions of interest of two successive frames. Once these matches are determined, they are used in the second step to estimate the local and the global motion by computing their displacement. Finally, the motion compensation step applies the inverse of the global motion model to adjust the frame to the correct position.

Indeed, two main motion types are identified in the video [7] one is global and the other is local. The global motion portrays the motion induced by the displacement of the camera. It affects the static and dynamic objects that make up the scene. The global motion in a frame is used to detect the background objects. This type of motion is used in the motion compensation step. As for the local motion, it depicts the motion of dynamic objects in the scene.

Within the same framework, Hsiao et al. [8] proposed a process of global motion estimation by computing the motions of the center block and the four corner blocks of the frame. Their method is based on the assumption that foreground objects are likely to be in the center of the frame. From this perspective, they attribute a greater weight to the central block than the corner blocks. Shen et al. [9] adopted the same approach as proposed in [8], but they applied circular blocks. Walha et al. [10] estimated motion through the application of the Scale Invariant Feature Transform (SIFT) algorithm [11] allowing the identification of the points of interest. Afterwards, the authors filtered the outliers resulting from the inaccuracy of the SIFT model using RANdom SAmple Consensus (RANSAC) [12]. In the matching step, they computed the Euclidean distance to obtain a set of motion vectors. Those vectors define the motion of the camera and the motion relative to the moving objects in the scene. To distinguish between these two motions, they assumed that the speed of the moving objects is much greater than that of the camera motion. Finally, a motion compensation is performed to obtain a stabilized video frame via affine transformations. In the same context, Minaeian et al. [13] proposed an approach for scene stabilization to detect the moving objects from UAV-captured video sequences. The feature extraction and tracking are performed through the Good Features To Track (GFTT) method. Thereafter, in order to achieve the image registration, they used the matched key points to generate a homography matrix between the reference frame t, and frames $t+\Delta t$ and $t+2\Delta t$. This would enable them to produce two adjusted images. Nonetheless, the values of all the parameters are fixed a priori except for the temporal interval. Carletti et al. [14] proposed a moving object detection module which consists of four steps: 1) camera compensation, 2) detection based on foreground mask, 3) foreground features extraction and clustering and 4) fusion of the foreground mask and the feature points. Nevertheless, using the frame differencing and the feature-based detection method affect the accuracy of the moving objects detection[15]. Within the scope of real-time human detection from UAV-captured video sequences, AlDahoul et al. [16] suggested a stabilization step that takes into account the real-time con-

straint. Thus, the authors proposed to estimate the direction and velocity of the motion vector between consecutive frames using the Horn-Schunck optical flow technique [17]. This procedure eliminates the motion resulting from the displacement of the UAV.

Taking into account the above-mentioned study, the methods of Hiso et al. [8] and Shen el al. [9] seem to be unrealistic as they rely on the assumption that foreground objects are more likely to be in the center of the image. Indeed, in the context of video surveillance both of the image center and ends are valued in the same way. As for the method of Walha et al. [10], they favored performance over computing time. Since, SIFT algorithm is computationally expensive, their method is not really suitable for video surveillance applications [10]. The Minaeian et al. [13] approach is appropriate for real-time onboard processing; however, it relies on the use of many parameters. The method proposed by AlDahoul et al. [16] takes into account the temporal constraint. Indeed, the use of the optical flow in the motion estimation step as well as the elimination of the motion compensation step made it possible to further optimize the computational time. Such a stabilization method is more suitable to our system. We choose, to perform a stabilization step which consists in estimating the motion through computing the optical flow. Therefore, we used the Lucas-Kande algorithm owing to its accuracy, simplicity and speed [18]. Since, we aimed to differentiate the motion related to objects from the motion induced by the displacement of the UAV, as Al-Dahoul et al. [16], we found it useless to carry out a motion compensation step.

2.2. Human detection

In literature, human detection methods can be classified into two categories: handcraft features extraction based methods and deep learning based methods.

Based on the handcraft features extraction, Dollar et al. [19] introduced the Fastest Pedestrian Detector in the West (FPDW) using the Histograms of Oriented Gradients (HOG) descriptor [20] at different scales. The authors proposed to rescale ($\frac{N}{K}$) the input image with K being a division factor rather than

N times. For each resized image its features are calculated. Therefore, these images features are used to approach the response of the features of the leftover $(N - \frac{N}{K})$ scales. Decreasing the number of image rescaling and subsequently the number of feature computations, has a positive impact on the computation time which drops considerably. Within the same vein of thought, [21] proposed a method for human detection. Although this method rests on the FPDW, it is distinguished by resizing the image during the training phase rather than in the test phase [19]. This method adopts the same idea as Viola and Jones [22]. The methods based on the handcraft features extraction have allowed to obtaining a satisfactory performance against the backdrop of human detection. However, these methods generally require prior extraction of the features and depend on the context of the application and the relevance of the used descriptor.

Human detection through using a deep learning method remains a hot research topic that is not fully explored. The deep learning methods are based on automatic feature extraction, using a convolutional neural networks. Within the same framework, AlDahoul et al. [16] proposed two deep learning models: The Supervised CNN (S-CNN), and the Pretrained CNN. They evaluated the performance of these models on the UCF-ARG dataset [23]. The recorded results by the second pretrained model outperform those of the first S-CNN model. This is due to the fact that the UCF-ARG dataset is not large enough. As a matter of fact, a transfer learning is necessary [16]. The authors [16] claim that the use of deep models is robust against various constraints such as: the variation of the altitude of the UAV, the variation in the acquisition point of view and the variation of the lighting conditions. Indeed, through the use of convolutional neural networks, as well as the variation of the learning sets, they succeeded in handling such constraints. On the other hand, the two proposed models dont handle the problem of close objects. Indeed, these models allow only classifying the regions of interest resulting from the computing of the optical flow and therefore do not detect the close objects located in the same region separately.

The study of the advantages and disadvantages of these methods has led us to

a pretrained deep neural network where we used the pretrained model AlexNet. Unlike AlDahoul et al. [16], who replaced the last layer of the AlexNet model with an SVM classifier, we opted for a softmax classifying layer. Moreover, compared to AlDahoul et al. [16], our main contribution is handling close up objects problem that are usually classified as a single object.

2.3. Human activity classification

In literature, two types of human activity classification methods are identified: handcraft features extraction based methods and deep learning based methods.

Within the scope of the handcraft features extraction based methods, Moussa et al. [24] proposed a human activity classification method based on spatial features. They used the Scale Invariant Feature Transform (SIFT) algorithm to identify the points of interest. Subsequently, the authors constructed a Bag of Word (BoW) using the K-means algorithm, which assigns each descriptor generated from the points of interest to the nearest visual word. They calculated the frequency histogram of the visual words. Finally, the SVM classifier was applied in order to determine the human activity class. Opposite to Moussa et al. [24] method, Sabri et al. [25] demonstrated that the human activity classification can be improved by using the spatio-temporal features. Therefore, they calculated the Histograms of Oriented Gradients (HOG) as well as the Histograms of the Optical Flow (HOF) to construct the features vectors. Within the same vein of thoughts, Burghouts et al. [26], diverged from the method [25] by using the Spatio Temporal Interest Points algorithm, which proved to be more discriminating for human activity classification. In [27] the authors introduced a new method based on the human body skeleton. Indeed, in order to extract the activity features, they applied the cumulative skeletonized images matrix. These methods provide satisfactory performance in the context of human activity classification. However, they are highly dependent on the choice of the extracted discriminative features [28].

Within the framework of deep learning based methods, Baccouche et al. [28]

proposed a deep learning model for human activity classification that consists of two steps. The first step of the model includes a 3D-CNN allowing the extraction and the learning of spatio-temporal features. The second step of the model uses the features resulting from 3D-CNN to train a recurrent neural network Long Short Term Memory networks (LSTM) to classify the video sequence. As for WANG et al. [29], they proposed to use a CNN to extract the spatial features of each frame. Then, they applied two LSTM types with an attention model to explore the temporal features between consecutive video frames. The use of two LSTM types allows the exploration of both of the output results of the convolutional layer and the fully connected layer. Hence, they used a joint optimization layer to merge the two types of temporal features. Sargano et al. [30] demonstrated that the methods proposed in the literature based on deep learning (3D-CNN, LSTM) require an extensive dataset to perform the learning. Therefore, the use of one of the previously cited models is not adequate in the cases of small dataset, because CNNs are likely to be subject to overfitting. Therefore, Sargano et al. [30] proposed to use a pretrained deep neural network. Thus, they applied the AlexNet model and replaced the softmax layer with an SVM-KNN hybrid classifier to adapt the model to their context. The authors evaluated the performance of the proposed model on the KTH and UCF-Sport datasets. Experimentally, the authors proved that it is better to use pretrained deep neural network, than to start learning from scratch in order to extract and learn spatio-temporal features.

Through this subsection, we reviewed the methods of the human activity classification: handcraft features extraction based methods and deep learning based methods. Within the first category, we concluded that extracting spatio-temporal features perform better than spatial features. However, these methods generally require a prior extraction of the features and depend on the relevance of the used descriptor. In the scope of the deep learning based method, we note that the methods based on learning the spatio-temporal features have obtained satisfactory results. However, they require a large data set in the offline phase, since there is not any pretrained neural network. As for the methods based on

learning spatial features using pretrained deep neural networks, the obtained results were promising. Although they do not deal with temporal features, the obtained results exceed the state-of-the-art reference methods since it takes advantage of the knowledge of the pretrained model. Based on this study, we proposed a method of human activity classification based on a pretrained deep neural network. Different from Sargano et al. [30], we chose another pretrained model that responds to our study context. In addition, our proposed method offers two classification scenarios: an instant classification of the video frames and an entire classification of the video sequences.

3. Proposed approach

In the context of human activity recognition from UAV-captured video sequences, we proposed a new approach that tends to improve the performance of human detection as well as the human activity classification. This approach consists of two phases; an offline phase and an inference phase (*cf.* figure 2). Within these two phases, the scene stabilization preprocessing was applied in order to determine the potential motion regions in the scene. The offline phase generates the human/non-human and human activity models using pretrained convolutional neural networks. The inference phase allows, through the generated models, the detection of humans and the classification of their activities. In the following section, each of these two phases was explored.

3.1. Offline phase

The offline phase integrates three steps: the scene stabilization, the human/non-human model generation and the human activity model generation.

3.1.1. Scene stabilization

In our context of study, the scene stabilization preprocessing is performed in order to detect the potential motion regions in the scene. Therefore, the motion resulting from the UAV will not be taken into consideration.

In our approach, we proceed to scene stabilization preprocessing in order to

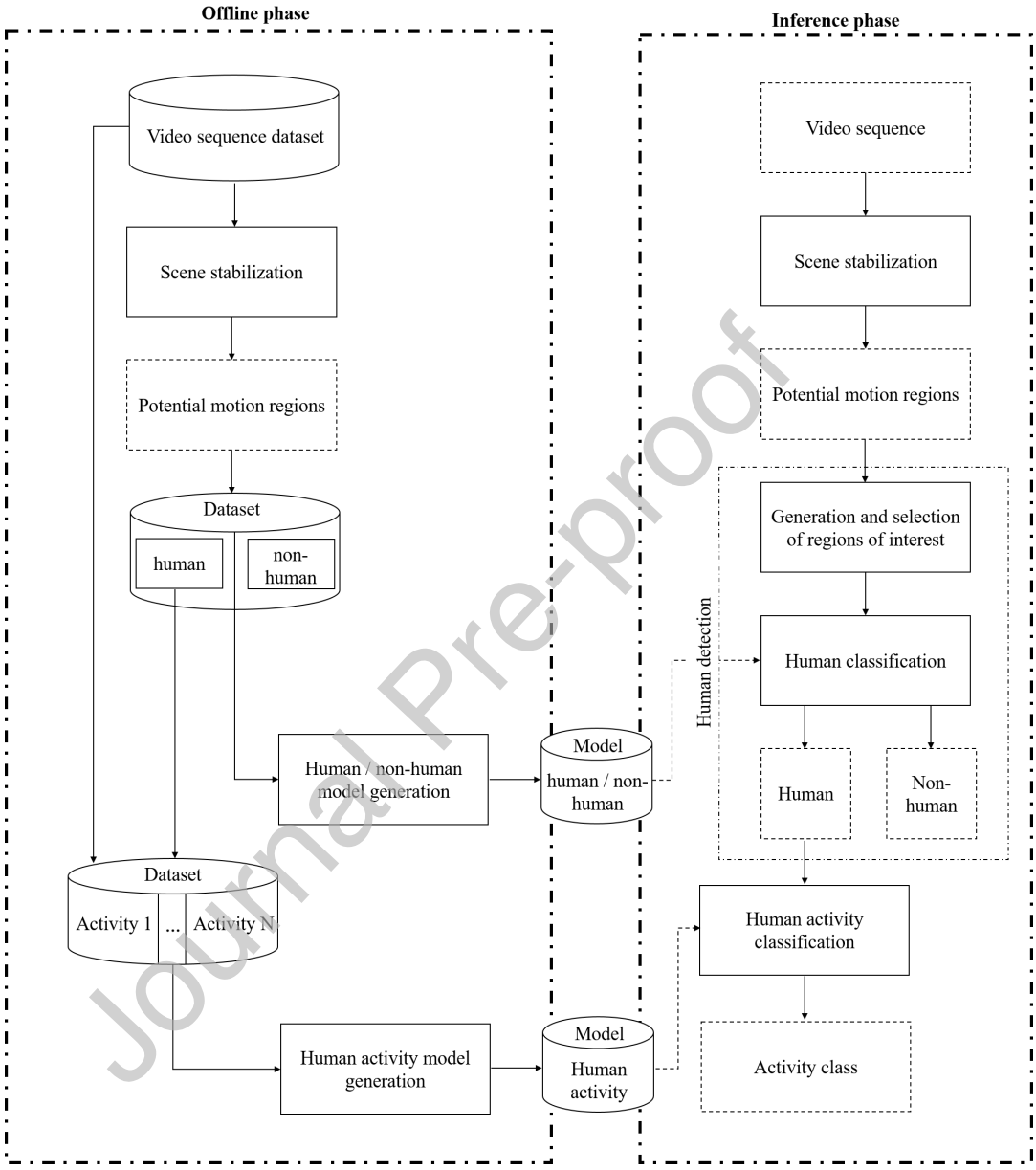

Figure 2: Proposed approach for human activity recognition from UAV-captured video sequences

identify the motions relative to the objects from the motions induced by the UAV. Therefore, the correspondence between two consecutive frames is computed relying on the following hypothesis: for a pixel of an image (Im_t), its corresponding in the image (Im_{t+dt})is in the same neighborhood and has the same color. Subsequently, during the motion estimation step, the image (Im_t) and the image (Im_{t+dt}) are transformed into grayscale images. The motion estimation is based on the assumption that the pixel intensity is constant between two consecutive images [18]. In order to determine the motion (dx, dy) of a pixel $I(x, y, t)$ after a time dt, the Taylor series was applied as follows:

$$I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \cdots \quad (1)$$

Assuming that the intensity of a pixel between two consecutive images is invariable, we get equation 2:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \quad (2)$$

Using equations 1 and 2, we can deduce equation 3:

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \cdots = 0 \quad (3)$$

By dividing equation 3 by dt, we compute the optical flow as follows:

$$-\frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial I}{\partial y}\frac{\partial y}{\partial t} \quad (4)$$

Where $\frac{\partial x}{\partial t} = V_x$ and $\frac{\partial y}{\partial t} = Vy$ are the field components of the optical flow $\overrightarrow{V}$, respectively, at the x and y coordinates. Thus, determining the optical flow for each pixel in the image sequence occurs through calculating the following

optical flow constraint equation:

$$-\frac{\partial I}{\partial t} = V_x \frac{\partial I}{\partial x} + V_x \frac{\partial I}{\partial y} \tag{5}$$

$$-I_t(p) = I_x(p)V_x + I_y(p)V_y \tag{6}$$

Equation 6 is made up of two known variables $I_x(p)$ and $I_y(p)$ and two unknown variables V_y et V_y. The determination of the two unknown variables V_x and V_y is at this step impossible since there is only one equation. To solve this problem, several methods such as Horn-Schunck [17], Lucas-Kanade [31], have been set forward. Although the Horn-Schunck method offers an entire solution for the optical flow calculation, it is based on an iterative calculation, which increases its computation time [18]. This temporal constraint is taken into consideration by Lucas-Kanade method, which assumes that all the neighbor pixels of a pixel p have similar motion [32]. Thus, Lucas-Kanades method is mostly suitable in our context of video surveillance. In the Lucas-Kanade method, all pixels belonging to the search window of size $(n \times n)$, of center p have the same motion as the pixel p [33]. According to this assumption, a system of n^2 equations was obtained:

$$\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \vdots & \vdots \\ I_x(p_{n^2}) & I_y(p_{n^2}) \end{bmatrix} \begin{bmatrix} V_x \\ V_y \end{bmatrix} = - \begin{bmatrix} I_t(p_1) \\ I_t(p_2) \\ \vdots \\ I_t(p_n^2) \end{bmatrix} \tag{7}$$

The system of equation 7 is an overdetermined system with more equations than unknowns. Thus, it can be solved using the least squares method [32, 34] by defining:

$$A = \begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \vdots & \vdots \\ I_x(p_{n^2}) & I_y(p_{n^2}) \end{bmatrix}, v = \begin{bmatrix} V_x \\ V_y \end{bmatrix} \text{ et } b = - \begin{bmatrix} I_t(p_1) \\ I_t(p_2) \\ \vdots \\ I_t(p_{n^2}) \end{bmatrix}$$

$$A^T A v = A^T b \tag{8}$$

$$v = (A^T A)^{-1} A^T b \tag{9}$$

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n^2} I_x(p_i)^2 & \sum_{i=1}^{n^2} I_x(p_i) I_y(p_i) \\ \sum_{i=1}^{n^2} I_x(p_i) I_y(p_i) & \sum_{i=1}^{n^2} I_y(p_i)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{i=1}^{n^2} I_x(p_i) I_t(p_i) \\ -\sum_{i=1}^{n^2} I_y(p_i) I_t(p_i) \end{bmatrix} \tag{10}$$

Further to the motion estimation of each pixel in the image, we obtain a set of motion vectors resulting from both of the UAV displacement and the potential motion regions in the scene. To differentiate these two motions, the speed of the motion regions is assumed to be greater than the speed of the motion relative to the mobile acquisition platform [18] (*cf.* figure 3 (a)). Thereafter, the motion vector norms are calculated for potential motion regions. Hence, a new image is obtained where each pixel receives an intensity value estimated by the computation of the norm of its motion vector divided by the maximum value of the calculated norms. Thus, an image whose pixels vary between [0..1] is obtained (*cf.* figure 3 (b)). This image is transformed into a binary image in

order to apply morphological operations (*cf.* figure 3 (c)). In order to eliminate the noise present in the image, the regions with a minimal number of pixels are removed (*cf.* figure 3 (d)). Then, morphological operations are applied (*cf.* figure 3 (e)). Finally, we can detect potential motion regions (*cf.* figure 3 (f)). Figure 3 illustrates the process of potential motion regions detection.

3.1.2. Human/non-human model generation

The generation of the human/non-human model is performed using a convolutional neural network. A pretrained model is applied as it is more promising in terms of performance than a model built from scratch [16]. Several pretrained convolutional neural networks have been proposed such as: AlexNet [35], ZFNet [36], VGGNet [37], GoogLeNet [38], ResNet [39]. Referring to AlDahoul et al. [16], we chose the pretrained convolutional neuronal network AlexNet which allows identifying the class of an image among 1000 categories. This choice is justified by the fact that human detection is considered as a binary classification issue (human/non-human) that does not require complex convolutional neural networks consisting of an important number of layers. Since the pretrained convolutional neuronal network AlexNet is composed of five convolutional layers followed by three fully connected layers, it is more suitable in our context of study. In contrast to AlDahoul et al. [16] who applied an SVM classifier, we adapted the AlexNet model by substituting the last classification layer by a new softmax layer in order to classify the objects into two categories, human and non-human.

3.1.3. Human activity model generation

To extract the human activities model, an automatic extraction of spatial features was performed. The extraction and learning of these features were achieved using a pretrained model. The choice of the pretrained convolutional neural network was based on a comparative study carried out by Kaiming et al. [39]. The authors compared the performance of pretrained CNNs (AlexNet [35],

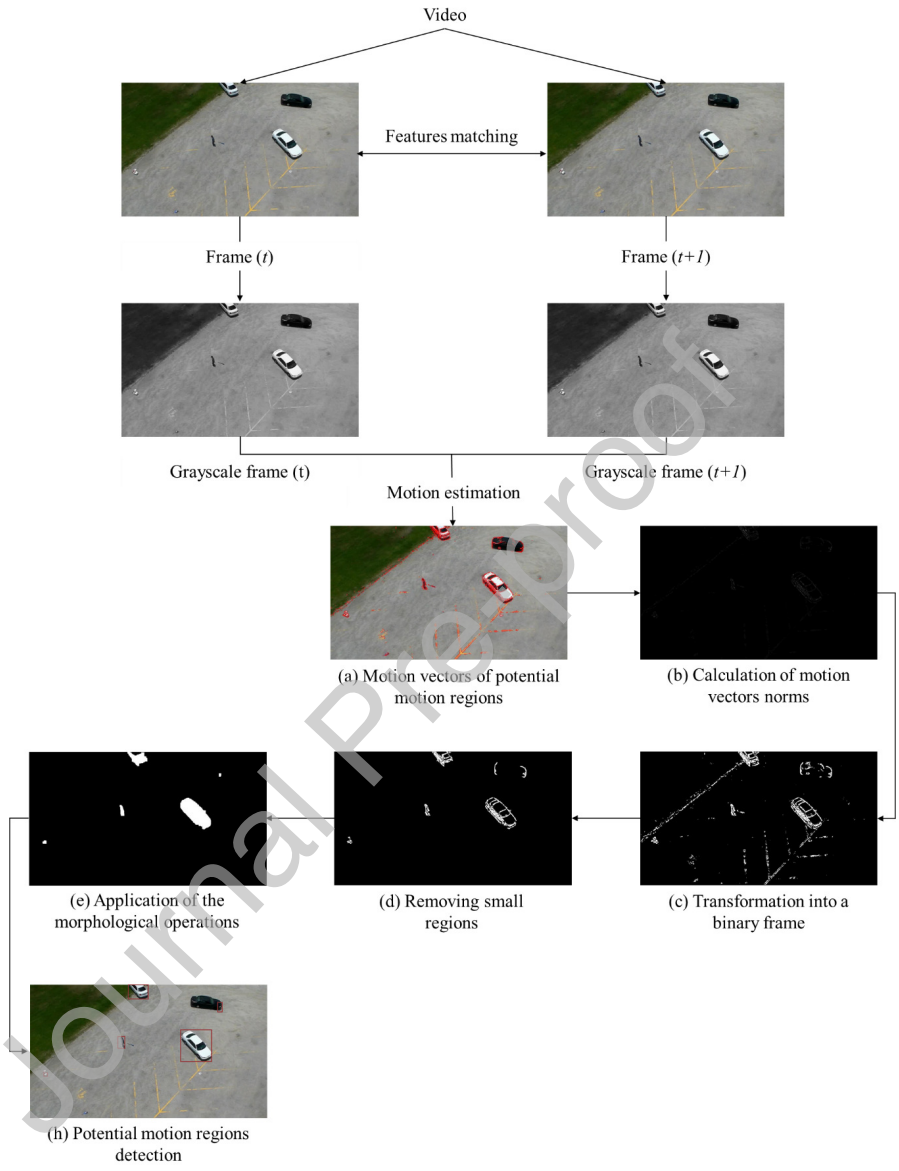

Figure 3: Process of potential motion regions detection

ZFNet [36], VGGNet [37], GoogLeNet [38], ResNet [39]) in terms of classification error rate and number of layers that affect the computation time. Based on this study, we noticed that ResNet has the lowest classification error rate, followed by

GoogLeNet. However, the number of layers in ResNet is seven times larger than GoogLeNet. According to this observation, we chose to proceed with GoogLeNet as it gives a good compromise between computation time and classification error rate. In addition, our choice is supported by the fact that GoogLeNet incorporates 9 Inception Modules which include convolutions at different sizes allowing the learning of features at different scales (*cf.* figure 4).

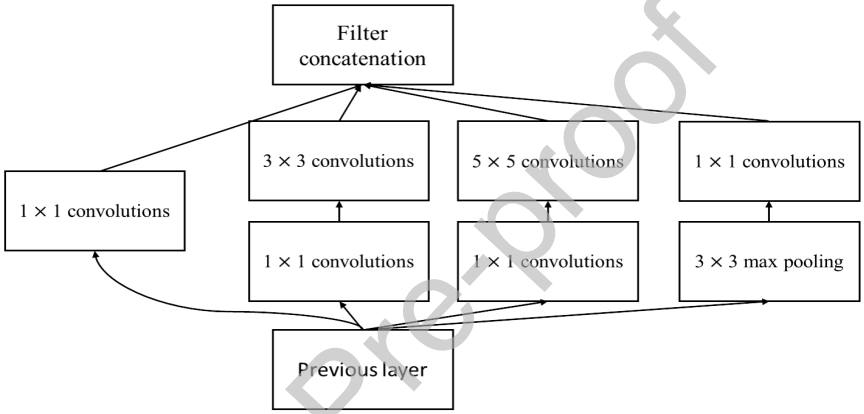

Figure 4: Inception module [38]

Furthermore, we note that in the GoogLeNet architecture, the penultimate Fully Connected Layer (FCL) was replaced by a pooling layer, having the Global Average Pooling as a strategy. Such a strategy reduces the size of the Feature Maps from $(n \times n \times nc)$ to $(1 \times 1 \times nc)$, where nc is the channel size of the input feature maps. Therefore, the total number of parameters is reduced, which decreases the computation time. In order to adapt GoogLeNet to the human activity classification problem, we replaced the softmax layer of the pretrained model by another softmax layer. The neurons number in the new softmax layer is equivalent to the required number of human activities. At the end of this step, we obtained a model that describes each of the human activities.

3.2. Inference phase

In the inference phase, we start by a scene stabilization to detect the potential motion regions. These regions are used as an input to the human detection step. Our contribution to the human detection method, consists in integrating the regions of interest generation in the selection module. Such a module has allowed the convolutional neural networks dedicated to classification tasks to address the problem of detection within a non-annotated dataset and consequently to separately locate close objects. In order to identify the human activity class, we classified it according to two scenarios: an instant video frames classification and an entire video sequence classification. In the following section, the human detection step as well as the human activity classification step were explored since the scene stabilization step was already investigated in section (3.1.1.).

3.2.1. Human detection

The module of the regions of interest generation and selection helps adapt the classic CNN to address the detection problem. We generate the regions of interest in order to classify those coming from a potential motion region rather than those the whole potential motion region. This process allows handling the problem of close up objects. Such objects usually appear in the same potential motion region (*cf.* figure 5) and are often classified as a single object.

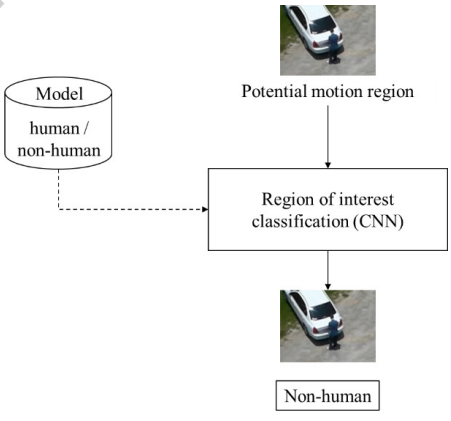

Figure 5: Human classification using a classic CNN

In literature, there is a wide range of methods of regions of interest generation such as: Edge boxes [40], Selective Search [41], BING [42], among others. Referring to the comparative study achieved by Zitnick et al. [40], we adopted the Edge boxes method which provides a good compromise between the recall rate and the computation time. In fact, the Edge boxes generates regions of interest from edges that provide a simplified but informative representation. This method is based on the hypothesis that the number of edges limited by a bounding box is proportional to the probability that this bounding box contains an object. The generated regions are then ranked according to a score calculated from the contours that are fully included in a bounding box. The Edge boxes method can generate various regions of interest for the same object. These RoIs vary slightly in position, shape or scale, which makes the selection step of the generated regions necessary. Accordingly, the NMS (Non Maximum Suppression) [40] algorithm is applied to elect the best generated regions of interest. The regions selected by NMS will be the input of the Human/non-human model generated during the offline phase. This model allows the extraction of features in order to classify these regions into two classes: human/non-human. Thereafter, the regions classified as a human are delimited in a 'bounding box'.

The regions of interest generation and selection module allow the classic CNN dedicated to classification tasks to address the problem of human detection. Figure 6 illustrates the advantage of such a method that not only treats and classifies close objects properly (human and/or car present in a potential motion region) but also detects (locates) humans.

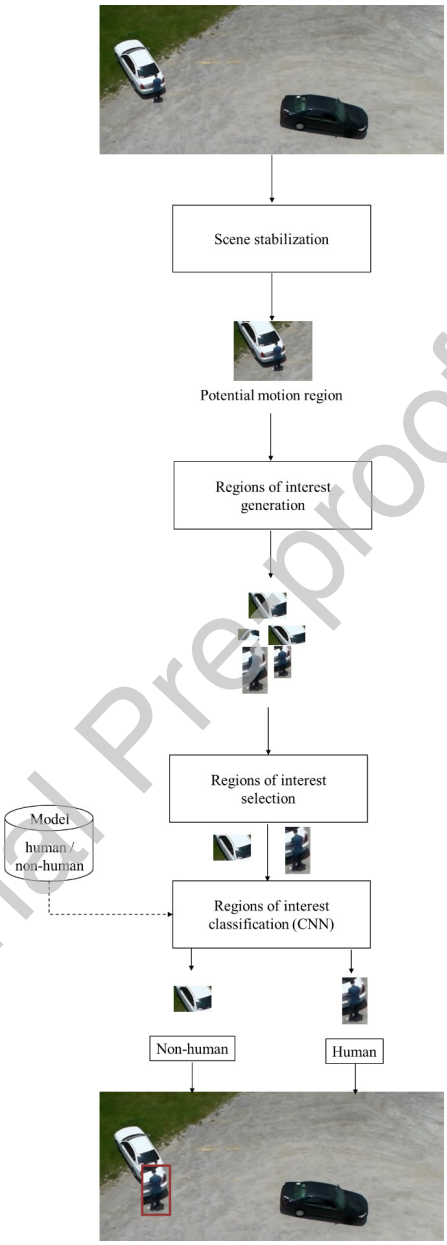

Figure 6: Proposed method for human detection using CNN

3.2.2. Human activity classification

The human activity model generated in the offline phase is used to extract the spatial features of a human activity and identify it. The classification is carried out through the softmax function, which converts the result of the last layer of the convolutional neural network into a vector of n real numbers. These numbers, whose sum is equal to 1, refer to the probability of each activity class. The activity class with maximum probability is assigned to the input human region. In this step, we propose a two-scenario classification method:

- An instant classification of each human region is performed in order to determine the activity class in each frame of the video sequence. Such a classification is suitable in video surveillance systems, where an alert is triggered in case of a suspicious activity. Figure 7 describes the instant classification process of a human activity.

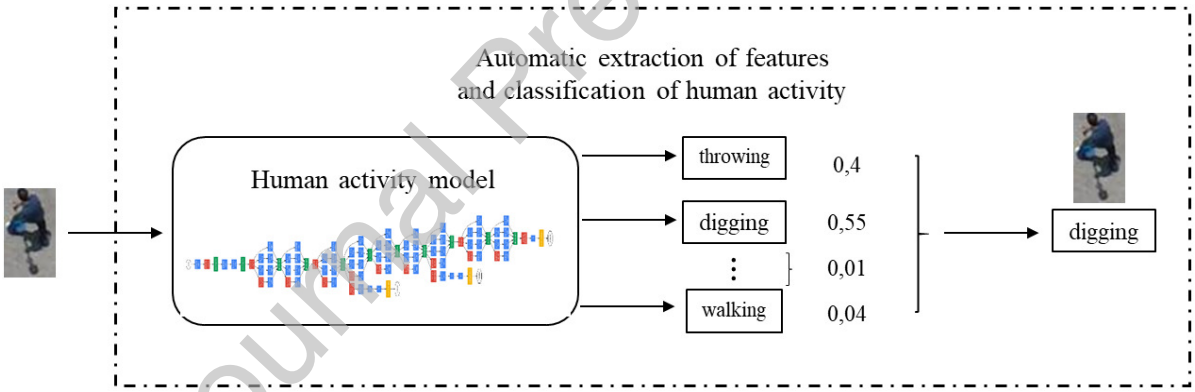

Figure 7: Instant human activity classification

- An entire classification consists of assigning a single activity label to the entire video sequence. Thereby, the activity class of each frame of the activity sequence is identified via an instant classification. Subsequently, the human activity sequence is assigned the label of the most frequent class activity. Figure 8 portrays this classification process.

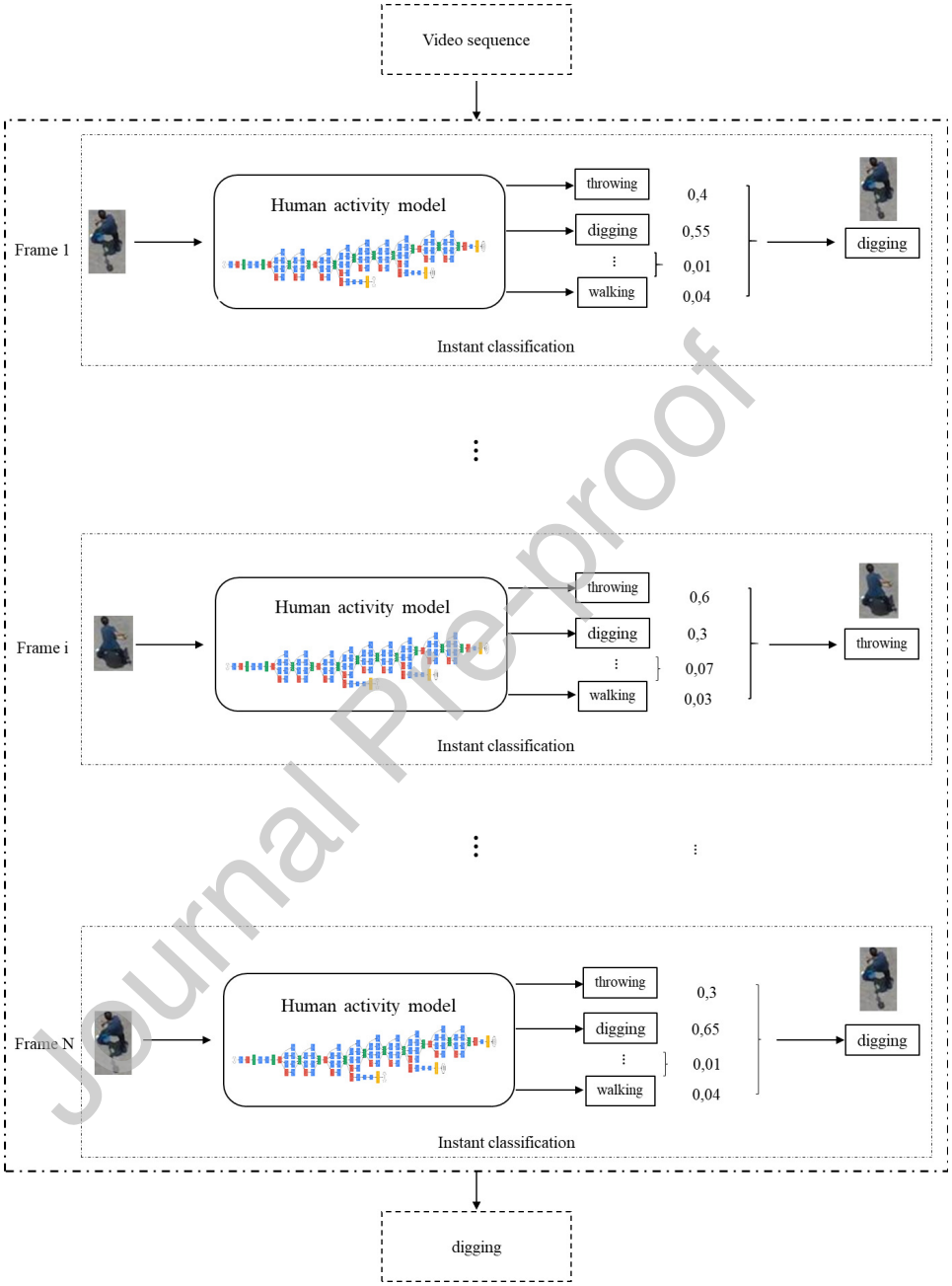

Figure 8: Entire human activity sequence classification

4. Experimental study

In this section, the performance of the proposed methods for human detection and human activity classification was assessed, the used dataset was described, and two series of experiments were carried out in order to compare and discuss the different obtained results. The first series of experiments evaluates the human detection method, while the second examines the performance of the human activity classification method.

4.1. Dataset Description

The evaluation of the proposed human detection and human activity classification methods was performed on the UCF-ARG dataset [23] captured by an aerial camera. The UCF-ARG proposes 10 human activities each of which is represented by 48 video sequences achieved by 12 people. This dataset is considered as one of the most complex datasets [10, 16, 43] because of the multiple constraints it exhibits such as: the severe ego-motion and UAV altitude variation, the point of view acquisition variation, the lighting conditions variation, and the intra and inter human activity-classes variability. In our study context and for a fair comparison with previous works, our experiments are in conformity with those of previous works [16, 43] focusing upon the following five activities: digging, throwing, running, walking and waving. Figure 9 displays the different classes of activities proposed in the UCF-ARG dataset.

4.1.1. First series of experiments: Human detection

For a fair evaluation of the proposed human detection method of Al-Dahoul et al. [16], we used a Leave-p-Out Cross-Validation (LpO CV), where $p = 4$ and defined a test configuration as follows {p1, p2, p3, p4} where p1, p2, p3 and p4 correspond to the 4 humans used in the test. Hence, the test configurations are as follows: {1, 2, 3, 4} {5, 6, 7, 8} {9, 10, 11, 12} {1, 3, 5, 7} {2, 4, 6, 8} {1, 4, 7, 10} {2, 5, 8, 11} {3, 6, 9, 12} {1, 5, 9, 12} {1, 6, 11, 12}. In this first series of experiments, our method was compared to the method proposed by AlDahoul et al. [16], (1) without and (2) with the regions of interest generation and

Figure 9: Samples from the UCF-ARG dataset: (a) digging, (b) throwing, (c) running, (d) walking, (e) waving

selection module, in terms of accuracy rate. Table 1 outlines this comparative study conducted on the UCF-ARG dataset.

Test set	AlDahoul et al. [16]	Proposed method	
		(1)	(2)
1_2_3_4	97.7170 %	99.7196 %	99.7286 %
1_3_5_7	97.8509 %	99.6975 %	99.6525 %
2_4_6_8	97.9344 %	99.7106 %	99.7056 %
1_4_7_10	98.4077 %	99.5909 %	99.6789 %
1_5_9_12	98.0556 %	99.7161 %	99.7079 %
2_5_8_11	98.0749 %	98.8032 %	98.8573 %
1_6_11_12	98.1257 %	99.6330 %	99.6031 %
3_6_9_12	98.5311 %	99.7968 %	99.8619 %
5_6_7_8	97.6115 %	99.5839 %	99.5036 %
9_10_11_12	98.5988 %	99.4356 %	99.5738 %
	98.0907 %	**99.5687** %	**99.5873** %

Table 1: Comparison of the performance of the proposed human detection method in terms of accuracy rate on the UCF-ARG dataset

Through this evaluation, we notice that the results of the proposed method (1) without the regions of interest generation and selection module outperform those obtained by AlDahoul et al. [16] in terms of accuracy rate. Such a result is justified by the fact that unlike AlDahoul et al. [16] who used an SVM classifier, we resorted to the use of the softmax layer during the classification step. Furthermore, we note that the average accuracy rate obtained by the proposed method (2) implementing the regions of interest generation and selection module overcomes the average accuracy rate obtained by the proposed method (1) as well as that of the AlDahoul et al. method [16]. These results are justified by the fact that the regions of interest generation and selection module correctly classifies the objects present in each potential motion region. For further evaluation, the classic rates of recall and precision were calculated on the UCF-ARG

dataset. Table 2 depicts this evaluation.

Test set	Proposed method			
	(1)		(2)	
	Recall	Precision	Recall	Precision
1_2_3_4	96.5597 %	98.7671 %	98.1196 %	97.5008 %
1_3_5_7	96.7343 %	98.9165 %	98.4007 %	97.0227 %
2_4_6_8	96.0756 %	98.5609 %	97.0967 %	97.7385 %
1_4_7_10	95.6740 %	98.2611 %	98.0787 %	97.1034 %
1_5_9_12	94.8131 %	99.5544 %	98.0614 %	96.8070 %
2_5_8_11	96.7913 %	91.7245 %	98.4309 %	90.9145 %
1_6_11_12	97.3259 %	98.1842 %	98.7619 %	96.6813 %
3_6_9_12	96.6720 %	99.2196 %	99.4835 %	98.1982 %
5_6_7_8	95.8946 %	98.8513 %	97.2435 %	96.5512 %
9_10_11_12	94.7707 %	97.2774 %	98.3658 %	96.3507 %
	96.1311 %	**97.9317** %	**98.2043** %	**96.4868** %

Table 2: Evaluation of the human detection method in terms of recall and precision rates on the UCF-ARG dataset

From Table 2, the improvement of the average recall rate highlights the contribution of using the regions of interest generation and selection module. Indeed, we record a gain equal to 2.0732% which is quite important in our context of video surveillance requiring the detection of all humans present in the coverage area of the UAV. However, our method (2) with the regions of interest generation and selection module has a slightly lower average precision rate (1.4449%) compared to that recorded in (1) without this module. Such a result does not affect the robustness of our method (2), as it would be better suitable in a context of video surveillance to trigger a false alert rather than triggering no alerts.

4.1.2. Second series of experiments: Human activity classification

For a fair comparison of our human activity classification method with that of Burghouts et al. [43], a Leave-One-Out Cross-Validation (LOO CV) was used, with $p = 1$. In this series of experiments, the proposed method of human activity classification was evaluated under the same experimental conditions as Burghouts et al. [43]. Thus, the accuracy rates obtained in the instant classification and the entire classification of the video sequences were calculated on the UCF-ARG dataset. Subsequently, the obtained results were compared to those recorded by Burghouts et al. [43] who apply an entire classification of the video sequence uniquely. Tables 3 and 4 illustrate this evaluation performed respectively by activity and by the test set.

Activity	Proposed method		Burghouts et al. [43]
	Instant classification	Entire classification	Entire classification
Digging	67 %	79 %	50 %
Throwing	51 %	69 %	33 %
Running	55 %	67 %	91 %
Walking	58 %	67 %	75%
Waving	51 %	56 %	33 %
	56 %	**68** %	**57**%

Table 3: Comparison of the performance of the human activity classification method per activity in terms of accuracy rate on the UCF-ARG dataset

Departing from this comparative study, we noted that the average accuracy rate of the proposed method applying an entire classification of the video sequence performs better than the method proposed by Burghouts et al. [43]. In fact, the pretrained CNN GoogLeNet allowed the extraction and learning of features at different scales and consequently the treatment of the intra-class and inter-class variability problems. The instant classification recorded lower rates than the entire classification rates. The average of its rates is comparable to

Test set	Proposed method		Burghouts et al. [43] Entire classification
	Instant classification	Entire classification	
1	62 %	65 %	
2	39 %	55 %	
3	60 %	75 %	
4	54 %	55 %	
5	70 %	85 %	
6	34 %	35 %	
7	52 %	60 %	
8	58 %	70 %	
9	47 %	60 %	
10	64 %	85 %	
11	59 %	75 %	
12	75 %	90 %	
	56 %	**68** %	**57**%

Table 4: Comparison of the performance of the human activity classification method per test set in terms of the accuracy rate on the UCF-ARG dataset

the method proposed by Burghouts et al. [43] who used a handcraft features extraction method. Although the instant classification is not as feasible as the entire one, it is more adequate to a video surveillance system.

5. Conclusion

The human activity recognition from UAV-captured video sequences stands as a promising field. In this context, we have proposed a new approach that consists of two phases: an offline phase and an inference one. Within these two phases, a scene stabilization based on the calculation of the optical flow was applied in order to detect the potential motion regions in the scene. In the offline phase, the human/non-human and the human activity models were generated

using pretrained convolutional neural networks. In the inference phase, the generated models were used to detect humans and classify their activities. Through the experimental evaluation of the proposed methods for the human detection and human activity classification on the UCF-ARG dataset, we managed to validate the performance of our methods compared to the existing ones. At this stage of analysis, we would assert that our research is just a step that might be further developed as it paves the way for different prospects for a constructive and fruitful contribution to the field. As future perspectives, we aspire to improve the human activity classification method using a probabilistic approach. In addition, we intend to tackle complex activities where people interact with one another and/or with objects.

References

[1] K. Huang, S. Wang, T. Tan, S. J. Maybank, Human behavior analysis based on a new motion descriptor, IEEE Transactions on Circuits and Systems for Video Technology 19 (12) (2009) 1830–1840.

[2] H. Wang, L. Wang, Learning content and style: Joint action recognition and person identification from human skeletons, Pattern Recognition 81 (2018) 23–35.

[3] H. Mliki, O. Arous, M. Hammami, Abnormal crowd density estimation in aerial images, Journal of Electronic Imaging 28 (1) (2019) 013047.

[4] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[5] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[6] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[7] K. Veon, Video stabilization using sift features, fuzzy clustering, and kalman filtering.

[8] J.-P. Hsiao, C.-C. Hsu, T.-C. Shih, P.-L. Hsu, S.-S. Yeh, B.-C. Wang, The real-time video stabilization for the rescue robot, in: 2009 ICCAS-SICE, IEEE, 2009, pp. 4364–4369.

[9] H. Shen, Q. Pan, Y. Cheng, Y. Yu, Fast video stabilization algorithm for uav, in: 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, Vol. 4, IEEE, 2009, pp. 542–546.

[10] A. Walha, A. Wali, A. M. Alimi, Video stabilization with moving object detecting and tracking for aerial video surveillance, Multimedia Tools and Applications 74 (17) (2015) 6745–6767.

[11] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[12] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.

[13] S. Minaeian, J. Liu, Y.-J. Son, Effective and efficient detection of moving targets from a uavs camera, IEEE transactions on intelligent transportation systems 19 (2) (2018) 497–506.

[14] V. Carletti, A. Greco, A. Saggese, M. Vento, Multi-object tracking by flying cameras based on a forward-backward interaction, IEEE Access 6 (2018) 43905–43919.

[15] A. Eltantawy, M. S. Shehata, An accelerated sequential pcp-based method for ground-moving objects detection from aerial videos, IEEE Transactions on Image Processing 28 (12) (2019) 5991–6006.

[16] N. AlDahoul, M. Sabri, A. Qalid, A. M. Mansoor, Real-time human detection for aerial captured video sequences via deep models, Computational intelligence and neuroscience 2018.

[17] B. K. Horn, B. G. Schunck, Determining optical flow, Artificial intelligence 17 (1-3) (1981) 185–203.

[18] S. D. Thota, K. S. Vemulapalli, K. Chintalapati, P. S. S. Gudipudi, Comparison between the optical flow computational techniques, International Journal of Engineering Trends and Technology 4 (10).

[19] P. Dollár, S. Belongie, P. Perona, The fastest pedestrian detector in the west.

[20] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: international Conference on computer vision & Pattern Recognition (CVPR'05), Vol. 1, IEEE Computer Society, 2005, pp. 886–893.

[21] R. Benenson, M. Mathias, R. Timofte, L. Van Gool, Pedestrian detection at 100 frames per second, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2903–2910.

[22] P. Viola, M. J. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance, International Journal of Computer Vision 63 (2) (2005) 153–161.

[23] A. Nagendran, D. Harper, M. Shah, UCF-ARG dataset, university of central florida, `http://crcv.ucf.edu/data/UCF-ARG.php` (2010).

[24] M. M. Moussa, E. Hamayed, M. B. Fayek, H. A. El Nemr, An enhanced method for human action recognition, Journal of advanced research 6 (2) (2015) 163–169.

[25] A. M. Sabri, J. Boonaert, S. Lecoeuche, E. Mouaddib, Caractérisation spatio-temporelle des co-occurrences par acp à noyau pour la classification des actions humaines, in: GRETSI'13, 2012.

[26] G. J. Burghouts, K. Schutte, Spatio-temporal layout of human actions for improved bag-of-words action detection, Pattern Recognition Letters 34 (15) (2013) 1861–1869.

[27] H. Mliki, R. Zaafouri, M. Hammami, Human action recognition based on discriminant body regions selection, Signal, Image and Video Processing 12 (5) (2018) 845–852.

[28] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: International workshop on human behavior understanding, Springer, 2011, pp. 29–39.

[29] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, J. Wu, Human action recognition by learning spatio-temporal features with deep neural networks, IEEE Access 6 (2018) 17913–17922.

[30] A. B. Sargano, X. Wang, P. Angelov, Z. Habib, Human action recognition using transfer learning with deep representations, in: 2017 International joint conference on neural networks (IJCNN), IEEE, 2017, pp. 463–469.

[31] B. D. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision.

[32] N. Sharmin, R. Brad, Optimal filter estimation for lucas-kanade optical flow, Sensors 12 (9) (2012) 12694–12709.

[33] J. L. Barron, D. J. Fleet, S. S. Beauchemin, Performance of optical flow techniques, International journal of computer vision 12 (1) (1994) 43–77.

[34] A. Khobragade, K. Kulat, C. Dethe, Motion analysis in video using optical flow techniques, International Journal of Information Technology and Knowledge Management 5 (1) (2012) 9–12.

[35] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[36] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.

[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[40] C. L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: European conference on computer vision, Springer, 2014, pp. 391–405.

[41] J. R. Uijlings, K. E. Van De Sande, T. Gevers, A. W. Smeulders, Selective search for object recognition, International journal of computer vision 104 (2) (2013) 154–171.

[42] M.-M. Cheng, Z. Zhang, W.-Y. Lin, P. Torr, Bing: Binarized normed gradients for objectness estimation at 300fps, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 3286–3293.

[43] G. Burghouts, A. van Eekeren, J. Dijk, Focus-of-attention for human activity recognition from uavs, in: Electro-Optical and Infrared Systems: Technology and Applications XI, Vol. 9249, International Society for Optics and Photonics, 2014, p. 92490T.

Hazar Mliki received her PhD in computer science from the University of Sfax in Tunisia. She is a researcher in the MIRACL Laboratory (Multimedia, Information Systems, and Advanced Computing Laboratory). She is currently an assistant professor at the National School of Electronics and Telecommunications (ENETCOM) of Sfax, Tunisia. Her research interests lie in Computer Vision, Pattern Recognition, and Machine Learning.

Fatma Bouhlel graduated with her masters degree in computer science from the University of Sfax, Tunisia. She is a researcher in the MIRACL Laboratory (Multimedia, Information Systems, and Advanced Computing Laboratory). Currently, she is preparing for her PhD at the University of Sfax, Tunisia (FSEGS). Her research interests lie in Computer Vision, Video Surveillance, Pattern Recognition, and Machine Learning.

Mohamed Hammami received his PhD in computer science from the Ecole Centrale at the Lyon Research Centre for Images and Intelligent Information Systems associated with the French research institution CNRS as UMR5205. He is currently an associate professor in the Computer Science Department at the Faculty of Science Sfax- Tunisia. He is a researcher in the MIRACL Laboratory (Multimedia, Information Systems, and Advanced Computing Laboratory). His current research interests lie in data mining and knowledge discovery in images and video, multimedia indexing and retrieval, face detection and recognition, and website filtering. He was a staff member at the RNTL-Muse project. He has served on technical conference committees and as reviewer for many international conferences.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: