Autonomous UAV Cinematography: A Tutorial and a Formalized Shot Type Taxonomy

Ioannis Mademlis[†], Nikos Nikolaidis[†], Anastasios Tefas[†], Ioannis Pitas[† °], Tilman Wagner[‡] and Alberto Messina[⋆]
[†]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[°]Department of Electrical and Electronic Engineering, University of Bristol, Bristol, UK
[‡]Deutsche Welle, Research and Cooperation Projects, Bonn, Germany
[⋆]RAI Radiotelevisione Italiana Centre for Research and Technological Innovation, Torino, Italy

***Abstract*—The emerging field of autonomous UAV cinematography is examined through a tutorial for non-experts, which also presents the required underlying technologies and connections with different UAV application domains. Current industry practices are formalized by presenting a UAV shot type taxonomy composed of framing shot types, single-UAV camera motion types and multiple-UAV camera motion types. Visually pleasing combinations of framing shot types and camera motion types are identified, while the presented camera motion types are modelled geometrically and graded into distinct energy consumption classes and required technology complexity levels for autonomous capture. Two specific strategies are prescribed, namely *focal length compensation* and *multidrone compensation*, for partially overcoming a number of issues arising in UAV live outdoor event coverage, deemed as the most complex UAV cinematography scenario. Finally, the shot types compatible with each compensation strategy are explicitly identified. Overall, this tutorial both familiarizes readers coming from different backgrounds with the topic in a structured manner and lays necessary groundwork for future advancements.**

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

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. UAVs have proven to be an affordable, flexible means for swiftly acquiring impressive aerial footage in diverse scenarios. They are suitable for movie/TV filming, 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. UAVs are expected to continue rising in popularity, for amateur and professional filmmaking alike [1] [2].

An illuminating example regarding the creative possibilities made available by UAV cinematography can be found in the introductory scene of the animated movie "The Secret Life of Pets" (2016). The film starts with a low-altitude aerial shot of the Statue of Liberty that climbs rapidly toward its head, bypassing it and then flying swiftly along a nearby suspension bridge. The virtual camera looks vertically down, towards the deck, from a constant altitude. Subsequently, the camera abandons the bridge and starts orbiting around a neighbouring skyscraper, until suddenly flying-by it to reveal an establishing shot of New York City. The camera then dives into the nearby Central Park, flying above the pedestrians, until the protagonist is seen riding a bicycle. Finally, the virtual camera starts tracking the linearly moving bicycle from the side, maintaining a steady distance and matching its speed. Although it could be argued that the above sequence mostly illustrates the artistic freedom inherent in 3D computer graphics, it is also true that UAVs are the best medium for filming such a sequence in live-action footage.

However, a number of challenges arise along with the new opportunities. Severe battery autonomy limitations, finite bandwidth in the wireless communication channel (e.g., WiFi, 4G cellular or radio link) and safety-motivated legal restrictions, complicate UAV usage and highlight issues that are not present in conventional filming. Legal restrictions typically require a direct line-of-sight between pilot and vehicle at all times, while also defining a maximum permissible flight altitude and a minimum distance from human crowds. Energy consumption restrictions are also important, given the UAV continuous flight time possible with current battery technology (typically, less than 25 minutes). Related limitations on processing power and payload weight derive from the fact that these are factors further reducing battery life.

Single-UAV filming 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 coverage. Initially, the director specifies the targets to be filmed, i.e., subjects or areas of interest within the scene. Then, 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.)[1]. The pilot and the cameraman, acting in coordination, attempt subsequently to implement this plan during mission execution. In such a

[1]We do not consider here other elements of cinematography, i.e., scene lighting and depth-of-field/focus, since we assume natural lighting in outdoor settings and fully focused images, so as to facilitate their analysis by computer vision algorithms.

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. They eliminate dead time intervals and maximize scene coverage, since the participating drones may simultaneously view overlapping portions of space from different positions. Due to the possibly large number of swarm members, a degree of cognitive autonomy would significantly ease their control, by lightening the burden on human operators. However, currently, drone swarms are mainly restricted to military applications, while full autonomy is still in embryonic stage. Existing swarms operate best at a human-oriented semi-autonomy level, i.e., a human operator regularly sends commands to a swarm leader (either pre-selected, or dynamically assigned) and receives back periodic status reports plus sensor information [3]. Additionally, swarm members interact with each other and with the leader UAV for purposes of coordination, task assignment and simple decision-making. Human supervision from the ground is not absolutely necessary, but typically facilitates more effective swarm behaviour, since humans tend to perform better at target identification and action prediction [4], while autonomous operation is superior in lower-level tasks, such as path planning and target following [5]. A leader-follower architecture combined with autonomous intra-swarm coordination allows easier operation by a single ground supervisor [6], permitting him to treat the swarm as a single entity.

The situation is currently different in civilian applications, where 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 filming 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 filming in concert. To achieve this, a formally standardized vocabulary of UAV cinematography building blocks and a systematic identification of UAV-related issues in the context of media production/broadcasting, along with possible solutions, is required. However, since UAV cinematography is an emerging topic, it has only recently been brought under research focus and, therefore, no clearly defined taxonomy of shot types achievable with a single UAV, or a swarm of multiple cooperating UAVs, has been specified.

The purpose of this article is both to familiarize readers coming from different backgrounds with the topic, as well as to lay exactly the necessary groundwork for future advancements. Thus, it attempts to formalize current practices, based on accumulated professional experience, in order to catalyze further research. Below, the recently formed field of autonomous UAV cinematography, firmly located at the intersection of aerial cinematography, aerial robotics, computer vision/machine learning and intelligent filming, is introduced and examined from multiple aspects. Background is tersely provided for readers who are not specialists in robotics or cinematography. Following preliminary research work ([7], [8], [9], [10], [11], [12], [13], [14], [15]), the main focus is on outdoor live event coverage, deemed as the most complex application of UAV cinematography. Different production scenarios involve either only subsets of the challenges presented here, or significantly more controlled filming settings (e.g., movie sets). Therefore, this tutorial is relevant to any kind of UAV media production.

The remainder of the article is organized in the following way. Section II briefly introduces the field of intelligent filming, paying particular attention to the use of UAV-mounted cameras. Section III reviews the underlying technologies which make autonomous UAV filming possible, including robotics and computer vision/machine learning modules. Additionally, connections with different, non-cinematography UAV applications are examined. Sections IV and V present an organized vocabulary of single-UAV cinematography, focusing on framing shot types (FSTs) and camera motion types (CMTs), respectively. Visually pleasing combinations of FSTs and CMTs are identified, while the presented camera motion types are mathematically modeled, clustered into four separate groups and graded into distinct energy consumption and required technology complexity levels. Thus, a complete UAV shot type taxonomy is detailed. Section VI extends this taxonomy to cinematographically meaningful multiple-UAV camera motion types. Section VII provides examples of how this formalized taxonomy can be exploited for facilitating UAV cognitive autonomy algorithms. Section VIII offers two specific strategies, namely *focal length compensation* and *multidrone compensation*, for partially overcoming a number of issues arising in UAV outdoor event coverage. The shot types compatible with each compensation strategy are explicitly identified. Section IX briefly presents conclusions drawn from the preceding discussion and prescribes future research directions.

II. Technologies for Intelligent Filming

Intelligent filming/editing is a recent research topic, with autonomous UAV filming being a currently emerging subfield. In general, the goal is to automate as much of the media production process as possible, while ensuring adherence to artistic and cinematographic constraints. Although a few low-hanging fruits have been grabbed, the general problem is still open and unsolved. Below, the relevant state-of-the-art is briefly reviewed, with an emphasis on UAV filming.

In [16] an optimization-based algorithm is presented that processes off-line a single high resolution video from a static

camera filming a staged event. It outputs a set of virtual pan-tilt-zoom (PTZ) moving camera feeds obeying cinematographic principles and user-provided constraints, regarding which actors should be contained and how they should be framed. The algorithm exploits face recognition/tracking sub-modules as a pre-processing step for the identification of actors, with no 3D perception of the stage involved.

In [17] the video input from a robotic camera visually tracking and physically following a pre-defined region (e.g., the current centroid of all players in a basketball game) is processed in real-time. It produces a virtual camera video output following a smooth, aesthetically pleasing trajectory, by employing a path planning algorithm and frame resampling. In contrast to most intelligent filming methods, this approach exploits special hardware.

In the case of UAV filming, the feasibility of manually designed drone trajectories with regard to vehicle physical limits is an important concern. The method in [18] re-times such a trajectory and outputs an optimized variant guaranteed to be feasible, without disturbing the intended visual content in the captured footage. A more general approach that is not specifically designed for cinematography applications is presented in [19], where custom high-level user goals are taken into account (e.g., codifying cinematography goals).

Intelligent filming/editing with multiple cameras presents additional challenges, highlighting the editing aspect. In [20] an optimization-based algorithm is presented for the computation of a single, aesthetically pleasing video, conforming to basic editing guidelines (such as the 180-degree rule and jump cut avoidance [21]), from raw feeds coming separately from multiple cameras. Operating also within a multi-camera context, the work in [22] approaches automated editing as a problem of camera selection over time and models it with a Partially Observable Markov Decision Process over temporal windows. A research effort oriented towards multi-camera UAV footage is [23], presenting an autonomous system that calculates the appropriate number of drones, in order to provide maximum coverage of targets from appropriate viewpoints.

End-to-end systems able to execute single-UAV filming missions have been developed as well. Such a system is presented in [24], 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. In [25], the authors present a tool that allows the user to implicitly specify the UAV path and the shot types to be filmed before executing a drone mission. This is done by prescribing desired "key-frames", i.e., actual, temporally ordered example video frames of the intended shot, so as to subsequently capture them autonomously during flight. In both cases, as well as in [26], the flight process is automated based on the cinematography plan, but no dynamic adaptation or active environment perception is involved.

A few commercial applications, also oriented towards outdoor single-UAV cinematography planning, have been released recently. Notably, *Skywand* [27] is a virtual reality system, allowing the user to aerially explore a 3D graphics model of the scene 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* [28] 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 in example key-frames, the computed flight paths are not on-the-fly adjustable and legal restrictions are not being considered.

Little effort has been expended towards investigating automated filming of dynamic scenes in unstructured environments using multiple cooperating UAVs, under battery autonomy, FoV/collision avoidance and flight zone restrictions. Notably, in [29], an on-line real-time planning algorithm is proposed that jointly optimizes feasible trajectories and control inputs for multiple UAVs filming a cluttered dynamic scene with FoV/collision avoidance, by processing user-specified aesthetic objectives and high-level cinematography plans. This method extends a previous, single-UAV method [30] that only optimizes local trajectory segments. Since both algorithms are designed for controlled indoor settings, UAV energy consumption is not taken into account and flight zone restrictions are not considered.

III. Technologies for Autonomous UAV Filming

Currently, commercial UAVs employed in media production are mostly manually controlled, with only a few rudimentary functionalities being performed autonomously. In state-of-the-art drones[2], such functionalities are obstacle avoidance, landing, physical target following or target orbiting (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. Both these basic functions and any future algorithms for more advanced, automated UAV flight and filming, require a number of underlying enabling technologies to be in place.

Below, these technologies are introduced and presented according to the functionality they provide, i.e., they are partitioned into the following subjects: *perception*, *planning and control*, *swarming* and *communications*. Within the first two of these subjects, each presented technology is further assigned a label of one out of two "complexity groups" we identified, based on the complexity of the required hardware. The two complexity groups are the *vision-* and the *3D-group*. The former one consists of visual Simultaneous Localization and Mapping (SLAM), 2D visual target detection, 2D visual target tracking and image-based visual servoing algorithms. In principle, it is feasible for these tasks to be performed in real-time by computer vision and machine learning algorithms, using only the monocular camera also employed for filming. The latter group consists in a set of methods and devices that allow functioning in global 3D Cartesian space. This is mainly achieved by the presence of Global Positioning System (GPS)

[2]E.g., the popular DJI Phantom 4, or Skydio R1

receivers [31] on-board the UAV, as well as (ideally) on the targets being filmed.

In general, the methods contained in the vision-group suffice for autonomously achieving physical target following and rudimentary cinematic coverage by the drone, as well as effective obstacle/collision avoidance and landing. However, technologies clustered under the 3D-group are essential to achieve more advanced, fully autonomous, non-trivial UAV cinematography, therefore it is imperative for them to start appearing in non-prototype drones. Hardware and software advancements are expected to allow this in the following years.

Finally, connections are made to different UAV application domains, such as surveillance, inspection, or rescue and recovery operations.

A. Perception

Visual SLAM [32] [33] [34] [35], a vision-group technology in its basic form, 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 [36], augmented by visual place recognition [37], graph-based map modelling [38] and loop closure [39] 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.

Despite the fact that visual SLAM can, in principle, be performed using a single camera, additional sensors may greatly enhance its effectiveness. Such sensors include a secondary stereoscopic camera and an Inertial Measurement Unit (IMU) [40] for more robust operation. Actually, these 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 can be used instead of stereoscopic 3D cameras in order to achieve increased accuracy and performance, as well as robustness to variable environmental lighting conditions [41]. 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 and significantly higher cost than a good camera, it is very likely that future technology improvements will increase their appeal.

LiDAR-based SLAM is also a high-end option for integrating obstacle and collision detection with generic environment mapping and self-localization. However, traditionally, separate, simple altimeter and ultrasound sensors can be employed to this end; such inexpensive sensors are found in virtually all professional drones.

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 fully automated UAVs are expected to rely solely on automatic visual target detection. Relevant state-of-the-art algorithms [42] [43] [44] [45] [46], based on deep neural networks [47], are impressively accurate and optimized for parallel execution on General-Purpose Graphical Processing Units (GP-GPUs) [48]. Such high-performance hardware has recently been commercialized in small, power-efficient form factor for embedded systems, ideal for on-board inclusion in UAVs[3]. 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 be 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 [49] [50]. 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 is expected to alleviate this issue. Novel research in lightweight neural networks [51] is a promising avenue to this end.

Obviously, 2D visual target detection and tracking are vision-group technologies. Assuming GPSs are available and operational, further possibilities are opened up. E.g., the 3D maps built by visual SLAM can be aligned with the common GPS coordinate frame, using a similarity transformation [52], and employed for assisting in global target, obstacle and UAV localization, leading to more robust operation exploiting multiple information sources.

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 filming process so as to autonomously capture specific, non-trivial shot types. Luckily, operating in global 3D Cartesian coordinates makes it meaningful to integrate a 3D visual target pose estimation algorithm into the vision-group pipeline. There are two main approaches to achieve this: a) 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 [53], or b) the machine learning approach, where the target's pose is directly regressed by a trained model that only uses the visual input [54] [55]. The first approach requires a 3D model of the target to be

[3]E.g., the NVIDIA Jetson TX2

known, while the second solution requires a regressor properly trained on a representative, fully annotated image dataset. The machine learning approach is not only more robust, but also, if a deep neural regressor is employed, such an algorithm may be incorporated into the 2D visual target detector and run entirely on 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 vision-group, using a deep neural network running on GP-GPU in real-time [56] [57]. Subsequently, the detected crowd ROI (in pixel coordinates) may be corresponded to the relevant terrain areas of the 3D map by perspective back-projection [58], so as to achieve a semantic annotation of the map [59]. 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. As in the case of 3D visual target pose estimation, such capabilities are entirely missing from current state-of-the-art UAVs.

Typically, the GPS signal is not available indoors and it may even be temporarily lost outdoors. Additionally, its usual position error is in the range of approximately up to 5 meters [31]. These problems can be bypassed by employing differential GPS units (accurate in the range of approximately 20 cm [60]), by IMU/GPS/visual SLAM fused localization and by replacing GPS with an Active Radio-Frequency IDentification (RFID) positioning system [61] in GPS-denied environments. These solutions, however, come with associated monetary and computational costs, which explains the fact that state-of-the-art commercial UAVs completely lack capabilities depending on the 3D-group, despite being universally equipped with simple GPS receivers.

B. Planning and control

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 [62] [63] 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) a high-level, long-term plan must be devised periodically, or when important events are detected, while b) a low-level plan can locally adjust that path during the intermediate intervals 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 minimized 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. Then, low-level replanning is reduced to simple reactive obstacle avoidance.

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) [64] or Linear-Quadratic Regulator (LQR) [65] controllers are employed for related tasks. The PixHawk/PX4 Autopilot [66], 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.

Besides traditional 3D path planning and motion control algorithms, a reinforcement learning approach can be alternatively employed for proper UAV trajectory planning and following, so as to capture desired target shots. Learning in a 3D context allows complex UAV/camera motion types to be implemented, such as a subset of the motion types described in Section V, but it has not yet been investigated for cinematography applications.

Beyond the 3D-group technologies presented above for planning and control, purely vision-group approaches can be employed that rely only on video input. Image-based visual servoing may 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 [67]) 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, similarly to how state-of-the-art commercial drones implement physical target following and orbiting. A number of UAV/camera motion types that can be autonomously implemented in such a manner are identified and described in Section V, while methodological examples of autonomous filming are briefly discussed in Section VII.

An alternative to image-based UAV motion control methods is reinforcement learning employing raw video input and outputting direct motor commands. 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. Deep neural networks have recently been employed in similar settings for UAV collision avoidance [68], indoor flight control in search and recovery operations [69] or high-level flight navigation [70]. An imitation learning variant has also been explored for drone racing [71], 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 are currently under research (nowhere near commercial deployment) and have not yet been investigated

for cinematography applications.

C. Swarming

No commercial civilian multiple-UAV platform exists, although semi-autonomous swarms of military fixed-wing UAVs, supported by automated intra-swarm coordination and task assignment, are regularly employed in the field (typically, under the leader-follower paradigm). Extending this model to civilian media production scenarios using VTOL drones is necessary, if autonomous multiple-UAV cinematography is to become a reality.

In general, typical swarm formation control methods [72] [73], aiming to enforce a synchronized group motion on all swarm members, are not readily applicable to cinematography applications for the entirety of a filming mission, since different swarm members may have to follow very different trajectories at a given time, in order to capture the desired shots and cover large areas. However, temporary swarm formations composed of a few UAVs may be employed for efficiently capturing individual multiple-UAV shots, such as the ones defined in Section VI. In such a scenario, it is not difficult to conceive the moving target being filmed acting as a reference point for intra-swarm position coordination. Although several paradigms for UAV swarm formation control exist, including leader-follower, behavioural and virtual structure approaches [74], there is no relevant research dedicated to cinematography applications.

Moreover, optimal autonomous task assignment based on directorial guidelines (namely, which UAV will cover each target, at what time interval and with what shot type) is a relatively unexplored area [8]. Any requirement for assignments to be able to dynamically change on-the-fly, based on detected semantic events (e.g., significant change in rank position of lead contestants during a sports race), further complicates the issue and highlights the importance of autonomous semantic event detection. In the simplest scenario, the latter may consist in a set of rules regarding the relative positions of targets in the 3D map, but more elaborate methods can be exploited (e.g., on-line activity recognition from the visual input [75]).

Multiple-UAV coordination in a swarm context also allows cooperative variants of algorithms from the vision and the 3D-group to be employed, e.g., cooperative visual SLAM [76] or cooperative path planning [77]. However, as in the case of formation control and task assignment, cinematography/media production applications impose constraints that have not been researched up to now.

In all the above methods, a choice has to be made regarding whether they will be implemented in a centralized or a distributed manner. In the first case, where a swarm member (or a ground control station) serves as "master", algorithm design is more efficient, the result is more optimal and the computational load is reduced at the "slave" members. However, a single point of failure and a possible communication/computational bottleneck, i.e., the master, are introduced to the swarm. The choice resulting in the optimal balance depends on the specific application and on the available hardware/infrastructure. This is one more aspect from which multiple-UAV cinematography/media production has not yet been systematically examined.

D. Communications

Communication issues in autonomous multiple-UAV cinematography can be seen as having both "data streaming" and "networking" aspects. In general, infrastructure for communications and related issues is critical for successful deployment of UAV swarms in practical scenarios [78], 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 filming 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 WiFi 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. It supports ad hoc routing and accounts 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 WiFi extensions must be developed, or falling back to LTE infrastructure is unavoidable, at the cost of increased latency.

E. Autonomous UAV Video Capture in Other Application Domains

Most of the technical issues and solutions described above also apply to any UAV application domain involving video capturing, besides cinematography applications. These include area surveillance and/or moving target monitoring [79], rescue and recovery operations, infrastructure inspection (e.g., of wind turbines [80], or agricultural production [81]), scientific exploration and 3D scene reconstruction tasks. In all of these scenarios, both UAV and target 3D localization, 2D target detection/recognition and tracking, 3D mapping, path planning and potential emergency landing site detection are directly relevant. Less significant issues in a non-cinematography setting include 3D target pose estimation, target central composition, human crowd detection and live data streaming.

For instance, with the possible exceptions of 3D scene reconstruction and surveillance applications, no human crowds are typically present near the UAV, while no framing shot type constraints are in place. Thus, autonomous cinematographic outdoor live event coverage arguably proves to be the most technically challenging single-UAV application overall, despite the existence of very narrowly defined, domain-specific problems (such as weed classification for agricultural inspection applications [82], or changes detection for area surveillance [79]).

On the other hand, swarming issues and methods are possibly more significant in non-cinematography applications. For instance, tight temporal deadlines and/or a very extended area of coverage, which are commonly found in all of the above scenarios, increase the relevance of UAV swarm deployment and coordination, as well as, in turn, of cooperative algorithms for SLAM, path planning, etc. In contrast, although multiple-UAV coverage is beneficial when filming events, as explained in Section I, the relative importance of swarm approaches is lower in cinematography applications, where employing multiple UAVs enriches the creative potential, but is not an absolute requirement for properly performing the desired task in time.

A special note must be made for the intelligent UAV filming systems and methods presented in Section II. Although the consideration of aesthetic criteria commonly found in such systems is only relevant to cinematography applications, their ability to pre-compute feasible drone trajectories for capturing desired footage is significant for all the different UAV application scenarios discussed above. However, typically, current intelligent UAV filming systems only operate in known environments under controlled settings. This is almost never the case in surveillance, rescue and recovery, or scientific exploration applications, rendering them almost useless for such tasks.

Nevertheless, future advances have the potential to change this situation. For instance, a search and rescue operation could benefit significantly from a similar system able to function in a partially unknown environment. The mission could be implicitly planned by simply specifying properties of the desired footage. Subsequently, actually capturing such footage in an autonomous manner would imply that the target has been detected and is being inspected from the requested view angles (e.g., to check for visible signs of damage or injuries).

IV. UAV Cinematography Framing Shot Types

The various shot types in UAV cinematography can be described using two complementary criteria: the framing shot type (FST) and the UAV/camera motion trajectory (CMT). Each CMT can be successfully combined with a subset of the possible FSTs, according to director's specifications, so as to achieve a pleasant visual result. FSTs are primarily defined by the relative size of the main subject/target being filmed (if any) to the video frame size. They are mostly derived/adapted from the ones found in traditional ground and aerial cinematography [21] [1] [2].

Based on visual inspection of sample UAV video coverage of outdoor events, we have defined eight FSTs in UAV cinematography: Extreme Long Shot, Long Shot, Medium Shot, Medium Close-Up, Close-Up, Two-Shot/Three-Shot and Over-the-Shoulder. Traditional cinematography also includes Extreme Close-Ups, a very "zoomed in" FST that typically depicts a human head from the lips to the forehead, in order to emphasize subject emotions. In practice, based on extensive visual inspection of professional and semi-professional UAV footage from multiple sources, we have not determined a corresponding FST to be regularly employed in UAV cinematography.

Table I summarizes the six basic FSTs using thresholds on the target ROI width/height to video frame width/height ratio ("coverage"). The other two FSTs (Two-Shot/Three-Shot, Over-the-Shoulder) can be seen as specific combinations of more basic ones. Verbal descriptions for all eight FSTs are provided below:

1) Extreme Long Shot (ELS): The target appears so distant from the camera that it may not even be visible (at least not in detail). The emphasis is on showing an expansive view of its surroundings. This FST typically provides scene context to the viewer and establishes the theater of action. In general, less than approximately 5% of the video frame width and/or video frame height is covered by the target. An example is depicted in Figure 1a.
2) Very Long Shot (VLS): The target is barely visible and perceived to be at a large distance from the camera, appearing small. The purpose of this FST is to firmly localize the target in his surroundings in the mind of the viewer. In general, approximately $5 - 20\%$ of the video frame width and/or video frame height is covered by the target. An example is depicted in Figure 1b.
3) Long Shot (LS): The entire target is visible and perceived to be at an intermediate distance from the camera. This FST depicts both the target and its surroundings clearly. In general, the target covers at most $20 - 40\%$ of

the video frame width and/or video frame height. An example is depicted in Figure 1c.

4) Medium Shot (MS): The target is perceived to be at a fairly short distance from the camera, appearing quite large. This FST places emphasis mainly on the target being filmed. In case the target is a person, a MS would depict him from the waist up. In general, the target covers at most $40-60\%$ of the video frame width and/or video frame height. An example is depicted in Figure 1d.
5) Medium Close-Up (MCU): Either the entire target, or a large portion of it is visible and perceived to be at a very short distance from the camera. This FST showcases the most interesting portion of the target. In case the target is a person, a MCU would depict him from the chest up. In general, the target covers at most $60-75\%$ of the video frame width and/or video frame height. An example is depicted in Figure 1e.
6) Close-Up (CU): The most interesting part of the target appears to be at a very short distance from the camera, spanning almost the entire foreground. This FST emphasizes a specific detail of the target's image. In case the target is a person, a CU would depict him from the neck up. In general, the target covers more than 75% of the video frame width and/or video frame height. An example is depicted in Figure 1f.
7) Two-Shot/Three-Shot (2S/3S): Two/three subjects appear simultaneously in the video frame, arranged so that they are equally visible. With regard to the perceived camera distance from the target, 2S/3S is typically a Long Shot or a Medium Shot. Examples are depicted in Figures 1g,h.
8) Over-the-Shoulder (OTS): The main target is clearly visible and perceived to be at a fairly short, or short, distance from the camera, while a secondary target is visible at the left or right edge of the video frame and appears to be at a significantly shorter distance from the camera. OTS can be regarded as a variant of the Two-Shot, with the main target being filmed as in a Very Long Shot, Long Shot or Medium Shot, and with the secondary target being filmed as in a Medium Shot, Medium Close-Up or Close-Up, respectively. Either the main or the secondary target can be a geographical landmark (e.g., a historical monument). This is an adaptation from the traditional OTS shot in movie/TV dialogue scenes, where the two targets are persons talking to each other and the secondary target is shown from behind. An example is depicted in Figure 1i.

Typically, the on-board camera is suspended from a *gimbal* that allows rapid, arbitrary camera rotation around its yaw, pitch and roll axis, within orientation limits prescribed by mechanical gimbal stops. Given a fixed, constantly visible target, the FST in UAV cinematography can be adjusted, in principle, by suitably modifying any combination of the following: the zoom level (controlled by the camera focal length f), the camera gimbal rotation and the UAV/camera world position. However, in most situations, simply altering f should be sufficient for achieving the desired FST. E.g., transitioning between any of the single-subject types (ELS, VLS, LS, MS, MCU, CU) is not only trivially performed by adjusting the zoom level, but such an approach is the least energy-consuming. When the multiple-subject FSTs are also considered (2S, 3S, OTS), transitions among FSTs may also require adjusting the camera gimbal rotation (e.g., so as to orient towards a midpoint between many subjects, instead of towards a single subject) and the UAV/camera position in world coordinates, in order for the proper filming angle to be achieved (especially in OTS shots). In the latter case, adherence to flight zone and legal restrictions must be ensured during the transition.

TABLE I: Basic FSTs and their corresponding target ROI to video frame width/height ratio percentage.

FST	Coverage
Extreme Long Shot (ELS)	$< 5\%$
Very Long Shot (VLS)	$5-20\%$
Long Shot (LS)	$20-40\%$
Medium Shot (MS)	$40-60\%$
Medium Close-Up (MCU)	$60-75\%$
Close-Up (CU)	$> 75\%$

Besides the above issues, single-subject FSTs are also affected by a directorial choice regarding the visual arrangement of the target within the video frame. The most common option is "central composition", where the image of the target being filmed is located at the center of the video frame, with the "Rule of Thirds" providing an alternative. According to the latter, a video frame is conceived as divided into a 9×9 rectangular grid, with the center of the target's image located at an intersection point of a vertical and a horizontal grid line [83]. In the more visually crowded multiple-subject FSTs, the "Rule of Thirds" should be the preferred composition, according to conventional cinematography guidelines.

Filming one or more moving targets imposes a demand for the above adjustments to be performed on top of, and in combination with, any gimbal or UAV parameter adaptations required by the target and/or UAV motion. Thus, the desired UAV/camera motion trajectory comes into play.

V. Single-UAV Camera Motion Shot Types

Several standard types of UAV/camera motion trajectories/types (CMTs) have emerged since the popularization of UAVs. As in the case of FSTs, most of them are derived/adapted from the ones found in traditional ground and aerial cinematography. A significant subset of these motion types ("target-oriented") are relative to a (still or moving) target being filmed, while the rest do not depend on a specific subject and emphasize capturing the scene ("scene-oriented"). Moreover, a small subset of the presented types do not involve actual UAV motion and have been included for reference purposes. In the sequel, they are referred to as "static shots", in contrast to "dynamic shots".

Below, a taxonomy of 26 CMTs is provided, complemented by mathematical modeling (details per CMT are in the Appendix). Verbal descriptions for a subset of them can be

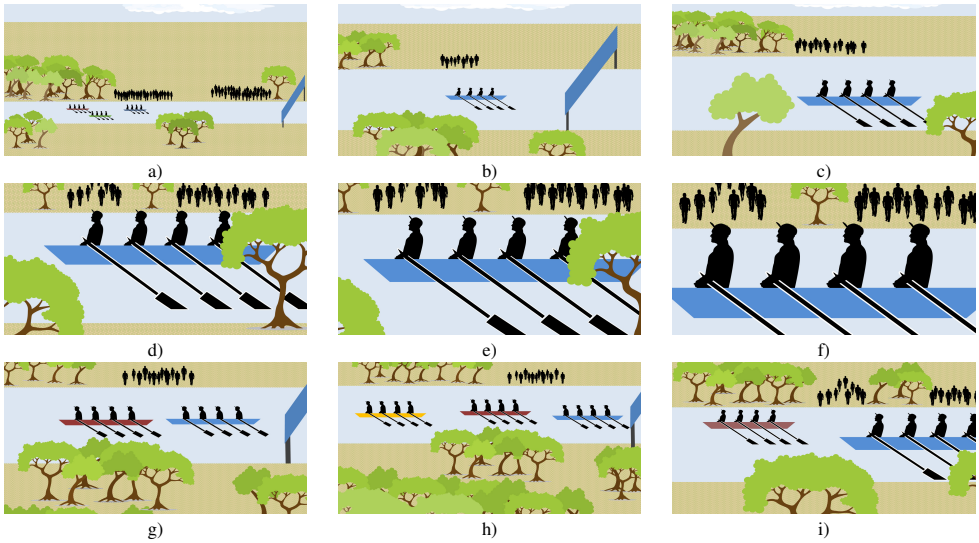

Fig. 1: Examples of different UAV shot types when filming boat targets from their side (Lateral Tracking Shot): a) Extreme Long Shot (ELS), b) Very Long Shot (VLS), c) Long Shot (LS), d) Medium Shot (MS), e) Medium Close-Up (MCU), f) Close-Up (CU), g) Two-Shot (2S), h) Three-Shot (3S) and i) Over-the-Shoulder (OTS). Notice that, in CU, the visual emphasis is on the boat crew.

found in recent photography/cinematography literature [1] [2]. The main focus is on coverage of outdoor events in dynamic, unstructured environments with moving targets/subjects (e.g., in live sports broadcasting), which represents the most challenging application. In different production scenarios, either only subsets of the shot types presented here are employed, or the scenes to be filmed consist in significantly more controlled settings (e.g., movie sets).

In terms of energy consumption, the following UAV operation ordering may be defined, from the least to most battery-intensive: camera operations (gimbal rotations, zoom), flying down, flying horizontally/hovering, flying up. In general, the direction of UAV flight dominates the energy-related behavior, with camera operations being relatively negligible. Thus, wherever possible, each presented CMT has been classified into one out of three possible energy consumption levels: "low", "medium" and "high".

In terms of framing, an assessment is provided regarding the FSTs compatible with each CMT. In general, the scene-oriented CMTs fit well only with ELS and VLS, since no real subject is being filmed, while the framing types that fit with target-oriented CMTs vary depending on the latters' purpose. For instance, in a Chase shot (where the UAV follows/leads a moving target from behind/from the front, maintaining a steady distance), the viewer is meant to experience a "simulation" of the target's motion within its environment, while the target is fully visible. Thus, a CU that excludes most of the surroundings from the video frame is an unsuitable FST. Moreover, in the extreme case of a Bird's Eye shot (a scene-oriented CMT, where the UAV flies vertically up, while the camera stays stable and focused vertically down), no specific FST fits well, since the distance from the ground starts small and increases steadily.

In terms of the required technology complexity level, each of the proposed CMTs is assigned to the minimum hardware/software combination necessary for autonomously capturing it in an unstructured, dynamic environment, based on the discussion in Section III. Three levels are possible: "minimum", "vision" and "vision+3D". The first one implies that no visual input analysis, GPS-based target localization, or any of the methods contained in the vision or the 3D-group is required for filming with this CMT. For instance, scene-oriented shots with simplistic UAV trajectories fall under this category. The second level refers to CMTs that require the vision-group methods to be available. The third level implies that methods contained in both the vision and the 3D-group are required, i.e., both on-target and on-drone GPS receivers (or equivalent positioning systems) must be active. Vision methods may be necessary even in scene-oriented dynamic shots, e.g., for assisting in mapping/localization and/or human crowd detection, therefore no pure "3D" (with no vision) designation has been given.

The presented mathematical treatment assumes that, given a camera frame-rate of T, time t is discrete, non-negative and proceeds in steps of $\frac{1}{T}$ seconds. The ijk-axes convention is employed for the 3D coordinate system. A separate timeline is employed for each shot type description, i.e., $t = 0$ indicates the start of a shot type filming session. At each time instance t the 3D positions of the UAV ($\tilde{\mathbf{x}}_t$) and the target ($\tilde{\mathbf{p}}_t$), as well as an estimated 3D target velocity vector ($\tilde{\mathbf{u}}_t$), are assumed known in a fixed, orthonormal, right-handed World Coordinate System (WCS), with its k-axis vertical to a local tangent plane

(hereafter shortened to "ground plane").[4] For instance, a local East-North-Up (ENU) coordinate system may be employed [84]. Additionally, at each time instance t, a current, orthonormal, right-handed target-centered coordinate system (TCS) is defined. Its origin lies on the current target position, its k-axis is vertical to the ground plane and its i-axis is the $\mathcal{L}_2$-normalized projection of the current target velocity vector onto the ground plane. In case of a still target, the TCS i-axis is defined as parallel to the projection of the vector $\tilde{\mathbf{p}}_0 - \tilde{\mathbf{x}}_0$ onto the ground plane. In both coordinate systems, the ij-plane is parallel to the ground plane and the k-component is called "altitude". Below, vectors expressed in TCS are denoted without the tilde symbol (e.g., $\mathbf{x}_t$, $\mathbf{p}_t$ and $\mathbf{u}_t$). In mobile robotics literature, an additional, vehicle-centered coordinate system is typically employed, having its origin located at a fixed distance from the UAV-mounted camera. Since the scope of this article does not include UAV control per se, we do not make use of such a coordinate frame and limit our analysis to cinematography issues.

Since the origin and the axes of TCS in terms of the WCS are fully defined, transforming between the two coordinate systems is trivial, allowing us to provide descriptions either in TCS or in WCS, depending on which approach is more succinct for each case. A subset of the presented motion types require pre-specification of motion parameters meant to adapt the UAV motion trajectory to concrete directorial guidelines (e.g., distance to be covered by the UAV).

We assume standard measurement units for the above quantities, i.e., distance is measured in meters, speed in meters per second and the frame-rate in frames per second. Moreover, in the mathematical description of scene-oriented motion types we assume a known "virtual target" (a 3D world point located in the visible scene) as a reference point for the TCS. Therefore, $\tilde{\mathbf{p}}_t$ and $\mathbf{p}_t$ are considered meaningful both in target and in scene-oriented motion types.

The 3D scene point at which the camera looks at time instance t is denoted by $\mathbf{l}_t$ (in TCS). It depends on the specific FST combined with each CMT. That is, for single-subject FSTs with central composition it holds that $\mathbf{l}_t = \mathbf{p}_t$, while in the case of a "Rule of Thirds" composition $\mathbf{l}_t$ has to be suitably adjusted. When a multiple-subject FST has been selected, $\mathbf{l}_t$ must be computed based both on $\mathbf{p}_t$ and the 3D positions of neighboring, secondary targets.

To simplify the following analysis, a single-subject FST with central composition is assumed (therefore, it usually holds that $\mathbf{l}_t = \mathbf{p}_t$ and $\mathbf{o}_t = -\mathbf{x}_t$). In several cases, the shot type is only meaningful if the target is moving linearly. Moreover, such an assumption is additionally made below in cases where the future position of the target or the UAV needs to be predicted, for reasons of modeling convenience (these cases are appropriately marked in the following analysis). Constant linear motion is assumed for both these scenarios, although extending the formulas for the case of constantly accelerated linear motion is trivial (assuming a target acceleration vector can be reliably estimated).

The 26 CMTs have been clustered into four groups containing similar motions. Geometrical description for each CMT is provided in the Appendix, while in Table III each one is assigned a list of compatible FSTs, an energy consumption grading and a required technology complexity level. In case the battery consumption varies depending on a parameter, the latter has been identified.

A. Static shots

Static shots are CMTs (either target-oriented, or scene-oriented) where there is no UAV motion. There are five CMTs falling under this category, with graphical examples provided in Figure 2:

1) *Static Shot (SS)* (scene-oriented), 2) *Static Shot of Still Target (SSST)* (target-oriented) and 3) *Static Shot of Moving Target (SSMT)* [2] (target-oriented) are the simplest, non-parametric CMTs. In all cases, the UAV simply hovers. In SS and SSST, the camera gimbal stays stable during filming, while in SSMT it rotates slowly so as to always keep the moving target properly framed. In SS, $\tilde{\mathbf{p}}_t$ refers to a fixed, virtual target position in WCS, selected by the director, while in SSST and SSMT it refers to the position of a real target in WCS.

4) *Static Aerial Pan (SAP)* is a scene-oriented, parametric CMT, where the camera gimbal rotates slowly (with respect to the yaw axis), in order to capture the scene context, while the UAV simply hovers [2]. $\tilde{\mathbf{p}}_t$ refers to a fixed, virtual target position selected by the director.

5) *Static Aerial Tilt (SAT)* is a scene-oriented, parametric CMT, where the camera gimbal rotates slowly (with respect to the pitch axis), in order to capture the scene context, while the UAV simply hovers [1] [2]. $\tilde{\mathbf{p}}_t$ refers to a fixed, virtual target position selected by the director.

B. Dynamic scene shots

Dynamic scene shots are CMTs where the UAV is moving, while the emphasis is on conveying the scene context to the viewer and/or achieving a visually pleasing cinematographic effect. Thus, the target is virtual. There are seven CMTs falling under this category, with graphical examples provided in Figure 3:

1) *Moving Aerial Pan (MAP)* and 2) *Moving Aerial Tilt (MAT)* [2] are two parametric CMTs, where the camera gimbal rotates slowly (with respect to the yaw/pitch axis, in MAP/MAT, respectively), in order to capture the scene context, while the UAV is slowly flying at a steady trajectory with constant velocity. $\tilde{\mathbf{p}}_t$ refers to a varying, virtual target position moving identically to the UAV, with initial $\tilde{\mathbf{p}}_0$ selected by the director. Therefore, during filming, the UAV position remains constant in TCS, but varies in WCS.

3) *Pedestal/Elevator Shot (PS)* [1] [2] and 4) *Bird's Eye Shot (BIRD)* [2] are parametric CMTs, where the UAV is slowly flying up or down, along the k-axis, with a constant velocity, while the camera gimbal remains stable. In PS the camera axis is parallel to the ground plane and $\tilde{\mathbf{p}}_t$ refers to a varying, virtual

[4]Following widespread convention, we employ the term "local tangent plane" for a plane parallel to the local sea level, while the term "terrain tangent plane" is reserved for the plane instantaneously tangent to the local terrain inclination.

TABLE II: UAV/Camera Motion Description Nomenclature.

$\tilde{\mathbf{p}}_t = [\tilde{p}_{t1}, \tilde{p}_{t2}, \tilde{p}_{t3}]^T$	The 3D target position in WCS, at time instance t
$\mathbf{p}_t = [p_{t1}, p_{t2}, p_{t3}]^T$	The 3D target position in TCS, i.e., the current TCS origin at time instance t. It is always equal to $[0, 0, 0]^T$.
$\tilde{\mathbf{x}}_t = [\tilde{x}_{t1}, \tilde{x}_{t2}, \tilde{x}_{t3}]^T$	The 3D UAV position in WCS, at time instance t
$\mathbf{x}_t = [x_{t1}, x_{t2}, x_{t3}]^T$	The 3D UAV position in TCS, at time instance t
$\tilde{\mathbf{u}}_t = [\tilde{u}_{t1}, \tilde{u}_{t2}, \tilde{u}_{t3}]^T$	The estimated 3D target velocity in WCS, at time instance t
$\mathbf{u}_t = [u_{t1}, u_{t2}, u_{t3}]^T$	The estimated 3D target velocity in TCS, at time instance t
$\tilde{\mathbf{v}}_t = [\tilde{v}_{t1}, \tilde{v}_{t2}, \tilde{v}_{t3}]^T$	The 3D UAV velocity in WCS, at time instance t
$\mathbf{v}_t = [v_{t1}, v_{t2}, v_{t3}]^T$	The 3D UAV velocity in TCS, at time instance t
$\tilde{\mathbf{l}}_t \in \mathbb{R}^3$	The 3D position at which the camera looks (known as the "LookAt point") in WCS, at time instance t
$\mathbf{l}_t \in \mathbb{R}^3$	The 3D position at which the camera looks in TCS, at time instance t
$\tilde{\mathbf{o}}_t = \tilde{\mathbf{l}}_t - \tilde{\mathbf{x}}_t$	The LookAt vector in WCS, at time instance t. It is a scalar multiple of the camera axis.
$\mathbf{o}_t = \mathbf{l}_t - \mathbf{x}_t$	The LookAt vector in TCS, at time instance t
$\tilde{\mathbf{i}}, \tilde{\mathbf{j}}, \tilde{\mathbf{k}}$	The WCS axes unit vectors
$\mathbf{i}, \mathbf{j}, \mathbf{k}$	The TCS axes unit vectors
T	The camera frame-rate

target position moving identically to the UAV (with initial $\tilde{\mathbf{p}}_0$ selected by the director), while in BIRD the camera axis is facing vertically down and $\tilde{\mathbf{p}}_t$ refers to a static virtual target position directly beneath the UAV. Therefore, in PS, the UAV position remains constant during filming in TCS, but varies in WCS, while in BIRD the UAV altitude constantly increases or decreases (in both coordinate systems).

5) *Moving Bird's Eye Shot (MOVBIRD)* and 6) *Survey Shot (SURVEY)* are two parametric CMTs where the camera gimbal remains stable, while the UAV is slowly flying in parallel to the terrain tangent plane with constant velocity. The camera is facing vertically down in the case of MOVBIRD, while in SURVEY it is facing ahead. $\tilde{\mathbf{p}}_t$ refers to a varying, virtual target position moving identically to the UAV, therefore the UAV position remains constant in TCS. In the case of MOVBIRD, the initial target position $\tilde{\mathbf{p}}_0$ depends on the initial UAV position $\tilde{\mathbf{x}}_0$. In SURVEY $\tilde{\mathbf{p}}_0$ is selected by the director, while the camera axis is approximately parallel to the terrain tangent plane.

7) *Fly-Through (FLYTHROUGH)* is a parametric CMT, where the camera gimbal remains stable, with the camera typically facing ahead, and the UAV is flying forward and through an opening/gap/hole with constant velocity [2]. $\tilde{\mathbf{p}}_t$ refers to a varying, virtual target position moving identically to the UAV. This is an aerial CMT only achievable with small-form UAVs, thus especially important for UAV cinematography.

C. Target tracking shots

Target tracking shots are CMTs where the UAV motion directly depends on the trajectory of a real target. There are eleven CMTs falling under this category, with graphical examples provided in Figure 4:

1) *Moving Aerial Pan with Moving Target (MAPMT)* and 2) *Moving Aerial Tilt with Moving Target (MATMT)* are parametric CMTs, where the camera gimbal rotates slowly (mainly with respect to the yaw/pitch axis, for MAPMT/MATMT respectively) so as to always keep the linearly moving target properly framed, while the UAV is slowly flying at a steady trajectory with constant velocity. $\tilde{\mathbf{p}}_t$ refers to the position of a real target, varying over time in such a manner that the target and the UAV trajectory projections onto the ground plane are approximately perpendicular/parallel to each other, for MAPMT/MATMT respectively.

3) *Lateral Tracking Shot (LTS)* [1] [2] and 4) *Vertical Tracking Shot (VTS)* are non-parametric CMTs, where the camera gimbal remains stable and the camera always focused on the moving target. In LTS, the camera axis is approximately perpendicular both to the target trajectory and to the WCS vertical axis vector $\tilde{\mathbf{k}}$, while the UAV flies sideways/in parallel to the target, matching its speed if possible. In VTS, the camera axis is perpendicular to the target trajectory and the UAV flies exactly above the target, matching its speed if possible. In both cases, $\tilde{\mathbf{p}}_t$ refers to a real, varying target position in WCS. During filming, the UAV position remains constant in TCS, but varies in WCS.

5) *Orbit (ORBIT)* A parametric CMT, where the camera gimbal is slowly rotating, so as to always keep the still or linearly moving target properly framed, while the UAV (semi-)circles around the target and, simultaneously, follows the latter's linear trajectory (if any) [1] [2]. During filming, the

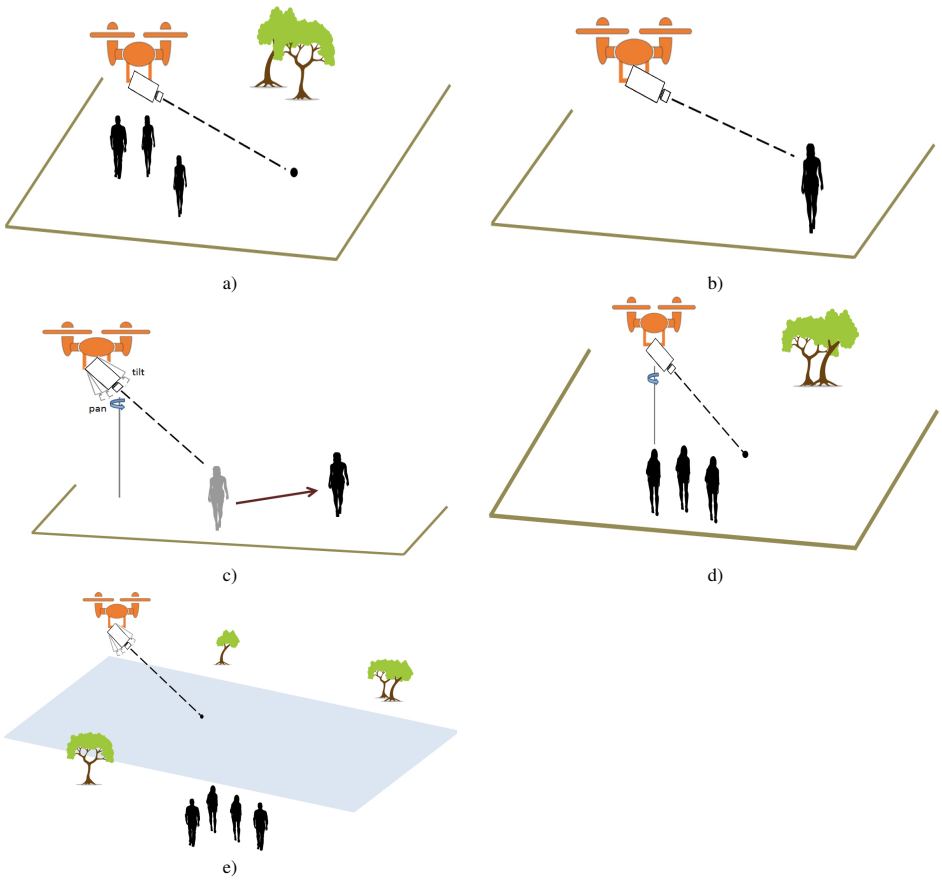

Fig. 2: Examples of different static UAV CMTs: a) Static Shot (SS), b) Static Shot of Still Target (SSST), c) Static Shot of Moving Target (SSMT), d) Static Aerial Pan (SAP) and e) Static Aerial Tilt (SAT).

UAV altitude remains constant in TCS, but may vary in WCS. $\tilde{\mathbf{p}}_t$ refers to a real, varying or static target position in WCS.

6) *Fly-Over (FLYOVER)* and 7) *Fly-By (FLYBY)* [2] are parametric CMTs, where the camera gimbal is slowly rotating (mainly along the pitch axis, in the case of FLYOVER), so as to always keep the still or linearly moving target properly framed. The UAV intercepts the target from behind/from the front (and to the left/right, in the case of FLYBY), at a steady altitude (in TCS) and with constant velocity, flies exactly above it/passes it by (for FLYOVER/FLYBY, respectively) and keeps on flying at a linear trajectory, with the camera still focusing on the receding target. The UAV and target trajectory projections onto the ground plane remain approximately parallel during filming. $\tilde{\mathbf{p}}_t$ refers to a real, varying or static target position in WCS.

8) *Descent (DESCENT)*, 9) *Descent Over (DESCENTOVER)* and 10) *Ascent (ASCENT)* are parametric CMTs, where the camera gimbal is slowly rotating (mainly along the pitch axis), so as to always keep the still or linearly moving target properly framed. The UAV linearly intercepts/backs away from the target (for DESCENT,DESCENTOVER/ASCENT, respectively) from behind or from the front, at a steadily decreasing/increasing TCS altitude (for DESCENT,DESCENTOVER/ASCENT, respectively), with constant velocity. In DESCENT, the shot ends with the UAV flying exactly above the target, in DESCENTOVER the UAV passes the target by, while in ASCENT the UAV keeps flying away from the target, with the camera still focusing on the latter. The UAV and target trajectory projections onto the ground plane remain approximately parallel during filming. $\tilde{\mathbf{p}}_t$ refers to a real, varying or static target position in WCS.

11) *Chase/Follow Shot (CHASE)* is a non-parametric CMT, where the camera gimbal remains stable and the camera always focused on the target [2]. The UAV follows/leads the target from behind/from the front, at a steady trajectory, steady distance and matching its speed if possible. $\tilde{\mathbf{p}}_t$ refers to a real, varying target position in WCS.

D. Dynamic target shots

Dynamic target shots are CMTs where the target is real, but UAV motion does not depend only on the target trajectory. In such scenarios, the FST can be either allowed to vary automatically according to the UAV-target distance at each time instance, or can be actively held fixed (via appropriately

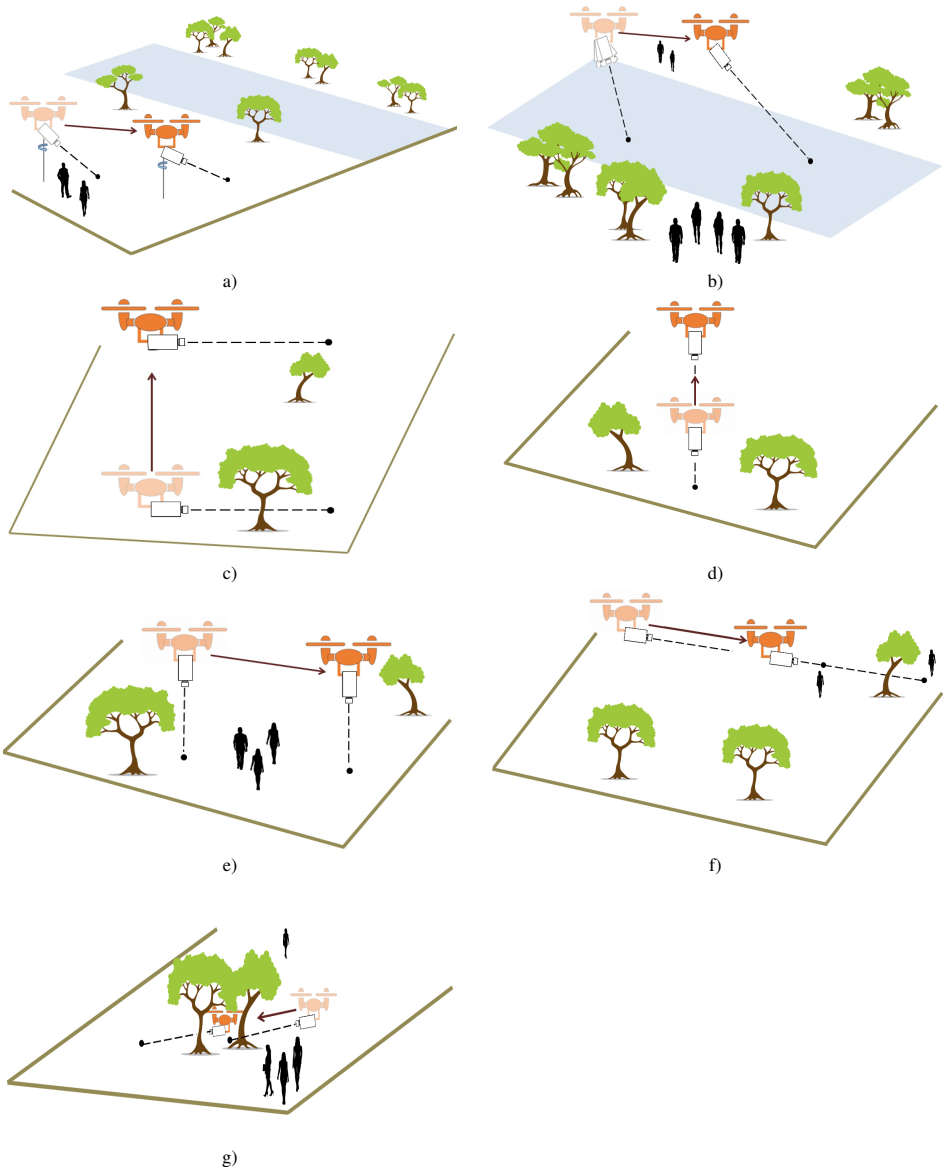

Fig. 3: Examples of different dynamic scene UAV CMTs: a) Moving Aerial Pan (MAP), b) Moving Aerial Tilt (MAT), c) Pedestal/Elevator Shot (PS), d) Bird's Eye Shot (BIRD), e) Moving Bird's Eye Shot (MOVBIRD), f) Survey Shot (SURVEY) and g) Fly-Through (FLYTHROUGH).

adapting the zoom level). There are three CMTs falling under this category, with graphical examples provided in Figure 5:

1) *Constrained Lateral Tracking Shot (CONLTS)* is a parametric CMT, where the camera gimbal remains stable and the camera always focused on the moving target. The UAV flies along the projection of the target trajectory onto a pre-defined "flight plane", vertical to the ground plane, while maintaining a constant TCS altitude during filming. This is relevant, for instance, in football match coverage, where the UAVs are allowed to fly only above the pitch sidelines. $\tilde{\mathbf{p}}_t$ refers to a real, varying target position in WCS.

2) *Pedestal/Elevator Shot With Target (PST)* is a parametric CMT, where the UAV is slowly flying up or down, along the k-axis, with constant velocity [1] [2]. The camera gimbal rotates slowly (mainly along the pitch axis), so as to always keep the linearly moving target properly framed. The projections of the camera axis and of the target trajectory on the ground plane are approximately lying on the same line during filming. $\tilde{\mathbf{p}}_t$ refers to a real, varying target position in WCS.

3) *Reveal Shot (RS)* is a parametric CMT, where the camera gimbal is stable, with the target initially out of frame (e.g., hidden behind an obstacle) [2]. The UAV flies at a steady

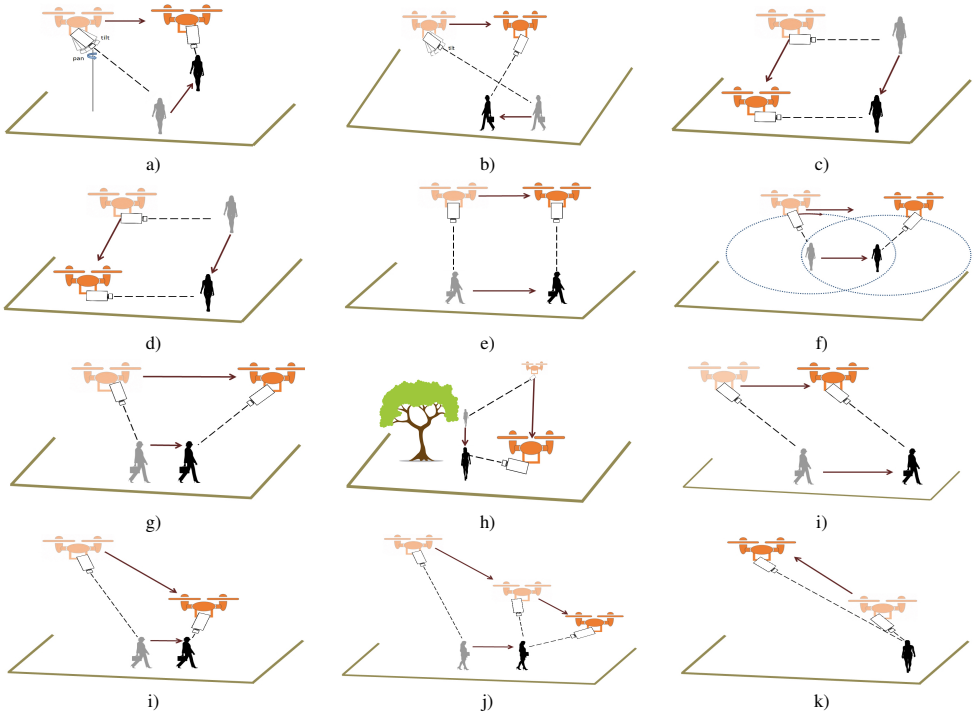

Fig. 4: Examples of different target tracking UAV CMTs: a) Moving Aerial Pan with Moving Target (MAPMT), b) Moving Aerial Tilt with Moving Target (MATMT), c) Lateral Tracking Shot (LTS), d) Vertical Tracking Shot (VTS), e) Orbit (ORBIT), f) Fly-Over (FLYOVER), g) Fly-By (FLYBY), h) Chase/Follow (CHASE), i) Descent (DESCENT), j) Descent Over (DESCENTOVER) and k) Ascent (ASCENT).

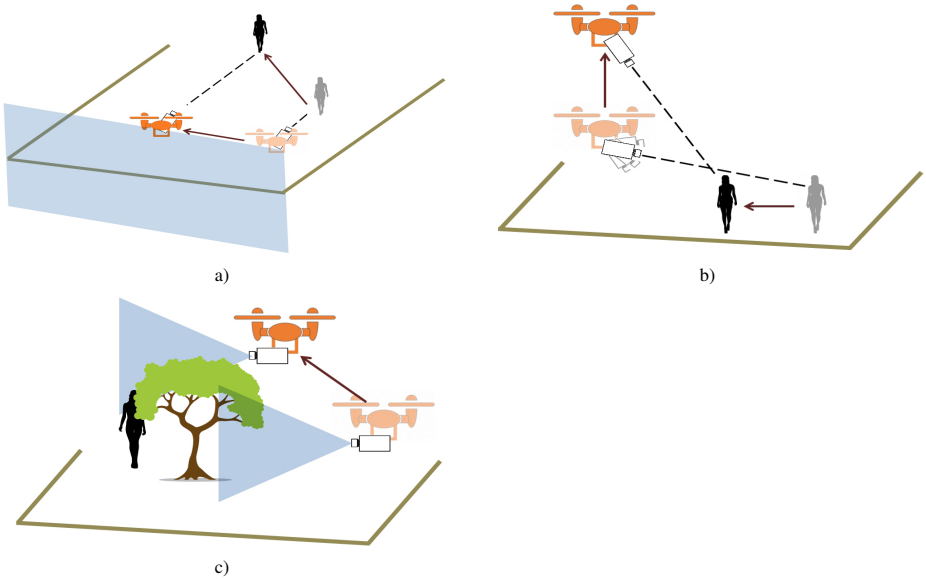

Fig. 5: Examples of different dynamic target UAV CMTs: a) Constrained Lateral Tracking Shot (CONLTS), b) Pedestal/Elevator Shot With Target (PST) and c) Reveal Shot (RS).

TABLE III: Single-UAV CMT properties: "Framing" refers to compatible FSTs, "Energy" refers to expected energy consumption, while "Technology" refers to required technology complexity level for autonomous capture. Light blue, yellow, green and red cells denote static, dynamic scene, target tracking and dynamic target shots, respectively. * denotes that energy consumption depends on UAV velocity direction, while ** denotes that energy consumption depends on target velocity direction. ↓ denotes a case where the energy consumption is low because the UAV is flying down, while ↑ denotes a case where the energy consumption is high because the UAV is flying up.

Camera Motion	Framing	Energy	Technology
SS	ELS, VLS	medium	minimum
SSST	All	medium	vision
SAP	ELS, VLS	medium	minimum
SAT	ELS, VLS	medium	minimum
SSMT	All	medium	vision
MAP	ELS, VLS	any *	vision+3D
MAT	ELS, VLS	any *	vision+3D
PS	ELS, VLS	↓ or ↑	minimum
BIRD	None	↓ or ↑	minimum
MOVBIRD	ELS, VLS	any *	vision+3D
SURVEY	ELS, VLS	any *	vision+3D
FLYTHROUGH	ELS, VLS	any *	vision+3D
MAPMT	LS, MS, MCU, OTS, 2S/3S	any *	vision+3D
MATMT	LS, MS, OTS, 2S/3S	any *	vision+3D
LTS	VLS, LS, MS, MCU, OTS, 2S/3S	any **	vision
VTS	VLS, LS, MS, MCU, 2S/3S	any **	vision
ORBIT	LS, MS, MCU, CU, 2S/3S	any **	vision
FLYOVER	LS, MS, MCU, CU, 2S/3S	any **	vision+3D
FLYBY	LS, MS, MCU, CU, 2S/3S	any **	vision+3D
DESCENT	LS, MS, MCU, CU, 2S/3S	low	vision+3D
DESCENTOVER	LS, MS, MCU, CU, 2S/3S	low	vision+3D
ASCENT	LS, MS, MCU, 2S/3S	high	vision+3D
CHASE	VLS, LS, MS, OTS, 2S/3S	any **	vision
CONLTS	LS, MS, MCU, OTS, 2S/3S	any **	vision
PST	LS, MS, 2S/3S	↓ or ↑	vision
RS	LS, MS, 2S/3S	any *	vision+3D

trajectory with constant velocity, until the target becomes fully visible. $\tilde{\mathbf{p}}_t$ refers to a real, varying or static target position in WCS.

VI. Multiple-UAV Camera Motion Types

Employing a swarm of cooperating UAVs for video coverage of outdoor events, not only offers great opportunities for novel cinematographic effects, but also a way to deal with many issues arising in single-UAV cinematography. As discussed in Section I, the main advantages are the ability to concurrently film the same target from multiple viewpoints and with multiple FSTs, elimination of dead time intervals due to UAV traveling and maximization of scene coverage.

The CMTs described in Subsection V can, in principle, be assembled in various combinations in order to produce an unlimited number of composite, multiple-UAV CMTs. However, only a percentage of these combinations are cinematographically meaningful and have the potential to significantly improve the resulting visual experience. In this Subsection, a minimal list of four specific, standardized relevant configurations is proposed. In all cases, the employed UAVs should stay out of each other's FoV at all times, in order to preserve transparency of the filming process for the viewer. Obviously, in movie/TV production or in advertising, this is not merely a recommendation, but an absolute requirement.

The multiple-UAV CMTs that have been identified are

detailed below. In all cases, the concurrently deployed UAVs may employ different FSTs, among the fitting ones, while the final broadcasted/edited video feed can alternate between the different UAV inputs, resulting in an exciting visual result. Graphical examples are provided in Figures 6 and 7.

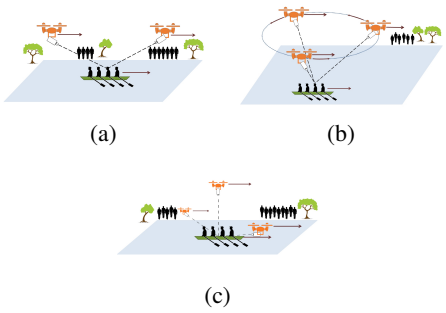

Fig. 6: Examples of three multiple-UAV CMTs: a) 2-UAV Chase (2CHASE), b) 3-UAV Orbit (3ORBIT) and c) 3-UAV Track (3TRACK).

1) 2-UAV Chase (2CHASE): A non-parametric CMT involving two UAVs, each one performing a CHASE on the selected moving target. The first drone leads, thus viewing the target from the front, while the second drone follows, thus viewing the target from behind. The distances between each of the UAVs and the target remains approximately constant during filming. However, the two distances need not be identical, i.e., the first drone can be a lot closer to the target than the second one, or vice versa. The FSTs compatible with 2CHASE are derived from single-UAV CHASE: VLS, LS, MS, OTS and 2S/3S. The mathematical description can be trivially derived from single-UAV CHASE.

In a 2CHASE scenario, alternating the active video feed between a frontal MS and a rear LS, for instance, has the potential to provide a novel, pleasing and dynamic visual experience.

2) 3-UAV Orbit (3ORBIT): A parametric CMT involving three UAVs, each one performing an ORBIT at the selected moving target with a common angular velocity. The (semi-)circular components of the three trajectories coincide, but the drones fly along it with a phase difference. Therefore, the three UAVs remain at all times at the vertices of a spinning triangle, that also moves linearly following the target motion. The FSTs compatible with 3ORBIT are derived from single-UAV ORBIT: LS, MS, MCU, CU and 2S/3S.

The mathematical description is easily derived from that of single-UAV ORBIT, under the following assumptions:

$$\theta_0^2 = \theta_0^1 + \frac{2\pi}{3}, \tag{1}$$

$$\theta_0^3 = \theta_0^1 + \frac{4\pi}{3}, \tag{2}$$

where θ_0^1, θ_0^2 and θ_0^3 are the initial angles of the first, the second and the third UAV, respectively.

3) 3-UAV Track (3TRACK): A non-parametric CMT involving three UAVs: two performing an LTS from opposing sides of the selected moving target ("lateral" UAVs), and one simultaneously performing a VTS ("vertical" UAV). Thus, the two lateral UAVs provide a comprehensive view of the target moving in its environment, while the vertical UAV provides on overview of the target motion from above, with all the cameras being "locked" on the target while simultaneously precisely "tracking" its trajectory. Both the mathematical description and the compatible FSTs can be trivially derived from single-UAV LTS and VTS.

Alternating the active video feed between these three views can provide an aesthetically pleasing and comprehensive shot of the moving target in its surrounding. It must be noted that, although the target itself will prevent the lateral UAVs from being visible to each other, FoV avoidance is more complex when examining whether the lateral UAVs are visible from the vertical one. It depends on the interplay between the vertical UAV's FST, its TCS altitude and the distance between each of the lateral UAVs from the target, thus possibly requiring careful coordination.

4) Dancing Drones (2DD): Dancing Drones, depicted in Figure 7, is a parametric CMT involving two UAVs, each one performing the first half of a FLYOVER and the second half of a FLYBY on the selected still or moving target. The first drone initially leads, thus viewing the target from the front and moving in the opposite direction from the latter, while the second drone initially follows, thus viewing the target from behind and moving to the same direction as the latter. Thus, the first step of 2DD corresponds to the first half of a double FLYOVER, with the two drones flying at very different speeds (the rear one must "catch up" with the moving target and get past it, while the frontal UAV flies towards it). When the two UAVs are about to pass exactly above the target, they avoid colliding by flying perpendicularly to their trajectory up to that point, in opposite directions, without slowing down or losing focus on the target. This intermediate step of 2DD lasts until a pre-specified distance d is covered by each UAV. Subsequently, the two drones turn 90^o once more and start flying in parallel to their original directions. This step of 2DD actually corresponds to the second half of a double FLYBY. Subsequently, each drone flies to its closest position lying upon the target trajectory and a new cycle of 2DD may begin, with the frontal and the rear UAV having exchanged roles.

During the entire filming session, both cameras stay focused on the target and the two UAVs remain at a steady altitude (in WCS). 2DD fits well with LS, MS, MCU, CU and 2S/3S FSTs. The mathematical description can be trivially derived from single-UAV FLYOVER and FLYBY, with pre-specified distance d also serving as the parameter d of FLYBY.

VII. Cognitive Autonomy Exploiting a UAV Shot Type Taxonomy

The presented UAV shot type taxonomy and the accompanying mathematical modelling may easily be employed for facilitating cognitive autonomy algorithms. Three examples are briefly described in this Section, in order to showcase the value

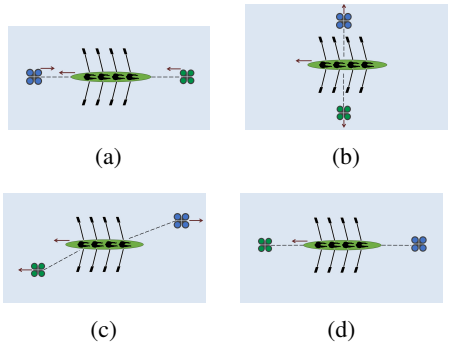

Fig. 7: One cycle of the Dancing Drones (2DD), depicted in consecutive steps: a) first half of a double FLYOVER, b) intermediate step, c) second half of a double FLYBY and d) final step that sets the stage for the next 2DD cycle.

of such a formalized taxonomy. In future research, each of these examples may be expanded to detailed, cinematography-aware algorithms.

A. Autonomous UAV Shot Capturing

According to Table III, all 26 single-UAV CMTs can be autonomously captured assuming a vision+3D operational environment, where 3D coordinates of both the target and the UAV are available at all times in global 3D Cartesian space. In fact, this is rather straightforward by employing black-box, low-level UAV/gimbal controllers, 3D baseline path-planning algorithms and the mathematical description of the CMTs that is provided in the Appendix. This is a task with immediate commercial impact, that clearly showcases the value of the proposed taxonomy.

Additionally, as shown in Table III, a subset of the CMTs could be implemented by relying only on vision technologies (i.e., no GPS-based localization). For instance, instead of relying on the formulas found in the Appendix, several target tracking CMTs can be alternatively defined as a set of requirements that relate 2D visual information, UAV trajectory and camera orientation. A good example would be CHASE, that can be described as the following set of requirements:

- The UAV velocity direction and the camera axis should always form and retain a yaw angle ω_ϕ equal either to 0 or π rad.
- The camera axis pitch angle should remain fixed to $\theta_C < 0$. This is an external parameter that implicitly determines the UAV altitude relative to the target.
- Camera focal length f should remain constant.
- Target 2D ROI area R_a should remain approximately fixed to d_a. d_a is an external parameter that implicitly determines the desired FST.
- Target ROI center $[R_x, R_y]^T$ should always remain at the video frame center/principal point.

Then, a vision-based UAV controller could be implemented which exploits the above requirements to form an error signal, driving a PID controller that controls instant UAV motion parameters. Thus, a target tracking shot could be executed without needing 3D target/UAV coordinates. The only requirement is for the target to initially be visible and detectable on the video frame.

As in the 3D-group-based capturing scenario, this approach is made possible only by the presence of a formalized UAV shot type taxonomy.

B. Object detection on video

A number of 2D object detection algorithms exploit spatiotemporal locality constraints found in video footage, instead of simply processing each video frame independently, in order to augment detection accuracy. Such algorithms may impose inter-frame spatial position constraints (e.g., encoding knowledge that the target ROI trajectory is smooth over time in the video footage), so as to better model expected apparent target motion on the video frames [85] [86] [87].

Assuming that the video being analyzed is derived from a known UAV shot type capture session, additional constraints can be inserted in order to further augment detection accuracy. For instance, in a LTS footage of a bicycle race, all visible bicycles are expected to mainly move horizontally across consecutive video frames (in pixel coordinates), or not move at all in the case of the specific target which the UAV physically tracks. Vertical apparent ROI motion should be negligible-to-none. This knowledge could be encoded as an additional constraint on the spatiotemporal detection algorithm.

Obviously, this would not be possible without a standardized UAV shot type taxonomy.

C. UAV video summarization

UAV video summarization methods have mostly been developed for post-processing geospatial aerial survey footage. This application directly maps to the SURVEY CMT, described in Section V, and typically leads to continuous, long-duration videos with a virtual target. In this case, summarization is a necessary analysis step that automatically selects the most interesting parts for human browsing. The most common algorithmic approach is first to construct a geo-registered video mosaic, either global one, or composed from multiple mini-mosaics that draw their content from different temporal segments of the original footage [88], and then detect objects and/or unusual activity patterns inside this material (e.g., by identifying outlying object trajectories that cannot be reconstructed well after sparse encoding [89]). Saliency-based scoring has also been employed for ranking visible object motion patterns within each video segment. Subsequently, all "interesting" object trajectories from each video segment are superimposed on the same background video frame [90].

However, if different CMTs from the proposed taxonomy are employed during video capturing, the entire process could be augmented with constraints deriving from knowledge of the cinematographic specifications. For instance, in target tracking CMTs, only specific video frames could be pre-selected for analysis, in order to reduce the computational overhead of

summarization. Thus, during an ORBIT circle around the target, only the 4 video frames captured when the UAV is directly behind, directly in front of and to the two sides of the target could be used. This is trivial when operating in a global 3D Cartesian space (e.g., if the targets are equipped with GPSs), but it may require advanced visual 3D target pose estimation algorithms otherwise.

An additional example would be a DESCENT CMT, where only the initial/final video segments at the start/end of the shot could be retained for further analysis, i.e., the video frames where the camera lies the farthest from the target and (almost) directly above it. In general, awareness of the CMT equips us with a priori knowledge regarding the most interesting parts of the footage.

VIII. Compensation Strategies

In practice, UAV cinematography involves issues outlined in Section I: battery autonomy limitations, finite bandwidth in the wireless video transmission channel, restricted flight zones arising from safety-motivated legal requirements, as well as collision/FoV avoidance (in case of multiple-UAV filming). In this Section, two general compensation strategies for alleviating a number of these issues are presented, i.e., *focal length compensation* and *multidrone compensation*. Specific scenarios where each of these strategies is applicable are also identified.

Focal length compensation refers to continuously varying the camera focal length (therefore, zoom level) while the UAV either hovers, or follows a different trajectory than the expected one, in order to partially compensate for an inability to fly along the trajectory specified by the selected shot type. The reasons for such an inability may be flight zone restrictions (e.g., a UAV is not permitted to fly over human crowds), or excessively high target speed in target tracking CMTs, given that maximum UAV speed is constrained. Limited battery autonomy may also be responsible, given that hovering is a less energy-consuming operation than flying up.

Multidrone compensation refers to on-line replacing one primary UAV with another ("auxiliary") in the middle of a continuous filming session. Transitioning the active video feed from the primary UAV to the auxiliary one must incur minimal disruption to the visual result. As before, the reasons for employing this strategy may be limited battery autonomy, i.e., the primary drone is expected to run out of power soon, or flight zone restrictions. In the latter case, focal length compensation may also be required to be employed on the primary UAV immediately before the transition.

In order for this strategy to work, a pool of auxiliary UAVs must be maintained available at all times, located at carefully selected scene positions dispersed throughout the scene to be covered. An auxiliary UAV must get notified and start flying to the appropriate location as soon as the filming session starts, so that it is already optimally placed during the transition.

Two concrete examples of focal length compensation are provided below. The first one is a CHASE from behind with a LS framing type, where a target moves uphill along an inclined plane. The expected UAV trajectory is also inclined upwards, following the terrain tangent plane. Due to energy consumption considerations, the UAV may stop physically following the target and start hovering. From that point in time on, the camera gimbal must begin rotating in order to keep the target properly framed (as in SSMT), while the focal length steadily increases (the camera zooms in) so as to retain the LS framing. The visual result obviously differs from that of a pure CHASE, due to different perspective, and the filming session may need to terminate early, since the target could hide behind an obstacle and the maximum focal length is limited. However, for a time, focal length compensation provides a good approximation of CHASE at a reduced energy cost. This example is illustrated in Figures 8a,b.

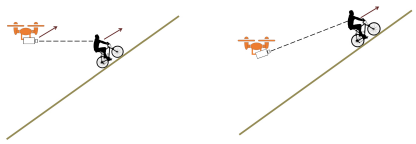

Fig. 8: Example of focal length compensation: a) The UAV shoots the target using a CHASE, b) The UAV starts hovering and transitions to a SSMT, while constantly increasing its focal length for as long as possible.

The second example is a semicircular ORBIT around the still target with a LS framing type and with a restricted flight zone in the middle part of the semicircular trajectory (e.g., a human crowd is present). Instead of following the expected trajectory, the UAV simply flies linearly from the initial point to the endpoint of the semicircle (thus performing a FLYBY), while continuously adjusting the focal length in the process, so as to maintain a LS of the target. As before, due to the different perspective, focal length compensation may only provide an approximation of ORBIT. This example is illustrated in Figures 9a,b.

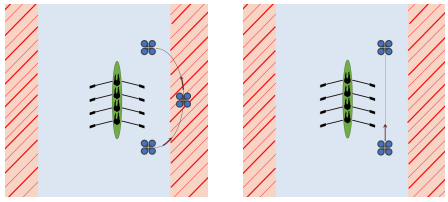

Fig. 9: Example of focal length compensation: a) the UAV shoots the still target using an ORBIT, but the main part of the semicircular trajectory is located within restricted flight area (denoted by light red color), b) the UAV flies linearly instead, actually performing a FLYBY, while constantly adjusting its focal length to maintain proper FST.

Two examples of multidrone compensation are also provided.

The first one is a LTS where the primary UAV is soon expected to enter a low-battery mode, requiring its emergency return to a recharging platform. An auxiliary UAV gets notified at the start of the filming session and begins travelling to the expected transition point with appropriate speed. When the transition between the two UAVs occurs, the active video feed is passed on to the auxiliary UAV, while the primary UAV stops filming and flies to the closest recharging platform. Since during the transition the two UAVs must keep a safety distance between them, the auxiliary UAV may initially, for a very short time interval, begin filming with SSMT CMT, before resuming the LTS. This example is illustrated in Figures 10a,b,c.

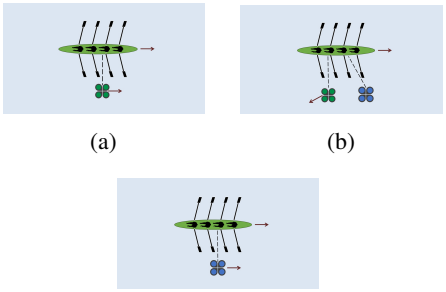

Fig. 10: Example of multidrone compensation: a) The primary UAV shoots the target using a LTS, b) During the transition, the primary UAV departs while the auxiliary UAV starts filming using an intermediate SSMT, c) The auxiliary UAV resumes the LTS.

The second multidrone compensation example is a MATMT where the primary UAV is soon expected to enter a restricted flight area. An auxiliary UAV gets notified at the start of the filming session and begins travelling to that point on the expected UAV trajectory that lies just beyond the end of the restricted flight area (position B). When the primary UAV reaches the restricted zone (position A) it starts hovering with focal length compensation, i.e., it keeps zooming on the approaching target, until the moment of transition. Then, the active video feed is passed on to the auxiliary UAV, while the primary UAV stops filming and returns to the recharging platform. Simultaneously, the auxiliary UAV resumes the MATMT from position B. This example is illustrated in Figures 11a,b,c.

Table IV depicts whether focal length compensation can be successfully employed for each of the single-UAV CMTs, in a cinematographically meaningful way. For instance, the purpose of MAP is to depict the scene context under a constantly changing perspective, therefore focal length compensation (although possible, with the UAV hovering) would not be a good strategy. Multidrone compensation may always be employed for reasons of limited battery autonomy, assuming a pool of auxiliary UAVs is available. In the case of flight zone restrictions, it can be reasonably used if focal length compensation is also compatible with the current shot type. The only requirement is that the restricted flight area is not so

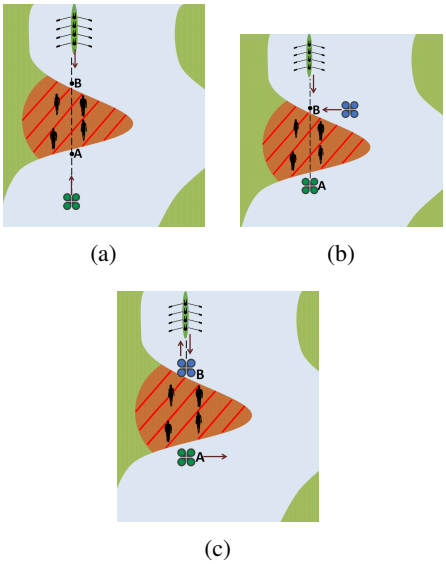

Fig. 11: Example of multidrone compensation: a) The primary UAV films the approaching target using a MATMT, but is unable to continue flying forward due to the restricted flight area just ahead (denoted by light red color), b) The auxiliary UAV is travelling to position B, while the primary UAV hovers at position A with focal length compensation, c) The auxiliary UAV resumes MATMT from position B, while the primary UAV stops filming and withdraws.

extended that the visual result would severely deviate from the expected one with the selected shot type.

IX. Discussion and 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 by the easy availability of advanced VTOL drones, but there is still a lot of room for improvements in multiple aspects. Directions for advancement derive from the currently limited UAV decisional and functional 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.

Outdoor event coverage in dynamic, unstructured environments is undeniably the most difficult and variable task relating to UAV media production. We can easily imagine an ideal scenario where a director gives high-level, concise event coverage instructions in near-natural language before the event. Subsequently, a fully autonomous UAV swarm would acquire the desired footage, while constantly and optimally adapting to the ever-changing situations arising within the event area, under the minimal oversight of a single flight supervisor. In a less ambitious variant, arguably more realistic at the current

TABLE IV: Compensation strategy compatibility with each of the single-UAV CMTs. Light blue, yellow, green and red cells denote static, dynamic scene, target tracking and dynamic target shots, respectively.

Camera Motion	Compensation Strategies	**Camera Motion**	Compensation Strategies
SS	✗	MATMT	✓
SSST	✗	LTS	✗
SAP	✗	VTS	✗
SAT	✗	ORBIT	✓
SSMT	✗	FLYOVER	✓
MAP	✗	FLYBY	✓
MAT	✗	DESCENT	✗
PS	✗	DESCENTOVER	✗
BIRD	✓	ASCENT	✓
MOVBIRD	✗	CHASE	✓
SURVEY	✗	CONLTS	✗
FLYTHROUGH	✓	PST	✗
MAPMT	✓	RS	✗

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, a deep understanding of UAV cinematography is required in order to realize them. Further advancements in sensor technology and computational hardware, as well as progress in UAV cognitive and functional autonomy, enabled by improvements in real-time image/video analysis and mobile networking, are expected to facilitate the process.

This tutorial serves both as an introduction to the topic, and as a step towards achieving a greater understanding of UAV cinematography, by exploiting accumulated industry experience. A wave of further research is needed towards realizing autonomous UAV swarms for dynamic, aerial media coverage requiring minimal human intervention from pilots or directors. Possible future directions include algorithms for on-line, real-time, optimal compensation strategy evaluation, fully autonomous filming (involving all UAV shot types) that considers optimal transitions between different shots, intra-swarm coordination and task assignment interacting with on-line semantic event detection, as well as tight integration of camera control with UAV path planning algorithms, under cinematography-aware guidelines. In all cases, energy efficiency, legal flight restrictions, collision/FoV avoidance and limited communication channel bandwidth are factors to be considered.

ACKNOWLEDGMENT

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). We are thankful to Chara Raptopoulou, Fotini Patrona and Panagiotis Ksanthos for assisting in the preparation of the figures included in this article. This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] E. Cheng, *Aerial Photography and Videography Using Drones*, Peachpit Press, 2016.

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

[3] A. Hocraffer and C. S. Nam, "A meta-analysis of human-system interfaces in unmanned aerial vehicle (UAV) swarm management," *Applied Ergonomics*, vol. 58, pp. 66–80, 2017.

[4] M. L. Cummings, A. Clare, and C. Hart, "The role of human-automation consensus in multiple unmanned vehicle scheduling," *Human Factors*, vol. 52, no. 1, pp. 17–27, 2010.

[5] M. L. Cummings, J. P. How, A. Whitten, and O. Toupet, "The impact of human-automation collaboration in decentralized multiple unmanned vehicle control," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 660–671, 2012.

[6] A. Kolling, P. Walker, N. Chakraborty, K. Sycara, and M. Lewis, "Human interaction with robot swarms: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 9–26, 2016.

[7] I. Mademlis, I. Mygdalis, C. Raptopoulou, N. Nikolaidis, N. Heise, T. Koch, J. Grunfeld, T. Wagner, A. Messina, F. Negro, S. Metta, and I. Pitas, "Overview of drone cinematography for sports filming," in *European Conference on Visual Media Production (CVMP) (short)*, 2017.

[8] A. Torres-González, J. Capitán, R. Cunha, A. Ollero, and I. Mademlis, "A multidrone approach for autonomous cinematography planning," in *Proceedings of Iberian Robotics Conference (ROBOT')*, 2017.

[9] 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.

[10] 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*, 2019.

[11] 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, no. 1, pp. 147–153, 2019.

[12] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "UAV cinematography constraints imposed by visual target tracking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018.

[13] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "Shot type feasibility in autonomous UAV cinematography," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[14] F. Patrona, I. Mademlis, A. Tefas, and I. Pitas, "Computational UAV cinematography for intelligent shooting based on semantic visual analysis," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

[15] A. Messina, S. Metta, M. Montagnuolo, F. Negro, V. Mygdalis, I. Pitas, J. Capitán, A. Torres, S. Boyle, and D. Bull, "The future of media production through multi-drones' eyes," in *International Broadcasting Convention (IBC)*, 2018.

[16] V. Gandhi and R. Ronfard, "A computational framework for vertical video editing," in *Proceedings of the Workshop on Intelligent Camera Control, Cinematography and Editing (WICED)*, 2015.

[17] P. Carr, M. Mistry, and I. Matthews, "Hybrid robotic/virtual pan-tilt-zom cameras for autonomous event recording," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2013.

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

[19] C. Gebhardt, B. Hepp, T. Nägeli, S. Stevšić, and O. Hilliges, "Airways: Optimization-based planning of quadrotor trajectories according to high-level user goals," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2016.

[20] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir, "Automatic editing of footage from multiple social cameras," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 81, 2014.

[21] B. Brown, *Cinematography: Theory and Practice: Image Making for Cinematographers and Directors*, Focal Press, 3rd edition, 2016.

[22] F. Daniyal and A. Cavallaro, "Multi-camera scheduling for video production," in *Proceedings of the IEEE Conference for Visual Media Production (CVMP)*, 2011.

[23] A. Saeed, A. Abdelkader, M. Khan, A. Neishaboori, K. A. Harras, and A. Mohamed, "On realistic target coverage by autonomous drones," *arXiv preprint arXiv:1702.03456*, 2017.

[24] 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.

[25] 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.

[26] Q. Galvane, J. Fleureau, F.-L. Tariolle, and P. Guillotel, "Automated cinematography with Unmanned Aerial Vehicles," in *Proceedings of the Workshop on Intelligent Camera Control, Cinematography and Editing (WICED)*, 2016.

[27] Amber Garage, "Skywand," https://skywand.com/.

[28] FreeSkies, "FreeSkies CoPilot," http://freeskies.co/.

[29] 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.

[30] T. Nägeli, J. Alonso-Mora, A. Domahidi, D. Rus, and O. Hilliges, "Real-time motion planning for aerial videography with dynamic obstacle avoidance and viewpoint optimization," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1696–1703, 2017.

[31] M. S. Grewal, L. R. Weill, and A. P. Andrews, *Global Positioning Systems, inertial navigation, and integration*, John Wiley & Sons, 2007.

[32] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robotics & Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[33] F. Fraundorfer and D. Scaramuzza, "Visual odometry: Part II: Matching, robustness, optimization, and applications," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 78–90, 2012.

[34] 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.

[35] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *arXiv preprint arXiv:1610.06475*, 2016.

[36] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Generic and real-time structure from motion using local bundle adjustment," *Image and Vision Computing*, vol. 27, no. 8, pp. 1178–1193, 2009.

[37] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[38] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 4, pp. 31–43, 2010.

[39] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Real-time visual loop-closure detection," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2008, pp. 1842–1847.

[40] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[41] J. Zhang and S. Singh, "LOAM: LIDAR odometry and mapping in real-time.," in *Proceedings of Robotics: Science and Systems*, 2014.

[42] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.

[44] D. Triantafyllidou and A. Tefas, "Face detection based on deep convolutional neural networks exploiting incremental facial part learning," in *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2016.

[45] D. Triantafyllidou, P. Nousi, and A. Tefas, "Lightweight two-stream convolutional face detection," in *Proceedings of EURASIP European Signal Processing Conference (EUSIPCO)*, 2017.

[46] D. Triantafyllidou, P. Nousi, and A. Tefas, "Fast deep convolutional face detection in the wild exploiting hard sample mining," *Big Data Research*, vol. 11, pp. 65 – 76, 2018.

[47] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, 2016.

[48] R. Couturier, *Designing scientific applications on GPUs*, CRC Press, 2013.

[49] 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.

[50] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2016.

[51] P. Nousi, E. Patsiouras, A. Tefas, and I. Pitas, "Convolutional neural networks for visual information analysis with limited computing resources," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018.

[52] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[53] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[54] N. Passalis and A. Tefas, "Concept detection and face pose estimation using lightweight convolutional neural networks for steering drone video shooting," in *Proceedings of EURASIP European Signal Processing Conference (EUSIPCO)*, 2017.

[55] N. Passalis and A. Tefas, "Bag-of-features pooling for deep convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[56] 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.

[57] M. Tzelepi and A. Tefas, "Graph-embedded convolutional neural networks in human crowd detection for drone flight safety," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2019.

[58] G. Verghese, "Perspective alignment back projection for monocular tracking of solid objects.," in *Proceedings of the British Machine Vision Conference (BMVC)*, 1993.

[59] 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.

[60] L. S. Monteiro, T. Moore, and C. Hill, "What is the accuracy of DGPS?," *The Journal of Navigation*, vol. 58, no. 2, pp. 207–225, 2005.

[61] X. Huang, R. Janaswamy, and A. Ganz, "Scout: Outdoor localization using Active RFID technology," in *Proceedings of IEEE Conference on Broadband Communications, Networks and Systems (BROADNETS)*, 2006, pp. 1–10.

[62] L. Yang, J. Qi, J. Xiao, and X. Yong, "A literature review of UAV 3D path planning," in *Proceedings of the IEEE World Congress on Intelligent Control and Automation (WCICA)*, 2014.

[63] S. Ragi and E.K.P. Chong, "UAV path planning in a dynamic environment via Partially Observable Markov Decision Process," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 49, no. 4, pp. 2397–2412, 2013.

[64] Celine Teuliere, Laurent Eck, and Eric Marchand, "Chasing a moving target from a flying UAV," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2011.

[65] Tayyab Naseer, Jürgen Sturm, and Daniel Cremers, "Followme: Person following and gesture recognition with a quadrocopter," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[66] L. Meier, P. Tanskanen, F. Fraundorfer, and M. Pollefeys, "Pixhawk: A system for autonomous flight using onboard computer vision," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[67] P. Serra, R. Cunha, T. Hamel, D. Cabecinhas, and C. Silvestre, "Landing of a quadrotor on a moving target using dynamic image-based visual servo control," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1524–1535, 2016.

[68] F. Sadeghi and S. Levine, "(CAD)2RL: Real single-image flight without a single real image," *arXiv preprint arXiv:1611.04201*, 2016.

[69] D.K. Kim and T. Chen, "Deep neural network for real-time autonomous indoor navigation," *arXiv preprint arXiv:1511.04668*, 2015.

[70] K. Kelchtermans and T. Tuytelaars, "How hard is it to cross the room? training (recurrent) neural networks to steer a uav," *arXiv preprint arXiv:1702.07600*, 2017.

[71] M. Mueller, V. Casser, N. Smith, and B. Ghanem, "Teaching UAVs to race using UE4Sim," *arXiv preprint arXiv:1708.05884*, 2017.

[72] H. Duan, Q. Luo, Y. Shi, and G. Ma, "Hybrid particle swarm optimization and genetic algorithm for multi-UAV formation reconfiguration," *IEEE Computational Intelligence Magazine*, vol. 8, no. 3, pp. 16–27, 2013.

[73] C. Ju and H. Son, "Multiple UAV systems for agricultural applications: control, implementation, and evaluation," *Electronics*, vol. 7, no. 9, pp. 162, 2018.

[74] K.-K. Oh, M.-C. Park, and H.-S. Ahn, "A survey of multi-agent formation control," *Automatica*, vol. 53, pp. 424–440, 2015.

[75] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P.J.M. Havinga, "A survey of online activity recognition using mobile phones," *Sensors*, vol. 15, no. 1, pp. 2059–2085, 2015.

[76] A. Nemra and N. Aouf, "Robust Cooperative UAV Visual SLAM," in *IEEE International Conference on Cybernetic Intelligent Systems (CIS)*, 2010.

[77] A. Tsourdos, B. White, and M. Shanmugavel, *Cooperative path planning of Unmanned Aerial Vehicles*, vol. 32, John Wiley & Sons, 2010.

[78] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on Unmanned Aerial Vehicle networks for civil applications: A communications viewpoint," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2624–2661, 2016.

[79] D. Avola, L. Cinque, G. L. Foresti, N. Martinel, D. Pannone, and C. Piciarelli, "A UAV video dataset for mosaicking and change detection from low-altitude flights," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, , no. 99, pp. 1–11, 2018.

[80] G. Morgenthal and N. Hallermann, "Quality assessment of unmanned aerial vehicle (UAV)-based visual inspection of structures," *Advances in Structural Engineering*, vol. 17, no. 3, pp. 289–302, 2014.

[81] W. Li, H. Fu, L. Yu, and A. Cracknell, "Deep learning-based oil palm tree detection and counting for high-resolution remote sensing images," *Remote Sensing*, vol. 9, no. 1, pp. 22, 2016.

[82] C. Hung, Z. Xu, and S. Sukkarieh, "Feature learning-based approach for weed classification using high resolution aerial images from a digital camera mounted on a UAV," *Remote Sensing*, vol. 6, no. 12, pp. 12037–12054, 2014.

[83] I. Tsingalis, A. Tefas, N. Nikolaidis, and I. Pitas, "Shot type characterization in 2D and 3D video content," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2014.

[84] H. Fourati and D.E.C. Belkhiat, *Multisensor Attitude Estimation: Fundamental Concepts and Applications*, CRC Press LLC, 2016.

[85] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[86] Y. Yuan, X. Liang, X. Wang, D.-Y. Yeung, and A. Gupta, "Temporal dynamic graph LSTM for action-driven video object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[87] F. Xiao and Y. Jae Lee, "Track and segment: An iterative unsupervised approach for video object proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[88] R. Viguier, C. C. Lin, H. AliAkbarpour, F. Bunyak, S. Pankanti, G. Seetharaman, and K. Palaniappan, "Automatic video content summarization using geospatial mosaics of aerial imagery," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, 2015.

[89] K. Pitstick, J. Hansen, M. Klein, E. Morris, and J. Vazquez-Trejo, "Applying video summarization to aerial surveillance," in *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*. International Society for Optics and Photonics, 2018.

[90] H. Trinh, J. Li, S. Miyazawa, J. Moreno, and S. Pankanti, "Efficient UAV video event summarization," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 2012.

APPENDIX: Mathematical Description of Single-UAV Camera Motion Types

The 26 CMTs, clustered into four groups, are described geometrically below according to the mathematical framework presented in Section V. Table V summarizes the relevant notation.

A. Static shots

1) *Static Shot (SS)* (scene-oriented), 2) *Static Shot of Still Target (SSST)* (target-oriented) and 3) *Static Shot of Moving Target (SSMT)*. The base mathematical description for all three is terse:

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1}, \quad \tilde{\mathbf{l}}_t = \tilde{\mathbf{p}}_t, \forall t \tag{3}$$

Additionally, for SS and SSST, it also holds that:

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_{t-1}, \forall t. \tag{4}$$

4) *Static Aerial Pan (SAP)*. The mathematical description (in TCS) employs $\mathbf{p}_t$ as a reference point marking the center of a line segment $S = (\mathbf{l}_{min}, \mathbf{l}_{max})$. The panning rotation consists in moving $\mathbf{l}_t$ along this line segment, thus $\mathbf{l}_{min}$ and $\mathbf{l}_{max}$ are defined based on an absolute maximum yaw camera rotation angle parameter θ (measured in degrees). When $\mathbf{l}_t = \mathbf{p}_t$, the current yaw camera rotation angle is zero and the camera axis is perpendicular to S. An additional angular velocity parameter ω (measured in degrees per second) affects how fast the panning rotation is performed. Therefore[5]:

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1}, \forall t \tag{5}$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_{t-1}, \forall t \tag{6}$$

$$n = \|\mathbf{x}_t\| \tan\theta \tag{7}$$

$$\mathbf{l}_{min} = n\frac{\mathbf{x}_t \times \mathbf{k}}{\|\mathbf{x}_t \times \mathbf{k}\|} \tag{8}$$

$$\mathbf{l}_{max} = -\mathbf{l}_{min} \tag{9}$$

$$\mathbf{l}_t = \mathbf{l}_{min} + p(t)(\mathbf{l}_{max} - \mathbf{l}_{min}) \tag{10}$$

$$p(t) \in [0,1], \quad p(0) = 0, \quad p(t) = p(t-1) + \frac{\omega}{2\theta T} \tag{11}$$

Figure 12 depicts the line segment connecting $\mathbf{l}_{min}$ and $\mathbf{p}_t$, along which $\mathbf{l}_t$ moves during the first half of a SAP shot.

5) *Static Aerial Tilt (SAT)*. The mathematical description (in TCS) is similar to that of SAP, but with the line segment $S = (\mathbf{l}_{min}, \mathbf{l}_{max})$ now being vertical to the ground plane. Parameters θ and ω should also be specified, as in SAP, with θ referring to absolute maximum pitch camera rotation angle. Therefore:

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1}, \forall t \tag{12}$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_{t-1}, \forall t \tag{13}$$

[5]Operator $\times$ denotes the cross product.

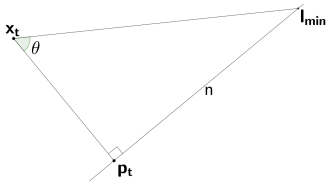

Fig. 12: Geometry of the Static Aerial Pan (SAP): the line segment connecting $\mathbf{l}_{min}$ and $\mathbf{p}_t$ is depicted. It is one half of the line segment $S = (\mathbf{l}_{min}, \mathbf{l}_{max})$, along which $\mathbf{l}_t$ moves during a SAP shot.

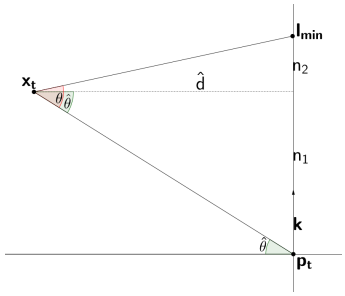

Fig. 13: Geometry of the Static Aerial Tilt (SAT): the line segment connecting $\mathbf{l}_{min}$ and $\mathbf{p}_t$ is depicted. As shown, its length is equal to $n_1 + n_2$. $\mathbf{k}$ is the vertical TCS unit axis vector.

$$\hat{d} = \sqrt{x_{t1}^2 + x_{t2}^2} \tag{14}$$

$$\hat{\theta} = \arcsin\frac{x_{t3}}{\|\mathbf{x}_t\|} \tag{15}$$

$$n_1 = x_{t3} \tag{16}$$

$$n_2 = \hat{d}\tan(\theta - \hat{\theta}) \tag{17}$$

$$n_3 = \hat{d}\tan(\theta + \hat{\theta}) - x_{t3} \tag{18}$$

$$n = n_1 + n_2 = x_{t3} + \hat{d}\tan(\theta - \hat{\theta}) \tag{19}$$

$$\mathbf{l}_{min} = n\mathbf{k} \tag{20}$$

$$\mathbf{l}_{max} = -n_3\mathbf{k} \tag{21}$$

$$\mathbf{l}_t = \mathbf{l}_{min} + p(t)(\mathbf{l}_{max} - \mathbf{l}_{min}) \tag{22}$$

$$p(t) \in [0,1], \quad p(0) = 0, \quad p(t) = p(t-1) + \frac{\omega}{2\theta T}, \quad \theta < \frac{\pi}{2} - \hat{\theta} \tag{23}$$

The line segment connecting $\mathbf{l}_{min}$ and $\mathbf{p}_t$, as well as the one connecting $\mathbf{l}_{max}$ and $\mathbf{p}_t$, are depicted in Figures 13 and 14, respectively. Thus, the entire line segment along which $\mathbf{l}_t$ moves during a SAT shot is visualized.

B. Dynamic scene shots

1) *Moving Aerial Pan (MAP)* and 2) *Moving Aerial Tilt (MAT)* [2]. The mathematical description of MAP is similar to that of SAP, but also incorporates synchronized UAV and virtual

TABLE V: UAV/Camera Motion Type description nomenclature.

$\tilde{\mathbf{p}}_t = [\tilde{p}_{t1}, \tilde{p}_{t2}, \tilde{p}_{t3}]^T$	The 3D target position in WCS, at time instance t
$\mathbf{p}_t = [p_{t1}, p_{t2}, p_{t3}]^T$	The 3D target position in TCS, i.e., the current TCS origin at time instance t. It is always equal to $[0, 0, 0]^T$.
$\tilde{\mathbf{x}}_t = [\tilde{x}_{t1}, \tilde{x}_{t2}, \tilde{x}_{t3}]^T$	The 3D UAV position in WCS, at time instance t
$\mathbf{x}_t = [x_{t1}, x_{t2}, x_{t3}]^T$	The 3D UAV position in TCS, at time instance t
$\tilde{\mathbf{u}}_t = [\tilde{u}_{t1}, \tilde{u}_{t2}, \tilde{u}_{t3}]^T$	The estimated 3D target velocity in WCS, at time instance t
$\mathbf{u}_t = [u_{t1}, u_{t2}, u_{t3}]^T$	The estimated 3D target velocity in TCS, at time instance t
$\tilde{\mathbf{v}}_t = [\tilde{v}_{t1}, \tilde{v}_{t2}, \tilde{v}_{t3}]^T$	The 3D UAV velocity in WCS, at time instance t
$\mathbf{v}_t = [v_{t1}, v_{t2}, v_{t3}]^T$	The 3D UAV velocity in TCS, at time instance t
$\tilde{\mathbf{l}}_t \in \mathbb{R}^3$	The 3D position at which the camera looks (known as the "LookAt point") in WCS, at time instance t
$\mathbf{l}_t \in \mathbb{R}^3$	The 3D position at which the camera looks in TCS, at time instance t
$\tilde{\mathbf{o}}_t = \tilde{\mathbf{l}}_t - \tilde{\mathbf{x}}_t$	The LookAt vector in WCS, at time instance t. It is a scalar multiple of the camera axis.
$\mathbf{o}_t = \mathbf{l}_t - \mathbf{x}_t$	The LookAt vector in TCS, at time instance t
$\tilde{\mathbf{i}}, \tilde{\mathbf{j}}, \tilde{\mathbf{k}}$	The WCS axes unit vectors
$\mathbf{i}, \mathbf{j}, \mathbf{k}$	The TCS axes unit vectors
T	The camera frame-rate

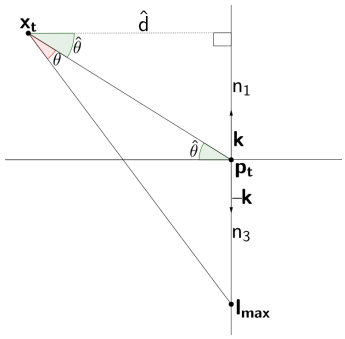

Fig. 14: Geometry of the Static Aerial Tilt (SAT): the line segment connecting $\mathbf{l}_{max}$ and $\mathbf{p}_t$ is depicted. As shown, its length is equal to n_3. $\mathbf{k}$ is the vertical TCS unit axis vector.

target motion prescribed by the UAV/target velocity vector $\tilde{\mathbf{v}}_t$. Because of this synchronized motion $\mathbf{x}_t$ remains constant in TCS for all time instances, while $\tilde{\mathbf{x}}_t$ in WCS varies over time according to $\tilde{\mathbf{v}}_t$. Similarly, $\mathbf{l}_{min}$ and $\mathbf{l}_{max}$ are constant in TCS but vary over time in WCS, since $\tilde{\mathbf{p}}_t$ changes according to $\tilde{\mathbf{v}}_t$. The parameters that must be specified are θ, ω and $\tilde{\mathbf{v}}_t$, with θ referring to absolute maximum yaw camera rotation angle. Therefore, the equations describing MAP are the Eqs. (7)-(11) from SAP, plus the following ones:

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1}, \forall t \tag{24}$$

$$\tilde{\mathbf{u}}_t = \tilde{\mathbf{v}}_t \tag{25}$$

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + \frac{\tilde{\mathbf{v}}_t}{T} t \tag{26}$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_0 + \frac{\tilde{\mathbf{u}}_t}{T} t \tag{27}$$

$$\mathbf{x}_t = \mathbf{x}_{t-1}, \forall t \tag{28}$$

The mathematical description of MAT is similar to that of MAP, but with the line segment $S = (\mathbf{l}_{min}, \mathbf{l}_{max})$ now being vertical to the ground plane. The parameters that must be specified are θ, ω and $\tilde{\mathbf{v}}_t$, with θ referring to absolute maximum pitch camera rotation angle. Therefore, the equations describing MAT are the Eqs. (14)-(23) from SAT, plus the Eqs. (24)-(28) from MAP.

3) *Pedestal/Elevator Shot (PS)* [1] [2] and 4) *Bird's Eye Shot (BIRD)*. The parameters that must be specified are d, i.e., the vertical distance to be traversed by the UAV during filming, and v_{t3}, i.e., the scalar speed of the UAV during filming (constant over time). The base mathematical description for both CMTs is the following:

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1} = v_{t3}\tilde{\mathbf{k}}, \forall t \tag{29}$$

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + t\frac{\tilde{\mathbf{v}}_t}{T} \tag{30}$$

$$\mathbf{l}_t = \mathbf{p}_t \tag{31}$$

$$t \in [0, \frac{Td}{|v_{t3}|}] \tag{32}$$

Additionally, the following hold for PS:

$$\tilde{\mathbf{u}}_t = \tilde{\mathbf{v}}_t \tag{33}$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_0 + t\frac{\tilde{\mathbf{u}}_t}{T} \tag{34}$$

$$\tilde{\mathbf{o}}_t^T \tilde{\mathbf{k}} \approx 0 \tag{35}$$

$$\mathbf{x}_t = \mathbf{x}_{t-1}, \forall t \tag{36}$$

$$\tilde{p}_{03} = \tilde{x}_{03} \implies x_{03} = 0 \tag{37}$$

and the following hold for BIRD:

$$\tilde{\mathbf{p}_0} = [\tilde{x}_{01}, \tilde{x}_{02}, 0]^T \tag{38}$$

$$\tilde{\mathbf{p}_t} = \tilde{\mathbf{p}_{t-1}}, \forall t \tag{39}$$

$$\tilde{\mathbf{o}}_t \times \tilde{\mathbf{k}} \approx \mathbf{0} \tag{40}$$

5) *Moving Bird's Eye Shot (MOVBIRD)* and 6) *Survey Shot (SURVEY)*. The parameter that must be specified is the direction and speed of flying, i.e., the velocity vector $\tilde{\mathbf{v}}_t$, lying on the terrain tangent plane. The latter differs from the ground plane in case of inclined terrain, otherwise they coincide. The base mathematical description for both CMTs is the following:

$$\tilde{\mathbf{u}}_t = \tilde{\mathbf{v}}_t \tag{41}$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_0 + t\frac{\tilde{\mathbf{u}}_t}{T} \tag{42}$$

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + t\frac{\tilde{\mathbf{v}}_t}{T} \tag{43}$$

$$\mathbf{l}_t = \mathbf{p}_t \tag{44}$$

Additionally, the following holds for MOVBIRD:

$$\tilde{\mathbf{p}}_0 = [\tilde{x}_{01}, \tilde{x}_{02}, 0]^T \tag{45}$$

and the following holds for SURVEY:

$$\tilde{\mathbf{o}}_t \times \tilde{\mathbf{v}}_t \approx \mathbf{0} \tag{46}$$

7) *Fly-Through (FLYTHROUGH)*. The parameters that must be specified are the time K (in seconds) until the gap is reached and the 3D position of the gap center ($\tilde{\mathbf{x}}_{KT}$) in WCS. The mathematical description is the following:

$$t \in [0, KT] \tag{47}$$

$$\tilde{\mathbf{d}} = \tilde{\mathbf{x}}_{KT} - \tilde{\mathbf{x}}_0 \tag{48}$$

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1} = \frac{\tilde{\mathbf{d}}}{K} \tag{49}$$

$$\tilde{\mathbf{p}}_0 = \tilde{\mathbf{x}}_{KT} \quad \tilde{\mathbf{u}}_t = \tilde{\mathbf{v}}_t \tag{50}$$

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_0 + t\frac{\tilde{\mathbf{u}}_t}{T} \tag{51}$$

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + t\frac{\tilde{\mathbf{v}}_t}{T} \tag{52}$$

$$\mathbf{l}_t = \mathbf{p}_t \tag{53}$$

C. Target tracking shots

1) *Moving Aerial Pan with Moving Target (MAPMT)* and 2) *Moving Aerial Tilt with Moving Target (MATMT)*. Parameter $\tilde{\mathbf{v}}_t$ must be specified. The base mathematical description for both is fairly simple:

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1} \forall t \implies \tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + \frac{\tilde{\mathbf{v}}_t}{T} t \tag{54}$$

$$\mathbf{l}_t = \mathbf{p}_t \tag{55}$$

Additionally, the following relation holds for MAPMT:

$$[\tilde{u}_{t1}, \tilde{u}_{t2}, 0][\tilde{v}_{t1}, \tilde{v}_{t2}, 0]^T \approx 0 \tag{56}$$

and the following relation holds for MATMT:

$$[\tilde{u}_{t1}, \tilde{u}_{t2}, 0]^T \times [\tilde{v}_{t1}, \tilde{v}_{t2}, 0]^T \approx \mathbf{0} \tag{57}$$

3) *Lateral Tracking Shot (LTS)* [1] [2] and 4) *Vertical Tracking Shot (VTS)*. The base mathematical description for both is fairly simple:

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{u}}_t \tag{58}$$

$$\tilde{\mathbf{o}}_t^T \tilde{\mathbf{u}}_t \approx 0 \tag{59}$$

$$\mathbf{x}_t = \mathbf{x}_{t-1}, \quad \mathbf{l}_t = \mathbf{p}_t \forall t \tag{60}$$

Additionally, the following relations hold for LTS:

$$\mathbf{o}_t \times \mathbf{j} \approx \mathbf{0}, \quad x_{03} \approx 0 \tag{61}$$

while the following relations hold for VTS:

$$\mathbf{o}_t^T \mathbf{j} \approx 0, \quad x_{03} > 0 \tag{62}$$

5) *Orbit (ORBIT)*. The parameters that must be specified are the desired 3D Euclidean distance $\lambda_{3D} = \|\tilde{\mathbf{x}}_t - \tilde{\mathbf{p}}_t\|_2 = \|\mathbf{x}_t\|_2$ (constant over time), the angle of the entire rotation to be performed around the target (θ) and the desired UAV angular velocity ω. Additionally, we can easily derive the initial angle θ_0 formed by the TCS i-axis (of time instance 0) and the vector from $\mathbf{p}_0$ to the projection of the known initial position $\mathbf{x}_0$ onto the TCS ij-plane. Then, ORBIT may be described in TCS using a planar circular motion:

$$t \in [0, \frac{T\theta}{\omega}] \tag{63}$$

$$\theta_0 = arctan\left(\frac{x_{02}}{x_{01}}\right) \tag{64}$$

$$x_{t3} = x_{03}, \forall t \tag{65}$$

$$\lambda = \sqrt{\lambda_{3D}^2 - x_{t3}^2} \tag{66}$$

$$\mathbf{x}_t = [\lambda \cos(t\frac{\omega}{T} + \theta_0), \lambda \sin(t\frac{\omega}{T} + \theta_0), x_{t3}]^T \tag{67}$$

$$\mathbf{l}_t = \mathbf{p}_t \tag{68}$$

The projection of the initial UAV position onto the TCS ij-plane ($\mathbf{x}_0$) is shown in Figure 15.

6) *Fly-Over (FLYOVER)* and 7) *Fly-By (FLYBY)*. The common parameter that must be specified is K, i.e., the time

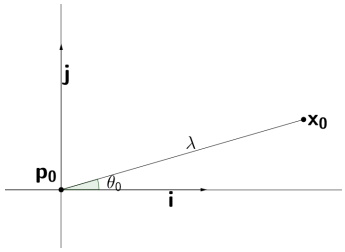

Fig. 15: Geometry of the Orbit (ORBIT): the projection of the initial UAV position onto the TCS ij-plane ($\mathbf{x}_0$) is depicted. $\mathbf{i}$ and $\mathbf{j}$ are two of the TCS unit axis vectors.

(in seconds) until UAV is located exactly above the target (for FLYOVER), or until the distance vector between the target and the UAV is minimized in Euclidean norm (for FLYBY). Additionally, d, i.e., the length of the projection of that minimal distance vector onto the ground plane, must be specified for FLYBY. Below, the target velocity is assumed constant for reasons of modeling convenience. The base mathematical description common to both CMTs is the following:

$$\mathbf{v}_0 = [\frac{u_{01}K - x_{01}}{K}, 0, u_{03}]^T \tag{69}$$

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1}, \quad \tilde{\mathbf{u}}_t = \tilde{\mathbf{u}}_{t-1}, \forall t \tag{70}$$

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + \frac{t}{KT}(\tilde{\mathbf{x}}_{KT} - \tilde{\mathbf{x}}_0) \tag{71}$$

$$[\tilde{u}_{t1}, \tilde{u}_{t2}, 0]^T \times [\tilde{v}_{t1}, \tilde{v}_{t2}, 0]^T \approx \mathbf{0} \tag{72}$$

$$\mathbf{l}_t = \mathbf{p}_t, \quad t \in [0, 2KT] \tag{73}$$

Additionally, the following hold for FLYOVER:

$$\tilde{\mathbf{x}}_{KT} = [\tilde{p}_{01} + \tilde{u}_{01}K, \tilde{p}_{02} + \tilde{u}_{02}K, \tilde{x}_{03} + \tilde{u}_{03}K]^T \tag{74}$$

$$x_{t2} \approx 0, \quad \mathbf{x}_t^T \mathbf{j} \approx 0, \forall t \tag{75}$$

and the following hold for FLYBY:

$$|x_{02}| = d > 0 \tag{76}$$

$$x_{t2} = x_{02}, \forall t \tag{77}$$

$$\mathbf{x}_{KT} = [0, x_{02}, x_{03}]^T \tag{78}$$

8) *Descent (DESCENT)*, 9) *Descent Over (DESCENTOVER)* and 10) *Ascent (ASCENT)*. The common parameter that must be specified is θ, i.e., the constant angle formed between the UAV and the TCS $\mathbf{i}$ axis. Additionally, K, i.e., the time (in seconds) until UAV is located exactly above the target, must be specified for DESCENT and DESCENTOVER. The smaller K is, the faster the UAV will move. Based on this observation, K may also be employed for parameterizing ASCENT, although in this motion type the UAV actually moves away from the target. Below, the target velocity is assumed constant for reasons of modeling convenience, while $\mathbf{R}_j$ refers to a 3×3 matrix that clockwise-rotates any vector multiplied with it along the j-axis. The base mathematical description common to DESCENT and DESCENTOVER is the following:

$$\mathbf{v}'_0 = [\frac{u_{01}K - x_{01}}{K}, 0, 0]^T \tag{79}$$

$$\mathbf{R}_j = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \tag{80}$$

$$\mathbf{v}_0 = \mathbf{R}_j \mathbf{v}'_0 \tag{81}$$

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1}, \quad \tilde{\mathbf{u}}_t = \tilde{\mathbf{u}}_{t-1}, \forall t \tag{82}$$

$$\mathbf{x}_{KT} = [\tilde{p}_{01} + \tilde{u}_{01}K, \tilde{p}_{02} + \tilde{u}_{02}K, \tilde{x}_{03} + \tilde{v}_{03}K]^T \tag{83}$$

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + \frac{t}{KT}(\tilde{\mathbf{x}}_{KT} - \tilde{\mathbf{x}}_0) \tag{84}$$

$$[\tilde{u}_{t1}, \tilde{u}_{t2}, 0]^T \times [\tilde{v}_{t1}, \tilde{v}_{t2}, 0]^T \approx \mathbf{0} \tag{85}$$

$$x_{t2} \approx 0, \quad \mathbf{x}_t^T \mathbf{j} \approx 0, \quad \mathbf{l}_t = \mathbf{p}_t, \forall t \tag{86}$$

Additionally, the following holds for DESCENT:

$$t \in [0, KT] \tag{87}$$

and the following holds for DESCENTOVER:

$$t \in [0, 2KT] \tag{88}$$

The mathematical description for ASCENT is similar. It is given by Eqs. (79), (80), (82), (85) and (86), while Eq. (81) is replaced by the following one:

$$\mathbf{v}_0 = -\mathbf{R}_j \mathbf{v}'_0 \tag{89}$$

11) *Chase/Follow Shot (CHASE)*. The mathematical description is the following:

$$\tilde{\mathbf{v}}_t \approx \tilde{\mathbf{u}}_t \tag{90}$$

$$x_{t2} = x_{02} \approx 0, \forall t \tag{91}$$

$$\mathbf{x}_t = \mathbf{x}_{t-1}, \forall t \tag{92}$$

$$\mathbf{l}_t = \mathbf{p}_t \tag{93}$$

D. Dynamic target shots

1) *Constrained Lateral Tracking Shot (CONLTS)*. The parameters that must be specified are $\tilde{\mathbf{n}} = [\tilde{n}_1, \tilde{n}_2, 0]^T$ and $\tilde{\mathbf{s}} = [\tilde{s}_1, \tilde{s}_2, \tilde{s}_3]^T$, i.e., a normal vector and a scene point jointly defining the flight plane in WCS. The mathematical description is based on determining the intersection of the flight plane with a line perpendicular to the plane, passing through $\tilde{\mathbf{p}}_t$:

$$a = \tilde{x}_{03} - \tilde{p}_{03} \tag{94}$$

$$d_t = \frac{(\tilde{\mathbf{s}} - \tilde{\mathbf{p}}_t)^T \tilde{\mathbf{n}}}{\tilde{\mathbf{n}}^T \tilde{\mathbf{n}}} \tag{95}$$

$$\tilde{\mathbf{x}}_t = d_t \tilde{\mathbf{n}} + \tilde{\mathbf{p}}_t + [0, 0, a]^T \tag{96}$$

$$\tilde{\mathbf{l}}_t = \tilde{\mathbf{p}}_t \tag{97}$$

2) *Pedestal/Elevator Shot With Target (PST)*. The parameters that must be specified are d, i.e., the vertical distance to be traversed by the UAV/target during filming, and v_{t3}, i.e., the

scalar speed of the UAV/target during filming (constant over time). The mathematical description includes the Eqs. (29) - (32) from PS, plus the following relation:

$$\mathbf{x}_t^T \mathbf{j} \approx 0 \tag{98}$$

3) *Reveal Shot (RS)*. The parameters that must be specified are K, i.e., the time (in seconds) until the target becomes fully visible, and $\tilde{\mathbf{x}}_{KT}$, i.e., a proper UAV position in 3D space from which the target will be fully visible in K seconds. The mathematical description is the following:

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_0 + \frac{t}{KT}(\tilde{\mathbf{x}}_{KT} - \tilde{\mathbf{x}}_0) \tag{99}$$

$$\mathbf{l}_t = \mathbf{p}_t, \quad t \in [0, KT] \tag{100}$$