ARTICLE

Vision-based human activity recognition via dispersion measures of spatiotemporal features

Anton Louise P. De Ocampo*[1] and Elmer P. Dadios[2]

[1]Electronics and Communications Engineering Department, De La Salle University, Manila, Philippines
[2]Manufacturing Engineering and Management Department, De La Salle University, Manila, Philippines

Human activity recognition (HAR) systems can be categorized as a sensor- or vision-based depending on the type of data it collects as inputs. The non-contact implementation of vision-based HAR is the confounding factor why such systems are preferred over systems using wearable sensors. Although remarkable feats have been achieved in the field of human activity recognition, one challenge remains the focus of recent researches –extraction and selection of suitable features. In this work, a novel approach in extracting and selecting feature vector for HAR implementation in smart farming is proposed. By exploiting the data distribution on stacks of difference maps, a new feature vector, which contains high discriminative attributes between activities performed by farmers in the field, is proposed. Using the k-NN classifier, the experiment obtained 98.89%, 98.69%, and 98.79% scores in precision, recall, and F1-measure in the classification of farmers' activities in the field.

KEYWORDS

Human activity recognition, difference map, feature extraction, unmanned aerial vehicles, smart farming

*Corresponding author
Email Address: anton_louise_deocampo@dlsu.edu.ph
Date received: June 17, 2020
Date revised: October 1, 2020
Date accepted: October 17, 2020

INRODUCTION

Over the years, the research efforts expended on human activity/action recognition (HAR) have increased significantly due to its importance in the fields of human-computer interaction (HCI), healthcare, and surveillance (Ann & Theng, 2014). Applications of HAR in the healthcare domain includes support to caring for elders (Schrader et al., 2020), assistance to motor-related rehabilitation, and assessment of various cognitive disorders (Zebin, Scully, & Ozanyan, 2017). Other applications of HAR may include identity recognition, community environment monitoring, and smart video surveillance (H. B. Zhang et al., 2019).

HAR systems can be categorized based on the type of data it uses to recognize the activities of the subject-of-interest. The use of inertial sensors attached to the body either as wearable devices or mobile edge computing (MEC) can be classifiers as sensor-based HAR. Vision-based HAR, on the other hand, relies solely on images or videos of persons to recognize the activities/actions performed.

Smartphones have been a global contributor to high-dimensional data due to the multi-functional sensors they carry. These sensors, together with the computing prowess of MEC devices, made HAR be implemented in mobile phones (Wan, Qi, Xu, Tong, & Gu, 2020). Aside from smartphones, application-specific wearables are having built-in inertial sensors that measure the movement of the subject wearing the device.

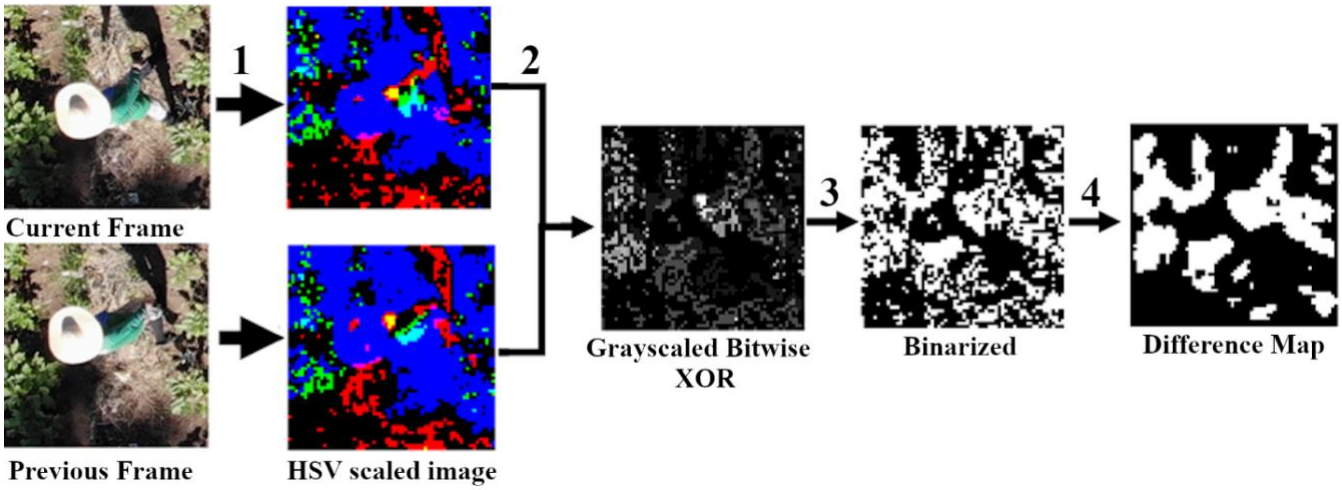

Figure 1: Difference map generation process includes conversion of RGB to HSV (1), bitwise XOR of the hue channel (2), image binarization (3), and median filtering (4).

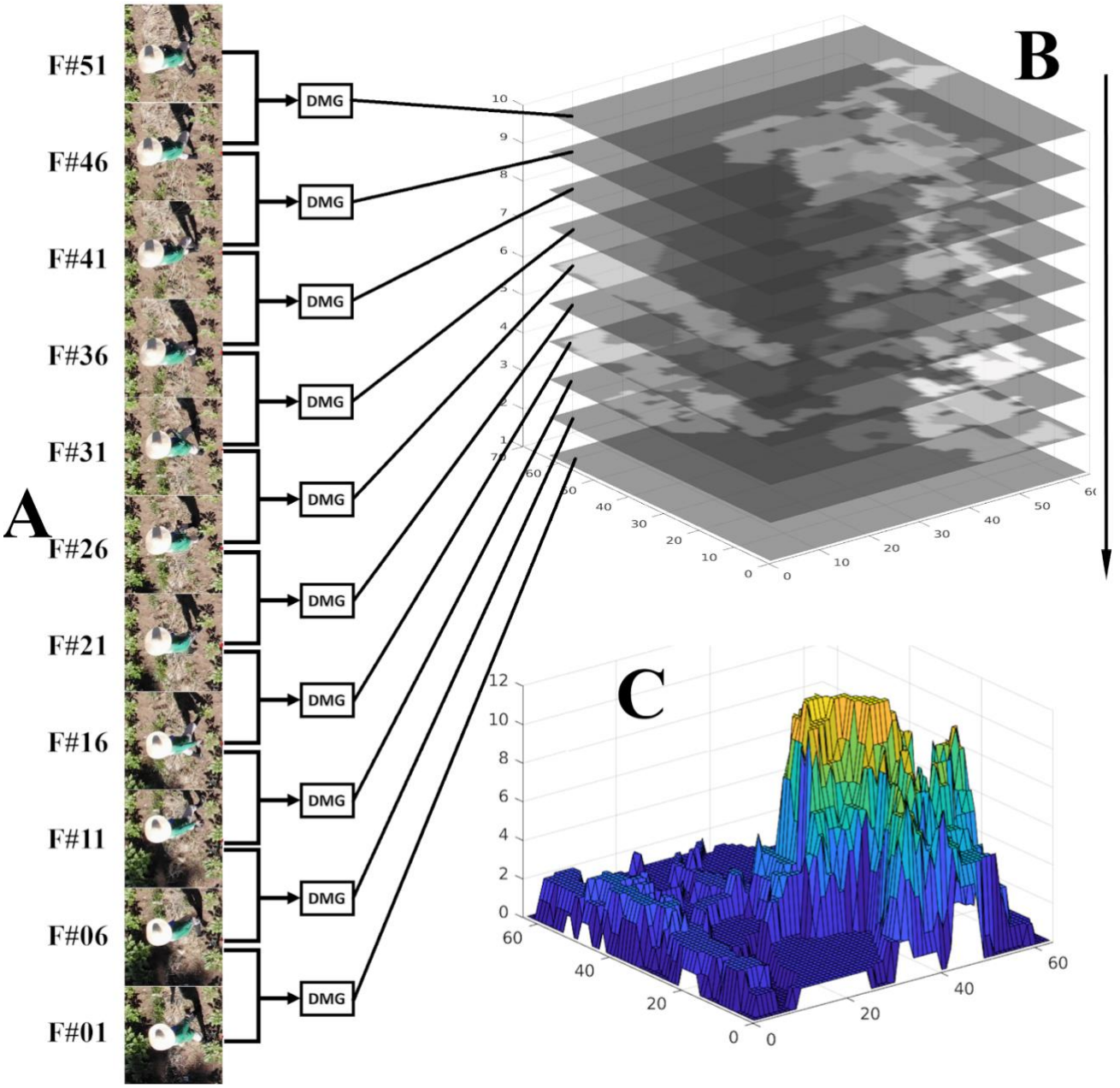

Figure 2: Eleven frames (A) in strides of 5 are necessary to generate 10 difference maps (B) of which the corresponding X-Y coordinates are summed up to create the volume feature map (C). "F#" = "Frame Number", "DMG" = "Difference Map Generator".

Table 1: Statistical measures derived from volume maps.

P	Measure	Equation	Details
1	Mean – overall	$\mu = \frac{\sum V(x_i, y_i)}{M}$	$V \stackrel{\text{def}}{=} volume\ map$ $M \stackrel{\text{def}}{=} size\ of\ V$
2	Std. deviation – overall	$\sigma = \sqrt{\frac{\sum (V(x_i, y_i) - \mu)^2}{M}}$	$\mu \stackrel{\text{def}}{=} mean$
3	Inter-quadrant Std. deviation	$\sigma_q = \sqrt{\frac{\sum_{k=1}^{4} (\mu_k - \mu)^2}{4}}$	$\mu_k = \frac{\sum_{i=\frac{lM}{4}+1}^{\frac{kM}{4}} V(x_i, y_i)}{\frac{M}{4}}$; $l = 0,1, \ldots, k-1$ $\mu_k = quadrant\ mean$
4 to 7	k[th] quadrant std. dev.	$\sigma_k = \sqrt{\frac{\sum_{i=\frac{lM}{4}+1}^{\frac{kM}{4}} (V(x_i, y_i) - \mu_k)^2}{M/4}}$	$k = quadrant$
8,12, 16, 20	k[th] quadrant, quartile 1	$Q_{k1} = \frac{1}{4}(n+1)^{th}\ term$	
9, 13, 17, 21	k[th] quadrant, quartile 2	$Q_{k2} = = \frac{2}{4}(n+1)^{th}\ term$	
10, 14, 18, 22	k[th] quadrant, quartile 3	$Q_{k3} = \frac{3}{4}(n+1)^{th}\ term$	
11, 15, 19, 23	k[th] Inter-quartile range	$IQR_k = Q_{k3} - Q_{k1}$	

P = position of the statistical measure on the feature vector.

Table 2: Classifiers evaluated for human activity recognition. Training used 75% of the dataset while the remain is used for validation

Classifier	Accuracy	Classifier	Accuracy
Fine Tree	97.1%	Coarse KNN	94.0%
Medium Tree	93.5%	Cosine KNN	97.6%
Coarse Tree	88.3%	Cubic KNN	97.6%
Linear Discriminant	92.2%	Fine KNN	99.5%
Linear SVM	94.4%	Medium KNN	97.9%
Quadratic SVM	98.3%	Weighted KNN	99.2%
Cubic SVM	99.4%	Boosted Trees Ensemble	97.1%
Fine Gaussian SVM	98.6%	Bagged Trees Ensemble	99.2%
Medium Gaussian SVM	98.3%	Subspace Discriminant Ensemble	91.7%
Coarse Gaussian SVM	92.7%	Subspace KNN Ensemble	99.3%
		RUS Boosted Tree Ensemble	92.2%

SVM = Support vector machine, KNN = kth-Nearest Neighbors, RUS = Random undersampling

Hospitals and rehabilitation centers catering to the needs of elderly people often use wearable devices to monitor patients' movement (Schrader et al., 2020). Aside from inertial sensors, temperature, ECG, and EEG sensors can also be used to analyze human action. In a smart healthcare framework, a combination of bodily sensors is used to generate data for analysis in a remote HAR system (Subasi, Khateeb, Brahimi, & Sarirete, 2020). Smartwatches are also proved to be an undisputed source of motion data for researches inclined to measure the physical activity of persons (Henriksen, Sci, Mikalsen, Sci, & Woldaregay, n.d.). The use of magnetic induction sensors to characterize movement is also categorized as sensor-based HAR (Golestani & Moghaddam, 2020).

On the other hand, vision-based HAR does not use any physical contact with the subject instead of images, and videos of movements are used for analysis. Cameras used in the acquisition of movement data are also referred to as ambient sensors (Schrader et al., 2020). In a comprehensive review performed by Zhang et al., both classical and recent approaches to vision-based HAR are discussed (H. B. Zhang et al., 2019). After summarizing massive literature on the subject, the review concluded that the data needed to capture the actions should be properly selected. Using cameras to capture RGB data or motion-sensing devices to capture depth in movements has been the trend in acquiring input for vision-based HAR (Mo, Li, Zhu, & Huang, 2016). However, several confounding factors deem recognition of human activity in real scenes very challenging. One factor emphasized in the review is the complexity and diversity of human poses.

In a review made by Jegham et al. on vision-based HAR, there are real-world challenges that remain as active domains in research, namely, anthropometric variation, multi-view variation, cluttered and dynamic background, intra-class variability and inter-class similarity, low-quality videos, occlusion, illumination variation, shadow, and scale variation, camera motion, insufficient data, and poor weather conditions (Jegham, Ben, Alouani, & Ali, 2020). Aside from the challenged imposed by human poses characterized by the intra-class variation and inter-class similarity between activities, activity recognition under real-world settings becomes more difficult due to the multitude of visual features that can be extracted from

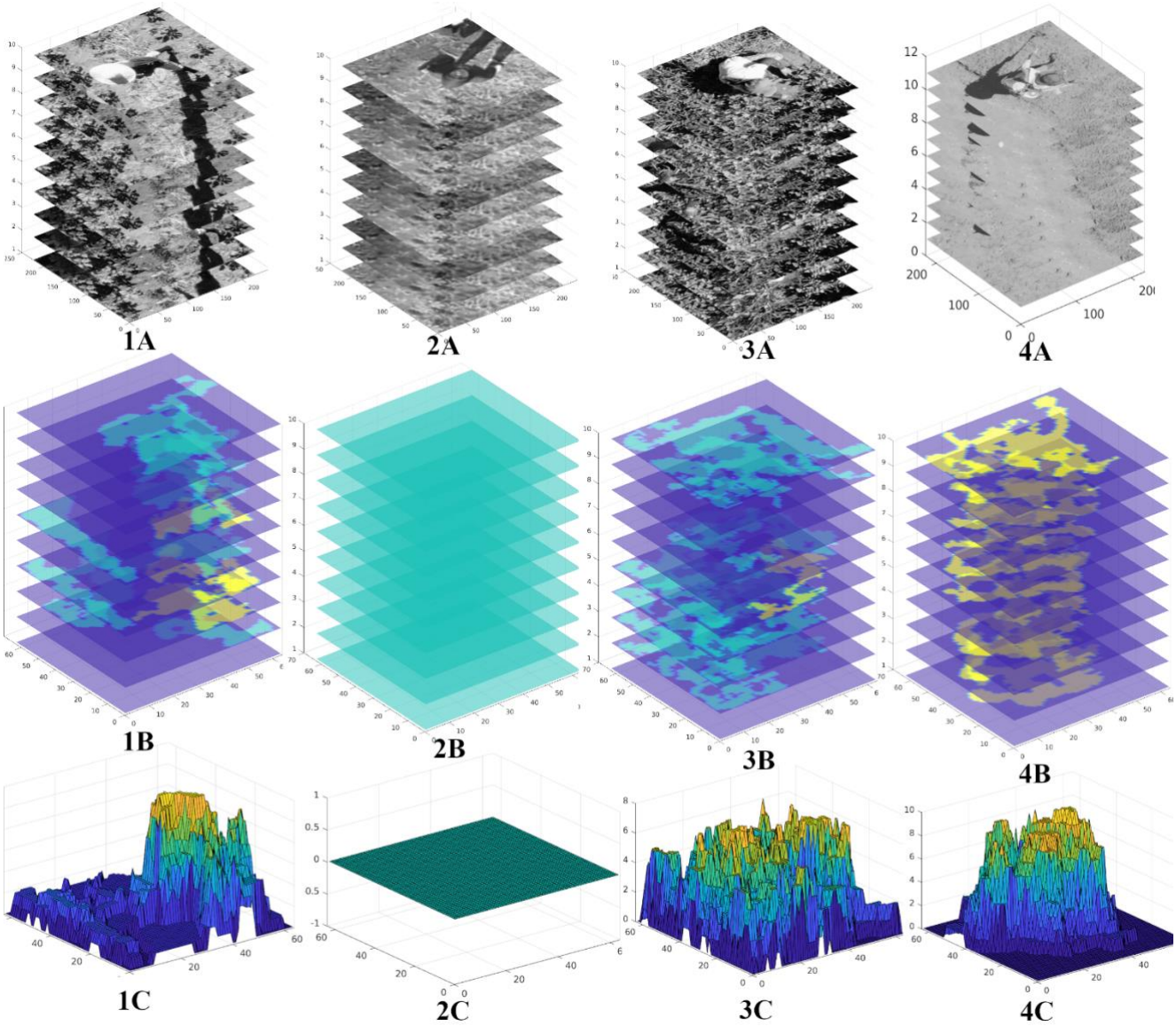

Figure 3: Stacked frames (A), difference maps (B), and volume feature maps (C) of four activity classes which are 'walking (1), 'standing / sitting still' (2), 'sitting while working' (3), and 'standing / walking while working' (4).

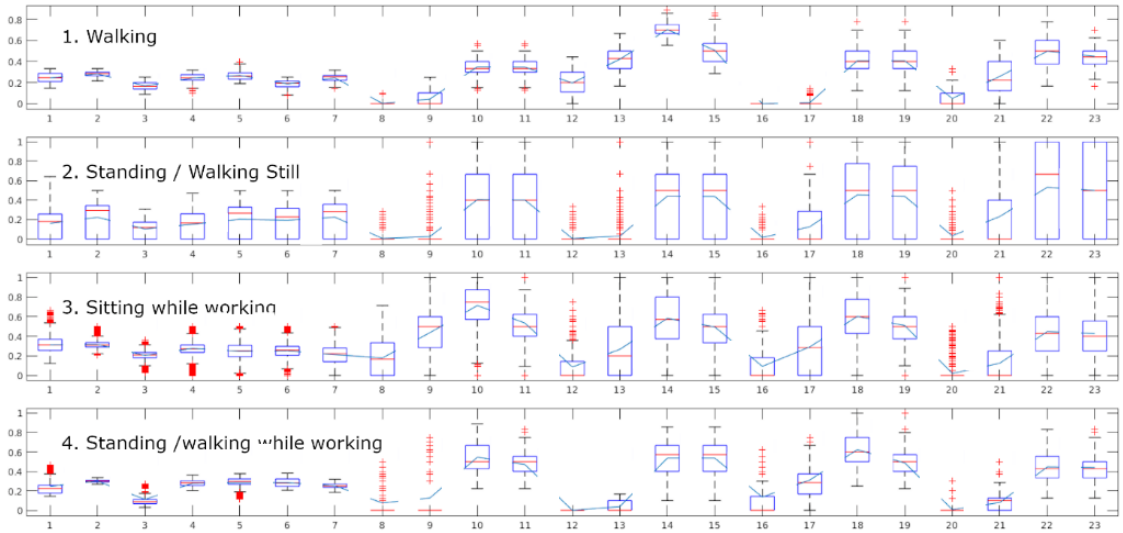

Figure 4: Consolidation of 23 statistical measures derived from 14,427 volume maps of different farmers performing at least one of the four activity classes. OverallAVG (1), OverallSTD (2), InterQuadStd (3), 1st QuadStd (4), 2nd QuadStd (5), 3rd QuadStd (6), 4th QuadStd (7), Q1Quartile1 (8), Q1Quartile2 (9), Q1Quartile3 (10), InterQuartileRange1 (11), Q2Quartile1 (12), Q2Quartile2 (13), Q2Quartile3 (14), InterQuartileRange2 (15), Q3Quartile1 (16), Q3Quartile2 (17), Q3Quartile3 (18), InterQuartileRange3 (19), Q4Quartile1 (20), Q4Quartile2 (21), Q4Quartile3 (22), InterQuartileRange4 (23).

EQUATIONS

Bitwise XOR

$$g_n(x,y) = hsv_{n-1}(x,y,1) \otimes hsv_n(x,y,1) \quad \text{(eq. 1)}$$

Where,

$$hsv_n(x,y,1) = hue\ channel\ of\ current\ frame$$
$$hsv_{n-1}(x,y,1) = hue\ channel\ of\ previous\ frame$$

Binarization

$$b(x,y) = \begin{cases} 1\ if\ g(x,y) \geq T \\ 0\ otherwise \end{cases} \quad \text{(eq. 2)}$$

Where,

$$g(x,y) = grayscale\ of\ XOR\ image$$

Difference map

$$d(x,y) = med(b(x-i, y-j), i, j, \in M) \quad \text{(eq. 3)}$$

Where,

$$d(x,y) = difference\ map$$
$$b(x,y) = binary\ image$$
$$M = filter\ size$$

Volume feature map

$$V(x,y) = \sum_{n=1}^{N} d_n(x,y) \quad \text{(eq. 4)}$$

Where,

$$N = number\ of\ frames$$

video recordings and captured images (S. Zhang et al., 2017)(Gong et al., 2019).

In any recognition problem that involves images, the selection of suitable features can be the demarcation line between successful classification and failure. Features are a set of attributes that present distinctive characteristics of different classes. Features must be presented in reduced volume than the raw data (Lim, Naguib, Dadios, & Avila, 2010) (De Ocampo, Bandala, & Dadios, 2019).

It is the object of this work to extract and select a suitable feature vector for recognition of human activities in real-world settings. In the end, the context of this work is focused on the (1) use of multi-level feature extraction of spatiotemporal attributes of the action video recordings, and the (2) use of such features in activity recognition problem applied to smart farming.

METHODS AND MATERIALS

Human activity recognition (HAR) based on visual attributes is a five-step process which includes: the extraction of a definite number of subsequent frames from the subject video (1); calculation of difference maps between two frames in comparison (2); coalescing difference maps into feature volume map (3); calculation of the dispersion measures of the feature volume map (4); and classification of the activity based on statistical measures of dispersion (5).

Data Acquisition

The primary sensor used in collecting the dataset for the experiment is a 12 megapixel 1/2.3" CMOS onboard an unmanned aerial vehicle (UAV) The UAV-based image acquisition device captured farmers working in the fields while flying at an altitude of 21m from the ground. Twenty-four videos, with a resolution and frame rate of 2280 x 4056 and 23fps respectively, are recorded at different dates but within the time interval of 8:00 am to 10:00 am.

Target Activity classes

The activities captured in the video set can be categorized into 4 general activities: 'walking' (1), standing/sitting still (2), sitting while working (3), and standing/walking while working (4). The generalization of the activities based on the day-to-day activities of the farmers and the nature of their work is done to reduce the intra-class variations. For instance, the removing of weeds and pruning the crops fall under the same activity class which is 'sitting while working'. Similarly, spraying pesticides and harvesting fall under the activity class 'standing/walking while working'.

Dataset

The dataset used in training and testing of the classifier is derived from the 24 video recordings of farmers' activities. The dataset is an excerpt from a larger dataset which is the Farmers' Dataset or FDS (De Ocampo & Dadios, 2019). Since the regions-of-interest, ROIs, are relatively small for the video resolution, localization of farmers from the video can be defined before the experimentations to allow faster processing of the samples. The collected data from the video recordings are counted to be 221, 2091, 11718, and 397 samples for each of the respective activity classes. It can be noticed that the activity class about working farmers while in sitting position dominates the dataset since most of their activities are in sitting or bent positions.

Feature Vector

The feature vector used to describe the activity of a farmer consists of the dispersion measurements from the volume maps. These volume maps are the aggregates of the difference maps that are derived from bitwise XOR of two corresponding ROIs from two subsequent frames. Suppose that feature vector generation used a stack of *N* video frames with a single ROI in each, then there would be *N-1* difference maps, which is equivalent to a single volume map or a single feature vector, that can be created. But for *N* video frames that contain *M* ROIs in each frame, then a total of *M* x *(N-1)* difference maps, which is equivalent to *M* volume maps or *M* feature vectors, can be created.

To generate a difference map, the corresponding ROIs in two subsequent frames are converted into scaled HSV images. Then, only the hue channels of the corresponding previous and current frames are subjected to bitwise XOR operation (eq. 1). The intuition behind the use of the hue channel amongst others is the observation that the activity of a person is manifested greatly in the movement of his garments – say true at least in the mentioned dataset. The result of bitwise XOR of the two ROIs is a grayscale representation of how the respective images differ. The grayscale bitwise XOR is converted into a binary image (eq. 2) which is then fed to a 3x3 2D median filter to reduce noise and allow smooth connections between blobs. The median filtering results in the difference map (Eq. 3) wherein the presence of blobs manifests dynamics between the two frames. The feature vector generation requires at least 10 difference maps equivalent to 11 frames in a stack to observe adequate variation between images. However, the videos in the dataset are captured in 23 frames per second so an activity would be barely visible in approximately 0.5 secs. Hence, a stride of 5 is employed in the stacking of frames for the volume map (Figure 2). The aggregation of difference maps to a volume map is straightforward, which is by summing up the values at the corresponding pixel coordinates (eq. 4). The striding denotes

that a video must have at least 51 frames before the first set of feature maps can be derived.

The initial size of the ROIs is 224 x 224 which is reduced to 64 x 64 by bicubic interpolation with a pixel neighborhood of 4 x 4 and 1-pixel overlap before the creation of difference maps. In representing the stack of difference maps, we used the notion of a volume where the third dimension is the intensity or the sum of corresponding pixels in the difference maps. Since the difference maps manifest variation in-between frames, the more disperse the volume map is, the more activity is present. The distribution of values in the volume map determines the activity type. In such a context, two measures of dispersion are calculated from the volume map: overall mean and standard deviation. To also consider the dynamics happening in different regions of the map, the volume map is divided into quadrants. From each quadrant, additional dispersion measures for each are calculated: quadrant standard deviation, quartiles, and inter-quartile range. Lastly, we also proposed the use of inter-quadrant deviation to account for the rotational or translational movement of the subject in the image relative to the camera. In total, twenty-three measures of dispersion are combined into the final feature vector used in human activity recognition (Table 1).

Classifier

In the selection of classifiers to be used for human activity recognition, 21 different models were evaluated. The classifier for human activity recognition is selected based on the training accuracy at a 25% holdout for validation. It is also important to take note that dispersion within the volume map is primarily the discriminating attribute of the feature vector.

RESULTS

A farmer's activity can be observed in subsequent video frames by the displacement of his garment due to the movements of arms and legs. Having a human figure in the center of each ROI suggests that if an activity is highly directional such as walking, higher color variations can be observed asymmetrical. The difference map tends to skew in a particular direction. The difference map generated from two ROIs shows the variation as blobs similar to what is shown in Figure 2. Human perception of movement can be attributed to color variation. Hence, we convert the ROIs from RGB to HSV to directly quantify any color change. The bitwise XOR of the hue channels results in an 8-bit grayscale image which manifests the difference between the two images. De-noising by use of the median filter highlights the area where the noticeable difference occurs.

Although a single difference map can show variation between ROIs, it lacks the temporal attributes to characterize an activity. Multiple and subsequent difference maps are needed to portray any visible activity. By aggregating all the generated difference maps (Figure 2A) and stacking them (Figure 2B), the sum of the corresponding pixels of the difference maps is represented as a volume map (Figure 2C). Here, the variations become more noticeable. The skewness suggests that the activity or movement happens in the quadrant where the volume peaks can be found, and where the movement of limbs is more prominent. This skewness of the volume map becomes the primary discriminating factor between the activity classes. The distribution of data in the volume map shows a strong correlation with the activity classes (Figure 3). When the object-of-interest is walking, the peaks of the volume map clutter in any of the image quadrants (Figure 3, 1C), while, there are no or fewer variations in the volume map when there is no noticeable activity (Figure 3, 2C). In the case of a working activity in sitting position, the peaks of the volume map are quite distributed (Figure 3, 3C) as opposed to working in a walking/standing position in which the peaks are near the centroid of the object-of-interest (Figure 3, 4C).

Because the distribution of data in the volume map is the key to characterize the activities, the feature vector for activity recognition must quantify the dispersion of data in the volume maps. This is achieved by using measures of dispersion as the final representation of the feature vector. By consolidating all the feature vectors derived from the available volume maps of a certain activity class and plotting it together with the other classes, the discriminating attributes of the feature vector become more evident (Figure 4). For example, in the activity class #2 which is sitting/standing still, the quartiles of most quadrants are zero or minimum which suggests that no noticeable activity is happening (Figure 4, plot 2). Besides, the range of values in most measurements in activity class #2 is quite wider than the other activities. This suggests that the volume map is flat, tantamount to no noticeable changes in frames are present.

Finally, the activity recognition classifier, i.e. k-NN, achieved 98.89%, 98.69%, and 98.79% scores in precision, recall, and F1-measure.

CONCLUSION

Feature extraction and selection remains an important aspect of activity recognition. The features derived from subsequent frames from an excerpt of a video proved to contain distinctive attributes about the activity being done. Eleven frames are used to extract difference maps and fuse them to a volume map. The manner of how data is distributed along the volume map represents the direction of the activity being done. The measures of dispersion of data in the volume map show promising distinctive attributes for classifiers to recognize. The extracted features, when used with a k-NN classifier, can provide 98.89% precision, 98.69% recall, and 98.79% F1-score.

ACKNOWLEDGMENTS

The authors would like to extend their gratitude to the management of the Department of Science and Technology, ERDT for the funding of this study.

CONTRIBUTIONS OF INDIVIDUAL AUTHORS

ALDO and EPS conceptualized the ideas. ALDO gathered and analyzed the data, prepared and finalized the manuscript, and conducted the literature review. EPS provided consultations and review of the algorithm and manuscript.

CONFLICTS OF INTEREST

The authors affirm that the study was done without any financial or commercial relationships that could be interpreted as a potential conflict of interest.

REFERENCES

Ann OC, Theng LB. Human activity recognition: A review. Proc IEEE Intl Conf Contr 2014; March : 389–393. https://doi.org/10.1109/ICCSCE.2014.7072750

De Ocampo ALP, Bandala AA, Dadios EP. Estimation of Triangular Greenness Index for Unknown PeakWavelength

Sensitivity of CMOS-acquired Crop Images. IEEE Int Conf Humanoid Nanotechnol Inf Technol Commun Control Environ Manage 2019 ; 1–5. https://doi.org/10.1109/hnicem48295.2019.9072796

De Ocampo ALP, Dadios EP. Radial greed algorithm with rectified chromaticity for anchorless region proposal applied in aerial surveillance. Int J Adv Intell Informatics 2019; 5(3) : 193–205. https://doi.org/10.26555/ijain.v5i3.426

Golestani N, Moghaddam M. Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks. Nat Commun 2020; 11(1). https://doi.org/10.1038/s41467-020-15086-2

Gong S, Liu C, Ji Y, Zhong B, Li Y, Dong H. Feature extraction and representation. Model Optim Sci Technol 2019. https://doi.org/10.1007/978-3-319-77223-3_4

Henriksen A, Sci C, Mikalsen MH, Sci C, Woldaregay AZ. Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research : Analysis of Consumer Wrist-Worn Wearables. J Med Internet Res 2018; 20(3):e110 . https://doi.org/10.2196/jmir.9157

Jegham I, Ben A, Alouani I, Ali M. Vision-based human action recognition : An overview and real world challenges. Forensic Sci Int 2020; 32(3) : 200901. https://doi.org/10.1016/j.fsidi.2019.200901

Lim LAG, Naguib RNG, Dadios EP, Avila JMC. Analysis of colonic histopathological images using pixel intensities and Hough Transform. Philipp Sci Lett 2010; 128–135.

Mo L, Li F, Zhu Y, Huang A. Human physical activity recognition based on computer vision with deep learning model. IEEE Instrum Meas Technol Conf 2016; July. https://doi.org/10.1109/I2MTC.2016.7520541

Schrader L, Vargas Toro A, Konietzny S, Rüping S, Schäpers B, Steinböck M, Bock T. Advanced Sensing and Human Activity Recognition in Early Intervention and Rehabilitation of Elderly People. J Popul Ageing 2020. https://doi.org/10.1007/s12062-020-09260-z

Subasi A, Khateeb K, Brahimi T, Sarirete A. Human activity recognition using machine learning methods in a smart healthcare environment. J Innov Health Inform 2020. https://doi.org/10.1016/b978-0-12-819043-2.00005-8

Wan S, Qi L, Xu X, Tong C, Gu Z. Deep Learning Models for Real-time Human Activity Recognition with Smartphones. Mob Netw Appl 2020; 25(2) : 743–755. https://doi.org/10.1007/s11036-019-01445-x

Zebin T, Scully PJ, Ozanyan KB. Human activity recognition with inertial sensors using a deep learning approach. Proc IEEE Sens 2017 ; (1). https://doi.org/10.1109/ICSENS.2016.7808590

Zhang HB, Zhang YX, Zhong B, Lei Q, Yang L, Du JX, Chen DS. A comprehensive survey of vision-based human action recognition methods. Sens 2019; 19(5) : 1–20. https://doi.org/10.3390/s19051005

Zhang S, Wei Z, Nie J, Huang L, Wang S, Li Z. A Review on Human Activity Recognition Using Vision-Based Method. J Healthc Eng 2017. https://doi.org/10.1155/2017/3090343