Article

Artificial Intelligence and Convolutional Neural Network for Recognition of Human Interaction by Video from Drone

Ghazal Shamsipour and Saied Pirasteh*

Department of Surveying and Geoinformatics, Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University (SWJTU), the Western Park of the Hi-Tech Industrial Development Zone, Chengdu, Sichuan 611756, China; gh.shamsipour@gmail.com (G.SH.); sapirasteh@swjtu.edu.cn (S.P.)

* Correspondence: sapirasteh@swjtu.edu.cn

Abstract: Recognition of the human interaction on the unconstrained videos taken from cameras and remote sensing platforms like a drone is a challenging problem. This study presents a method to resolve issues of motion blur, poor quality of videos, occlusions, the difference in body structure or size, and high computation or memory requirement. This study contributes to the improvement of recognition of human interaction during disasters such as an earthquake and flood utilizing drone videos for rescue and emergency management. We used Support Vector Machine (SVM) to classify the high-level and stationary features obtained from Convolutional Neural Network (CNN) in key-frames from videos. We extracted conceptual features by employing CNN to recognize objects from first and last images from a video. The proposed method demonstrated the context of a scene, which is significant in determining the behaviour of human in the videos. In this method, we do not require person detection, tracking, and many instances of images. The proposed method was tested for the University of Central Florida (UCF Sports Action), Olympic Sports videos. These videos were taken from the ground platform. Besides, camera drone video was captured from Southwest Jiaotong University (SWJTU) Sports Centre and incorporated to test the developed method in this study. This study accomplished an acceptable performance with an accuracy of 90.42%, which has indicated improvement of more than 4.92% as compared to the existing methods.

Keywords: drone video; human action recognition; CNN; Support vector machine (SVM)

1. Introduction

Thousands of people are killed because of disasters annually. The research gap is on how the technology and advanced machines help to rescue people and survive human's life in dangerous situations. Smartest machines do not have enough vision technology to understanding and managing the content of videos, because they are blind. In this regard, researchers have tried to use computer vision techniques to empower devices [1-6]. Therefore, the present study aims to contribute a method of artificial intelligence and CNN for recognition of human interaction by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

video from drone where possibly apply in disasters management.

In many applications, including disaster, we utilize artificial intelligence and computer vision. One of the essential goals is the development of automation machines that can analyze and understand their environmental interactions with a human. These smartest machines have been taught to see as a human to understanding objects and surroundings. For instance, to understand a scene, they need to identify human, the geospatial location of objects, background clutter, and understanding of human behaviour. They may use video from camera drones or normal cameras and smartphones. Among the mentioned above, the recognition and understanding of human behaviour for surveillance, control, and security have received special attention from the computer vision at the time of the birth of this subject [7-9]. The importance of this application is on how we detect human or objects. In this regard, several algorithms have been proposed to detect and recognize human behaviour [10-13].

Recognition of human behaviour refers to identifying a video label according to human action, and it has been started from a couple of decades ago [7]. In general, the process can be divided into three steps: (i) Detection, including segmentation and feature extraction, (ii) tracking, and (iii) understanding action includes classifier and action label. A simple system of a teaching process overview depicts in Figure 1. Generally, the video frames are taken by a camera such as an iPhone or a camera drone. Further, the segmentation performs on every single frame to identify the main area and the human body. Later, the human body is tracked in all frames to detect the trajectory of the human motion. Finally, by applying a classification, each trajectory match to an appropriate class and the label of action is produced [14,15]. This process requires analyzing and tracking human in all video frames. Hence, the consequence is big data with high computation therefore, it requires a large capacity memory for extracting features. Extraction of features requires a sophisticated technique of action understanding. Figure 1 illustrates features extraction which, plays a key role in the recognition of human behaviour methods.

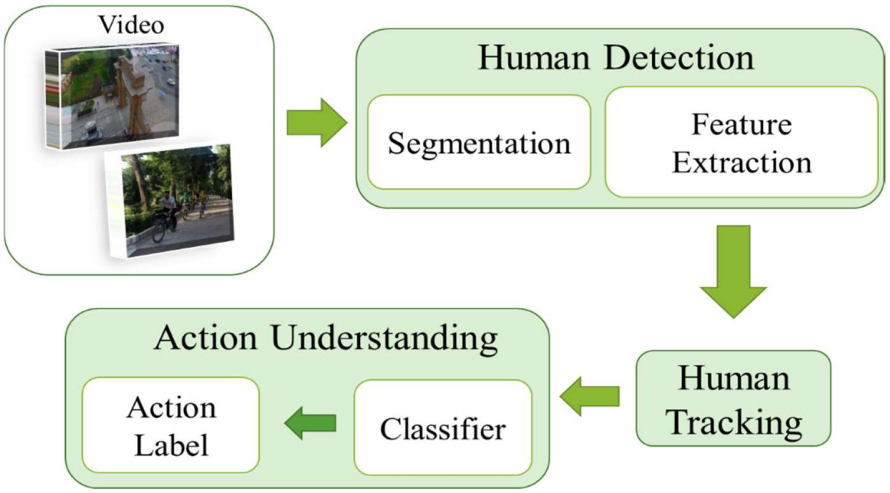


Figure 1: System overview

Now, according to the feature extraction, recognition of human behaviour can be categorized into classical models and deep models [16]. Classical models use a local feature, global feature, or combination of them to address the problem of human behaviour recognition. Global features describe the full image to generalize the entire object, whereas the local features describe the image patches of an object. Global features include contour representations, shape descriptors, and texture features and local features

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

represent the texture in an image patch. Although global features are suitable for image retrieval, object detection, classification, and identifying the human body; however, they are very sensitive to noise such as Gaussian noise and salt-and-pepper. Local features also divide the image into small pieces, which is very costly and has a very high computational process [17,18]. In other hands, deep models use neural networks to address the problem of human behaviour recognition. These models are successful in image analysis tasks like recognition and segmentation [16]. One of the successful deep models for image classification is CNN [19,20]. Nevertheless, most developed models cannot perform as classical models to use in videos processing and solve issues such as background clutter, occlusion, and diversity of the human body and structure. Furthermore, they are tested on datasets which are largely limited to take in constrained settings and shot by a professional camera under good brightness like Kungliga Tekniska högskolan (KTH) [9,21]. Table 1 illustrates several popular models and algorithms that are developed by researchers for human action recognition. These models are well-defined computations, and they are formed as a result of algorithms.

Table 1: Presents the results of methods

Author	*Year*	*Datasets*	*Methods*
Laptev et al.[22]	2008	KTH and Realistic samples	Employed spatial and temporal features and generalize spatial pyramids.
Wang et al.[23]	2009	KTH, UCF Sport, and Hollywood2	Evaluated and compare previously proposed local features.
Ikizler-Cinbis et al.[24]	2010	YouTube datasets	Proposed a multiple instance learning, where the training set is ambiguous and the training labels are associated with bags of instances.
Ji et al. [25]	2012	KTH, TRECVID 2008	Represented a 3D CNN which is an extension of 2D ConvNet in the uncontrolled environment for videos.
Simonyan et al.[26]	2014	UCF-101 and HMDB-51	Proposed a two-stream deep Convolutional Networks (ConvNet) architecture, which incorporates spatial and temporal variations to capture appearance and motion.
Ravanbakhsh et al.[16]	2015	UCF Sport, KTH, and UCF-11	Introduced a hierarchical structure, which enables to capture sub-actions from a complex action.
Wang et al.[27]	2015	HMDB51 and UCF101	Presented a representation model for a video called trajectory pooled deep-convolutional descriptor (TTD).
Chen et al.[28]	2016	MSR-Action3D dataset	Used depth motion maps (DMMs).
Liu et al.[29]	2017	Northwestern-UCLA, UWA3DII, NTU, and MSRC-12	Presented an enhanced skeleton visualization method with multi-stream CNN.
Rahmani et al.[30]	2017	IXMAS, UWA3D, N-UCLA, and UCF Sports	Proposed a Non-Linear Knowledge Transfer Model which is a a deep fully-connected neural network that transfers knowledge of human actions from any unknown view.
Kamel et al.[31]	2018	Microsoft action 3-D, Multimodal action, and the University of	Two input descriptors are used for action representation (depth motion and moving joints descriptor) and to better feature extraction for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

	Texas at Dallas-multimodal human action	classification, three CNN channels are trained.

Nevertheless, in this study, we developed an advanced method for human behaviour recognition in video employing CNN, where ImageNet trains the network. In this study, the human behaviour has been divided into four categories: (1) actions, (2) activities, (3) interactions, and (4) group activity [14]. Then we parsed human interactions into small actions through deep models. Moreover, besides the extraction of features and recognition of objects from videos in this proposed method, we also incorporated the potential of camera drone videos [32-34,5]. The use of drone vides is to find the human body. It has different challenges like motion blur, noise, and distance of drone from ground, and brightness. However, the potential of this study is to recognize the human body during disasters for rescue and emergency management. It will build a platform for future study on artificial geospatial intelligence (GeoAI) and smarter map. This study improved the performance of the previous algorithms, which was developed by Ravanbakhsh et al. (2015) [16]; Weinzaepfel et al. (2015)[35], Shamsipour et al. (2017)[15]. In this study, we utilized image processing techniques on videos taken from Da-Jiang Innovations (DJI) drone. We collected drone videos as well as the ground platform videos. These videos were captured by cameras and then employed a pre-processing method to enhance the image resolution in order to achieve better information. A pre-trained CNN extracts the vectors with conceptual features that specify the image objects. In this method, SVM determines the relationships between the frame objects and labels the videos. However, this study contributes (a) a new method and algorithm coded for recognition of human-object interaction. (b) Improving the performance method of action recognition with 4.92% accuracy as compared to the existing method mentioned above. (c) Developing the pre-processing component enhances the image resolution for better information, and (d) created opportunities to develop GeoAI for the smarter map and disaster emergency as well as rescue applications beyond in future studies.

Finally, this study explained the proposed method with simulation results and comparison with other techniques in Section 2. Section 3 presents the experimental running algorithm we coded on videos, and then it follows with results and validation. Section 4 presents results and discussion, and then Section 5 concludes the study with some recommendation for future activities.

2. Data and method

In this study, the authors used the dataset of Olympic sports [36], UCF Sports [37,38]. These videos are collected from the UCF (https://www.crcv.ucf.edu/data/UCF_Sports_Action.php)[39] and Olympic Sports (http://vision.stanford.edu/Datasets/OlympicSports) [40]. A video was also taken by the DJI drone from the SWJTU Sports Centre, Chengdu, China to test the proposed method. We used DJI camera drone with the following specifications (Table 2) to capture videos (Figure 2). This study compares the performance of the proposed method with the existing techniques applying for the same videos from UCF and Olympic Sports. In addition to this comparison, we applied the proposed method to a drone video to test the performance and potential of utilization for disaster management and rescue applications. The proposed method recognizes human behaviour by implementing CNN. We attempted three steps (1) selection of key-frames of videos, (2) creating an object representation, and (3) supervised classification and action labelling.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

Figure 3 describes the flowchart of processing steps in this study. Step (1) extract the video frame and select key-frames, step (2) employ adaptive windows and compute CNN features for two images to map a probability vector with 1000 elements. Step (3) Train SVM to predict the label of action.

Figure 2: DJI drone and onboard cameras

Table 2: Specification of drone used in this study

DRONE	
Dimensions	1668 mm × 727 mm × 727 mm with propellers frame arms and GPS mount unfolded (including landing gear) 437 mm × 402 mm × 553 mm with propellers, frame arms and GPS mount folded (excluding landing gear)
Package Dimensions	525 mm × 480 mm × 640 mm
Weight (with six TB47S batteries)	9.5 kg
Hovering Accuracy (P-GPS)	Vertical: ±0.5 m, Horizontal: ±1.5 m
Max Pitch Angle	25°
Max Speed	40 mph / 65 kph (no wind)
Operating Temperature	14° F to 104° F (-10° C to 40° C)
REMOTE CONTROLLER	
Operating Frequency	920.6 MHz to 928 MHz (Japan); 5.725 GHz to 5.825 GHz; 2.400 GHz to 2.483 GHz
Video Output Port	HDMI, SDI, USB
Operating Temperature	14°F to 104°F (-10° C to 40° C)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

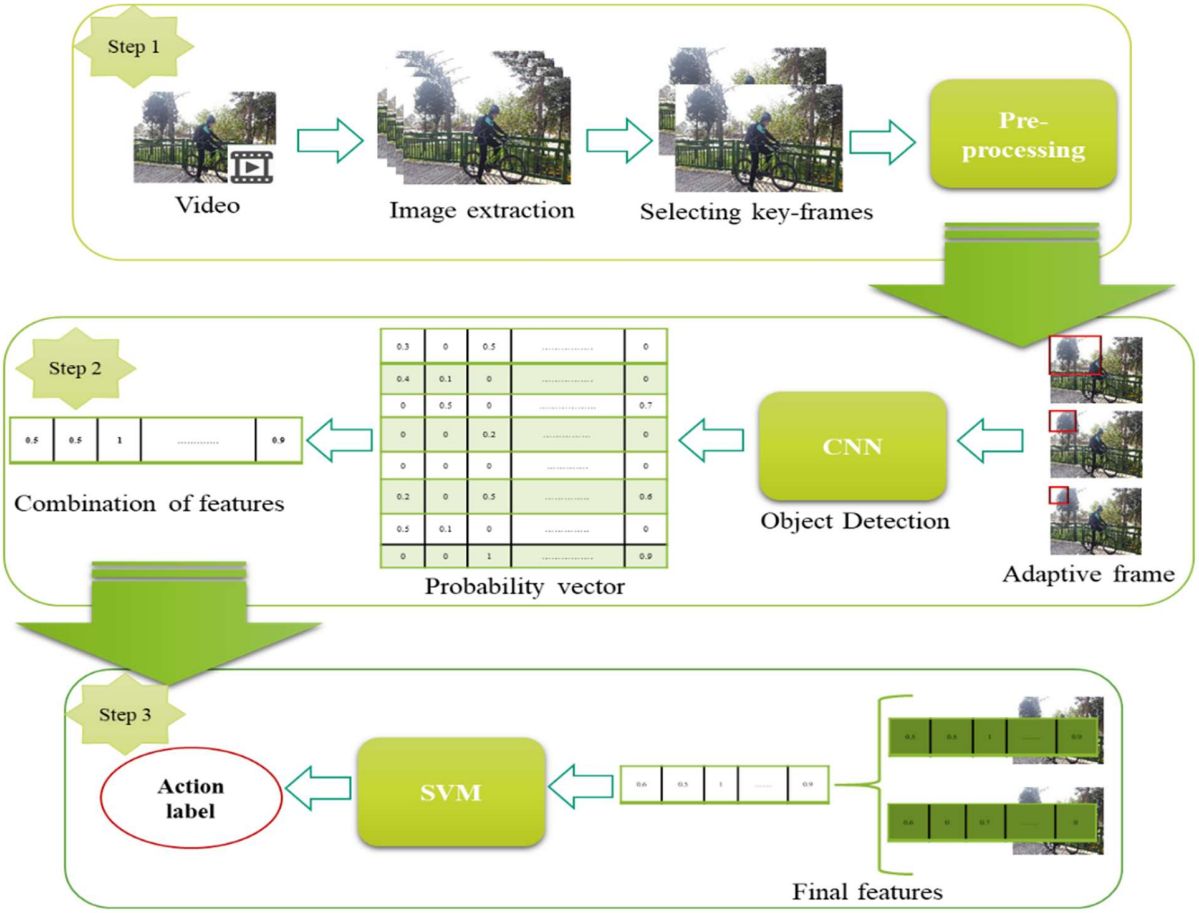


Figure 3: Overview of the method flowchart

2.1. Selection of key-frames of videos

We take raw videos of the camera as an input. The videos consist of temporal and spatial features. The most developed methods use temporal features to track human body and capture motion. Since we do not need human tracking in this proposed method, we eliminated the temporal features. As a result, we split a video into images with a high framerate and then extract the first and last image to find all objects in the video. It is because this process allows removing occlusion and preserving objects. The advantage of this selection without tracking human is the reduction of computation complexity. With this, the process continues to select each image from a video and to improve the resolution of images.

Although technology has been enhanced in the development of cameras and more efficient equipment are available, but sometimes the videos do not have a very good resolution. Converting images from a high resolution to low resolution is easy; however, it is very challenging on the contrary. A simple and very effective way of solving this issue is the super-resolution methods [41-43]. The purpose of these methods is to reconstruct the images with low quality into a high quality. The image redesign with super-resolution methods eliminates effects such as motion blur and noise (Figure 4). In the proposed technique, it is a pre-processing component that enhances the image resolution to acquire better information. This super-resolution learns a mapping between the low and high-resolution image. The mapping can be built from a hierarchical structure process [43]. It takes the low-resolution image as the input and generates the high-resolution image as an output product.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

Figure 4. Imaging system [43]

2.2. Creating image representation

In this step, the question is on how we represent the objects and human action in each image extracted from the video. To represent spatial features, we employed the pre-trained CNN with ImageNet [20]. Krizhevsky et al. (2017) [20] used CNN to classify 2.1 million high-resolution images into 1000 classes. It processed 400 images per second, and it detected the label of one object in each process of per-image. However, in the proposed method, we use CNN to generate a probability vector for each image. This probability vector represents the probability of the existence of the different object in each image. For example, if an element i of a vector $\vec{r}$ has value n for a soccer field image, we can say that the probability of the existence of an object x in the image is n.

The designed CNN by Krizhevsky et al. (2017) [20] has seven layers of fully connected. The last layer of the CNN is classification layer; however, in this study, we eliminated this layer and used the output of the previous layer as an image descriptor for recognition [44]. Instead of the discarded layer, the SVM has been applied to predict the label of human interaction. This layer is known as *fc7*, and we can track the changes in the *fc7* output to distinguish between survival human and the stuck one. We also applied Principle Component Analysis (PCA) to every window from the image (Figure 3) in each CNN process to decrease the dimensions of the probability vectors to 1×1000 dimensions [45,16].

Usually, an image consists of different objects with different sizes. These objects may be buildings, soccer field, trees, horses, and balls and even simple things such as a cat. They can present an infinite number of variations in structure or body shape. Also, an object from a video depends on the distance from the camera and drone shooting to object. However, humans appear in different dimensions. Here, we took the partition of each image by the adaptive filter on nine different windows and sizes. It processes the extraction of human body and object from step 2 of the flowchart process method (Figure 3).

We selected a window size with $\frac{1}{3} \times \frac{1}{3}$, $\frac{1}{3} \times \frac{1}{4}$, and $\frac{1}{3} \times \frac{1}{5}$ of the full image size (Figure 5). In the next stage, the window size is $\frac{1}{4} \times \frac{1}{3}$, $\frac{1}{4} \times \frac{1}{4}$, and $\frac{1}{4} \times \frac{1}{5}$, respectively (Figure 6). In the last processing stage, we selected the window size with a length of $\frac{1}{5} \times \frac{1}{3}$, $\frac{1}{5} \times \frac{1}{4}$, and $\frac{1}{5} \times \frac{1}{5}$, respectively (Figure 7).

This process of partition continues in the full image by applying CNN for every window mentioned

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

above size until the iteration of windows sizes partition are completed. In this case, big objects are detected on small windows, and small objects are detected on large windows. Each window overlaps with its neighbour with 50% to recognize and localize objects. Adaptive frame, CNN object detection, probability vector, and combined of features process apply on every extracted of the first and last image from the selection key-frames of video, separately.

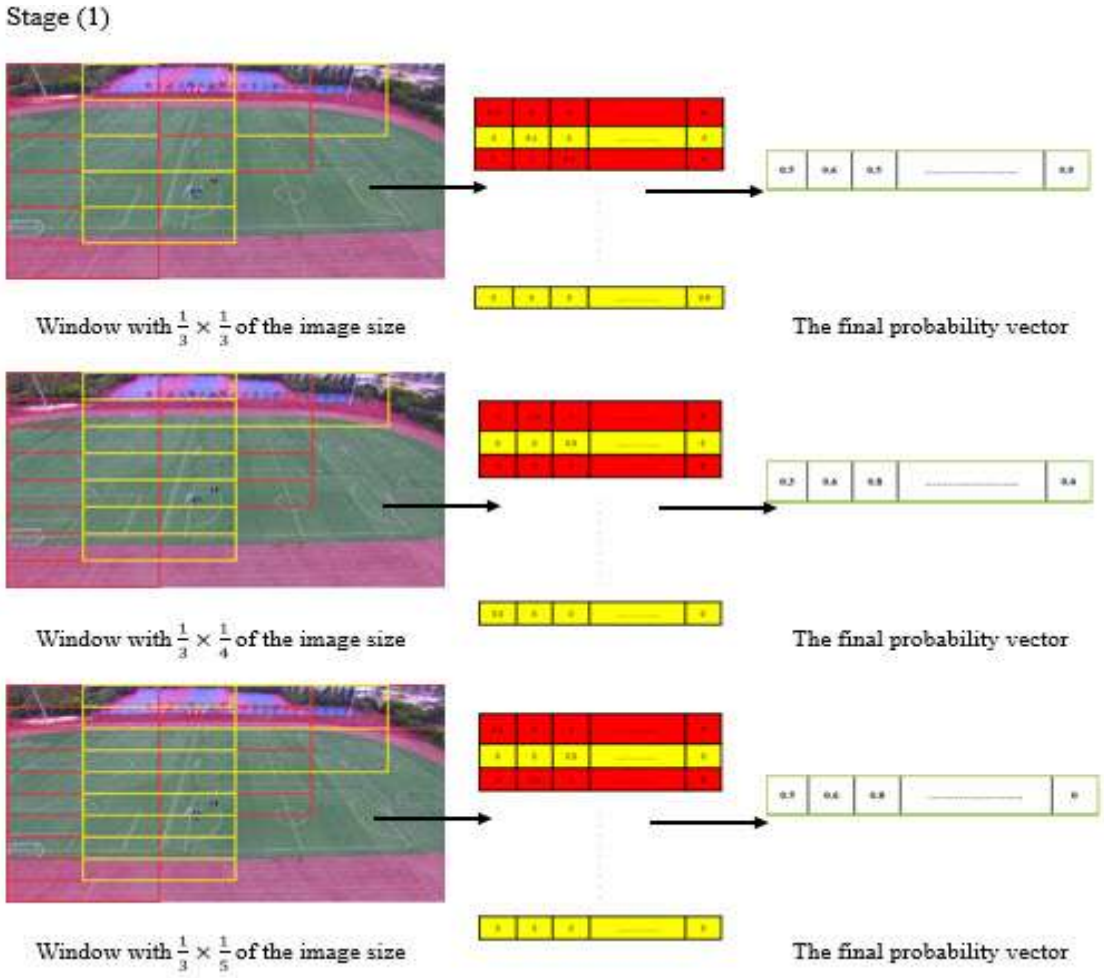


Figure 5: Process of image partition with $\frac{1}{3} \times \frac{1}{3}$, $\frac{1}{3} \times \frac{1}{4}$, and $\frac{1}{3} \times \frac{1}{5}$ window

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

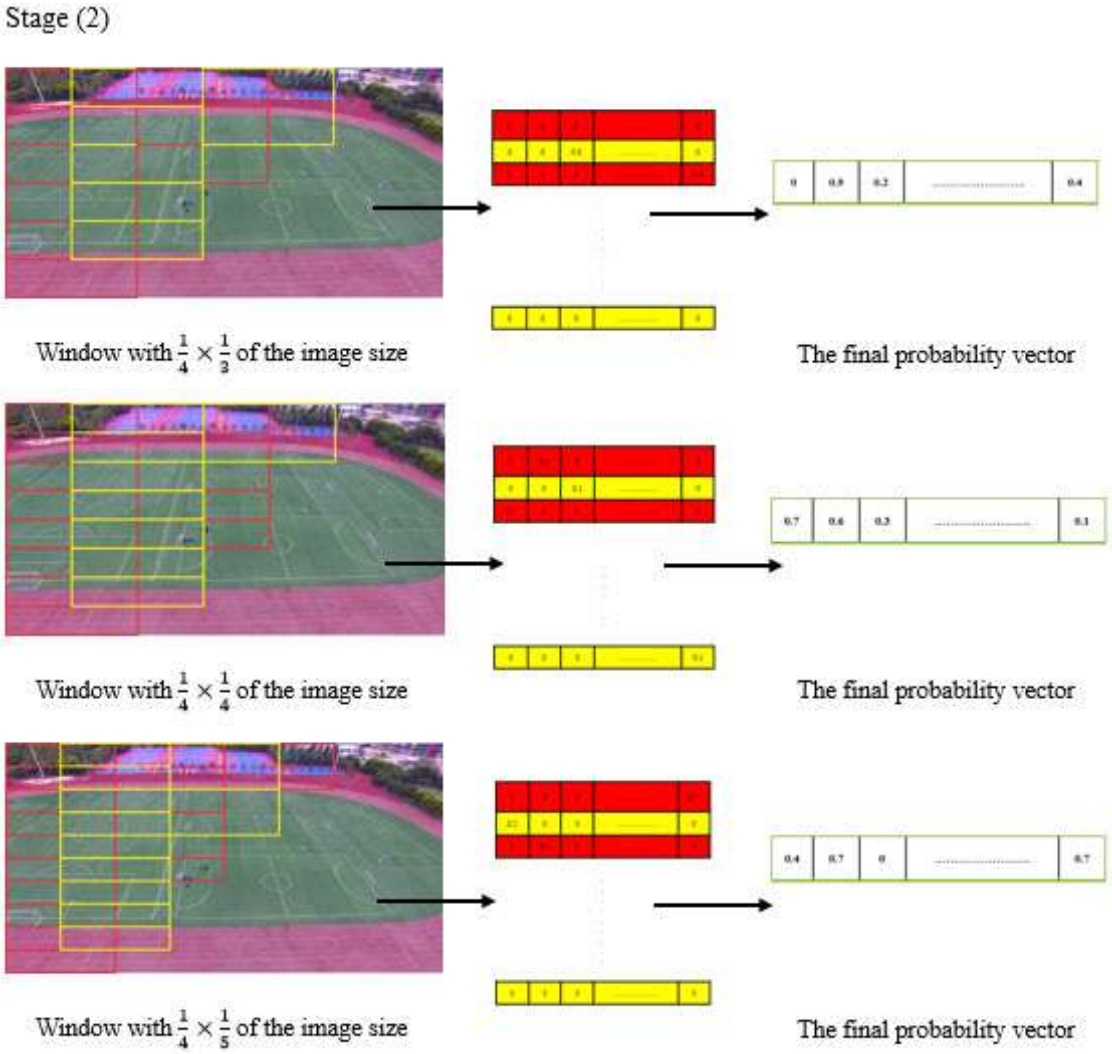


Figure 6: Process of image partition with $\frac{1}{4} \times \frac{1}{3}$, $\frac{1}{4} \times \frac{1}{4}$, and $\frac{1}{4} \times \frac{1}{5}$ window

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

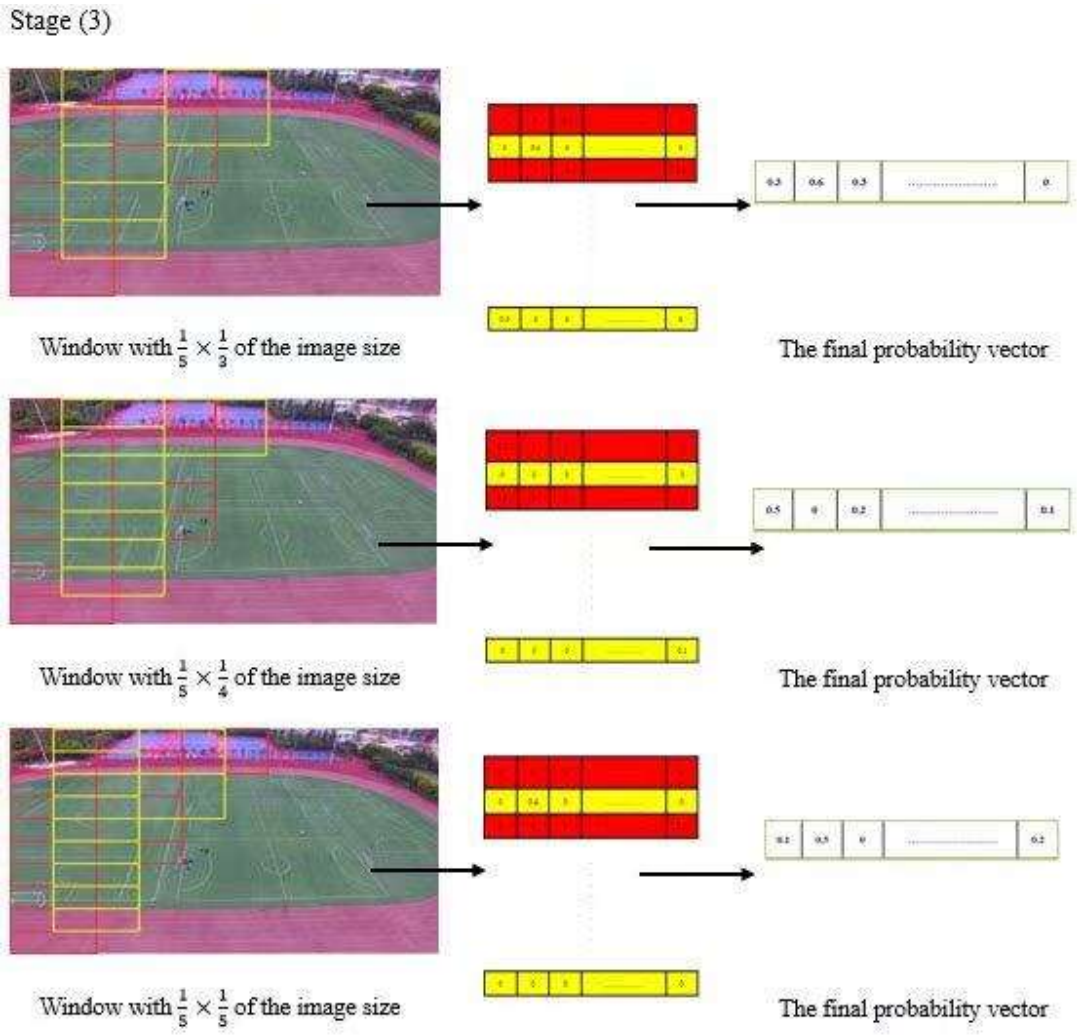


Figure 7: Process of image partition $\frac{1}{5} \times \frac{1}{3}$, $\frac{1}{5} \times \frac{1}{4}$, and $\frac{1}{5} \times \frac{1}{5}$ window

The output of the CNN for each window is a probability vector, and the desired result is the maximum probability vector. Therefore, in this process, we compared each vector with the previous vector to find out the maximum probability. For instance, for the trees detection scenario in DJI camera drone video if the probability vector in the first window shows up n% of probability and the second window shows the m% (i.e., where $m>n$) desired result is m% of the probability.

2.3. Supervised classification and action labeling

Building on the previous steps, this study recognized all objects on the images. We created a probability vector of length l for each image. For example, the position probability vector $\vec{r}$ of the image has units of length. Then we can define vector $r\hat{}$ *as* $r\hat{}=\vec{r}/|\vec{r}|$. Both $\vec{r}$ and $|\vec{r}|$ have units of length and therefore, $r\hat{}$ has units $l / l =1$, which means that it is dimensionless. Finally, to improve the efficiency of the supervised classification, the two probability vectors of the first and last image of the video frame are converted into a vector by union function.

When we extract feature vectors, a classification framework can be used in order to recognize the human interaction and types of actions in videos. To classify the data, we employed machine learning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

techniques. Machine learning techniques are categorized into supervised, non-supervised, and semi-supervised learning [46,47]. In semi-supervised learning, some of the data (X) is labelled (Y), and it sets between both supervised and non-supervised. In unsupervised learning, we only have input data (X) and no labelled or correct answers. Finally, in supervised learning, we provided the model with labelled data (Y) so that the model can learn the mapping function based on those label samples. Because the process of learning from the training data can be thought of as a supervisor, it is called supervised learning. Most of the practical machine learning have used supervised learning. An example of a supervised learning algorithm is SVM used in classification and regression [48,49,22]. For each group of data, the SVM generates two parallel lines. It separates the space in a single pass to generate flat and linear partitions. Divide the 2 categories by a clear gap that should be as wide as possible. The partitioning process was done by a plane called hyperplane. The SVM creates hyperplanes with the largest margin in high-dimensional space to separate given data into classes. To find the optimal hyperplane, the following equation can be used:

$$a.x+b=0 \qquad (1)$$

Where, a.x is the scalar product of a and x.

We used non-linear SVM to compute cluster and recognize human interaction in the videos [48,50]. SVM algorithms can use mathematical functions and they are defined as kernel. The function of the kernel is to take data as input and transforms it into the required form. The polynomial kernel equation (2) used in this study for image processing.

$$k(x_i, x_j) = (x_i . x_j + 1)^d \qquad (2)$$

Where d is the degree of the polynomial.

We trained primal SVM and used standard quadratic programming techniques. The dataset was split by assigning 70% of the videos to the training set, 10% to a validation set and 20% for testing purpose. The training data is used to learn the mapping function for SVM. The cluster sets its parameters with the training data that has a specific output. The validation data is used to minimize overfitting and evaluate the performance of the model in an unseen. The testing data is used to assess the performance of the SVM [47].

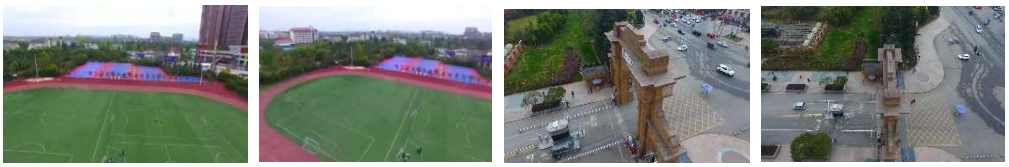
3. Experiments

As we mentioned before, the purpose of this study is to develop a method for extracting human interaction in videos, including camera drone videos for disasters management and rescue. For this, the images extracted from videos are partitioned into different sizes. Once the feature vector has been determined for each video, the label can be discovered by using SVM. In the proposed method, the authors have evaluated the accuracy of results as compared to other methods. For this, we applied the proposed algorithm on the Olympic and UCF Sports action and then described the transfer learning experiments on the drone videos from SWJTU Sports Center (Figure 8) to test the potential of the proposed algorithm for disaster management and rescue applications.

(a)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

(b)



(c)

Figure 8: The example of datasets. (a) UCF sports action, (b) Olympic sports action, and (c) Drone videos from SWJTU sports center. (Source for (a) and (b): https://www.crcv.ucf.edu/data/UCF_Sports_Action.php and http://vision.stanford.edu/Datasets/OlympicSports/) [39,40].

3.1. Dataset

The Olympic and UCF Sports action [51] consist of 150 videos annotated with more than 20 classes. The classes are arranged in a taxonomy that contains intra- and inter-class variation such as Team Sports, Winter Sports, and Ball Sports. For example, these datasets contain 14 different types of diving-side, 6 different types of lifting and 12 types of riding horse.

The drone video is taken by DJI camera drone to capture videos (see Table 2). We changed the resolution of these video from full HD (i.e. 1080p) into 240p, 320p, and 480p. Figure 8c depicts an example of these datasets.

3.2. An experimental setting

We designed the evaluation procedure in three cases. (1) Partitioned lengths, (2) overlapping windows, and (3) a number of images. The setup describes for each case in details as follows.

3.2.1. Partitioned lengths

The first experimental case is the lengths of the windows, which can recognize all objects with different sizes. In the partitioned technique, we employed the windows with nine different sizes for each image. According to the results we obtained, the best recognition of the objects is achieved on length by:

$$\frac{1}{3} \times \frac{1}{3}, \frac{1}{3} \times \frac{1}{4}, \text{and } \frac{1}{3} \times \frac{1}{5} \text{ windows}$$

$$\frac{1}{4} \times \frac{1}{3}, \frac{1}{4} \times \frac{1}{4}, \text{and } \frac{1}{4} \times \frac{1}{5} \text{ windows}$$

$$\frac{1}{5} \times \frac{1}{3}, \frac{1}{5} \times \frac{1}{4}, \text{and } \frac{1}{5} \times \frac{1}{5} \text{ windows}$$

3.2.2. Overlapping windows

The second experimental case is based on the rate of overlapping for each window. We applied three experiments to find the most appropriate size for overlapping. In this regard, we attempted 20%, 50%, and 80% of overlapping. According to the result and computation complexity, the best overlapping window is 50% pixels.

3.2.3. A number of images

The third experiment is based on the number of key-frame. The performance of recognition of human behaviour depends on the number of images which is selected among the video frames. Although with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

increasing the number of images, the performance of the proposed method rises; however, by increasing the number of images we found that the amount of memory requirement and complexity of the calculations are increased. Therefore, this study employed only two frames of each video (i.e., first and last images) for the recognition process as compared to the existing method using three images randomly [15]. In the experimental, we applied the first and last images of video with five more images. The experimental result shows in Table 3.

This experimental case aimed to evaluate the proposed method and to realize the fact that this method captures the information regardless of visual appearance and background clutter. Hence, we considered only visual information for feature extraction. We employed fixed-sized frames for each video (90-frames) for visual representation and then applied PCA to reduce the size of the feature. The first and last frame of each video indicates a key-frame. We applied several experiments to find the appropriate number of images. According to the experimental results, the best performance can be achieved from the first and last image of the video. The experiments were implemented in Matlab R2018b under Windows 10 and were run on an ASUS core i7 with CPU 2.0 GHz with 8 GB RAM. Figure 9 illustrates the confusion matrix of per action class in the Olympic and UCF Sports dataset [52].

Nevertheless, to support the proposed method, Table 4 illustrates the summary of the results of the comparison methods with the UCF and Olympic Sports datasets. Besides, Table 3 shows the results of the proposed approach on drone videos.

Table 3: The result of the proposed method on SWJTU drone video

Drone videos			
The resolution	Accuracy for 2 images %	Accuracy for 3 images %	Accuracy for 7 images %
240p	94.83	95.01	95.93
320p	95.51	95.74	96.71
480p	95.75	95.87	96.82

Table 4: Comparison of the proposed method on UCF Sports, Olympic Sports

UCF Sports		Olympic Sports	
Method	Accuracy %	Method	Accuracy %
Ravanbakhsh,et al.,2015 [16]	88.10	Gaidon, et al., 2014[53]	85.00
Weinzaepfel, et al.,2015[35]	90.50	Wang,, et al., 2016 [3]	85.80
Shamsipour et al., 2017 [15]	94.40	Shamsipour et al., 2017[15]	86.60
The proposed method	**93.67**	The proposed method	**90.42**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

	Diving	Golf-swinging	Kicking	Swinging on High Bar	Riding	Running	Swinging on Bench	Lifting	Skateboarding	Total	Recall
Diving	100	0	0	0	0	0	0	0	0	100	100%
Golf- swinging	0	92	3	0	0	3	0	0	2	100	92%
Kicking	0	2	96	0	0	2	0	0	0	100	96%
Swinging on High Bar	3	0	1	89	0	0	4	3	0	100	89%
Riding	0	0	0	0	99	0	1	0	0	100	99%
Running	0	3	5	0	1	90	0	1	0	100	90%
Swinging on Bench	3	0	2	0	0	5	87	03	0	100	87%
Lifting	0	0	0	0	0	0	0	100	0	100	100%
Skateboarding	0	0	0	1	0	0	0	0	90	100	90%
Total	106	97	107	90	100	109	92	107	92	900	
Precision	94.34%	94.84%	89.72%	98.89%	99%	82.57%	94.56%	93.46%	97.83%		

(a)

	Basketball Layup	Discus Throw	Hammer Throw	Shot Put	Bowling	Jump	Tennis Serve	Javelin Throw	Vault	Pole Vault	Dive	snatch	Total	Recall
Basketball Layup	97	0	0	0	0	0	3	0	0	0	0	0	100	97%
Discus Throw	0	80	4	9	0	2	0	3	2	2	0	0	100	80%
Hammer Throw	0	9	79	7	0	0	0	3	2	0	0	0	100	79%
Shot Put	0	4	7	82	0	0	0	4	3	0	0	0	100	82%
Bowling	0	0	0	0	100	0	0	0	0	0	0	0	100	100%
Jump	0	2	0	4	0	86	0	1	0	1	6	0	100	86%
Tennis Serve	0	0	0	0	0	0	100	0	0	0	0	0	100	100%
Javelin Throw	2	4	4	2	0	2	0	82	4	0	0	0	100	82%
Vault	0	4	0	2	0	1	0	2	90	1	0	0	100	90%
Pole Vault	0	0	0	0	0	6	0	0	5	89	0	0	100	89%
Dive	0	0	0	0	0	0	0	0	0	0	100	0	100	100%
snatch	0	0	0	0	0	0	0	0	0	0	0	100	100	100%
Total	99	103	94	106	100	97	103	95	106	91	106	100	1200	
Precision	97.98%	77.67 %	84.04%	77.36 %	100%	88.66 %	97.09 %	86.32 %	84.91 %	97.8 %	94.3 4%	100%		

(b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

	Walking	Standing	Play football	Riding bicycle	Total	Recall
Walking	97	3	0	0	100	97%
Standing	4	98	0	0	100	96%
Play football	1	1	98	0	100	98%
Riding bicycle	0	0	0	100	100	100%
Total	102	100	98	100	400	
Precision	95.10%	96%	100%	100%		

(c)

Figure 9: (a) A confusion matrix for UCF Sports datasets, and (b) a confusion matrix of Olympic Sports datasets. (c) a confusion matrix for SWJTU sports center drone video

3.3. Validation

We performed three validation approaches to evaluate the performance of the proposed method. The first validation is Recall, which is the fraction of relevant instances that have been retrieved over the total amount of related instances. Recall, also known as sensitivity, and it is calculated by the following equation [54]:

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

Where TP is True Positive, and FN is False Negative.

The second validation is the Precision, and it is the fraction of relevant instances among the retrieved instances [55,56]. Precision is also known as a positive predictive value, and the following equation calculates it:

$$\text{Precision}=\frac{TP}{TP+FP} \quad (4)$$

And the third validation is accuracy, which is calculated by the following equation:

$$(TP+TN)/(TP+TN+FP+FN) \quad (5)$$

Where the TP is the number of positive cases correctly identified, FP is the number of negative cases incorrectly classified as positive, and FN is the number of positive cases incorrectly classified as negative.

4. Results and discussion

This study compares the performance of the proposed method with the existing techniques applying in the same videos from UCF and Olympic Sports. Besides we apply the proposed method to a drone video to test the performance and potential of utilization for disaster management and rescue applications. This study attempted the partitioning process for images to increase the efficiency of identification of different object and sizes. This study shows that the best results of partitioning process and recognition of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 August 2019 doi:10.20944/preprints201908.0289.v1

objects are $\frac{1}{3} \times \frac{1}{3}$, $\frac{1}{3} \times \frac{1}{4}$, and $\frac{1}{3} \times \frac{1}{5}$ windows, $\frac{1}{4} \times \frac{1}{3}$, $\frac{1}{4} \times \frac{1}{4}$, and $\frac{1}{4} \times \frac{1}{5}$ windows, $\frac{1}{5} \times \frac{1}{3}$, $\frac{1}{5} \times \frac{1}{4}$, and $\frac{1}{5} \times \frac{1}{5}$ windows of the image size. The image which is captured with these windows is the input of CNN. It also shows that the appropriate overlapping percentage of the partition for images extracted from a video frame is 50%. This study reveals that we can use the first and last frame of each video because of less memory usage and processing complexity without reducing the efficiency of the performance.

The confusion matrix of the datasets from UCF Sports, Olympic Sports, and SWJTU videos implied to per action class indicates different accuracy for each detected human behaviour (Figure 9). When we apply confusion matrix to the drone video of SWJTU for a few human behaviours such as walking, standing, play football, and riding a bicycle, the accomplishment of accuracy is more significant than UCF Sports and Olympic Sports videos. The accuracy of each above-mentioned class of human behaviour for the drone video is 97%, 98%, 98%, and 100%. This confusion matrix indicates that the proposed method is a powerful approach to camera drone videos for recognizing human behaviours.

The proposed method applied to the UCF Sports videos indicates that the accuracy is 93.67%. It reveals a better result than two existing methods [16,35]. It also seems that the proposed method does not have a significant difference with the previous method [15]. Similarly, the proposed method applied to Olympic Sports videos reveals a significant improvement of accuracy to 90.42% as compared to the three mentioned existing methods indicated in Table 4. Moreover, the accomplishment of the proposed method in camera drone a video with 240p, 320p, and 480p resolution in 2 images, 3 images, and 7 images indicate acceptance and a better accuracy. Therefore, this study can be probably applied to improve drone images extracted from camera drone videos for various applications such as disaster management and rescue.

The result of the proposed method discusses the scene and the objects, which are highly influential on human behaviour. Therefore, by detecting objects and the scene, we solved issues such as background clutter, occlusion, and light. When we remove the tracking process and temporal features, we can reduce complicated computation and memory usage from videos on a highly challenging dataset during recognition of human behaviour. The results reveal that the use of learning methods can apply in the classification part and SVM. It also implies in the feature extraction when we utilized CNN, it improves the technique with 4.92% accuracy as compared to the previous study [15]. The performance of pre-trained CNN with eliminating the classification layer and utilization of the output of the last layer as an image descriptor generate a probability vector for each image. The SVM predicts the label of human interaction. This study achieved consolidated experiments of supervised pre-training. We expect that it will help obtaining enough computational power and increase the amount of labelled data in the videos and images significantly.

5. Conclusions and recommendation

In this study, we introduced a technique for recognizing human behaviour in videos that outperforms methods in two benchmarks and drone videos. The feature representations are achieved by the hierarchical structure of pre-trained CNN, where the classification layer is eliminated, and the output of the previous layer is used as an image descriptor. Inspired by object recognition, we introduced a method where it determines merely human behaviour. It is based on the recognition of objects and a scene without tracking human in all video frames. Furthermore, the experimental results concluded that if we employ a

learning method in the feature extraction part, we can improve the performance by more than 4.92% on action recognition accuracy as compared to other reported methods. The authors' findings can probably advance the field of research to build infrastructure for creating a smarter map of human-object interaction in future studies implementing the GeoAI by videos and images. Finally, we recommend using very large windows sizes from each drone video frame. It allows the spatial structure to provide helpful information for disaster management and rescue, that is, missing or far less detectable in still images.

Author Contributions: Conceptualization, G.Sh. and S.P.; Data curation, G.Sh. and S.P.; Formal analysis, G.Sh.; Funding acquisition, S.P.; Investigation, S.P.; Methodology, G.Sh. and S.P.; Project administration, S. P.; Resources, S.P.; Software, G.Sh. and S.P.; Supervision, S.P.; Validation, G.Sh.; Visualization, S.P.; Writing – original draft, S.P. and G.Sh.; Writing – review & editing, S.P and G.Sh.

Funding: This research received funds for publication of this research outcomes from the start up funds, Faculty of Geosciences and Environmental Engineering (FGEE), Southwest Jiaotong University (SWJTU), China.

Conflicts of Interest: The authors declare no conflict of interest.

ORCID ID

Saied Pirasteh http://orcid.org/0000-0002-3177-037

References

[1] Hofmann, T. Unsupervised learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, **2001**, 42, 177-196.

[2] Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep Learning for Visual Understanding: A review, *Neurocomputing*, **2016** ,187, 27–48.

[3] Wang, H.; Oneata, D.; Verbeek, J.; Schmid, C. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, **2016**, 119(3), 219-238.

[4] Knauf, K.; Memmert, D.; Brefeld, U. Spatio-temporal Convolution Kernels, *Machine Learning*., **2016**,102, 247–273.

[5] Shih, HC. A survey on content aware video analysis for Sports, *IEEE Transactions on Circuits and Systems for Video Technology*, **2017**, 99(9):1212 – 1231. DOI: 10.1109/TCSVT.2017.2655624.

[6] Laaribi, A.; Peteres, L. GIS and the 2020 Census: Modernizing Official Statistics, ISBN-13: 978-1589485044264, *Publisher Esri*, San Francisco, USA, **2019**, pp.264.

[7] Poppe, R. A survey on vision-based human action recognition. *Image and vision computing*, **2010**, 28(6):976-990.

[8] Ke, S.R.; Thuc, H.; Lee, Y.J.; Hwang, J.N.; Yoo, J.H.; Choi, K.H. A review on video-based human activity recognition. *Computers*, **2013**, 2(2): p. 88-131.

[9] Aggarwal, J.; Xia L. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*. **2014**, 48:70-80.

[10] Wu, W.; Chen, X.; Yang, J. Detection of text on road signs from video. *IEEE Trans. Intell. Transp. Syst.*, **2005**, 6(4):378–390.

[11] Jiménez,P.G.; Maldonado,S.; Gómez,H.; Lafuente-Arroyo,S.; López, F. Traffic sign shape classification and localization based on the normalized FFT of the signature of blobs and 2D homographies. *Signal Process*, **2008**, 88:2943–2955.

[12] Escalera, S.; Baró, X.; Pujol, O.; Vitrià, J.; Radeva, P. Traffic-Sign Recognition Systems. Springer Science and Business Media. *Springer London Dordrecht Heidelberg New York*. **2011**, ISBN 978-1-4471-2244-9. 96 pages.

[13] Niebles, J.C.; Chen, C.W.; Fei-Fei, L. Modeling temporal structure of decomposable motion segments for activity classification. *European conference on computer vision*, Springer, Berlin, Heidelberg, **2010**, pp. 392-405.

[14] Vishwakarma, S.; Agrawal A.J.T.V.C. A survey on activity recognition and behaviour understanding in video surveillance. *The Visual Computer*, **2013**, 29(10): 983-1009.

[15] Shamsipour, G.; Shanbehzadeh J.; Sarrafzadeh H. Human action recognition by conceptual features. International MultiConference of Engineers and Computer Scientists (*IMECS*) **2017**. March 15 – 17, Hong Kong, **2017**, ISBN: 978-988-14047-3-2. Vol.I.

[16] Ravanbakhsh, M.; Mousavi, H.; Rastegari, M.; Murino, V.; Davis, L.S. Action recognition with image based CNN features. **2015**, arXiv preprint arXiv:1512.03980.

[17] Tsai, C.-F.J.I.A.I. Bag-of-words representation in image annotation: A review. ISRN *Artificial Intelligence*. **2012**, Article ID 376804, 19 pages.

[18] Xu, X.; Tang, J.; Zhang, X.; Liu, X.; Zhang, H.; Qiu, Y. Exploring techniques for vision based human activity recognition: Methods, systems, and evaluation. *Sensors*, **2013**, 13(2):1635-1650.

[19] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*. Boston, MA, USA. Electronic ISBN: 978-1-4673-6964-0. **2015**, pp. 1-9. DOI: 10.1109/CVPR.2015.7298594.

[20] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks, *Communications of the ACM*, **2017**, 60(6):84-90. DOI:10.1145/3065386.

[21] Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, **2013**, 24(5):971-981.

[22] Laptev, I.; Marszałek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. *IEEE Conference on Computer Vision and Pattern Recognition*. **2008**,DOI: 10.1109/CVPR.2008.4587756.

[23] Wang, H.; Ullah, M.M.; Klaser, A.; Laptev, I.; Schmid, C. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference*. **2009**, DOI : 10.5244/C.23.124.

[24] Ikizler-Cinbis, N.; Sclaroff, S. Object, scene and actions: Combining multiple features for human action recognition. *European Conference on Computer Vision*. Springer. **2010**, pp. 494-507.

[25] Ji S.; Wei X.; Yu K. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*. **2012**, 35(1): 221-231.

[26] Simonyan, K.; Zisserman A. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*. **2014**, pp. 568-576.

[27] Wang, L.; Qiao Y.; Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7-12 June, Boston, MA, USA. **2015**, pp. 4305-4314. DOI: 10.1109/CVPR.2015.7299059.

[28] Chen, C.; Liu K.; Kehtarnavaz, N.J.J. Real-time human action recognition based on depth motion maps. *Real-Time Image Processing*. **2016**, 12(1): 155-163.

[29] Liu, M.; Liu, H.; Chen, C.J.P.R. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, **2017**, 68:346-362.

[30] Rahmani, H.; Mian, A.; Shah, M. Learning a deep model for human action recognition from novel viewpoints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2017**, 40(3):667-681.

[31] Kamel, A.; Sheng, B.; Yang, P.; Li, P.; Shen, R.; Feng, D.D. Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. **2018**, 49 (9): 1806-1819. DOI: 10.1109/TSMC.2018.2850149.

[32] Blondel, P.; Potelle, A.; Pégard, P.C.; Lozano, P.R. How to improve the HOG detector in the UAV context. *IFAC Proceedings Volumes*. **2013**, 46(30): 46-51.

[33] Liu, B.; Wu, H.; Su, W.; Sun, J. Sector-ring HOG for rotation-invariant human detection. *Signal Processing: Image Communication*, **2015**, 54: p.1-10.

[34] Pirasteh, S.; Rashidi, P.; Rastiveis, H.; Huang, S.; Zhu, Q.; Liu, G.; Li, Y.; Li, J.; Seydipour, E. Developing an algorithm for buildings extractions and determining changes from airborne LiDAR point clouds. *Remote Sensing*, **2019**, 11, 1272; doi:10.3390/rs11111272.

[35] Weinzaepfel, P.; Harchaoui, Z.; Schmid, C. Learning to track for spatio-temporal action localization. in Proceedings of the IEEE International Conference on Computer Vision, pp. 3164-3172.

[36] Nebiker, S., Cavegn, S., Loesch, B., 2015. Cloud-Based geospatial 3D image spaces—a powerful urban model for the smart city. *ISPRS International Journal of Geo-Information*, **2015**, 4(4): 2267-2291.

[37] Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage, AK, USA. **2008**, 1(1):6. Print ISSN: 1063-6919, DOI: 10.1109/CVPR.2008.4587727.

[38] Soomro, K.; Zamir, A.R. Action recognition in realistic sports videos, in Computer vision in sports. *Springer International Publishing Switzerland*. **2014**, Chapter 9, p. 181-208. DOI 10.1007/978-3-319-09396-3_9.

[39] https://www.crcv.ucf.edu/data/UCF_Sports_Action.php

[40] http://vision.stanford.edu/Datasets/OlympicSports

[41] Park, S.C.; Park, M.K.; Kang, M.G. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, **2003**, 20(3):21-36.

[42] Nasrollahi, K.; Moeslund, T.B. Super-resolution: a comprehensive survey. *Machine Vision and Applications*, **2014**, 25(6):1423-1468.

[43] Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern analysis and Machine Intelligence*. **2016**, 38(2):295-307.

[44] Bourdev, L.D. Pose-aligned networks for deep attribute modeling. *Google Patents*. 2016, No 9,400,925, issued July 26.

[45] Pirasteh, S.; Ziaei, H.; Rizvi, S.M. Comparative study of OIF and Crosta methods on ETM+ 2002: using remote sensing techniques in arid and semi-arid environment Esfahan Iran, *Indian Petroleum*. **2005**, 14 (1):67-79.

[46] Brownlee, J. Supervised and unsupervised machine learning algorithms. *Machine Learning Mastery*,

2016, 16(03).

[47] Brownlee J. What is the Difference Between Test and Validation Datasets. *Machine Learning Mastery*. **2017**, https://machinelearningmastery.com/difference-test-validation-datasets/.

[48] Meyer, D.; Wien, F.T. Support vector machines. The Interface to libsvm in package e1071, p.8.

[49] Abe, S., 2005. Support vector machines for pattern classification. London: *Springer*. **2015**, 2, 44.

[50] Vapnik, V. The nature of statistical learning theory. *Springer*, New York, NY. **2013**, ISBN:978-1-4419-3160-3. DOI:org/10.1007/978-1-4757-3264-1.

[51] Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*. **2013**, 117(6): 633-659.

[52] Landis, J.R.; Koch G.G.J. The measurement of observer agreement for categorical data. *Biometrics*. **1977**, p. 159-174.

[53] Gaidon, A.; Harchaoui, Z.; Schmid, C. Activity representation with motion hierarchies. *International Journal of Computer Vision*, **2014**, 107(3):219-238.

[54] Arroll, B.; Jackson, R.; Beaglehole, R. Validation of a three-month physical activity recall questionnaire with a seven-day food intake and physical activity diary, *Epidemiology*, **1991**, 2(4):296-9.

[55] Kruve, A.; Rebane, R.; Kipper, K.; Oldekop, M.-L.; Evard, H.; Herodes, K.; Ravio, P.; Leito, I. Tutorial review on validation of liquid chromatography–mass spectrometry methods: Part I. *Analytica Chimica Acta*, **2015**, 870, 29-44.

[56] Evard, H.; Kruve, A.; Leito, I. Tutorial on estimating the limit of detection using LC-MS analysis, part II: Practical aspects. *Analytica Chimica Acta*, **2016**, 942, 40-49.