Autonomous UAV Filming in Dynamic Unstructured Outdoor Environments

Ioannis Mademlis[†], Nikos Nikolaidis[†], Anastasios Tefas[†], Ioannis Pitas[†], Tilman Wagner[‡] and Alberto Messina[⋆]

[†]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

[‡]Deutsche Welle, Research and Cooperation Projects, Bonn, Germany

[⋆]Radiotelevisione Italiana (RAI), Centre for Research and Technological Innovation, Torino, Italy

Abstract—**Recent mass commercialization of affordable Unmanned Aerial Vehicles (UAVs, or "drones") has significantly altered the media production landscape, allowing easy acquisition of impressive aerial footage. Relevant applications include production of movies, television shows or commercials, as well as filming outdoor events or news stories for TV. Increased drone autonomy in the near future is expected to reduce shooting costs and shift focus to the creative process, rather than the minutiae of UAV operation. This short overview introduces and surveys the emerging field of autonomous UAV filming, attempting to familiarize the reader with the area and, concurrently, highlight the inherent signal processing aspects and challenges.**

Keywords*—*UAV cinematography, intelligent shooting, autonomous drones

I. Introduction

The rapid popularization of commercial, battery-powered, camera-equipped, Vertical Take-off and Landing (VTOL) Unmanned Aerial Vehicles (UAVs, or "drones") during the past five years, has already affected media production and coverage. UAVs have proven to be an affordable, flexible means for swiftly acquiring impressive aerial footage in diverse scenarios, such as movie/TV shooting, outdoor event coverage for live or delayed broadcast, advertising or newsgathering, partially replacing dollies and helicopters. They offer fast and adaptive shot setup, the ability to hover above a point of interest, access to narrow spaces, as well as the possibility for novel aerial shot types not easily achievable otherwise, at a minimal cost. They are expected to continue rising in popularity, for amateur and professional filmmaking alike [1].

However, a number of challenges arise along with the new opportunities. Severe battery autonomy limitations (typically, less than 25 minutes of flight time), finite bandwidth in the wireless communication channel (e.g., Wi-Fi, 4G/LTE cellular or

The research leading to these results has received funding from the European Union's European Union Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE).

radio link) and safety-motivated legal restrictions, complicate UAV usage and highlight issues that are not present when filming with conventional means. Legal restrictions typically include a requirement for a pilot maintaining direct line-of-sight with the vehicle at all times (fully autonomous civilian drones are illegal), maximum permissible flight altitude and minimum distance from human crowds. Energy consumption restrictions are also important, given the UAV continuous flight time possible with current battery technology, as well as related limitations on processing power and payload weight; the latter are factors that further reduce battery life.

Single-UAV shooting with a manually controlled drone is the norm in media production today, with a director/cinematographer, a pilot and a cameraman typically required for professional filming. Initially, the director specifies the targets to be filmed, i.e., subjects or areas of interest within the scene. Then, (s)he designs a cinematography plan in pre-production, composed of a temporally ordered sequence of target assignments, UAV/camera motion types relative to the current target (e.g., Orbit, Fly-By, etc.) and framing shot types (e.g., Close-Up, Medium Shot, etc.), which the pilot and the cameraman, acting in coordination, attempt subsequently to implement during shooting. In such a setting, each target may only be captured from a specific viewpoint/angle and with a specific framing shot type at any given time instance, limiting the cinematographer's artistic palette. Moreover, there can only be a single target at each time, restricting the scene coverage and resulting in a more static, less immersive visual result. Finally, the "dead" time intervals required for the UAV to travel from one point to another, in order to shoot from a different angle, aim at a different target, or return to the recharging platform, impede smooth and unobstructed filming.

Swarms/fleets of multiple UAVs, composed of many cooperating drones, are a viable option for overcoming the above limitations, by eliminating dead time intervals and maximizing scene coverage, since the participating drones may simultaneously view overlapping portions of space from different positions. Due to the possibly large number of fleet members, a degree of decisional and functional autonomy would significantly ease their control, by lightening the burden on human operators.

However, in civilian applications, it is typical for a human pilot per UAV to be legally required, due to safety considerations and lack of reliable vehicle autonomy. In media production, employing a pilot and a cameraman per drone may increase filming costs prohibitively. Additionally, the cooperation of multiple UAVs inherently gives rise to various coordination challenges, such as that the swarm members need to avoid collisions between them and stay out of each other's field-of-view (*FoV avoidance*), in order for the shooting process to be transparent.

Facing the above issues without prohibitive resource expenditure or human intervention, as well as in a manner that takes into account UAV-specific concerns (e.g., battery autonomy limitations, FoV/collision avoidance, restricted flight zones, etc.), requires intelligent algorithms for automating UAV flight and shooting in concert. Thus, the area of autonomous UAV filming has recently been formed, firmly located at the intersection of aerial cinematography, aerial robotics, computer vision/machine learning and intelligent shooting. Within it, applied signal processing has a significant role to play, especially in the form of real-time image/video analysis and in overcoming communication challenges.

An introduction to this emerging field follows, along with an assessment of its current state and possible directions of future progress. Conforming to recent research [2], the main focus is on outdoor live event coverage, whose challenges include filming over long distances, with possibly vast maps, in (at least partially) unscripted and uncontrolled settings. Different production scenarios involve either only subsets of the challenges presented here, or significantly more controlled shooting settings (e.g., movie sets). Indoor filming comes with its own set of problems, due to more problematic UAV positioning/self-localization, with localization errors giving rise to heightened safety concerns in cluttered environments. However, accurate modern indoor positioning systems exist which mitigate such issues.

II. Intelligent UAV Shooting

Intelligent UAV shooting is a currently emerging research area with significant industry potential. In general, the goal is to automate as much of the media production process as possible, while ensuring adherence to artistic and cinematographic constraints. A few low-hanging fruits have been grabbed, but the general problem is still open and unsolved.

For UAVs, the feasibility of manually designed drone trajectories with regard to vehicle physical limits is an important concern. Recent methods (e.g., [3]) re-time such a trajectory and output an optimized variant guaranteed to be feasible, without disturbing the intended visual content in the captured footage. More importantly, end-to-end systems able to execute single-UAV shooting missions have been developed [4] [5]. Such systems are capable of guiding an UAV outdoors so as to autonomously capture high-quality footage based on cinematographic rules. Static shots and transitions between them are computed automatically, based on well-established visual composition principles and a list of canonical shots. Typically, the user implicitly specifies the UAV path and the shot types to be filmed before executing a drone mission, by prescribing desired “key-frames”, i.e., actual, temporally ordered example video frames of the intended shot within a virtual scene representation, so as to subsequently capture them autonomously during flight. The flight process is automated based on the cinematography plan.

A few commercial applications of similar nature have been released recently. Notably, *Skywand* is a virtual reality system, allowing the user to aerially explore a 3D graphics model of the scene (s)he wants to cover and identify/place desired key-frames within the virtual environment. The system then computes the real UAV trajectory, as well as the corresponding sequence of camera rotations, required for a smooth shot containing these key-frames to actually be filmed. *Freeskies CoPilot* is a mobile software suite, offering similar functionality but with a simple 3D map instead of a VR interface. In both cases, the resulting drone autonomy and environment perception is minimal, the cinematography plan consists simply of example desired key-frames which are cumbersome to define, the computed flight paths are not on-the-fly adjustable and legal restrictions are not being considered.

Although there are examples of algorithms that simply calculate the appropriate number of drones, so as to provide maximum coverage of targets from appropriate viewpoints, in general, little to no effort has been expended towards investigating automated shooting of dynamic scenes in unstructured environments using multiple cooperating UAVs, under battery autonomy, FoV/collision avoidance and flight zone restrictions. Notably, in [6], an on-line real-time planning algorithm is proposed that jointly optimizes feasible trajectories and control inputs for multiple UAVs filming a cluttered dynamic indoor scene with FoV/collision avoidance, by processing user-specified aesthetic objectives and high-level cinematography plans.

III. Autonomous UAV Filming

Automated UAV flight and filming require a number of underlying enabling technologies to be in place, if they are to operate in a satisfactory manner. Below, the relevant state-of-the-art is clustered into three groups: the 2D group, the 3D group and the video capture/communication group.

A. The 2D Group

The first required technology group, hereafter called "2D group", heavily involves image/video processing and semantic analysis operating on the image plane. It consists of a combination of 2D visual target detection, 2D visual target tracking and image-based visual servoing. In principle, it is feasible for all the above tasks to be performed in real-time by computer vision and machine learning algorithms, using only the monocular camera also employed for shooting.

2D visual target detection is necessary for localizing the target's image (i.e., the Region-of-Interest, or ROI) on a video frame, so that the system knows exactly how to rotate the camera in order to achieve central composition framing. Additionally, visual target detectors can also be exploited for identifying a possible obstacle or an on-ground UAV landing site. The extracted ROI is a rectangle (described in pixel coordinates) that encloses the target's image. In currently available drones, similar methods are already employed to better adjust a manually pre-specified ROI, based on the video content. In the future, more automated UAVs are expected to rely solely on automatic visual target detection. Relevant state-of-the-art algorithms, based on deep neural networks, are impressively accurate and optimized for parallel execution on General-Purpose Graphical Processing Units (GP-GPUs). Such high-performance hardware has recently been commercialized in small, power-efficient form factor for embedded systems, ideal for on-board inclusion in UAVs[1]. However, current processing power and energy consumption restrictions limit what is possible on a UAV, in comparison to desktop computers.

2D visual target tracking tracks a pre-specified ROI on the consecutive frames of a video sequence, by taking advantage of spatiotemporal locality constraints, and updates the ROI pixel coordinates at each video frame. Although tracking can be performed by simply re-detecting the target at each video frame, a better approach is to periodically re-initialize the ROI using a 2D visual target detector and employ a separate visual tracker for the intermediate intervals. Correlation filter-based trackers are suitable for real-time operation [7]. Although it is very difficult to achieve top accuracy in real-time with current state-of-the-art 2D visual detectors and trackers, given the processing power limitations of UAV hardware, future progress in reducing their computational requirements, e.g., by novel research in lightweight neural networks, is expected to alleviate this issue.

Image-based visual servoing can be used for properly rotating the camera and sending suitable motion commands to the UAV motors, so as to achieve a specific cinematography (e.g., maintaining central composition framing) or control (e.g., landing) purpose in an autonomous manner. In essence, it is a visual feedback control loop that only requires a target ROI, possibly automatically derived from 2D visual detection/tracking, as input. More advanced visual servoing can also be employed for controlling UAV motion so as to autonomously capture a number of desired shot types based solely on visual input.

An alternative to image-based visual servoing is reinforcement learning employing raw video input and motor command output. Thus, any need for accurate vehicle or environment models is bypassed and the resulting controller is more adaptive to dynamic situations, at the cost of losing precise, analytic solutions and requiring advanced robotics simulator software and/or large properly annotated image datasets for training. Deep neural networks have recently been employed in similar settings for UAV collision avoidance, indoor flight control in search and recovery operations or high-level flight navigation [8]. An imitation learning variant has also been explored for drone racing [9], where a neural network learns to map video input to proper motor control commands in a supervised setting, using datasets obtained by employing human pilots in a photorealistic simulator. However, such approaches have not yet been investigated for cinematography applications.

In general, the methods contained in the 2D group suffice for autonomously achieving physical target following and rudimentary cinematic coverage by the drone, as well as effective landing.

B. The 3D Group

The second required technology group, hereafter called “3D group”, operates on top of the first one and consists of a set of methods and devices that allow functioning in global 3D Cartesian space. These technologies are essential to achieve fully autonomous, non-trivial UAV filming with safe and

[1] E.g., the NVIDIA Jetson series

effective obstacle/collision avoidance. This is mainly achieved by employing Visual Simultaneous Localization and Mapping (SLAM), as well as by the presence of Global Positioning System (GPS) receivers on-board the UAV and (ideally) on the targets being filmed.

Visual SLAM [10] can be used to detect and avoid obstacles during flight time, by mapping the immediate environment and localizing the drone with respect to that 3D map. Localization includes an estimation for both the position and the orientation of the UAV-mounted camera at each time instance. Visual SLAM performs an incremental 3D scene reconstruction based on the camera feed, using a real-time, on-line variant of Structure-from-Motion algorithms, augmented by visual place recognition, graph-based map modelling and loop closure modules. The computed map is typically a 3D point cloud, either sparse, semi-dense or dense, with the first estimated location of the UAV employed as the arbitrary origin of the map coordinate system. However, since a point cloud cannot distinguish between unobserved and observed-to-be-empty space, different approaches are typically employed for safe map representation in autonomous vehicles (Octomap [11], an octree-based 3D occupancy grid, is a popular choice).

Despite the fact that Visual SLAM-based obstacle detection can, in principle, be performed using a single camera, additional sensors may greatly enhance the algorithm effectiveness. Such sensors can include an altimeter and an ultrasound module for assisting in obstacle avoidance, as well as a secondary stereoscopic camera and an Inertial Measurement Unit (IMU) for more robust Visual SLAM. Actually, altimeters, IMUs and ultrasonic sensors constitute standard equipment for all professional drones. On the other hand, Light Detection and Ranging sensors (LIDARs) are more rarely employed visual sensors that may be used instead of stereoscopic 3D cameras in order to achieve increased accuracy and performance, as well as robustness to variable environmental lighting conditions, for tasks such as SLAM. Their main strength derives from the dense 3D scene reconstructions of unmatched quality they can provide. Although, currently, top LIDARs have lower refresh rate, lower resolution, lack of color perception, greater weight and significantly higher cost than a good camera, it is very likely that future technology improvements will increase their appeal.

The 3D maps built by Visual SLAM (preferably, by jointly exploiting stereoscopic 3D camera and IMU inputs) can be aligned with the common GPS coordinate frame, using a similarity transformation, and employed for assisting in global target, obstacle and UAV localization, leading to more robust operation exploiting multiple information sources.

The dynamic 3D map built and constantly maintained by the drone can then serve as input to a 3D path planning algorithm. Such algorithms for UAVs are currently able to deal with complex dynamic and kinematic constraints in real-time, resulting in nearly-optimal collision-free paths being computed on-line. Thus, everything seen by the camera may be mapped onto a common 3D world coordinate system and elaborate UAV motion trajectories can be planned, so as to autonomously capture any cinematic shot type desired. Due to the dynamic nature of the environment, path planning may take place in two levels: a high-level long-term plan must be devised periodically, or when important events are detected, while during the intermediate intervals a low-level plan can locally adjust that path according to the current situation (e.g., in case a moving target suddenly changes motion direction) or cinematography requirements. The need for such a partitioning, however, can be reduced (to a degree) if the vehicle paths are always being planned in a variable, target-centered coordinate system, thus outputting a set of temporally ordered waypoints relative to the target. Subsequently, at each time instance during the actual execution of the path plan, the next relative waypoint can be located on-the-fly in the global 3D map, by utilizing the known, current target 3D position in the GPS coordinate frame.

Low-level motion control is an issue directly related to path planning, since it involves the actual execution of the current path plan. For VTOL UAVs, such as quadrotors, motion control relying on GPS-IMU fusion is already a mature technology. In general, Proportional-Integral-Derivative (PID) or Linear-Quadratic Regulator (LQR) controllers are employed for related tasks. The PixHawk/PX4 Autopilot, a popular low-level flight trajectory control system, offers a commercial off-the-shelf PID cascade control solution for UAVs that allows vehicle steering at various levels, ranging from designating path waypoints to directly feeding raw motion commands to the motors.

The fusion of IMU, GPS and Visual SLAM information, in principle, allows accurate, real-time, global UAV localization in both position and orientation. Targets, on the other hand, can only be localized with regard to their position. However, target orientation must be known in order to accurately steer the UAV and guide the shooting process so as to autonomously capture a number of non-trivial shot types (for instance, consider the cinematographic requirement of filming a subject from a very specific view angle). Luckily, operating in global 3D Cartesian coordinates makes it meaningful to integrate a 3D visual target pose estimation algorithm into the vision processing pipeline, thus bringing image/video analysis to the forefront once more. There are two main approaches to achieve this: the computer vision approach, where predefined landmark points are detected/tracked on the target's image and used to solve the Perspective-n-Point problem, or the machine learning approach, where the target's pose is directly regressed by a trained model that only uses the visual input. The first approach requires a 3D model of the target to be known, while the second solution requires a regressor properly trained on a representative, fully annotated image dataset. The machine learning approach, in case a deep neural regressor is employed, allows integration with the 2D visual target detector and execution on a GP-GPU in real-time, as a unified neural network. However, no commercial UAV offers such capabilities yet.

The existence of the global, dynamic 3D map also makes it meaningful to detect human crowds in the 2D visual input. This process can also be integrated into the 2D group, using a deep neural network running on GP-GPU in real-time [12]. Subsequently, the detected crowd ROI (in pixel coordinates) may be mapped to the relevant terrain areas of the 3D map by perspective back-projection, so as to achieve a semantic annotation of the map. This is important, due to legal regulations restricting UAV flight above human crowds. A similar process can be followed for recognizing and localizing potential emergency landing sites and flying towards them if needed.

Typically, the GPS signal is not available indoors and it may even be temporarily lost outdoors. Additionally, its usual position error is up to 5 meters. These problems can be bypassed by employing differential GPS units (accurate in the range of approximately 20 cm), by IMU/GPS/Visual SLAM fused localization and by replacing GPS with an Active Radio-Frequency IDentification (RFID) or a Wireless Positioning System (WPS) solution in GPS-denied environments. These approaches, however, come with associated monetary and computational costs, which explains the fact that state-of-the-art commercial UAVs lack several capabilities derived from the 3D group, despite being universally equipped with simple GPS receivers.

C. Video Capture and Communication Group

Infrastructure for communications and related issues is critical for successful deployment of UAV swarms in practical scenarios, especially in live event media coverage applications. Even in single-UAV missions it is challenging to stream high-resolution video (especially 4K UHD, i.e., the norm in media production) down to a ground station with Quality-of-Service (QoS) guarantees, while simultaneously executing all of the previously described algorithms in real-time. Video acquisition, compression, synchronization and transmission are procedures easily implemented using professional cameras and open-source software, although the lack of media production-quality camera models with Camera Serial Interface (CSI) connectivity (allowing rapid and stable capture for reliable on-line processing) is an existing practical issue. However, they jointly consume significant processing power and energy, on a computing platform already strained in these resources. The issue cannot simply be solved by dedicated hardware, since the latter would come with additional energy consumption, monetary and weight overhead. Therefore, at the current stage of technology, a trade-off has to be made between the broadcast video resolution, the hardware cost and the level of vehicle cognitive autonomy.

In simpler, non-live coverage, i.e., when filming for deferred broadcast, or shooting a scripted sequence, on-the-fly video transmission is not required (video may simply be stored on-board and retrieved later). In fact, if all processing is performed on-board in a completely autonomous manner, there is not even need for networking. However, communications are required in all other cases, including the non-live single-UAV filming where a subset of the less critical algorithms previously described, e.g., crowd/landing site detection and high-level path planning, are executed on a computationally powerful ground station, at the cost of significant latency (at best, about one hundred milliseconds). In general, a private QoS-guaranteeing 4G/LTE infrastructure suffices for the task, given the high mobility of the UAVs and the possibly long distances that need to be covered in outdoor event filming. Traditional Wi-Fi is a less costly, suboptimal alternative with higher latency and significantly smaller range, while public LTE networks are not reliable due to the lack of a way to prioritize UAV communications over telephony. The main challenge lies in live broadcasting; even private LTE will not allow consistent 4K UHD video streaming, unavoidably leading to a fall back on FullHD resolution.

If a swarm of multiple cooperating UAVs is employed, additional issues arise. Most importantly, in live coverage, the available bandwidth may not be enough to support live FullHD video streaming from all drones concurrently, resulting in a hard upper limit on the number of drones (a simple linear relation exists between the required total bandwidth and the number of employed UAVs). Furthermore, if direct coordination between the drones themselves is required (so as to autonomously capture a multiple-UAV shot, to execute distributed variants of algorithms such as SLAM, or simply for redundancy/fault tolerance), then an intra-swarm Flying Ad Hoc Network (FANET) should be employed, supporting ad hoc routing and accounting for high node mobility, long distances and rapidly varying network topology. Despite recent advances, FANETs are not yet a mature technology; for actual deployment, either custom, optimized Wi-Fi extensions must be developed, or falling back to LTE infrastructure is unavoidable, at the cost of increased latency.

IV. Autonomous Features in Current Commercial UAVs

The employed algorithms in current commercial drones do not cover the entire range of the research methods presented in Section III. For instance, instead of pure image-based visual servoing, more traditional optimal control methods are typically employed, where control signals are computed by explicitly constructing trajectories through configuration space, subject to costs formulated in image space. Other tasks, such as 3D target pose estimation or human crowd detection, are not being performed at all, while learnt control policies (e.g., via reinforcement or imitation learning) are not commonly utilized outside laboratory settings. Advances in processing hardware (e.g., using the NVIDIA Jetson TX2 board, or a future model) and algorithm efficiency/performance are expected to reduce the gap between research and commercial implementations/capabilities of autonomous UAV features.

The presented technologies are visualized in Figure 1, where the ones currently appearing only in research settings are clearly separated from methods already employed in commercial UAVs. The methods in the 2D and 3D group are further examined in Figure 2, where the input/output exchanges between them and the most important sensors are visible.

The two most popular commercial state-of-the-art UAVs for videography purposes are DJI Phantom IV Pro (employing the Intel Movidius Myriad 2 Vision

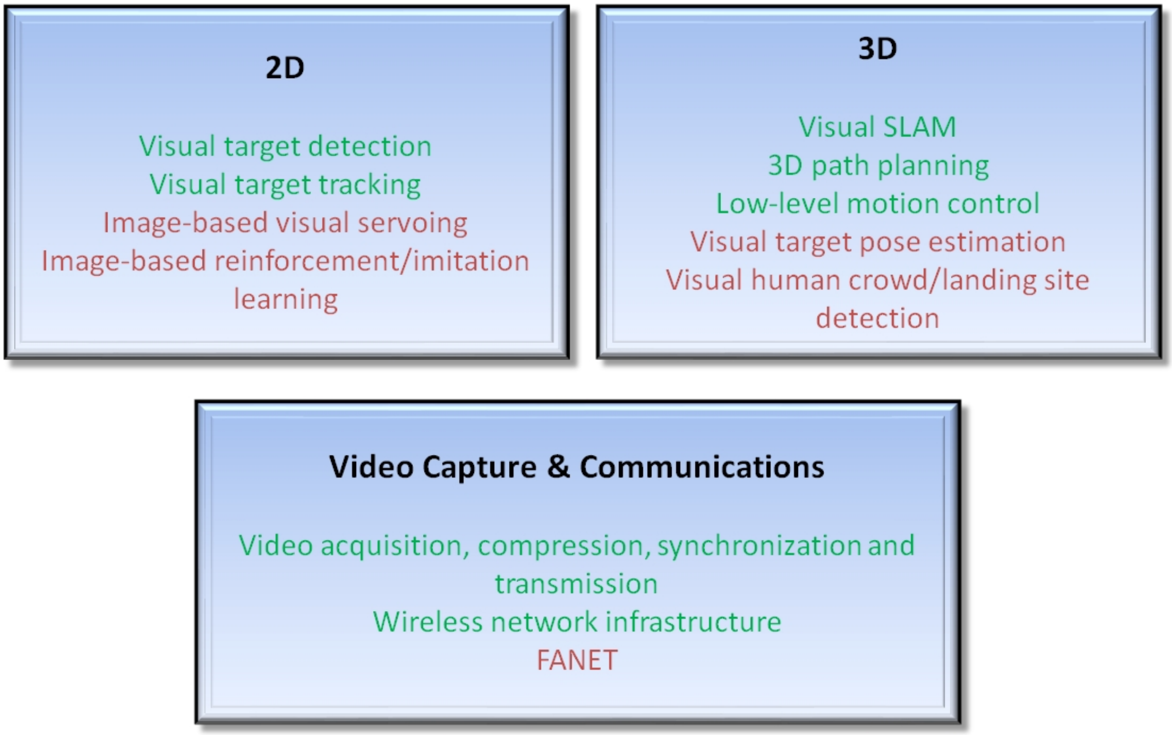
2D
Visual target detection
Visual target tracking
Image-based visual servoing
Image-based reinforcement/imitation learning
3D
Visual SLAM
3D path planning
Low-level motion control
Visual target pose estimation
Visual human crowd/landing site detection
Video Capture & Communications
Video acquisition, compression, synchronization and transmission
Wireless network infrastructure
FANET

Fig. 1: A visualization of the presented technologies, clustered in three groups. Within each group, the methods currently only appearing in research settings are written in a red font, while the methods currently employed in commercial UAVs are written in a green font.

Processing Unit) and the more recent Skydio R1 (built around the more powerful NVIDIA Tegra X1 System-on-a-Chip). They offer similar autonomous capabilities, such as obstacle detection and avoidance, automated landing, physical target following/target orbiting enabled by visual target tracking (for low-speed, manually pre-selected targets), as well as automatic central composition framing, i.e., continuously rotating the camera so as to always keep the pre-selected target properly framed at the center.

However, Skydio R1 is a more advanced platform due to the more capable computing hardware and the multiple pairs of stereoscopic cameras, cooperating to build a 3D occupancy volume as an environment map. It integrates improved Visual SLAM, path planning and deep learning object detection functionalities. Its main selling point is the impressive obstacle avoidance behaviour, even in highly cluttered spaces. However, the resulting footage is typically lacking in cinematic quality, since the encoded knowledge about cinematography is rudimentary and there is no integration with intelligent shooting algorithms.

V. Future Prospects

During the 21st century, UAVs have evolved from remotely controlled curiosities with purely military applications into a technological revolution, taking multiple industries by storm and paving the way for massively available embodied autonomous agents. Aerial cinematography has already been transformed

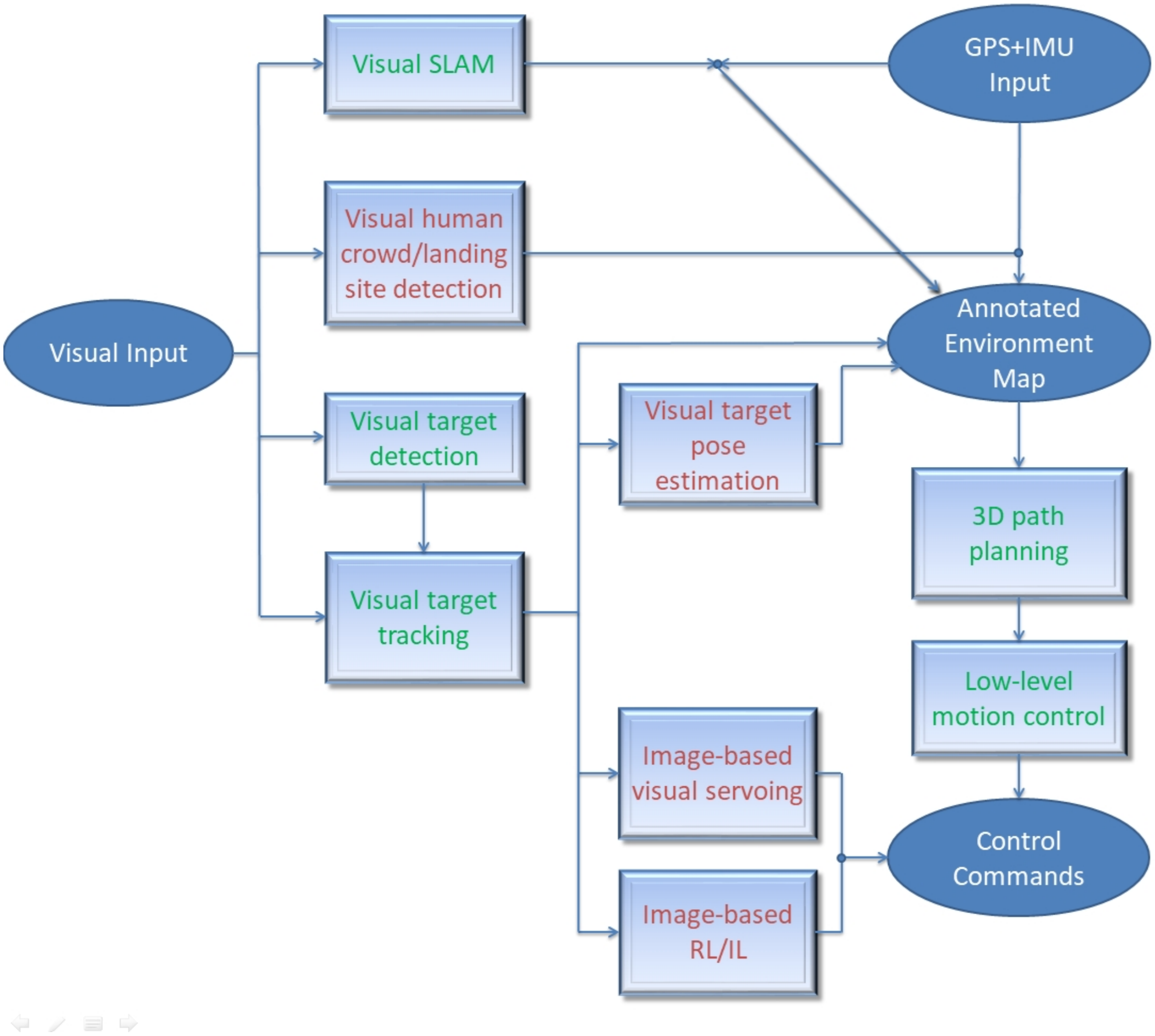

Fig. 2: A visualization of the input/output exchanges between the presented technologies from the 2D/3D groups and the most important sensors.

by the easy availability of advanced VTOL drones, but there is still a lot of room for improvements in multiple aspects. The currently limited UAV autonomy, the lack of commercial off-the-self cooperative UAV swarm platforms, the multitude of complications arising from legal or technological restrictions, as well as the absence of multiple-UAV cinematography expertise, are all issues prescribing directions for advancement.

We can easily imagine an ideal scenario where a director gives high-level, concise cinematography instructions in near-natural language before filming. Subsequently, a fully autonomous UAV swarm would acquire the desired footage, while constantly and optimally adapting to the ever-changing situations arising within the shooting area, under the minimal oversight of a single flight supervisor. In a less ambitious variant, arguably more realistic at the current level of technology, the director would come up with a detailed cinematography plan and, if deemed necessary, would be able to manually intervene during production.

For both scenarios, further advancements are required in order to realize them. Beyond upgrades in sensor technology and computational hardware, progress in UAV cognitive and functional autonomy, enabled by improvements in real-time image/video analysis and mobile networking, respectively, have to be attained in the near future.

References

[1] C. Smith, *The Photographer's Guide to Drones*, Rocky Nook, 2016.

[2] 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.

[3] M. Roberts and P. Hanrahan, "Generating dynamically feasible trajectories for quadrotor cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 61, 2016.

[4] N. Joubert, M. Roberts, A. Truong, F. Berthouzoz, and P. Hanrahan, "An interactive tool for designing quadrotor camera shots," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 238, 2015.

[5] N. Joubert, D. B. Goldman, F. Berthouzoz, M. Roberts, J. A. Landay, and P. Hanrahan, "Towards a drone cinematographer: Guiding quadrotor cameras using visual composition principles," *arXiv preprint arXiv:1610.01691*, 2016.

[6] T. Nägeli, L. Meier, A. Domahidi, J. Alonso-Mora, and O. Hilliges, "Real-time planning for automated multi-view drone cinematography," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 132:1–132:10, 2017.

[7] O. Zachariadis, V. Mygdalis, I. Mademlis, N. Nikolaidis, and I. Pitas, "2D visual tracking for sports UAV cinematography applications," in *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017.

[8] A. Carrio, C. Sampedro, A. Rodriguez-Ramos, and P. Campoy, "A review of deep learning methods and applications for unmanned aerial vehicles," *Journal of Sensors*, vol. 2017, 2017.

[9] G. Li, M. Mueller, V. Casser, N. Smith, D. L. Michels, and B. Ghanem, "Teaching UAVs to race with observational imitation learning," *arXiv preprint arXiv:1803.01129*, 2018.

[10] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[11] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[12] M. Tzelepi and A. Tefas, "Human crowd detection for drone flight safety using convolutional neural networks," in *Proceedings of EURASIP European Signal Processing Conference (EUSIPCO)*, 2017.