DroneSURF: Benchmark Dataset for Drone-based Face Recognition

Isha Kalra, Maneet Singh, Shruti Nagpal, Richa Singh, Mayank Vatsa, and P.B. Sujit
IIIT-Delhi, India

***Abstract*— Unmanned Aerial Vehicles (UAVs) or drones are often used to reach remote areas or regions which are inaccessible to humans. Equipped with a large field of view, compact size, and remote control abilities, drones are deemed suitable for monitoring crowded or disaster-hit areas, and performing aerial surveillance. While research has focused on area monitoring, object detection and tracking, limited attention has been given to person identification, especially face recognition, using drones. This research presents a novel large-scale drone dataset, DroneSURF: Drone Surveillance of Faces, in order to facilitate research for face recognition. The dataset contains 200 videos of 58 subjects, captured across 411K frames, having over 786K face annotations. The proposed dataset demonstrates variations across two surveillance use cases: (i) active and (ii) passive, two locations, and two acquisition times. DroneSURF encapsulates challenges due to the effect of motion, variations in pose, illumination, background, altitude, and resolution, especially due to the large and varying distance between the drone and the subjects. This research presents a detailed description of the proposed DroneSURF dataset, along with information regarding the data distribution, protocols for evaluation, and baseline results.**

I. Introduction

Drones or Unmanned Aerial Vehicles (UAVs) can be controlled remotely and pre-programmed to complete a specific task. Recently, UAVs have been used for a wide variety of applications such as photography, active monitoring [24], automated delivery systems [1], disaster relief [9], [21], vehicle detection [2], traffic and motion analysis [18], forest fire monitoring [28], human action recognition [3], and gesture and pose based drone control [22]. Since drones can be controlled remotely, they are often used to access regions which are otherwise challenging to navigate through or are inaccessible to humans. This makes the drone technology an ideal candidate for monitoring remote locations and extremely crowded places, without causing any disruption. Drone usage can significantly reduce the manual load and enable law enforcement agencies to search for missing persons, and locate criminals or wanted individuals.

To the best of our knowledge, limited attention has been given to the challenging yet important application of drone based face recognition. Automating the process of face recognition or tracking using drones can greatly benefit surveillance and remote monitoring scenarios [20]. It can improve the efficiency of security systems, specifically in crowded scenarios such as stadiums or festivals (Fig. 1(a)), and disaster relief operations in regions which are difficult to access by humans, or in monitoring remote locations with strenuous terrains. Fig. 1(b) presents a flood hit region being surveyed using a drone camera. The fundamental design and properties of drones make them a suitable and viable device for passively surveying such regions, or actively monitoring a specific area or individual until physical help is available.

(a) Monitoring crowded scenarios

(b) Surveying disaster-hit or inaccessible areas

Fig. 1: Sample scenarios where drone based face detection or recognition will prove to be useful. Images taken from the Internet: https://tinyurl.com/yayk7qsv; https://tinyurl.com/ydx4v9q8

A. Existing Drone-based Datasets

Drone based face recognition brings with it a new set of challenges such as the effect of motion, pose, illumination, background, and height. The presence of these challenges, along with the relatively lower resolution of captured faces, varying distance between the drone and the subjects, the problem of drone based face recognition is thus rendered further challenging. To the best of our knowledge, limited research has been performed to automate drone based face recognition (Table I and Fig. 2). Recently, Bindemann *et al.* [5] analyzed the human performance for person identification in videos captured using a drone. The authors established the challenging nature of the problem, and poor performance for face identification. Hsu and Chen [12], [13] proposed

TABLE I: Literature review of recent publicly available drone based datasets.

Dataset (Year)	Purpose	Drone Mobility	Dataset Size	Annotated Faces
MRP Drone dataset (2014) [17]	Person re-identification	Yes	∼16,000 frames	No
MiniDrone dataset (2015) [6]	Area monitoring	Yes	22,860 frames	No
Stanford Drone dataset (2016) [26]	Human trajectory prediction	Yes	929,499 frames	No
DroneFace (2017) [13]*	Face recognition	No	2,057 images	No
VisDrone2018 (2018) [30]	Object detection and tracking	Yes	179,264 frames, 10,209 images	No
IJB-S (2018) [16]**	Person identification	Yes	10 videos, 1,487 images	No
Proposed DroneSURF (2018)	**Face detection and recognition**	**Yes**	**200 videos, 411,451 frames**	**Yes**

*Drone scenario simulated using stationary GoPro camera **Information listed above is only about the UAV component of the dataset. Complete dataset contains 350 surveillance videos and 202 enrollment videos.

Fig. 2: Sample images from recent drone/UAV based datasets, along with the proposed DroneSURF dataset.

the *DroneFace* dataset, and evaluated the performance of existing techniques and commercial systems. It is important to note that the *DroneFace* dataset simulates data captured by a drone by using a stationary GoPro camera. The subjects are stationary, with neutral expressions, and have been asked to look in a single direction, without their glasses, in order to only capture the camera's height effect. The study provides an overview of the challenging nature of the problem, however it does not depict a true picture of drone based face recognition due to the relatively constrained settings. Recently, Kalka *et al.* [16] proposed the IJB-S dataset containing a component of 10 UAV based videos for face recognition. Apart from the above mentioned research, different army and law enforcement organizations utilize UAVs to guard borders for surveillance, and are exploring options to use drone based technology for rescue missions.

B. Contributions

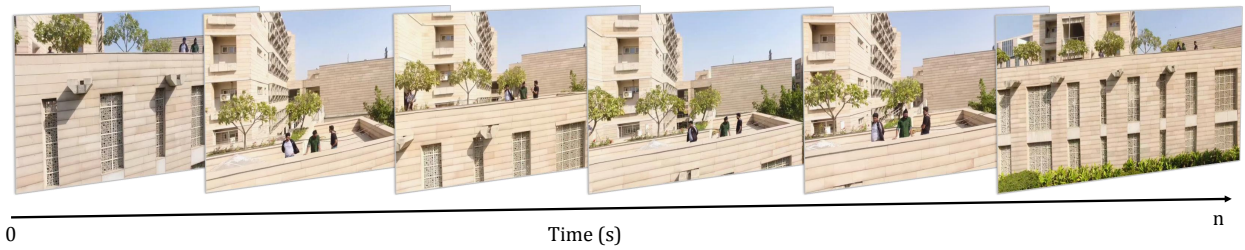
To address the limited availability of drone-based datasets for face recognition, this research presents a benchmark dataset, termed as *DroneSURF: Drone SUrveillance of Faces*. The proposed dataset contains 200 videos of 58 subjects captured using a drone camera, with variations across use-case, location, and acquisition time. In order to simulate real world scenarios, each video contains a group of individuals. The dataset contains a total of 411k frames, with over 786k face annotations. To the best of our knowledge, this is the first dataset introduced specifically for research of drone based face recognition, which will be made publicly available to the research community. This research also presents the experimental protocols for face detection and recognition, along with the baseline results. Results for face detection have been reported with two state-of-the-art face detectors, while face recognition results are presented with four features (hand-crafted and deep learning based), along with a commercial-off-the-shelf system. It is our assertion that the availability of the proposed DroneSURF dataset will facilitate research in this direction.

II. Proposed DroneSURF Dataset

As demonstrated in Table I and Fig. 2, there are few publicly available UAV/drone datasets for person identification. Specifically, for face recognition, there exists no database collected using drones, which simulates a real world scenario, except the recently proposed IJB-S dataset which has a small subset of 10 UAV videos. Most of the datasets are collected for general-purpose monitoring or activity recognition, with limited focus on person identification or face recognition. As discussed and observed in the literature [5], surveillance using drones is a challenging problem owing to several factors such as the quality and movement of the drone, subject to be captured, and the environment. For example, a subject's unrestricted movement and distance from the drone often results in high pose and resolution variations. Coupled with the movement of the drone, its altitude, and the varying environmental factors, the task of drone face recognition is rendered extremely challenging. In order to model such variations, the proposed DroneSURF dataset contains data captured across eight different settings, with varying location, surveillance scenario, and acquisition time. The proposed DroneSURF dataset contains a total of 200 videos featuring 58 subjects across 411K frames. Details about the surveillance scenarios modeled in the dataset and the statistics are discussed in the following subsections.

A. Surveillance Settings

Drones can be used for performing two types of surveillance: (i) active and (ii) passive. In order to simulate the real world scenarios, the proposed DroneSURF dataset contains videos captured for both these surveillance settings. Fig. 3 presents sample frames of videos captured in two different



(a) Active Surveillance

(b) Passive Surveillance

Fig. 3: Sample frames from four videos for active and passive surveillance demonstrating the progression of the video from 0^{th} second to the last (n^{th}) second of the video.

locations with active and passive surveillance settings. The application and dataset details for each of the two settings is discussed below:

- **Active surveillance** is useful in situations where a specific subject or set of subjects are actively monitored by a UAV/drone. For the DroneSURF dataset, users were asked to walk from point A to point B, while the drone captured their movement from the front. The drone actively monitors the user by flying a few meters ahead of him/her. It is important to note that user co-operation is not required for active surveillance; the drone is required to follow the movements of the subject.
- **Passive surveillance** corresponds to the scenarios wherein a drone is used to monitor an area or event, without explicit focus on a particular subject or object. For the DroneSURF dataset, users were asked to roam about in a particular area, have discussions, or simply walk, while the drone captured the entire area. In this case, the movement of the drone is independent to that of the subjects; its aim is to just record the events of a particular region, as opposed to monitoring specific individuals.

The proposed dataset contains 100 videos each for active and passive surveillance. The availability of videos captured under different settings makes the proposed dataset challenging, and simulates real world surveillance scenarios.

B. Dataset Statistics

The DroneSURF dataset is captured in eight different settings with variations across the surveillance setting, location, and time of capture. Details regarding the different surveillance settings (active and passive) have already been provided above. Data has been captured at two outdoor locations: (i) at the ground level, where subjects are asked to walk in a park-like environment, and (ii) at the terrace of a building. For each location and surveillance scenario, data is captured twice: (i) during the morning and (ii) during the evening, before sunset. Videos captured in the morning are well illuminated, while the evening videos contain comparatively lower illumination. However, since the evening videos are captured before sunset, it is ensured that the videos contain sufficient illumination, thus eliminating the need for any additional source of illumination. Fig. 4 showcases some of the covariates present in the dataset.

Fig. 4: Variations across pose, illumination, occlusion, and resolution observed in the proposed DroneSURF dataset.

TABLE II: Statistics of the DroneSURF dataset.

Characteristic	Surveillance Setting		
	Active	Passive	Total
Videos	100	100	200
Subjects	58	58	58
Mean (secs.)	57	79	68
Min. (secs.)	18	34	18
Total Frames	172,263	239,188	411,451
Annotated Faces	332,693	379,504	786,813

For each combination of location, surveillance use case, and time of capture, there exist 25 videos featuring 58 subjects. Each subject belongs to the age bracket of (18, 40) years, and each video contains subjects appearing in groups of 2-3. For a particular setting combination, one

group appears in only one video. Since there exist eight combinations of location, surveillance scenario, and time of capture, each group of subjects occur in eight videos, thus resulting in a total of 200 videos. The subjects and pairings remain consistent across different settings. Table II presents the statistics of the DroneSURF dataset, where active surveillance contains over 172K frames, while passive surveillance contains over 239K frames spread across 100 videos each.

Along with the videos, the DroneSURF dataset also contains high resolution gallery images of each subject. These images are captured in constrained scenarios with good illumination, high resolution, and with four different poses. This is done in order to simulate a real world law enforcement scenario, wherein data captured via a drone will be matched against a pre-acquired database of high quality images. The UAV/drone videos have been captured using the DJI Phantom 4[1] which is one of the high-end, entry-level professional drones available in the market. The videos have been captured at a frame rate of 30fps, and at a resolution of 720p. For the high resolution gallery images, smart phones with 12 mega-pixel camera have been used. The dataset will be released for research purposes in order to facilitate research in this direction. Details regarding the dataset nomenclature and data distribution are provided in the following subsection.

C. Nomenclature and Data Distribution

Other than the videos and high resolution gallery images, the dataset also contains the annotated face images and bounding box coordinates for each face region in a given frame. The bounding boxes are obtained by using the Medianflow [15] and Boosting [11] trackers, coupled with manual inspection and annotations. Fig. 5 presents the distribution of annotated face image dimensions of the proposed dataset. A large portion of the face images are smaller than 64×64, thereby resulting in a challenging set of low resolution samples.

For releasing the dataset, videos are divided into directories based on the surveillance scenarios, i.e. active or passive, which are further divided based on time of capture, followed by the location. As mentioned previously, the dataset contains 58 subjects, each of whom have been given a unique identifier: a number in the range of $1-58$. For each group of subjects, there exists a video for a combination of surveillance use case, time of capture, and location. For each combination, there are 25 videos, each of which have been placed in a separate folder.

Each video folder is named as '$Subject_1$,$Subject_2$' or '$Subject_1$,$Subject_2$,$Subject_3$', depending on whether a given video has two or three subjects. Here $Subject_i$ corresponds to the unique identifier of the i^{th} subject present in the video. For example, if a video contains subjects 23 and 24, the video will be named '*23,24*'. Each folder contains the original video, and subfolders containing the ground truth face images, and a text file containing the bounding box coordinates for the subject. Each face is stored as '*FrameN.jpg*', where N is the frame number. The bounding boxes are available in a text file which provide the frame number, and the top left and bottom right coordinates of the bounding boxes.

[1]https://www.dji.com/phantom-4

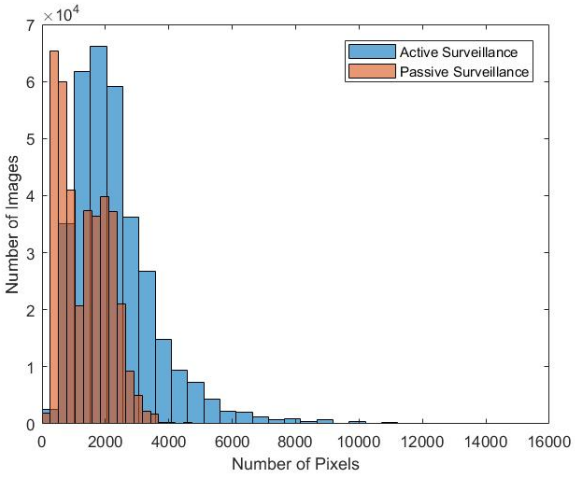


Fig. 5: Histogram of ground truth annotated faces with respect to their size. A large majority of images have less than 4000 pixels, that is images are smaller than 64×64.

III. Protocols and Baseline Results

Protocols and baseline results have been provided for the task of face detection and face recognition. Results are reported for both the surveillance use cases: active and passive surveillance.

A. Protocol

The proposed DroneSURF dataset has been partitioned into subject disjoint training and testing sets. Videos pertaining to 40% of the subjects (24 subjects) are used for testing, while videos of the remaining 34 subjects form the training set. Since each video features 2-3 subjects, this results in the training and testing set containing 120 and 80 videos, or equivalently 252,205 and 159,246 frames, respectively.

B. Face Detection

Baselines for face detection have been computed with two state-of-the-art face detectors: Viola Jones [29] and Tiny Face [14] on the test set of the proposed DroneSURF dataset. In order to obtain the performance of the face detectors, comparison has been performed with the annotated ground truth faces. Both the face detectors provide a bounding box for the detected face, which is classified as a True Positive if it has $> 50\%$ overlap with the ground truth face. For the two face detectors, Viola Jones and Tiny Face, Table III presents the precision and recall values for both scenarios of active and passive surveillance. Fig. 6 also presents the number of false positives with different values of True Positive Rate (TPR) for Tiny Face detector. Some of the key findings are: **(i) Performance of Face Detectors:** Best precision values of 96.52% and 95.36% are obtained for active and passive

TABLE III: Precision (%) and recall (%) obtained for face detection on both use cases of active and passive surveillance.

Algorithm	Active Surveillance		Passive Surveillance	
	Precision	Recall	Precision	Recall
Viola Jones [29]	22.60	27.50	2.15	1.15
Tiny Face [14]	96.52	94.59	95.36	78.80

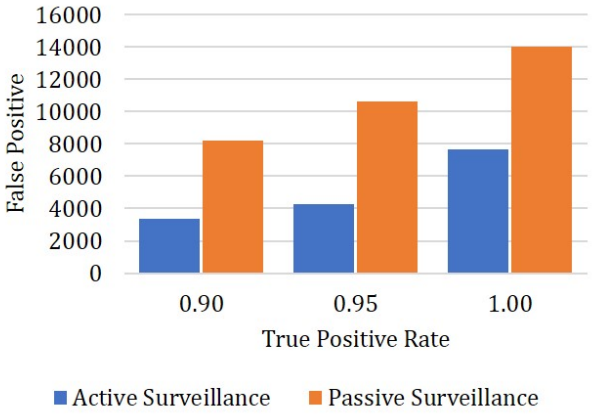
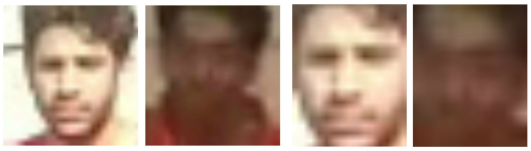


Fig. 6: True Positive Rate versus number of False Positives obtained using TinyFace detector.

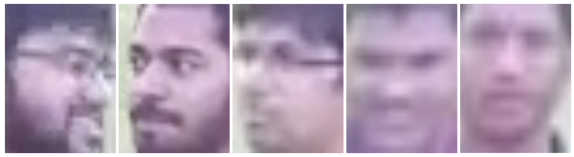
surveillance, respectively, using Tiny Face detector. Similarly, best recall values of 94.59% and 78.80% are obtained with the Tiny Face detector, respectively. Since Tiny Face detector is pre-trained explicitly for detecting faces of lower resolution, it achieves enhanced performance on the proposed DroneSURF dataset. Fig. 7 presents sample face images detected correctly by both the detectors, along with the ones correctly detected by Tiny Face detector only. Frontal face images with minor pose variations are often correctly detected by both the detectors.

(ii) Analysis of Face Detection: Tiny Face and Viola Jones detected a total of 131K and 64K faces for active surveillance, while the ground truth annotated faces are a little over 125K. For passive surveillance, Tiny Face and Viola Jones detected a total of 136K and 35K faces, respectively, for the ground truth annotated faces of over 155K. For Tiny Face, the total number of detected faces is relatively higher than the ground truth faces in the use case of active surveillance. Therefore, increasing the possibilities of false positives at the detection stage. On the other hand, the total number of detected faces by Viola Jones detector is less than half of the total annotated faces.

(iii) Active versus Passive Surveillance: As can be observed from Table III, face detection performance is higher in the use case of active surveillance as compared to passive surveillance, with both the face detectors. In passive surveillance, since the drone does not *actively* monitor the subject's movement, it results in images with high variations in pose, illumination, occlusion, and resolution. The movement of the drone coupled with the independent movement of subjects results in a challenging set of videos. Specifically, face images of the active surveillance use case have an inter-eye distance in the range of (5, 25) pixels, whereas face images



(a) Detected by Viola Jones (b) Detected by TinyFace



(c) Detected by TinyFace but not by Viola Jones

Fig. 7: Sample faces detected by the two face detectors.

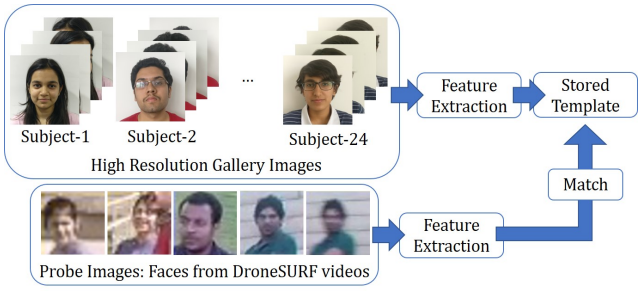


Fig. 8: Protocol for face recognition: high resolution gallery images are matched with faces from the DroneSURF videos.

of passive surveillance have an inter-eye distance in the range of (3,12) pixels. As observed from Fig. 6, with the Tiny Face detector, the number of false positives are much higher for passive surveillance, as compared to active surveillance. In order to obtain a TPR of 0.90, around 3.2K false positives are also processed for active surveillance, on the other hand, the false positives are almost 8K for a TPR of 0.90 for passive surveillance.

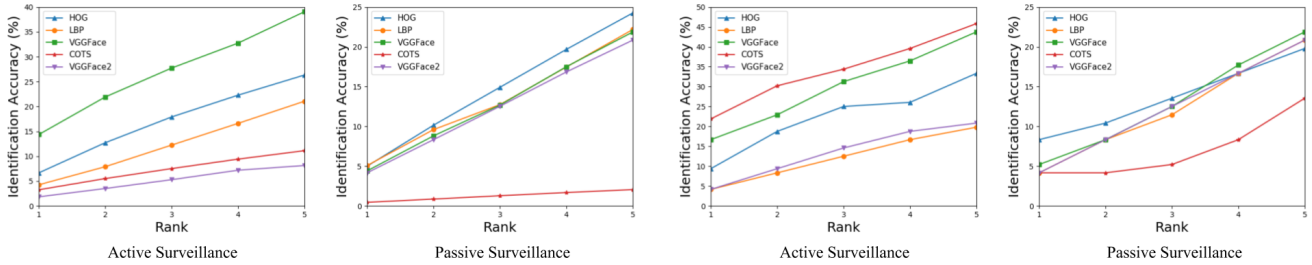
C. Face Recognition

Baselines for face recognition have been computed with two hand-crafted features: **(i)** Histogram of Oriented Gradients (HOG) [8], **(ii)** Local Binary Pattern (LBP) [23], two deep learning based feature extractor: **(iii)** VGG-Face [25], **(iv)** VGG-Face2 [7], and **(v)** a Commercial-Off-The-Shelf system (COTS). As shown in Fig. 8, probe faces of the DroneSURF videos are matched with the high resolution gallery images. Here, probes correspond to the annotated face regions of the DroneSURF videos. Feature extraction is performed on the probe images, using each of the above mentioned techniques, followed by Euclidean distance based matching with the features of the high resolution gallery. Baseline results have been presented for frame-wise identification as well as video-wise identification. In both the protocols, probe corresponds to the annotated face images captured by the drone, and the gallery corresponds to the high resolution face images.

1) Frame-wise Identification: Table IV presents the frame-wise rank-1 identification accuracy obtained for ac-

TABLE IV: Rank-1 accuracy (%) for frame-wise and video-wise identification.

Algorithm	Frame-wise Identification			Video-wise Identification					
	All frames	Frame Selection (500)		All frames		Frame Selection: Alternate		Frame Selection: Quality	
		Alternate	Quality	Min fusion	Mean fusion	Min fusion	Mean fusion	Min fusion	Mean fusion
Active Surveillance									
HOG	6.66	6.19	6.65	8.33	6.25	5.20	5.20	9.37	5.20
LBP	4.26	4.19	4.17	4.16	4.16	4.16	4.16	4.16	4.16
VGGFace	14.36	14.36	16.78	13.54	13.54	9.37	12.50	15.62	16.67
VGGFace2	4.47	4.65	4.83	4.16	4.16	4.16	4.16	4.16	4.16
COTS	3.26	3.04	3.05	5.21	10.42	11.46	13.54	11.46	21.88
Passive Surveillance									
HOG	5.00	5.05	4.45	7.30	4.16	6.25	4.16	8.33	4.16
LBP	5.08	4.12	4.12	4.16	4.16	4.16	4.16	4.16	4.16
VGGFace	4.41	4.66	4.95	2.08	5.20	4.16	5.20	5.20	4.16
VGGFace2	3.86	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16
COTS	0.46	0.41	0.29	1.04	2.08	4.16	2.08	4.16	2.08



(a) Frame-wise Identification

(b) Video-wise Identification

Fig. 9: CMC curves for frame-wise and video-wise identification for both active and passive surveillance.

tive and passive surveillance. Accuracy has been computed frame-wise, for each frame, as well as on frames chosen by means of a frame selection technique. For frame selection, 500 frames are chosen from each video for identification. Two techniques have been used for frame selection: alternate frame selection, where each n^{th} frame is chosen, such that the total selected frames from a video are 500, and selecting the top 500 frames based on a no-reference quality metric, Naturalness Image Quality Evaluator (NIQE) [19]. Some key observations are as follows:

(i) Face Recognition Performance: From the frame-wise identification results reported in Table IV, it can be observed that for all frames, best rank-1 identification performance of 14.36% is obtained with VGG-Face feature descriptor for active surveillance, while an accuracy of 5.08% is achieved with LBP features for passive surveillance. On active surveillance, algorithms other than VGGFace obtain less than 7% rank-1 identification performance, while in case of passive surveillance, all algorithms perform less than 6%.

(ii) Active versus Passive Surveillance: In case of passive surveillance, since the drone does not actively follow the movements of the subjects, it results in videos having high pose and resolution variations, occlusions, and unconstrained movement. The effect of this variation is observed in the identification accuracy, where the performance for passive surveillance is lower than that on active surveillance, for almost all cases of features and frame selection techniques. Fig. 9(a) presents the Cumulative Match Characteristic (CMC) curves on all frames. At rank-5, VGG-Face descriptor achieves the best performance of around 39% for active surveillance, while the highest performance for passive surveillance is only around 24%. Reduced performance for the use case of passive surveillance demonstrates the challenging nature of the given problem.

(iii) Effect of Frame Selection: Table IV also presents the performance of different algorithms after performing frame selection. With VGG-Face, quality-based frame selection results in increased performance for both the scenarios (14.36% to 16.78%, and 4.66% to 4.95%), as compared to alternate frame selection. However, it is interesting to note that the effect of quality-based frame selection on accuracy is not consistent across features and use cases. It is our hypothesis that since the face images are of very low resolution and poor quality, frame selection does not always result in an increased identification performance.

2) Video-wise Identification: Along with frame-wise identification, results have also been computed for video-wise identification. Feature extraction is performed on the manually annotated face regions, followed by Euclidean distance based classification and score-level fusion. Table IV presents the rank-1 identification accuracy (%), for all frames and with frame selection techniques, via two score-level fusion techniques [27]: minimum and average. The results obtained are similar to those of frame-wise identification; some key observations are as follows:

(i) Face Recognition Performance: COTS obtains the best rank-1 performance (21.88%) for active surveillance, while HOG features obtain the best performance (8.33%) for passive surveillance. Fig. 9(b) presents the CMC curves for the

five algorithms. The best combination of frame selection and fusion technique has been plotted for each feature extractor.

(ii) Active versus Passive Surveillance: As observed for the protocol of frame-wise identification as well, the rank-1 performance of active surveillance is better as compared to passive surveillance, across different features and frame selection techniques. These results further strengthen the more challenging nature of face recognition for the use case of passive surveillance. From Fig. 9(b), it is interesting to note that the COTS performs best at rank-5 for active surveillance, reporting an accuracy of around 43%, while VGGFace achieves the best rank-5 performance for passive surveillance (around 22%).

(iii) Effect of Frame Selection: In case of active surveillance, it is interesting to note that 'alternate frame selection', which selects a total of 500 frames from a given video does not improve the recognition performance for the three feature extractors. On the other hand, intelligent frame selection, based on the quality (NIQE) results in an improved rank-1 performance. This demonstrates the requirement of an intelligent system, at the pre-processing stage of frame selection as well [4], [10]. Moreover, consistent with previous results, owing to the bad quality, low resolution, high occlusions, and varying pose of data captured for passive surveillance, a consistent increase in accuracy is not observed upon applying the frame selection techniques.

IV. Conclusion

Drone based aerial monitoring and surveillance has garnered significant attention over the past few years. Recent research has mostly focused on person detection, object detection, or area monitoring. This research presents a novel drone videos dataset, *DroneSURF: Drone Surveillance for Faces*, with specific applicability to face recognition. The proposed dataset contains 200 videos of 58 subjects, captured across different use-cases, locations, and times of the day. Data is captured across 411K frames, having over 786K face annotations. Experimental protocols and baseline results have been provided for the task of face detection and recognition. We believe that the availability of DroneSURF dataset will enable researchers to further explore the problem of face recognition in aerial videos, thereby facilitating the utility of drone based recognition in real world scenarios.

V. Acknowledgement

We thank Praveen Kumar Jhanwar for helping with the data collection. We also thank the volunteers for taking part in the data collection. This research is partially supported through the Infosys Center for Artificial Intelligence, IIIT-Delhi. S. Nagpal is supported via the TCS PhD fellowship.

References

[1] Amazon Air Prime. {https://www.amazon.com/Amazon-Prime-Air/b?ie=UTF8&node=8037720011}.

[2] J. H. Anna Gaszczak, Toby P. Breckon. Real-time people and vehicle detection from UAV imagery. In *SPIE*, volume 7878, pages 7878 – 7878 – 13, 2011.

[3] M. Barekatain, M. Marti, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger. Okutama-Action: an aerial view video dataset for concurrent human action detection. In *IEEE CVPRW*, pages 2153–2160, 2017.

[4] S. Bharadwaj, M. Vatsa, and R. Singh. Biometric quality: a review of fingerprint, iris, and face. *EURASIP Journal on Image and Video Processing*, (1), 2014.

[5] M. Bindemann, M. C. Fysh, S. S. Sage, K. Douglas, and H. M. Tummon. Person identification from aerial footage by a remote-controlled drone. *Scientific Reports*, 7(1), 2017.

[6] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi. Privacy in mini-drone based video surveillance. *IEEE FG*, 2015.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*, pages 67–74, 2018.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society CVPR*, pages 886–893, 2005.

[9] P. Doherty and P. Rudol. A UAV search and rescue scenario with human body detection and geolocalization. In M. A. Orgun and J. Thornton, editors, *Advances in Artificial Intelligence*, 2007.

[10] G. Goswami, M. Vatsa, and R. Singh. Face verification via learned representation on feature-rich video frames. *IEEE T-IFS*, 12(7):1686–1698, 2017.

[11] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *BMVC*, pages 6.1–6.10, 2006.

[12] H.-J. Hsu and K.-T. Chen. Face recognition on drones: Issues and limitations. In *ACM DroNet*, 2015.

[13] H.-J. Hsu and K.-T. Chen. DroneFace: an open dataset for drone research. In *ACM MMSys*, pages 187–192, 2017.

[14] P. Hu and D. Ramanan. Finding tiny faces. In *IEEE CVPR*, pages 1522–1530, 2017.

[15] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *ICPR*, pages 2756–2759, 2010.

[16] N. D. Kalka, B. Maze, J. A. Duncan, K. A. OConnor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. IJB–S: IARPA Janus Surveillance Video Benchmark. In *IEEE BTAS*, 2018.

[17] R. Layne, T. M. Hospedales, and S. Gong. Investigating open-world person re-identification using a drone. In *ECCV*, pages 225–240, 2014.

[18] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer. UAV-enabled intelligent transportation systems for the smart city: Applications and challenges. *IEEE COMML*, 55(3):22–28, 2017.

[19] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE SPL*, 20(3):209–212, 2013.

[20] N. H. Motlagh, M. Bagaa, and T. Taleb. UAV-Based IoT Platform: A Crowd Surveillance Use Case. *IEEE COMML*, 55(2), 2017.

[21] N. H. Motlagh, T. Taleb, and O. Arouk. Low-altitude unmanned aerial vehicles-based internet of things services: Comprehensive survey and future perspectives. *IEEE Internet of Things Journal*, 3(6):899–922, 2016.

[22] J. Nagi, A. Giusti, G. A. Di Caro, and L. M. Gambardella. Human control of UAVs using face pose estimates and hand gestures. In *ACM/IEEE HRI*, pages 252–253, 2014.

[23] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE T-PAMI*, 24(7):971–987, 2002.

[24] R. Palmer. Duchess of cambridge protected by high-tech police drones during royal visit, 2017.

[25] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, 2015.

[26] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. *ECCV*, pages 549–565, 2016.

[27] M. Singh, R. Singh, and A. Ross. A comprehensive overview of biometric fusion. *Information Fusion*, 2019.

[28] W. Staff. Fighting forest fires before they get bigwith drones. {http://www.wired.com/2015/06/fighting-forest-fires-get-big-drones/}, 2015.

[29] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[30] P. Zhu, L. Wen, X. Bian, L. Haibin, and Q. Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018.