Batteries, camera, action!
Learning a semantic control space for expressive robot cinematography

Rogerio Bonatti[1,2], Arthur Bucker[3], Sebastian Scherer[2], Mustafa Mukadam[1] and Jessica Hodgins[1]

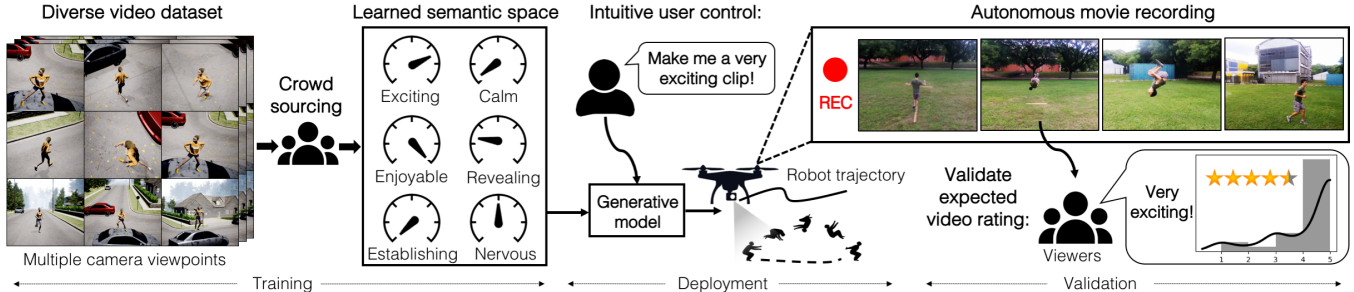[1]Facebook AI Research, [2]Carnegie Mellon University, [3]University of São Paulo

Fig. 1: Our framework uses a diverse set of video clips to learn a semantic descriptor space from crowdsourced ratings. During deployment, a user can intuitively manipulate the robot camera motion with descriptors instead of tuning low-level robot positioning parameters.

Abstract— Aerial vehicles are revolutionizing the way film-makers can capture shots of actors by composing novel aerial and dynamic viewpoints. However, despite great advancements in autonomous flight technology, generating expressive camera behaviors is still a challenge and requires non-technical users to edit a large number of unintuitive control parameters. In this work, we develop a data-driven framework that enables editing of these complex camera positioning parameters in a semantic space (*e.g.* calm, enjoyable, establishing). First, we generate a database of video clips with a diverse range of shots in a photo-realistic simulator, and use hundreds of participants in a crowd-sourcing framework to obtain scores for a set of semantic descriptors for each clip. Next, we analyze correlations between descriptors and build a semantic control space based on cinematography guidelines and human perception studies. Finally, we learn a generative model that can map a set of desired semantic video descriptors into low-level camera trajectory parameters. We evaluate our system by demonstrating that our model successfully generates shots that are rated by participants as having the expected degrees of expression for each descriptor. We also show that our models generalize to different scenes in both simulation and real-world experiments. Data and video found at: **https://sites.google.com/view/robotcam**.

I. INTRODUCTION

Aerial vehicles have become an important tool for supporting human creativity and expressiveness, fundamentally altering the way both professional and amateur users produce media content for movies, sports and virtual / augmented reality. Much of the impact of aerial cameras stems from their capability to compose aerial and dynamic viewpoints that are infeasible using traditional devices such as hand-held cameras and dollies [1]. In addition, recent developments both in industry [2, 3] and academia [4]–[6] now allow drones to detect, track and follow objects of interest autonomously while maintaining safety in cluttered environments.

However, a major limitation of today's cinematography platforms is the difficult and unintuitive interface for camera control. A scene can be captured in myriad ways, and each camera path provides a different visualization of the story [7], causing distinct impressions on the viewers. Users are not able to directly control the emotional expression of the final footage. Instead, they need to carefully tune the camera's position, velocities and angles to achieve a desired expression. An added challenge is the complex interaction between camera parameters, which produces a combinatorial complexity that drives the viewer's perception. This task is cumbersome to users not only due to the large parameter search space, but also because it requires intimate knowledge of cinematography rules [8]–[10], which take years to master.

Our work aims to fill in the gap in existing controls for autonomous cameras. As seen in Fig. 1, our framework enables users to determine the desired shot types based on intuitive and perceptually meaningful parameters (*e.g.* exciting, enjoyable, establishing). We employ a photo-realistic simulator to generate a database of video clips with a diverse set of shot types, and use crowd-sourced perceptual experiments to obtain semantic scores for each video. We then learn the two-way mapping between low-level shot parameters and semantic descriptors. We take inspiration from previous works that created semantic control spaces for domains such as cloth animations [11], object and body shapes [12, 13], and robot motion [14]. Our contributions are threefold:

1) Perceptual experiments: First, we conduct a series of experiments to determine the minimal perceptually valid step sizes for different shot parameters. We then build a dataset of 200 videos using variations along these units, and use an interactive crowd-sourcing platform to obtain numerical scores for different subjective semantic descriptors;

2) Semantic control space: Using the perceptual scores, we learn a regression model that provides an intuitive semantic control space to re-parameterize the desired video expression into low-level shot parameters to guide the aerial camera;

3) Experimental validation: We validate the learned models in a series of experiments in simulation and real-world tests. We show that shots generated from the semantic space are rated by participants as having the expected degrees of expression for each attribute, and that the model generalizes to different actors, activities, and background compositions.

II. RELATED WORK

Autonomous aerial cinematography: We find multiple lines of work on the use of drones for cinematography. For instance, [15] and [16] use optimization methods for navigation on or close to pre-defined trajectories segments. We also find works that use pre-selected high-level cinematographic guidelines such as distances and angles relative to actors [9, 10] as an input for trajectory optimization in unscripted scenes among obstacles [4, 17]. Alternatively, [18] control the camera based on image projection features.

Autonomous artistic reasoning: Autonomous reasoning about how camera movements can provide artistic value to videos has been a topic of interest in multiple contexts. For instance, [19] and [20] train policies to control pan-tilt-zoom cameras in basketball and soccer games by imitating human demonstrations and using video features. For drone cinematography, [21] use reinforcement learning to train a policy that switches between four basic shot types to maximize human-provided rewards. We also find works that aim to imitate different camera motion styles. For example, [22] use trajectory optimization so that a flying camera can imitate the idiosyncrasies of hand-held camera motions. [23] use supervised learning to predict camera actions for different shot types, and [24] build a latent space of shot types out of movie examples, which is coupled with a generative model.

Sentiment analysis: Unlike the previous works which were focused on *imitating* a particular behavior, sentiment analysis focuses on *understanding* how humans perceive and interpret visual content [25]. Most video sentiment models are based on psychology literature and use a combination of the Valence-Arousal-Dominance state model [26] with semantic descriptors grouped by similarities [27]. Several authors explore the relationship between image features and sentiments [28]–[30], and train models for movie recommendation, summarization, and information retrieval [29, 31].

Learning a perceptual control space: Our approach is motivated by previous works that edit processes using semantic or context-specific attributes, as found in the fields of garment simulations [11], 3D shapes [12, 13], and images [32]. Closest to our domain in robotics, [14] developed a data-driven interface for semantic design of expressive robot motion. All these past works share a common approach: they first build a dataset connecting instances to their intuitive semantic labels (usually via crowd-sourcing), and then use this data to learn a generative model mapping the semantic descriptors back to unintuitive low-level parameters.

Our work is the first to develop a perceptual control space for designing camera trajectories. We allow users without any domain expertise in aerial vehicles or cinematography to capture expressive camera behaviors. Unlike previous works

Fig. 2: Experimental setup: a) Photo-realistic simulator with an animated character and aerial vehicle; b) Shot parameters that define the vehicle's position relative to the actor, in spherical coordinates.

that imitate styles, our system does not require the user to search for specific examples of videos they want to indirectly emulate. As we detail next, we show that we can generate any footage by directly selecting its desired semantic descriptors.

III. PERCEPTUAL EXPERIMENTS

Our overarching goal is to learn a perceptually meaningful space for aerial camera control. We focus our study on single-actor shots, and assume the existence of sparse sets of obstacles such as the ones found in suburban environments (*e.g.* trees, telephone poles, traffic signs). All simulated data is recorded using a drone in the photo-realistic environment AirSim [33] (Fig. 2a) coupled with a custom ROS interface [34]. We limit our clips to 15 second segments, which is a reasonable unit of length for individual shots [35, 36]. As seen in the supplementary video, the main scene consists of an animated character running down a street, avoiding a car parked on the road, and finally jumping on top of another vehicle and dancing. We chose this actor motion because it contains both static and dynamic segments, and causes a relatively neutral emotional impression on the viewer.

We employ the base motion planner from [17], which uses a trajectory optimization method to avoid collisions and occlusions with the environment. Within this framework, a shot is parameterized by the positions and velocities of the drone relative to the actor. We define a shot using spherical coordinates (Fig. 2b): $\Omega_{shot} = [\rho, \dot{\rho}, \theta, \dot{\theta}, \phi, V_z]^T \in \mathbb{R}^6$, where ρ and $\dot{\rho}$ are the distance and velocity towards the actor, θ and $\dot{\theta}$ are the horizontal angle and angular velocity, ϕ is the tilt angle and V_z the vehicle's vertical speed.

A. Minimal Perceptual Units for Shot Parameters

Our approach requires us to build a mapping between shot parameters and semantic descriptors. To do so, we must sample the manifold containing all possible shot variations. Within the cinematography literature [9, 10] we find canonical sets of values for parameters like the distance to the actor ρ (close-up, medium, and long shots) and tilt angles (variations every $45°$). However, a naive approach for sampling the remaining parameters results in a prohibitively large state space given that some of its parameters such as angle rates are virtually unbounded. In addition, parameter steps that are too small may render imperceptible changes in the final video, resulting in redundant samples.

To address these issues, we learn a minimally perceptible unit of measure along each dimension in the shot parameter space. A major challenge arises because this is a high-dimensional space, and metrics are not globally consistent
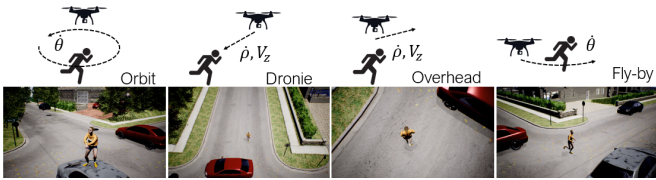
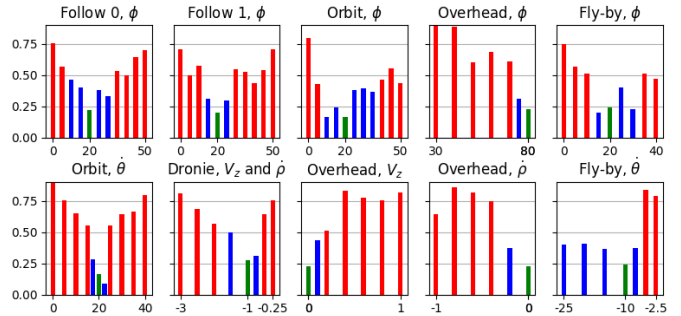Fig. 3: Examples of dynamic shot presets used in the study.

Fig. 4: Minimal perceptual deltas for each parameter. The preset value is shown in green, red indicates statistically significant perceptual variations ($p = 0.05$), and blue are the insignificant variations. Heights display the percentage of videos rated as being different than their preset.

(*e.g.* variations in angular velocity will be more perceptible when the camera is closer to the actor).

In order to make our experiments tractable we adopt the simplifying assumptions of [11], and evaluate local axis-aligned perceptual steps around a set of meaningful shot preset configurations. In Table I we define six shot presets based on common aerial shot types: the static shots (*Follow 0/1*) maintain a constant relative position between camera and actor, and Fig. 3 depicts the dynamic shots.

TABLE I: Shot presets for perceptual units study. We only vary the parameters in bold for each preset.

Unit Preset	ρ	θ	ϕ	$\dot{\rho}$	$\dot{\theta}$	V_z
Unit	[m]	[°]	[°]	[m/s]	[°/s]	[°/s]
Follow 0	8	0	**20**	0	0	0
Follow 1	8	135	**20**	0	0	0
Orbit	5	0	**20**	0	**20**	0
Dronie	25	0	**45**	**-0.5**	0	**-0.5**
Overhead	8	180	**85**	**0**	0	**0**
Fly-by	15	150	**20**	0	**-8**	0

For each parameter ϕ, $\dot{\rho}$, $\dot{\theta}$, V_z, we sample approximately ten linearly distributed variations around its preset value over the maximum range. We only allow changes that keep the shot within its original type. Table I highlights the variable parameters in bold. We create a total of 84 videos, and for each variation we perform a two-sided t-test analyzing if the resulting video is perceived as the same or different from the preset video (with $p = 0.05$ significance).

For this first web survey (WS1) we recruited over 200 participants using Amazon Mechanical Turk (MTurk) [37]. This research received a waiver from our Institutional Review Board, and participants were compensated for their time. After being approved on a short qualifying task, each participant viewed a total of 12 pairs of videos, one being the preset and the other being either a variation or the preset against itself. Videos were played asynchronously three times (only one video played at a time), and after watching at least once, participants answered: "*Is the camera perspective the same or different in the two clips shown?*" Each clip was compared 30 times against its preset.

Figure 4 displays our results showing the minimal perceptual units around each preset and parameter. As expected, we find that not all noticeable variations are symmetrical around the presets, and deltas are only locally consistent. For example, a change larger than $\Delta\phi = 5°$ in tilt is noticeable for *overhead* shots, while a larger delta $\Delta\phi = 10°$ is required for *follow* shots, where the preset angle is lower. Humans are surprisingly good at noticing variations in angular speed: deltas as small as $\Delta\dot{\theta} = 2.5°$ were noticeable.

B. Obtaining Semantic Scores for Videos

Once we are able to generate a diverse dataset of videos with the minimal perceptual units of each shot parameter, the next step is to devise a method to obtain numerical scores for the subjective semantic descriptors representing each clip (our choice of descriptors is detailed in Section III-C). To do so, we use *TrueSkill* [38], which is a relative ranking algorithm based on pairwise comparisons, similar to the Elo chess-player rating algorithm [39]. Pairwise comparisons have been shown to be more consistent and render more accurate results than absolute scales for subjective scores [28, 40]–[42]. We model the score for each semantic descriptor d_i as a Gaussian distribution $d_i = \mathcal{N}(\mu_i, \sigma_i)$, and update both mean and variance of each pair of samples after each comparison.

C. Building a Semantic Descriptor Space via Crowd-Sourcing

We initially compiled a list of 15 semantic descriptors (Table II) which are commonly used to refer to subjective qualities of images and videos in the cinematography and psychology literature [9, 10, 25, 28, 29, 42]. Because our scenes do not include dialog or soundtrack, we do not include descriptors which are uniquely associated with audio features.

A camera control space consisting of 15 descriptors is still not an intuitive interface for non-expert users. Therefore, in order to reduce the cognitive load on the operator, our first semantic study targeted reducing the space to a smaller number of descriptors. To do so, we generated a database of 50 distinct 15-second clips using each shot preset's minimal perceptual units. We randomly sampled parameter variations for all axes simultaneously, and allowed positive and negative deltas using multiples of $\{0, 0.5, 1.0, 1.5, 2.0\}$ from the perceptual units.

To obtain semantic scores for each shot, we designed a second web survey (WS2) on MTurk. This time, each user was shown one pair of clips at a time, which played synchronously three times. After watching the videos at least once, users answered five questions of the type "*Which video is more ___?*", where ___ denotes the different descriptors. Each user watched a total of 12 video pairs, and passed a qualifying task before the survey. We processed answers from a new set of 200 users using *TrueSkill* to obtain a descriptor vector

$d \in \mathbb{R}^{15}$ containing the relative scores of each semantic descriptor for each clip. We analyzed the similarity between all pairs of descriptors using correlation coefficients, as seen in Figure 5.

TABLE II: 7 Clusters of the 15 original descriptors shown within brackets [], and divided by emotion axis and direction. We place no descriptor as representing negative valence, but low values of *interesting* and *enjoyable* can span this emotion.

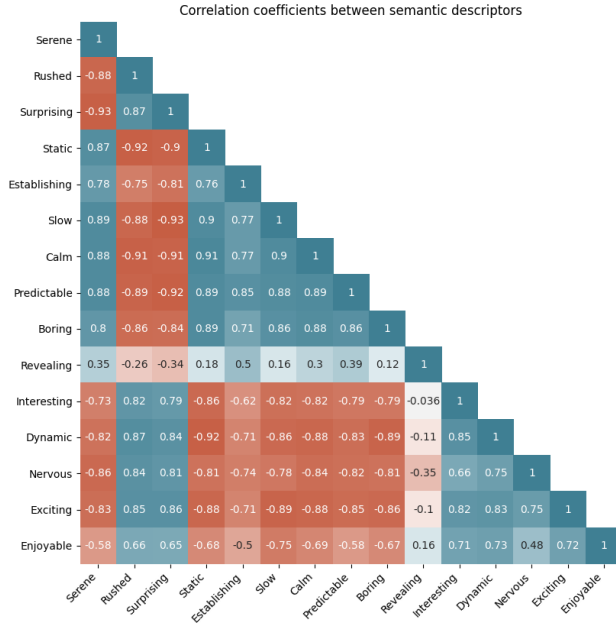| Axis	Arousal	Valence	Dominance
Positive	[**Exciting**, Surprising, Rushed, Dynamic]	[**Interesting**], [**Enjoyable**]	[**Establishing**], [**Revealing**]
Negative	[**Calm**, Slow, Predictable, Boring, Serene, Static]	NA	[**Nervous**]

Fig. 5: Correlations between a set of 15 semantic descriptors within a diverse dataset of 50 videos. Data processed from 200 user surveys.

We found that some groups of descriptors present large positive and negative correlations with one another. For instance, videos labeled as being very *calm* also present high scores for *slow*, and low scores for *dynamic*. We use the Affinity Propagation algorithm [43] to cluster groups of descriptors with high correlation. As seen in Table II, we reduce the dimensionality of the descriptor space by building a total of seven groups of descriptors, with one descriptor representing the group for subsequent studies. The choice of the number of clusters is ambiguous, and any value between 1 and 15 clusters would have been possible by varying the *preference value* in the Affinity Propagation algorithm. Our choice of seven clusters supported two criteria: (i) it presents a small enough number of parameters that a user can interact within an interface, and (ii) the resulting seven clusters can be grouped to span all axes of the Valence-Arousal-Dominance [26] space, which is widely used in affective content analysis [28].

Next, we generated the final dataset for learning the mapping between semantics and shot parameters, now using a larger set of 200 randomly generated videos sampled around the preset values. Analogously to WS2, we deployed a third perceptual rating survey (WS3) to compute a 7-dimensional descriptor vector for each video ($d \in \mathbb{R}^{7}$) using answers from a new set of 500 participants who passed a qualification test. In order to verify our initial assumption that the resulting clusters are able to span the full space of the three emotions axis, we again calculate a correlation matrix between descriptors. This time we also artificially create negated scores by mirroring each descriptor around its mean value, and then use a multidimensional scaling method [44] to find the best 3D coordinates that fit our data, treating correlation scores as distances between points. Figure 6 shows that our experimental data in fact is able to span a 3D affective space, as seen by our fitted Valence-Arousal-Dominance basis vectors. The vectors are not fully orthogonal, indicating correlations among emotion axes, which is an expected result based on the psychology literature [42].
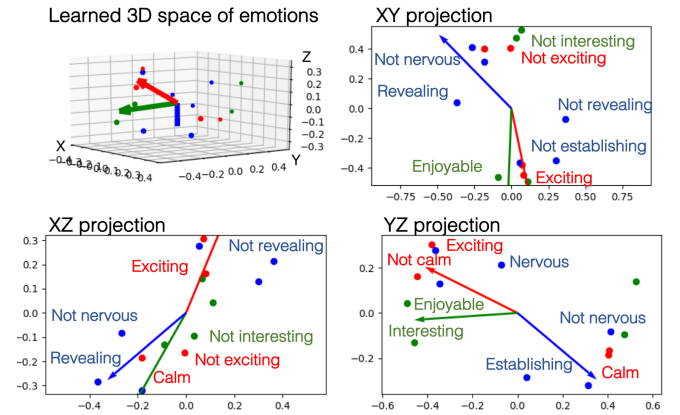

Fig. 6: Learned 3D space of emotions computed from a survey with 200 videos and 500 participants. We display seven semantic descriptors and their negated values (*e.g.* Calm, Not calm), and fit a emotion basis vector along each major emotion axis: Arousal (red), Valence (green) and Dominance (blue).

IV. LEARNING A SEMANTIC CONTROL SPACE

Given the dataset of shot parameters $\Omega_{shot} \in \mathbb{R}^{6}$ and semantic descriptors $d \in \mathbb{R}^{7}$, we wish to learn a function $f : d \to \Omega_{shot}$ that provides a mapping from a set of desired descriptors to the set of shot parameters that produces a clip with such emotional impression on the viewer (D2P model). We are also interested in the inverse question: given a shot, can we infer its emotional expression as a mapping $f^{-1} : \Omega_{shot} \to d$ (P2D model)?

A. Training Details

We explored two approaches to learn such functions: linear regression (LR) and deep neural networks (DNN). We employ Lasso regression [45] for our linear model, as it includes an additional loss to reduce the $L1$ norm of coefficients. We tested different DNN architectures, and after a series of tests using 5-fold cross validation on our dataset, we settled with a fully connected network with 3 hidden layers containing 32, 16, and 8 neurons respectively. We augment the shot parameter vector with 5 additional variables corresponding to a one-hot encoding of the shot type (follow, orbit, dronie,

overhead, fly-by), which is used to process the model's output into one of the canonical drone shot types. All values are normalized to the $[-1, 1]$ range for training.

B. Model Results

Both LR and DNN presented very similar performance on the D2P model using a dataset of 200 videos, with $R^2_{LR} = 0.19$ and $R^2_{DNN} = 0.22$. Given this similarity and the fact that LR allows us to interpret the values of its coefficients more easily, the rest of our analysis here uses linear models. We note, however, that we trained both models with a relatively small quantity of examples due to the high cost of obtaining labeled data. With more data, a DNN model would very likely achieve superior performance.

Figure 7a shows the normalized coefficients for the D2P model. We notice intuitive relationships between parameters which were learned entirely from user evaluations. For instance, *establishing* videos are linked to larger distances ρ to the actor, matching the cinematographic concept of establishing shots. Other strong relationships we identify are of *calm* videos having smaller angular velocities $\dot{\theta}$, and enjoyable clips tending to veer towards lower tilt angles ϕ, away from overhead shots.

We also investigate the P2D i.e. inverse model, which maps shot parameters to descriptor values. Figure 7b shows the relationships between variables. The distance between camera and actor causes the largest impact on viewer perception. Other shot parameters have a selective influence on descriptors. For instance, larger tilt angles ϕ produce less revealing and less interesting videos, but have less influence on how calm a video is.

V. EXPERIMENTAL VALIDATION

A. Semantic Control Space

We validated the quality of the learned semantic control space in a series of user experiments. Using the D2P model, we generated a set of videos per semantic descriptor by linearly sampling increasingly low to high desired expression values over the range of $[-2, 2]$ standard deviations. The choice of values for the descriptor vector is not entirely arbitrary, as several of the attributes present positive and negative correlations. For instance, if the user is only interested about the value of a single descriptor and sets a high score for *exciting*, the *calm* score must naturally go down. Therefore, we treat the descriptor vector d as a multivariate normally distributed variable $d = [d_1, d_2]^T$, where $d_2 \in \mathbb{R}^n$ are the n user-specified scores. We then calculate the remaining scores d_1 using the means and covariances between groups of variables: $d_1 = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(d_2 - \mu_2)$.

In contrast to our previous experiments, here we created a new survey on MTurk (WS4) where users viewed each clip individually, and scored them using a 5-point Likert scale (1 being *Not ___ at all*, and 5 being *Extremely ___*). We averaged scores from 15 users per clip. An absolute scale is preferred for this study because our goal is to understand how average users perceive the resulting clips, and not just to create a relative video ranking.
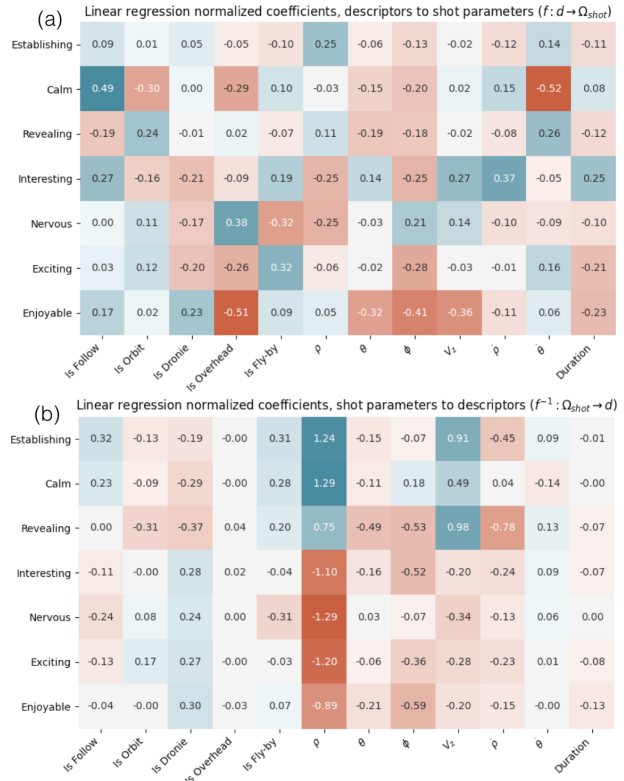

Fig. 7: Normalized linear model coefficients: a) mapping from semantic descriptors to shot parameters (D2P model), and b) mapping from shot parameters to semantic descriptors (P2D model).
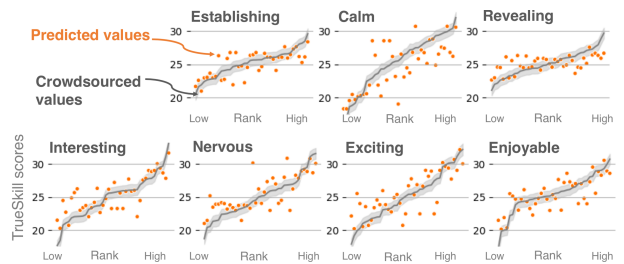

Fig. 8: Comparison of semantic scores predicted from shot parameters using the learned model (orange) against the original crowdsourced values (gray). Shaded gray area displays TrueSkill score uncertainty. Videos are ordered in ascending order of scores.

As seen in Figure 9, we test our model in two simulation environments using 42 videos (6 per descriptor axis). The first environment is identical to the one employed for learning the models, and the second one tests how well the D2P model generalizes to a new actor type (soccer player dribbling a ball), scene background (grass field with an audience cheering), and shot duration (20s, against 15s earlier).

In addition, we also validate our model in real-world experiments, collecting a total of 28 clips (4 per descriptor axis). We use a Parrot Bebop 2 drone and control it with the real-time autonomy stack developed by [4] to visually localize the actor and plan smooth aircraft motions. We collected additional shots (Fig. 11) shown in the supplementary video.

Figure 9 shows that our D2P model is able to successfully generate shots that are rated by participants as having the expected degrees of expression for each descriptor.
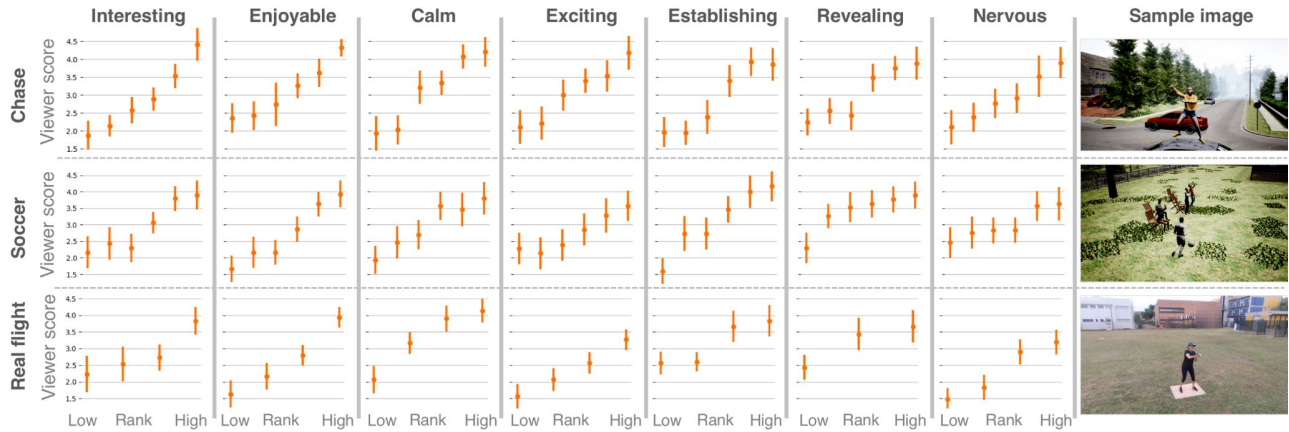
Fig. 9: User study results validating generative shot model for different simulated and real-world scenes. For each descriptor we generated videos ranging from low to high desired scores, and averaged the perception score from 15 users on a 5-point absolute Likert scale. Perceived scores match the desired ascending ranking.
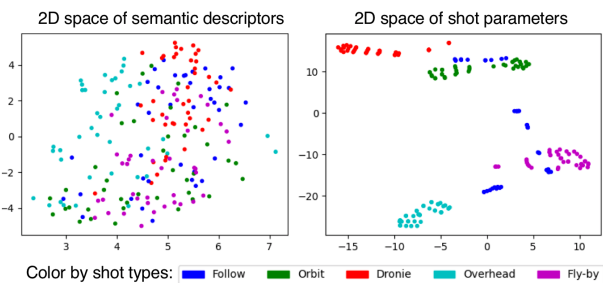

Fig. 10: Reduction of semantic descriptor space and shot parameter space into 2D compressions using T-SNE. Colors indicate shot types.
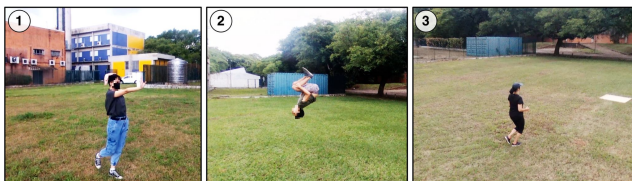

Fig. 11: Examples of shots taken with our autonomous drone setup: 1) Dance moves; 2) Parkour; 3) Running scenes. The full sequences are shown in the supplementary video.

Furthermore, the model generalizes well to other simulated scenes and to real-world footages, which strongly suggests that our semantic control space is not overly attached to specific features of the training environment or to a single set of actor motions.

B. Semantic Shot Classification and Latent Space Analysis

We also validate the quality of the P2D model on a held-out test set with 20% of the 200 videos. Fig. 8 shows the quality of the model's predictions per emotion ($R^2 = 0.69$).

Figure 10 provides further insights into the mappings between shot parameters and descriptors. We use T-SNE [46] to visualize a 2D projection of higher-dimensional spaces, and color the data by shot types. Based on the 2D visualization and on the different R^2 score values, we argue that learning a model that maps shot parameters to descriptors is an easier task, as it requires learning a surjective function (many-to-one). The inverse mapping is harder since one point in the descriptor space can correspond to multiple set of shot parameters (one-to-many).

VI. CONCLUSION AND DISCUSSION

In this paper, we present a semantic control space for aerial cinematography that allows users to guide camera motions using intuitive controls instead of selecting low-level shot parameters. We use crowdsourcing with hundreds of users to learn the mapping between shot parameters and semantic descriptors in a diverse dataset of 200 videos. We validate our semantic generative shot model in multiple experiments, and show that it successfully generalizes to new scenes in simulation and in real-world experiments with an aerial robot. Additionally, we provide insights into the space of emotions our model can represent, and investigate the relationship between shot parameters and descriptors to understand what our model is effectively learning.

Our framework targets non-technical users, and can generate shot parameters directly from a semantic vector. However, expert users can easily adapt it to gain more control over the model's outcome. For example, one can learn separate generative models for individual shot types and gain more control over the system's inputs / outputs.

We find multiple directions for future work in the area. For instance, we are interested in employing a larger set of parameters to control the shots, such as lens zoom [47], and potentially even soundtracks. In addition, we would like to extend our framework to accept environment features into the generative model. Another limitation we face is regarding the shot time horizon: currently our algorithm generates a single shot at a time. However, there is great potential in developing algorithms that can reason over longer durations, and infer emotional expression over sequences of shots.

REFERENCES

[1] V. Santamarina-Campos and M. Segarra-Oña, "Introduction to drones and technology applied to the creative industry. AiRT project: An overview of the main results and actions," in *Drones and the Creative Industry*. Springer, 2018, pp. 1–17.

[2] DJI. (2018) Dji mavic, https://www.dji.com/mavic. [Online]. Available: https://www.dji.com/mavic

[3] Skydio. (2018) Skydio r1 self-flying camera, https://www.skydio.com/technology/. [Online]. Available: https://www.skydio.com/technology/

[4] R. Bonatti, W. Wang, C. Ho, A. Ahuja, M. Gschwindt, E. Camci, E. Kayacan, S. Choudhury, and S. Scherer, "Autonomous aerial cinematography in unstructured environments with learned artistic decision-making," *Journal of Field Robotics*, 2020.

[5] B. F. Jeon, D. Shim, and H. J. Kim, "Detection-aware trajectory generation for a drone cinematographer," *arXiv preprint arXiv:2009.01565*, 2020.

[6] C. Huang, F. Gao, J. Pan, Z. Yang, W. Qiu, P. Chen, X. Yang, S. Shen, and K.-T. T. Cheng, "Act: An autonomous drone cinematography system for action scenes," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7039–7046.

[7] N. T. Proferes, *Film Directing Fundamentals: see your film before shooting*. Taylor & Francis, 2017.

[8] R. H. Stanley, *The movie idiom: film as a popular art form*. Waveland Press, Inc., 2011.

[9] C. J. Bowen and R. Thompson, *Grammar of the Shot*. Taylor & Francis, 2013.

[10] D. Arijon, "Grammar of the film language," 1976.

[11] L. Sigal, M. Mahler, S. Diaz, K. McIntosh, E. Carter, T. Richards, and J. Hodgins, "A perceptual control space for garment simulation," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, 2015.

[12] M. E. Yumer, S. Chaudhuri, J. K. Hodgins, and L. B. Kara, "Semantic shape editing using deformation handles," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, 2015.

[13] S. Streuber, M. A. Quiros-Ramirez, M. Q. Hill, C. A. Hahn, S. Zuffi, A. O'Toole, and M. J. Black, "Body talk: Crowdshaping realistic 3d avatars with words," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, 2016.

[14] R. Desai, F. Anderson, J. Matejka, S. Coros, J. McCann, G. Fitzmaurice, and T. Grossman, "Geppetto: Enabling semantic design of expressive robot behaviors," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

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

[16] 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 (TOG)*, vol. 36, no. 4, p. 132, 2017.

[17] R. Bonatti, C. Ho, W. Wang, S. Choudhury, and S. Scherer, "Towards a robust aerial cinematography platform: Localizing and tracking moving targets in unstructured environments," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[18] C. Huang, Z. Yang, Y. Kong, P. Chen, X. Yang, and K.-T. T. Cheng, "Through-the-lens drone filming," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4692–4699.

[19] J. Chen and P. Carr, "Mimicking human camera operators," in *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 215–222.

[20] J. Chen and J. J. Little, "Where should cameras look at soccer games: Improving smoothness using the overlapped hidden markov model," *Computer Vision and Image Understanding*, vol. 159, pp. 59–73, 2017.

[21] M. Gschwindt, E. Camci, R. Bonatti, W. Wang, and S. Scherer, "Can a robot become a movie director? learning artistic principles for aerial cinematography," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[22] A. Ashtari, S. Stevšić, T. Nägeli, J.-C. Bazin, and O. Hilliges, "Capturing subjective first-person view shots with drones for automated cinematography," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 5, pp. 1–14, 2020.

[23] C. Huang, Y. Dang, P. Chen, X. Yang *et al.*, "One-shot imitation filming of human motion videos," *arXiv preprint arXiv:1912.10609*, 2019.

[24] H. Jiang, B. Wang, X. Wang, M. Christie, and B. Chen, "Example-driven virtual cinematography by learning camera behaviors," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 45–1, 2020.

[25] E. Cambria and A. Hussain, "Sentic computing," *Cognitive Computation*, vol. 7, no. 2, pp. 183–185, 2015.

[26] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[27] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[28] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[29] L. Canini, S. Benini, and R. Leonardi, "Affective analysis on patterns of shot types in movies," in *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2011, pp. 253–258.

[30] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92.

[31] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE transactions on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[32] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, 2014.

[33] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics*. Springer, 2018, pp. 621–635. [Online]. Available: https://arxiv.org/abs/1705.05065

[34] ROS. (2018) Robot operating system (ros). [Online]. Available: http://www.ros.org/

[35] B. Mahadani and B. MahadaniI, "How film shot length decrease in last century," Feb 2015. [Online]. Available: http://bhushanmahadani.com/film-shot-length-decreased-in-last-century/

[36] J. E. Cutting, J. E. DeLong, and C. E. Nothelfer, "Attention and the evolution of Hollywood film," *Psychological science*, vol. 21, no. 3, pp. 432–439, 2010.

[37] Amazon. Amazon mechanical turk. [Online]. Available: https://www.mturk.com

[38] R. Herbrich, T. Minka, and T. Graepel, "Trueskill™: a bayesian skill rating system," in *Advances in neural information processing systems*, 2007, pp. 569–576.

[39] A. E. Elo, *The rating of chessplayers, past and present*. Arco Pub., 1978.

[40] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[41] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 4, pp. 762–774, 2010.

[42] P. A. Russell and C. D. Gray, "Ranking or rating? some data and their implications for the measurement of evaluative response," *British journal of Psychology*, vol. 85, no. 1, pp. 79–92, 1994.

[43] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[44] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[45] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[46] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[47] P. Pueyo, E. Cristofalo, E. Montijano, and M. Schwager, "Cinemairsim: A camera-realistic robotics simulator for cinematographic purposes," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 1186–1191.