BAYESIAN FUSION OF MULTIVIEW HUMAN CROWD DETECTIONS FOR AUTONOMOUS UAV FLEET SAFETY

Efstratios Kakaletsis, Ioannis Mademlis, Nikos Nikolaidis, Ioannis Pitas

Department of Informatics, AIIA Lab, Aristotle University of Thessaloniki
GR-54124 Thessaloniki, GREECE
Emails:{nnik, pitas}@csd.auth.gr

ABSTRACT

In this paper, a Bayesian method for fusing multiple visual human crowd detections (in the form of heatmaps) under an autonomous UAV fleet deployment setting is proposed, aiming at enhanced vision-assisted human crowd avoidance in line with common UAV safety regulations. 2D crowd heatmaps are derived using deep neural human crowd detectors on multiple UAV camera streams covering the same large-scale area over time (e.g., when each drone tracks a different target). Then, these heatmaps are back-projected onto the 3D terrain of the navigation environment. The projected crowd heatmaps are fused by exploiting a Bayesian filtering approach that favors newer crowd observations over older ones. Thus, during flight, an area is marked as crowded (therefore, a no-fly zone) if all, or most, UAV-mounted visual detectors have recently and confidently indicated crowd existence on it. Empirical evaluation on synthetic multiview video sequences depicting human crowds in outdoor environments verifies the efficiency of the proposed method against the no-fusion case.

***Index Terms*—** Multiview, Crowd Detection, Bayesian fusion, Linear Opinion Pool, Drone Safety

1. INTRODUCTION

Camera-equipped Unmanned Aerial Vehicles (UAVs, or "drones") are widely employed for a variety of applications, including media production, search and rescue operations, infrastructure inspection, etc. Cognitive autonomy functionalities, such as visual object/target detection and tracking [1] [2] [3], are gaining more and more traction in current commercial UAVs, since they facilitate significantly easier drone deployment and operation [4] [5] [6] [7]. However, safety concerns constitute an obstacle to more widespread adoption of autonomous UAVs, mainly due to the risk they pose to humans in case of malfunction [8] [9].

Drone flight regulations postulate human crowd avoidance: drones are typically not allowed to fly over a crowded area and must maintain a certain safety distance from the crowd. Thus, in autonomous UAVs, the on-board cognitive functionalities should be partially devoted to implementing these policies. To this end, 2D crowd regions can be detected on video frames using heatmaps, derived through crowd detection approaches that rely on embedded Convolutional Neural Networks (CNNs) [10]. Each heatmap is a grayscale image that spatially corresponds to the RGB video frame it was derived from, but where the luminance value of each pixel represents the probability it depicts human crowd. Subsequently, the crowd regions identified on each 2D heatmap (in pixel coordinates) can be back-projected by raycasting onto the 3D map of the navigation environment [11], to indicate crowd gathering locations that define no-fly zones on this 3D map. In the case of multiple drones being deployed in a fleet setting, with each UAV carrying its own camera, the 3D crowd regions derived from the camera stream of each vehicle can simply be accumulated over time and combined using an OR operator, as in [11]. However, this is suboptimal due to 2D visual crowd detection noise and the possibility of dynamically moving human crowds.

This paper presents a novel method for fusing 3D human crowd detections in an autonomous UAV fleet setting. The algorithm aims at producing semantic 3D map annotations denoting localized human crowds for enhanced, vision-assisted crowd avoidance, since the semantic annotations may be exploited by a path planner to form safe UAV trajectories. The problem setup is as follows: a certain 3D area can be viewed by more than one UAV-mounted cameras, at coinciding or different time instances. The video frames of each camera stream are fed to a 2D visual human crowd detector generating 2D crowd heatmaps [10]. This information is back-projected onto the 3D navigation map using prior knowledge of all drone-mounted camera parameters, independently for each UAV [11] (as shown in Figure 1). The main contribution of this paper is a proposed Bayesian method of centralized multiview 3D crowd detection fusion as a subsequent post-processing step, so as to increase 3D crowd localization accuracy by combating 2D visual crowd detection noise and taking into account the possibly dynamic nature of the detected groups of people.

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 731667 (MULTIDRONE). The authors would like to thank Mr.Charalampos Symeonidis for providing the synthetic dataset.

2. RELATED WORK

Although several works utilize deep CNNs for crowd analysis and understanding, e.g. [12, 13, 14], research on crowd detection in drone-captured images is an uncharted territory. One reason might be that the aerial point-of-view bears additional challenges (e.g., small person size, occlusions etc.), in comparison to a ground point-of-view. Since the crowd first needs to be detected on-frame, relevant algorithms must be capable of efficiently distinguishing between crowded and non-crowded video frames. An application-tailored deep Convolutional Neural Network is presented in [10], where a pre-trained model is finetuned for the task of crowd detection. Moreover, in [15], the authors propose a novel crowd detection method for drone safe landing, based on an extremely lightweight and fast Fully Convolutional Neural Network.

Despite the prominence of Bayesian methods in general data fusion, literature on Bayesian multiview visual information fusion is rather limited. A Bayesian filtering approach for multiview head pose estimation is presented in [16]. The method fuses neural network outputs from multiple camera views. Bayesian multiview filtering was also used for robust multiple camera pedestrian detection and fake pedestrian detection removal in [17]. Finally, Bayesian multiview filtering for body orientation estimation based on silhouette information in a smart room environment is presented in [18].

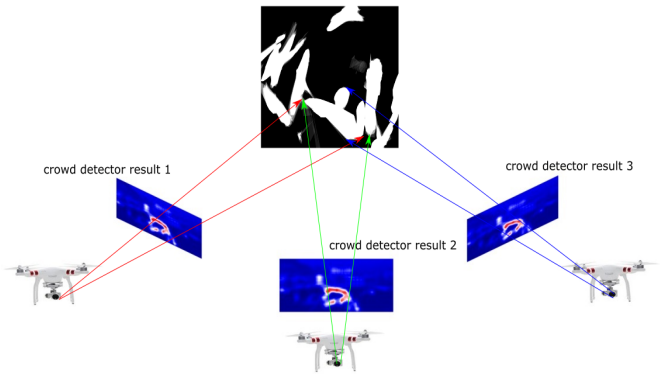

Fig. 1: Multiview human crowd detection/map annotation setup.

3. PROPOSED METHOD

Let us assume that there are N UAVs equipped with the same camera (mounted on a 3DoF gimbal) and 2D visual crowd detector system. Timestamps can be used to synchronise the video streams from the UAV cameras. For each video frame, the crowd detector outputs an estimated probability that there is indeed human crowd visible in each pixel, hence resulting in a crowd heatmap having real pixel values in the domain $[0, 1]$ and a resolution identical to the initial RGB image. A pixel value of $1.0/0.0$ implies absolute detector confidence that human crowd is/is not depicted there, respectively, while a pixel value of 0.5 implies maximum detector uncertainty.

Each pixel can be back-projected on the known 3D terrain using UAV position and camera parameters. Such an approach is presented in [11], that details the implementation in ROS (Robotic Operating System [19]) of a raycasting method to annotate the 3D occupancy grid of an Octomap with crowd information. Octomap is a 3D map format representing the 3D terrain as a regular grid of voxels. The intrinsic and extrinsic camera parameters at each time instance are required to be known and temporally synchronized for each UAV-mounted camera.

Let us also define the following Bernoulli-distributed (binary event space) random variables (RVs):

A: random variable denoting crowd presence in a voxel of the Octomap.

B_i: random variable denoting crowd detection in the corresponding pixel on the heatmap of the i-th UAV.

The posterior probability of a voxel actually containing crowd can be derived according to Bayes theorem and Linear Opinion Pool [20]:

$$P\left(A=1|B_1=b_1,...,B_N=b_N\right) = \sum_{i=1}^{N} w_i \frac{P(B_i=b_i|A=1)P(A=1)}{P(B_i=b_i|A=1)P(A=1)+P(B_i=b_i|A=0)P(A=0)}, \quad (1)$$

where $b_i \in \{0, 1\}$ (non-crowd/crowd) is the detector's binary observation from the i-th UAV, derived by thresholding its raw probability output o_i. In the above:

- $P(B_i = 1|A = 1)$ is the detector's True Positive Rate
- $P(B_i = 1|A = 0)$ is the detector's False Positive Rate
- $P(B_i = 0|A = 1)$ is the detector's False Negative Rate
- $P(B_i = 0|A = 0)$ is the detector's True Negative Rate. All these four probabilities can be evaluated experimentally for the specific crowd detector being viewed.
- $P(A = 1)$ is the a-priori probability of a certain location (voxel) to be occupied by crowd. Under the assumption that human crowds would only gather in potential UAV landing sites [8], due to their flat ground morphology, this probability can be considered equal to the percentage of the area surface that is suitable for landing.

Weights w_i encode the degree to which we take each detector's decision into account regarding the currently processed voxel. This degree should be higher for confident detector outputs (either positive or negative) and lower for more uncertain ones. Given the above, w_i can be determined as the absolute difference between the original detector's output percentage m_i from 0.5:

$$w_i = \frac{|m_i - 0.5|}{\frac{1}{2}N}, \quad (2)$$

where, for voxels on which crowd has been projected, m_i is their raw probability output (derived by averaging the o_i values of the corresponding heatmap pixels), while for voxels

where no crowd has been projected it holds that $m_i = 0$. The denominator in Eq. (2) is required so as to keep all probabilities in Eq. (1) valid (i.e., lying within the interval $[0, 1]$).

Gradually forgetting old detections for each voxel can be introduced by properly weighting over time the contributions of each UAV (e.g., by using a very slow exponential decay), if no newer observation are made regarding that specific voxel. Any new observations can either override the old ones (a new posterior probability is computed from scratch, referred to as Temporal Fusion Policy 1), or be merged with them by properly extending Eq. (1) with additional terms in the sum (Temporal Fusion Policy 2). Both temporal fusion policies compute an aggregate/fused probability $\tilde{P}_t$ of a voxel actually containing crowd at current time instance t, using the following probability blending formula:

$$\tilde{P}_t = (1 - \alpha\gamma_t)\beta_{\Delta t}\tilde{P}_{t'} + \\ +\alpha\gamma_t P\left(A = 1 | B_1 = b_1, ..., B_N = b_N\right)_t, \quad (3)$$

where t' is the last time instance the voxel was visible through any of the UAVs/cameras, $\Delta t = t - t'$, λ is a temporal decay rate hyperparameter, $\beta_{\Delta t} = e^{-\lambda\Delta t}$ and $\tilde{P}_{t'}$ is the last stored aggregate probability that the voxel in question contains crowd, as computed in previous time instance t'. Also, $\alpha \in (0, 1]$ is a hyperparameter regulating the degree to which the current/newest observations override older ones. $\gamma_t \in \{0, 1\}$ is not a parameter, but a binary value denoting whether the voxel being currently processed is visible by at least one UAV at the present time instance t. Thus, $[1 - \alpha\gamma_t]$ evaluates to $1 - \alpha$ when at least one new observation is currently available ($\gamma_t = 1$) and to 1 otherwise ($\gamma_t = 0$).

The current fused probability of crowd existence is computed according to Eq. (1), by employing the detections of current time instance t (if the corresponding area is visible by at least one UAV). Overall, setting α to 1 results in Temporal Fusion Policy 1, while setting it to a real value in the interval $(0, 1)$ results in Temporal Fusion Policy 2.

Using the above setup, the presence of human crowd in an area is considered to be certain if and only if all N UAV-mounted visual detectors concurrently detect it at the present moment with high probability. In case only a subset of the UAVs detect crowds and/or a significant time interval has lapsed since last crowd detection in that area, the output aggregate/fused probability falls dramatically.

4. EMPIRICAL EVALUATION

4.1. Dataset

In order to evaluate the proposed crowd detection approach, synthetic multiview video sequences depicting human crowds in outdoor environments and captured by simulated UAVs were constructed using the UAV simulator Microsoft AirSim, built on top of the real-time 3D graphics engine Unreal Engine 4 (UE4). A UE4 mountainous terrain model was selected and multiple crowds were placed on the sides of a road serving as the path of a bicycle race. The crowds were designed to move slowly along the road while keeping their cohesion. Using AirSim, three UAVs were deployed to follow three cyclists (one drone per cyclist), being quite far apart for one another. Overall, three annotated video sequences (one per drone), each containing more than $K = 1056$ video frames, were created. The annotations included the center of each crowd region (in pixel coordinates) and a per-pixel ground truth segmentation map containing the following five classes: crowd, ground, sky, road, cyclist. An example video frame, along with its ground-truth segmentation map, is shown in Figure 2.

(a) RGB image. (b) Ground truth segmentation map.

Fig. 2: Samples of the multiview human crowd aerial image dataset.

Crowd presence ground-truth on the 3D terrain was derived from the available per-frame segmentation maps, where pixels of the crowd class were processed to form connected components by merging the individuals silhouettes and filling the gaps. These connected components (regions) were then projected to the 3D terrain. The rates being used in Eq. (1) (TPR, FPR, FNR, TNR) were pre-computed by assessing the performance of the employed 2D visual crowd detector algorithm on the annotated dataset it was trained on [10].

4.2. Evaluation Procedure

Performance of the proposed multiview crowd heatmap fusion was measured by Intersection-over-Union (IoU) on the projected crowd detections.We measure the mean IoU over all video frames, so as to observe the performed annotation accuracy of the crowd regions.

The mean IoU is given by:

$$IoU_{mean} = \frac{1}{K}\sum_{i=1}^{K}\frac{Overlap_i}{Union_i}, \quad (4)$$

where K is the total number of video frames, while $Overlap_i$, $Union_i$ are the overlap and the union area, respectively, between the ground-truth information and the projection of crowd prediction regions onto the terrain.

The comparison of the ground-truth and prediction area is conducted as follows: every drone "sees" at each time instance/video frame only part of the 3D world. Thus, we cannot always compare the predicted crowded area with the entire ground-truth map, since at each time instance there may

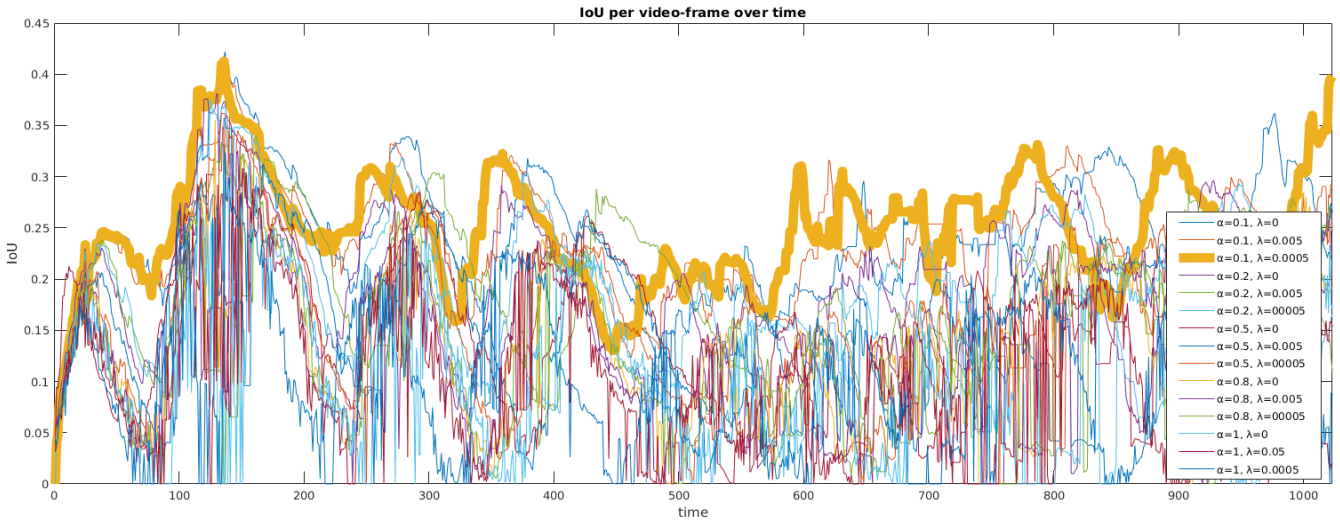

Fig. 3: IoU per video frame. The best-performing variant of the proposed method is highlighted in bold.

exist map areas which none of the UAVs has ever seen up to now. In order to account for this effect during IoU computations, at each video frame we only consider the union of the projections of the camera fields-of-view on the map from all UAVs at all time instances from mission start up to now, instead of the entire map.

Table 1: Empirical evaluation results (mean IoU).

Method	**mIoU** ($\pm std$)
no-fusion [11]	0.1271 ($\pm$0.0872)
multiview $a = 0.1, \lambda = 0$	0.2403 ($\pm$0.0612)
multiview $a = 0.1, \lambda = 0.05$	0.2394 ($\pm$0.0578)
multiview $a = 0.1, \lambda = 0.0005$	**0.2488** ($\pm$0.0562)
multiview $a = 0.2, \lambda = 0$	0.1992 ($\pm$0.0723)
multiview $a = 0.2, \lambda = 0.05$	0.1825 ($\pm$0.0719)
multiview $a = 0.2, \lambda = 0.0005$	0.1957 ($\pm$0.691)
multiview $a = 0.5, \lambda = 0$	0.1434 ($\pm$0.0769)
multiview $a = 0.5, \lambda = 0.05$	0.1532 ($\pm$0.0771)
multiview $a = 0.5, \lambda = 0.0005$	0.1582 ($\pm$0.0790)
multiview $a = 0.8, \lambda = 0$	0.1336 ($\pm$0.0742)
multiview $a = 0.8, \lambda = 0.05$	0.1286 ($\pm$0.0759)
multiview $a = 0.8, \lambda = 0.0005$	0.1281 ($\pm$0.0764)
multiview $a = 1, \lambda = 0$	0.1062 ($\pm$0.0767)
multiview $a = 1, \lambda = 0.05$	0.1069 ($\pm$0.0748)
multiview $a = 1, \lambda = 0.0005$	0.1053 ($\pm$0.0745)

The empirical evaluation results are illustrated in Table 1 and in Fig. 3. Table 1 depicts the mean IoU and the standard deviation of the IoU values. Evidently, the proposed method outperforms the no-fusion method [11] (which accumulates the 3D detections/annotations derived from all drones over time and simply combines them using an OR operator) for the optimal values of α and λ. The maximum mean IoU value is obtained in the case of probability blending with $\alpha = 0.1$, which essentially means that the current 3D detections evaluated using (1) contribute by 10%, whereas the last available 3D detections contribute by the remaining 90%. Increasing α beyond this value (including $\alpha = 1$, which is equivalent to Temporal Fusion Policy 1) leads to algorithm performance deterioration. Regarding λ, one can notice that this parameter has a much less pronounced effect and that the best results are typically obtained by using a very small value. Similar conclusions can be drawn from Fig. 3, depicting a plot of per-frame IoU over time. Each per-frame IoU value is computed only within the union of all drone cameras' fields-of-view, after their projection on the area map, aggregated from mission start up to current video frame. Thus, map areas which have never been seen by any drone until the current time instance, are excluded from IoU computation.

5. CONCLUSION

In this paper, an efficient Bayesian approach for multiview fusion of visual human crowd detections in an autonomous camera-equipped UAV fleet setting is presented. The proposed method aims to increase the accuracy of the crowd location annotations onto a 3D world map in a multiview context, involving convolutional neural visual crowd detectors. Empirical evaluation on a synthetic dataset reliably indicates the superiority of the proposed method in comparison to the no-fusion scenario. Additionally, the presented approach can also be used without significant modifications for fusing other types of detections (also coming in the form of CNN-derived heatmaps), such as potential landing sites. Future work may concentrate on studying such extensions, as well as on evaluating the performance of the proposed approach on real-world multiview crowd datasets.

6. REFERENCES

[1] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "Shot type constraints in UAV cinematography for autonomous target tracking," *Information Sciences*, vol. 506, pp. 273–294, 2020.

[2] R. Cunha, M. Malaca, V. Sampaio, B. Guerreiro, P. Nousi, I. Mademlis, A. Tefas, and I. Pitas, "Gimbal control for vision-based target tracking," in *European Signal Processing Conference, Satellite Workshop (EUSIPCOW)*, 2019.

[3] P. Nousi, I. Mademlis, I. Karakostas, A. Tefas, and I. Pitas, "Embedded UAV real-time visual object detection and tracking," in *Proceedings of the IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2019.

[4] I. Mademlis, V. Mygdalis, N. Nikolaidis, and I. Pitas, "Challenges in autonomous UAV cinematography: an overview," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[5] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments," *IEEE Signal Processing Magazine*, vol. 36, pp. 147–153, 2018.

[6] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and I. Pitas, "High-level multiple-UAV cinematography tools for covering outdoor events," *IEEE Transactions on Broadcasting*, vol. 65, no. 3, pp. 627–635, 2019.

[7] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous UAV cinematography: A tutorial and a formalized shot type taxonomy," *ACM Computing Surveys*, vol. 52, no. 5, pp. 105, 2019.

[8] E. Kakaletsis and N. Nikolaidis, "Potential UAV landing sites detection through Digital Elevation Models analysis," in *European Signal Processing Conference, Satellite Workshop (EUSIPCOW)*, 2019.

[9] C. Symeonidis, I. Mademlis, N. Nikolaidis, and I. Pitas, "Improving neural non-maximum suppression for object detection by exploiting interest-point detectors," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[10] M. Tzelepi and A. Tefas, "Human crowd detection for drone flight safety using Convolutional Neural Networks," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*. IEEE, 2017.

[11] E. Kakaletsis, M. Tzelepi, P. Kaplanoglou, C. Symeonidis, N. Nikolaidis, A. Tefas, and I. Pitas, "Semantic map annotation through UAV video analysis using deep learning models in ROS," in *Proceedings of the International Conference on Multimedia Modeling (MMM)*. Springer, 2019.

[12] L. Boominathan, S. Kruthiventi, and R. Venkatesh Babu, "Crowdnet: a deep convolutional network for dense crowd counting," in *Proceedings of the ACM Multimedia Conference*. ACM, 2016.

[13] J. Shao, K. Kang, C. Change Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[14] D. Babu Sam, S. Surya, and R. Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] G. Castellano, C. Castiello, C. Mencar, and G. Vessio, "Crowd detection for drone safe landing through Fully-Convolutional Neural Networks," in *International Conference on Current Trends in Theory and Practice of Informatics*. Springer, 2020.

[16] M. Voit, K. Nickel, and R. Stiefelhagen, "A Bayesian approach for multi-view head pose estimation," in *Proceedings of the IEEE International Conference on Multi-sensor Fusion and Integration for Intelligent Systems*, 2006.

[17] P. Peng, Y. Tian, Y. Wang, J. Li, and T. Huang, "Robust multiple cameras pedestrian detection with multi-view Bayesian network," *Pattern Recognition*, vol. 48, no. 5, pp. 1760–1772, 2015.

[18] L. Rybok, M. Voit, H. K. Ekenel, and R. Stiefelhagen, "Multi-view based estimation of human upper-body orientation," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 2010.

[19] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "ROS: an open-source Robot Operating System," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, 2009.

[20] M. Stone, "The opinion pool," *The Annals of Mathematical Statistics*, pp. 1339–1342, 1961.