Feedback from Pixels: *Output Regulation via Learning-Based Scene View Synthesis*

Murad Abu-Khalaf MURAD@CSAIL.MIT.EDU
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA

Sertac Karaman SERTAC@MIT.EDU
MIT Laboratory for Information and Decision Systems
Cambridge, MA 02139, USA

Daniela Rus RUS@CSAIL.MIT.EDU
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA

Editors: A. Jadbabaie, J. Lygeros, G. J. Pappas, P. A. Parrilo, B. Recht, C. J. Tomlin, M. N. Zeilinger

Abstract

We propose a novel controller synthesis involving feedback from pixels, whereby the measurement is a high dimensional signal representing a pixelated image with Red-Green-Blue (RGB) values. The approach neither requires feature extraction, nor object detection, nor visual correspondence. The control policy does not involve the estimation of states or similar latent representations. Instead, tracking is achieved directly in image space, with a model of the reference signal embedded as required by the internal model principle. The reference signal is generated by a neural network with learning-based scene view synthesis capabilities. Our approach does not require an end-to-end learning of a pixel-to-action control policy. The approach is applied to a motion control problem, namely the longitudinal dynamics of a car-following problem. We show how this approach lend itself to a tractable stability analysis with associated bounds critical to establishing trustworthiness and interpretability of the closed-loop dynamics.

Keywords: Pixels, Feedback Control, View Synthesis, Visual Servoing, Car-Following, Stability

1. Introduction

Our aim is to investigate the integration of visual signals into feedback loops for the purpose of controller synthesis and analysis, and without requiring a perception module in the loop. We treat the camera as a high-dimensional sensor and propose a principled approach grounded in mathematical control theory to investigate stability and associated theoretical limitations of the closed-loop performance.

In this paper, we consider output regulation class of problems where the output measurement includes a pixelated image. We feel the contribution of this paper is as follows:

- We treat each RGB pixel as a measurement and do not attempt to grayscale or threshold the image and can handle an arbitrary image size or resolution.
- Compared to visual servoing approaches, our work does not involve hand-crafted geometrical feature extractions, correspondence or matching, pose estimation, or an interaction matrix.

- Unlike most existing approaches, we integrate vision into reactive low-level control without a need for a perception module, end-to-end imitation learning, the estimation of states or similar latent representations.
- Our approach works for moving targets and non-stationary environments.
- Embedded in our controller is an internal model of the tracked visual reference. This is achieved by incorporating a view synthesizer in the loop at inference or execution time.
- We show a systematic way to synthesize static output feedback controllers, such as a proportional controller, via necessary and sufficient conditions in the literature.
- Our approach does not require discretizing the action space or the state space, and works in continuous-time synthesis and analysis.
- Our work is amenable to stability analysis.
- In the car-following example, our approach maintains physically interpretable representations of the underlying dynamics, *e.g.* state-space variables from first principles.

In Section 1.1, we provide a context to our contribution by reviewing related work. Section 1.2 covers notational remarks. Section 2 introduces the problem statement concisely in the context of an application domain, while Section 3 presents the main result. In Section 4 we provide conclusions and future directions. Appendix B shows simulations using CARLA from Dosovitskiy et al. (2017).

1.1. Related Work

Several recent results for vision-in-the-loop control attempt to leverage learning-based approaches via end-to-end learning, mainly imitation learning, to essentially map pixels to actions via a static map as in Bojarski et al. (2016) and Amini et al. (2018) in the context of driving. Another body of work attempts to first get a latent representation of the underlying dynamics of the process from visual input as in Watter et al. (2015), Banijamali et al. (2018), Hafner et al. (2019) and structured latent representations as in Johnson et al. (2016). In Zhang et al. (2019), such latent representations are used in model-based reinforcement learning in the context of manipulation.

In Collewet and Marchand (2011), geometric feature extraction or matching was alleviated by using the luminance of all pixels in 2D direct visual servoing. However, such methods require computing explicitly an interaction matrix and solving a nonlinear optimization problem resulting in a small region of convergence. Therefore, in Saxena et al. (2017) and Bateux et al. (2018), the relative pose error is learned from a current and reference images for the purpose of posed-based visual servoing. While these methods alleviate the need for camera parameters and scene geometry, servoing is done towards a non-moving target.

In Amini et al. (2020), a data-driven simulator is used to train a policy via reinforcement learning from an initial stable policy provided by a human driver. The simulator generates perturbations along an initial policy by taking a 2D image captured along the initial trajectory, creating a depth map and a 3D point cloud, applying a desired viewpoint transformation on the 3D data, then synthesizing a novel 2D view of the scene. The view synthesizer is not deployed at inference time; only the learned policy is.

A different body of work leverages video prediction in the form of visual foresight and scene view synthesis Hirose et al. (2018) and Hirose et al. (2019a) along with model predictive control as in Hirose et al. (2019b) in the context of robot navigation.

Closed-loop stability is emphasized in Nagai and Sakai (2013) in the context of sloshing dynamics, where no geometric feature extraction is done. Instead, a single-input multi-output system

identification is used to approximate and map linearly the input to a matrix representing a reduced grayscale image of the liquid surface. The linear time-invariant (LTI) system is then converted to a port-Hamiltonian system where a passivity-based controller is applied. To reduce computational intensity, Sakai and Ando (2014) applies model reduction on the matrix space to reduce output size then performs LQG control, while Sakai and Sato (2014) uses feature extraction to map the liquid surface to polynomial space.

Another recent approach by Dean et al. (2020a) proposes to learn a perception map from high-dimensional data, the image, to a low dimensional latent representation as a state or partial state observation. Robust control is applied on the low dimensional latent representation, resulting in a dynamic output feedback controller and stability is shown under specific conditions, and extended by Dean and Recht (2020) and Dean et al. (2020b) to show safety.

In Suh and Tedrake (2020), Lyapunov stability to a target set is shown for an approach based on image visual foresight using linear models to solve a quasi-static pile manipulation problem. The state represents a grayscale image of the pile and image-to-image transitions are learned via switched-linear models. The action space is discrete and switching among actions corresponds to switching among linear models.

1.2. Notation

R denotes the real line. Given multidimensional array $Y \in \mathsf{R}^{p\times q\times r}$, $vec(\cdot)$ orderly stacks the $q \times r$ columns of Y one slice at a time until r. An all one-entries $n\times m$ matrix is denoted by $\mathbf{1}_{n\times m}$, and by $\mathbf{1}$ when the size is context-dependent. A continuous vector function f of dimension m that is a function of an n dimensional vector is represented by $\mathcal{C}(\mathsf{R}^{n\times 1}, \mathsf{R}^{m\times 1}) = \{f : \mathsf{R}^{n\times 1} \to \mathsf{R}^{m\times 1} | f \in \mathcal{C}\}$. $I_{Cam} \in \mathsf{R}^{W\times H\times C}$ denotes an RGB image from a camera, and $I_{Syn} \in \mathsf{R}^{W\times H\times C}$ is an RGB image from a synthesizer, of width, height and channel sizes denoted by W, H and C respectively.

2. Problem Formulation — Car-Following

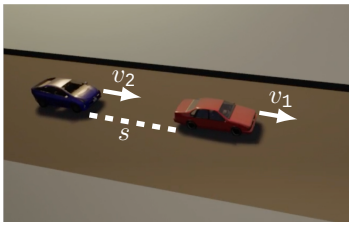

Figure 1: Car-following.

$v_1(t) \in \mathsf{R}$	leader speed,	error signals:
$v_2(t) \in \mathsf{R}$	follower speed,	$x_1(t) = \tilde{v}_1(t) = \bar{v}(t) - v_1(t)$,
$f_2(t) \in \mathsf{R}$	follower force,	$x_3(t) = \tilde{v}_2(t) = \bar{v}(t) - v_2(t)$,
$s(t) \in \mathsf{R}$	spacing,	$x_2(t) = \tilde{s}(t) = \bar{s} - s(t)$,
$\bar{v}$	leader desired speed,	$u(t) = \tilde{f}_2(t) = \bar{f}_2(t) - f_2(t)$,
$\bar{s}$	desired spacing,	$m_1, m_2 > 0$ mass of vehicles,
$\bar{f}_2$	steady-state force.	$\alpha_1, \alpha_2 > 0$ drag coefficients.

We formulate the problem in the context of a concrete example from the application domain of autonomous driving, namely car-following as depicted in Figure 1. In this case, the objective is for the autonomous blue car to follow a leading red car by matching its speed and keeping a desired longitudinal inter-vehicle spacing. The error dynamics can be written as follows:

$$\dot{x}_1(t) = -\tfrac{\alpha_1}{m_1} x_1(t), \quad (1a) \qquad \dot{x}_2(t) = x_1(t) - x_3(t), \quad (1b)$$

$$\dot{x}_3(t) = -\tfrac{\alpha_2}{m_2} x_3(t) + \tfrac{1}{m_2} u, \quad (1c) \qquad y(t) = I_{Cam}(\bar{s} - x_2, \Theta, \Omega), \quad (1d)$$

$$e(t) = \bar{y} - y = I_{Cam}(\bar{s}, \Theta, \Omega) - I_{Cam}(\bar{s} - x_2, \Theta, \Omega). \quad (1e)$$

Equations (1a) to (1c) follow from Levine and Athans (1966). Equation (1d) is a measurement model where $I_{Cam}(s, \Theta, \Omega)$ represents an image captured by a front-facing camera attached to the follower, and where Θ represents specific parameters of the leader, while Ω represents specific parameters of the driving environment background. Moreover, $I_{Cam}(\bar{s}, \Theta, \Omega)$ in (1e) is a reference image for the same Θ and Ω had the spacing been the desired spacing $\bar{s}$. In some sense $I_{Cam}(\bar{s}, \Theta, \Omega)$ can be thought of as imagined instead of measured unlike $I_{Cam}(s, \Theta, \Omega)$ which is measured. This builds on neuroscientific concepts of *analysis-by-synthesis* where it is believed that mental imagery plays a role in human vision Yildirim et al. (2020). We show in Section 3.2 how to obtain $I_{Cam}(\bar{s}, \Theta, \Omega)$.

Assumption 1 Background Invariance*: Assume that*

$$e(t) = I_{Cam}(\bar{s}, \Theta, \Omega) - I_{Cam}(\bar{s} - x_2, \Theta, \Omega) = H(x_2, \bar{s}, \Theta). \tag{2}$$

This says that $e(t)$ is invariant to background changes Ω.

Assumption 2 Null Space*: For $H(x_2, \bar{s}, \Theta)$ in (2), assume that $H(x_2, \bar{s}, \Theta) = \mathbf{0} \iff x_2 = 0$. Then it follows that for a given $\bar{s}, \Theta$*

$$ker(e(x)) = \{x \in \mathsf{R}^{n\times 1} : x_2 = 0\}. \tag{3}$$

Assumption 3 Error Direction*: Let $h(x_2, \bar{s}, \Theta) = vec(H(x_2, \bar{s}, \Theta))$. Without loss of generality,*

$$\bar{s} - s \geq 0 \iff \mathbf{1}^| \cdot h(\bar{s} - s, \bar{s}, \Theta) \geq 0. \tag{4}$$

Assumption 4 Monotonic*: For a given $\bar{s}$ and Θ, consider $h(x_2, \bar{s}, \Theta)$ in (4). If $\beta \geq \alpha \geq 0$ or $-\beta \geq -\alpha \geq 0$, then*

$$h(\beta, \bar{s}, \Theta)^| h(\beta, \bar{s}, \Theta) \geq h(\alpha, \bar{s}, \Theta)^| h(\alpha, \bar{s}, \Theta). \tag{5}$$

Assumption 5 Locally Quadratic*: For a given $\bar{s}$ and Θ, consider $h(x_2, \bar{s}, \Theta)$ in (4). We assume that over a local domain $D \subset \mathsf{R}^{n\times 1}$, where $x = \mathbf{0} \in D$, that*

$$h(x_2, \bar{s}, \Theta)^| h(x_2, \bar{s}, \Theta) \approx c^2(\bar{s}, \Theta) {x_2}^2, \tag{6}$$

for some nonzero *constant $c(\bar{s}, \Theta) \in \mathsf{R}$.*

Definition 1 Uniformly Ultimately Bounded (UUB) *Khalil (2002): A solution of $\dot{x}(t) = f(t, x)$ is said to be UUB with an ultimate bound of ϵ if $\exists \epsilon > 0, \Delta > 0$ such that $\forall \delta \in (0, \Delta)$, $\exists T(\delta, \epsilon) \geq 0$:*

$$\|x(t_0)\|_2 \leq \delta \implies \|x(t)\|_2 \leq \epsilon, \forall t \geq t_0 + T(\delta, \epsilon).$$

Problem 1 Output Regulation Solvability*: Consider the car-following dynamics (1). Determine the existence of a static policy*

$$u = F(y, \bar{y}), \tag{7}$$

such that the regulated output $vec(e(t))$ is asymptotically stable with x_1, x_2, and x_3 bounded.

Note that the static control policy (7) does not require knowledge of $\bar{v}$.

Problem 2 Learning-Based Output Regulation*: Find a static policy (8) for the dynamics (1), where $\hat{\bar{y}}$ is learned to approximate $\bar{y}$, such that $vec(\hat{e}(t)) = vec(\hat{\bar{y}} - y)$, x_1, x_2, and x_3 are UUBs:*

$$u = F(y, \hat{\bar{y}}). \tag{8}$$

3. Main Result

In Section 3.1 we show the existence of solutions to Problem 1 by casting the problem as a static output feedback problem. In Section 3.2, we show the architecture of a view synthesizer that will be used to provide reference images needed to compute the tracking error and regulated output (1e). In Section 3.3, we show a block diagram of the proposed controller, and discuss how to treat the RGB values so that generality is not lost as stated in Assumption 3. Later in Section 3.4, we show closed-loop stability with the camera and the reference view synthesizer in the loop.

3.1. Existence of Solutions

To address Problem 1, we first note that (7) is a static policy. One direction to follow is therefore to reduce Problem 1 into the following problem.

Problem 3 Static Output Feedback*: Consider the car-following dynamics* (1)*. Determine the existence of a static policy* (9) *such that* x_1*,* x_2*, and* x_3 *are asymptotically stable:*

$$u(t) = F(e(t)). \tag{9}$$

Problem 3 is a state-regulation problem. The next theorem shows the existence of a solution to this state-regulation Problem 3, and thus to the output regulation Problem 1.

Lemma 1 *For a fixed* $\bar{s}$*,* Θ*, consider writing* (1a)*,* (1b) *and* (1c) *in the form* $\dot{x} = f(x) + g(x)u(x)$ *and* $h(x) = h(x_2, \bar{s}, \Theta)$*. There exists* $V(x) = x^\intercal P x$ *with* $P = P^\intercal \geq \mathbf{0}$ *and* $G(x) \in \mathcal{C}(\mathsf{R}^{n\times 1}, \mathsf{R}^{m\times 1})$ *such that over a domain* $D \subset \mathsf{R}^{n\times 1}$*, where* $x = \mathbf{0} \in D$*:*

$$0 = \frac{dV(x)}{dx}^\intercal f(x) - \frac{1}{4}\frac{dV(x)}{dx}^\intercal g(x)g^\intercal(x)\frac{dV(x)}{dx} + h^\intercal(x)h(x) + G^\intercal(x)G(x), \tag{10a}$$

$$0 = \frac{dV(x)}{dx}^\intercal f(x), \qquad \forall x \in ker(h(x)). \tag{10b}$$

Theorem 1 *A static output feedback policy* (9) *exists that solves Problem 3, and thus Problem 1.*

Proof First, if Problem 3 has a solution, this implies that Problem 1 is solvable because $u(t) = F(y, \bar{y}) = F(e(t))$ and $\lim_{t\to\infty} x_2(t) = 0 \implies \lim_{t\to\infty} H(x_2(t), \bar{s}, \Theta) = \mathbf{0}$ by Assumption 2 and local continuity from Assumption 5. From Lemma 1 there exists a positive semi-definite solution to (10) over a domain $D \subset \mathsf{R}^{n\times 1}$. It follows from Astolfi and Colaneri (2002) and Astolfi and Colaneri (2001) that there exists a stabilizing state-feedback policy

$$u(x) = G(x) - \frac{1}{2}g^\intercal(x)\frac{dV(x)}{dx}, \tag{11}$$

and using the rank theorem, (11) can be written as a static output feedback policy $u(h(x)) = F(e(t))$ over a region around the equilibrium point. Thus Problem 3 has a solution. ■

3.2. Reference View Synthesis

We show how to synthesize an imagined reference image $\bar{y} = I_{Cam}(\bar{s}, \Theta, \Omega)$ that places the leading car at the desired inter-vehicle spacing $\bar{s}$ as would be viewed by the following car. In doing so, $\bar{y}$ needs to ideally satisfy Assumption 1. To do so, we consider an approach based on appearance flow Zhou et al. (2016) which has been proposed in the context of 3D view transformation Tatarchenko et al. (2016). However, our objective herein is not to transform the entire view, but rather to generate

a view that corresponds to moving an object in the scene closer to or farther away from the observer through a frozen background. Moreover, unlike other work subsequent to Zhou et al. (2016), namely Park et al. (2017), we do not worry about occlusion issues that are more relevant in the rotation of 3D objects and the need to inpaint the hidden sides of the object by hallucinating a view completion.

Our reference view synthesizer is shown in Figure 2 which takes as input a raw camera RGB image and the desired inter-vehicle spacing, and generates as output a view placing the leading vehicle at the desired spacing away from the following vehicle. The raw camera image is an input to an autoencoder that is trained to generate an appearance flow as its output based on $\bar{s}$, where this appearance flow determines which pixels from the camera raw image to copy from as opposed to generating pixels from scratch. The generated appearance flow and the raw camera image are both fed to a bilinear sampler and the output is an RGB image representing the synthesized view. Note that the camera raw image is an input to *both* the autoencoder, and the bilinear sampler. The bilinear sampler is differentiable for backpropagation purposes as shown in Jaderberg et al. (2015).

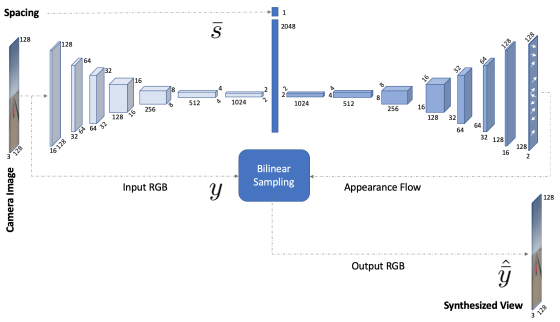

Figure 2: Reference View Synthesizer.

The encoder is constructed from 8 convolutional neural networks (CNNs) each followed by a rectified linear unit (RELU) and with the last layer flattened. All have a stride of 2, padding of 1 and kernel of 4 except for the first layer which has a kernel size of 3, stride of 1 and padding of 1.

The decoder is constructed from 7 convolutional transpose neural networks with stride 2, padding 1 and kernel of 4, each followed by a RELU and a CNN with kernel 3, stride 1 and padding 1 followed by a tangent hyperbolic function. The last layer clearly outputs values between -1 and 1, representing the appearance flow. The input to the decoder is the flattened output of the encoder in addition to the desired spacing $\bar{s}$.

The bilinear sampler takes as input the raw camera RGB tensor and the appearance flow-field tensor which acts on an identity sampling grid to form a modified sampling grid. The modified sampling grid determines, for each output pixel, the location of the input pixels to copy from. Almost all background pixels are copied from their original locations as is to ensure background invariance, while the pixels representing the current location of the leading car and the desired location are impacted. The reference image can therefore be represented as

$$\hat{\hat{y}} = I_{Syn}(\bar{s}, I_{Cam}(\bar{s} - x_2, \Theta, \Omega)), \tag{12}$$

which will be assumed to satisfy Assumptions 1 to 5, and the following assumption.

Assumption 6 View Synthesis Error*: For a given* Θ *and* $\bar{s}$*,* $\exists \epsilon_1 > 0$ *such that*

$$\|vec(\underbrace{I_{Cam}(\bar{s}, \Theta, \Omega)}_{y} - \underbrace{I_{Syn}(\bar{s}, I_{Cam}(s, \Theta, \Omega))}_{\hat{\hat{y}}})\|_2 \leq \epsilon_1. \tag{13}$$

3.3. Block Diagram of the Feedback Loop

We start by showing a block diagram of the proposed controller with a camera and a reference view synthesizer in the loop as shown in Figure 3.

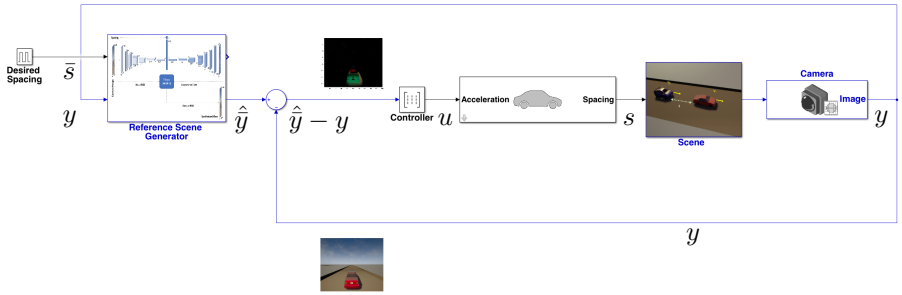

Figure 3: Block Diagram of Feedback Loop.

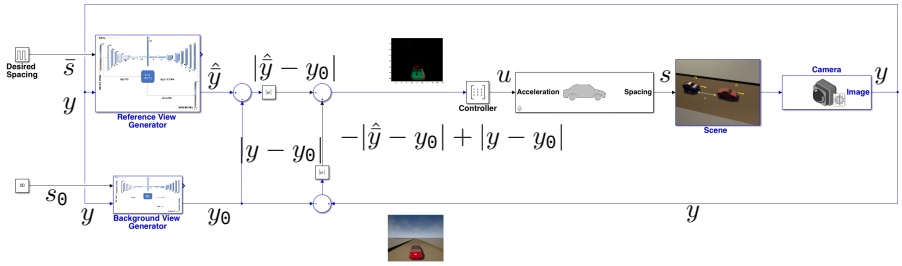

Figure 4: Block Diagram of a Generalized Feedback Loop.

As straightforward as this may seem, the block diagram of Figure 3 may loose the generality Assumption 3 states. To see this, consider a case where $\bar{s} - s \geq 0$ and the following two 3-by-3 pixel images where we show a single color channel only, *e.g.* Green:

$$\bar{z} = \begin{bmatrix} B & O & B \\ B & B & B \\ B & B & B \end{bmatrix}, \qquad (14a) \qquad z = \begin{bmatrix} B & B & B \\ B & B & B \\ O & O & O \end{bmatrix}. \qquad (14b)$$

The reference image $\bar{z}$ has 1 pixel in the first row denoted by the letter O representing the color of an object traversing a background of color denoted by B. The image z has 3 pixels in the last row representing the same observed object at a closer distance to the observer thus occupying more pixels. From (4), we get

$$\mathbf{1}^{|} \cdot vec(\bar{z} - z) = (O - B) + 3(B - O) = 2(B - O). \qquad (15)$$

If the object is black moving in a green background, then we have $O = 0$ and $B = 1$ and thus $\mathbf{1}^{|} \cdot vec(\bar{z} - z) \geq 0$, otherwise if the object is green and moving through a black background, then $O = 1$ and $B = 0$ and thus $\mathbf{1}^{|} \cdot vec(\bar{z} - z) \leq 0$.

To enforce the generality of Assumption 3, we need an expression that is invariant to the polarity of $(B - O)$, in other words a function of $|B - O|$. Consider a 3-by-3 pixel image z_0 representing the background only whose elements are all B values. By adding and subtracting z_0 to (15) and taking absolute values, we get the following

$$\mathbf{1}^{|} \cdot vec(-|z_0 - \bar{z}| + |z_0 - z|) = -|B - O| + 3|B - O| = 2|B - O|, \qquad (16)$$

which is the desired expression. Equation (16) provides a clear breakdown to how the error signal can achieve the desired error directionality and magnitude. We therefore reorganize the block diagram in Figure 3 as shown in Figure 4 to ensure the generality of Assumption 3 is not lost.

Note that we may use the view synthesizer to generate a background by choosing s_0 to be a large value, thus the leading car essentially is vanishing from the view.

3.4. Stability Analysis of the Learning-Based Controller

The stability analysis will be discussed for the block diagram of Figure 3. We treat the following nonlinear controller which has a proportional gain acting on a neural network based error signal

$$u = vec(K)^{\intercal} \cdot vec(\hat{\bar{y}} - y) = vec(K)^{\intercal} \cdot vec(\hat{e}), \tag{17}$$

which relates to (7) and mainly (9); and where $\hat{\bar{y}} - y = I_{Syn}(\bar{s}, I_{Cam}(s, \Theta, \Omega)) - I_{Cam}(s, \Theta, \Omega)$. Note that $\hat{\bar{y}}$ reflects an internal model principle. $\hat{e}$ enables background invariance, thus generalization to backgrounds. Generalization and sample efficiency are key performance issues Chen et al. (2020) and Sax et al. (2019).

We first note that the dynamical system (1) can be decomposed into two subsystems, a stable uncontrollable subsystem governing the dynamics of $x_1(t)$ and a controllable subsystem governing the dynamics of $x_2(t)$ and $x_3(t)$. By decoupling the stable uncontrollable subsystem, we have:

$$\dot{x_2}(t) = -x_3(t), \tag{18a}$$

$$\dot{x_3}(t) = -\frac{\alpha_2}{m_2} x_3(t) + \frac{1}{m_2} u, \tag{18b}$$

$$y(t) = I_{Cam}(\bar{s} - x_2, \Theta, \Omega), \tag{18c}$$

$$\hat{e}(t) = \hat{\bar{y}} - y = I_{Syn}(\bar{s}, I_{Cam}(s, \Theta, \Omega)) - I_{Cam}(\bar{s} - x_2, \Theta, \Omega). \tag{18d}$$

The following theorem demonstrates stability and thus addresses Problem 2.

Theorem 2 *Consider controller* (17) *and let* $K = \mathbf{1}$. *The dynamics* (18) *for a fixed* Θ *is UUB.*

Proof We first construct an appropriate Lyapunov function candidate. Let $u^*(x_2) = vec(K)^{\intercal} \cdot vec(H(x_2, \bar{s}, \Theta))$. Consider the following positive definite function for subsystem (18)

$$V(x_2, x_3) = w_1 {x_2}^2 + w_2 x_2 x_3 + w_3 {x_3}^2 + w_4 \int_0^{x_2} u^*(z) dz, \tag{19}$$

where $w_1 > 0$ is arbitrary, and $w_3 > 0$, and w_2 are chosen appropriately and such that $w_1 x_1^2 + w_2 x_2 x_3 + w_3 x_3^2$ is positive definite in x_2 and x_3. Moreover $w_4 > 0$ will be chosen appropriately noting that the integral term is nonnegative due to Assumption 3 and $K = \mathbf{1}$.

From Assumption 5, $u^*(z)$ is locally continuous in z. Differentiating $V(x_2, x_3)$ along the trajectories of (18), we get

$$\begin{aligned} \dot{V}(x_2, x_3) =& 2w_1 x_2 \dot{x_2} + w_2 \dot{x_2} x_3 + w_2 x_2 \dot{x_3} + 2w_3 x_3 \dot{x_3} + w_4 \dot{x_2} u^*(x_2), \\ =& \left(-2w_1 - \frac{\alpha_2}{m_2} w_2\right) x_2 x_3 + \left(-w_2 - 2\frac{\alpha_2}{m_2} w_3\right) x_3^2 + \frac{w_2}{m_2} x_2 u + \left(2\frac{w_3}{m_2} u - w_4 u^*\right) x_3. \end{aligned} \tag{20}$$

Adding and subtracting $\frac{w_2}{m_2}x_2u^*$ to (20), we get

$$\begin{aligned}\dot{V}(x_2,x_3) = &\left(-2w_1 - \frac{\alpha_2}{m_2}w_2\right)x_2x_3 + \left(-w_2 - 2\frac{\alpha_2}{m_2}w_3\right)x_3^2 \\ &+ \frac{w_2}{m_2}x_2(u-u^*) + \frac{w_2}{m_2}x_2u^* + \left(2\frac{w_3}{m_2}u - w_4u^*\right)x_3.\end{aligned} \tag{21}$$

Choosing $w_2 = -2\frac{m_2}{\alpha_2}w_1$ to cancel the x_2x_3 term, and choosing $w_4 = 2\frac{w_3}{m_2}$ we get

$$\begin{aligned}\dot{V}(x_2,x_3) = &\left(2\frac{m_2}{\alpha_2}w_1 - 2\frac{\alpha_2}{m_2}w_3\right)x_3^2 - 2\frac{w_1}{\alpha_2}x_2(u-u^*(x_2)) - 2\frac{w_1}{\alpha_2}x_2u^*(x_2) \\ &+ 2\frac{w_3}{m_2}x_3(u-u^*(x_2)).\end{aligned} \tag{22}$$

We finally choose $w_3 > (\frac{m_2}{\alpha_2})^2w_1$ to force the coefficient of the first term in the right-hand side of (22) to be negative. We therefore write (22) as follows

$$\begin{aligned}\dot{V}(x_2,x_3) &\leq \left(2\frac{m_2}{\alpha_2}w_1 - 2\frac{\alpha_2}{m_2}w_3\right)|x_3|^2 + 2\frac{w_3}{m_2}|x_3||u-u^*(x_2)| \\ &\quad - 2\frac{w_1}{\alpha_2}|x_2||u^*(x_2)| + 2\frac{w_1}{\alpha_2}|x_2||u-u^*(x_2)|, \\ &= -|x_3|\left(\left(2\frac{\alpha_2}{m_2}w_3 - 2\frac{m_2}{\alpha_2}w_1\right)|x_3| - 2\frac{w_3}{m_2}|u-u^*(x_2)|\right) \\ &\quad - 2\frac{w_1}{\alpha_2}|x_2|\left(|u^*(x_2)| - |u-u^*(x_2)|\right).\end{aligned} \tag{23}$$

From Assumption 6, it can be shown that $\exists\epsilon_2 > 0$ such that $|u-u^*(x_2)| < \epsilon_2$, which when substituted in (23) we get

$$\dot{V}(x_2,x_3) \leq -|x_3|\left(\left(2\frac{\alpha_2}{m_2}w_3 - 2\frac{m_2}{\alpha_2}w_1\right)|x_3| - 2\frac{w_3}{m_2}\epsilon_2\right) - 2\frac{w_1}{\alpha_2}|x_2|\left(|u^*(x_2)| - \epsilon_2\right). \tag{24}$$

It can be shown that $\exists r > 0$ and a ball $B([x_2,x_3],r)$ around the origin such that if $[x_2,x_3] \notin B([x_2,x_3],r)$ then $\dot{V}(x_2,x_3) \leq 0$. ■

4. Conclusion

We demonstrated that stable feedback control directly from raw pixels is plausible and promising, and that introduced assumptions hold reasonably well for the application domain considered within a simulator environment. For further improvements and scalability, we need to investigate approaches to relax strong assumptions and have the theory encompassing of more practical scenarios and different types of motions and tracked objects, and to further provide quantitative and qualitative assessments on generalization and sample complexity. The method generalized well to different driving backgrounds that have not been seen before due to the ability of the synthesizer to be reasonably invariant to background changes. The approach provides a more clear path to apply control theory directly to pixels and establish safe and trustworthy dynamical systems that are more interpretable compared to purely end-to-end learning approaches. The approach can extend to various automatic control applications where a cheap camera sensor can be deployed for motion control.[1]

1. Code is available at `https://github.com/abukhalaf/FeedbackFromPixels_L4DC2021`

Acknowledgments

Toyota Research Institute provided funds to support this work.

Appendix A. Proof of Lemma 1

Proof Note that $h(x) = h(x_2, \bar{s}, \Theta)$, and therefore from Assumption 2, it follows that $ker(h(x)) = \{x \in \mathsf{R}^{n\times 1} : x_2 = 0\}$. Moreover, by writing $f(x) = Ax$, $g(x) = B$, and locally $h^{|}(x)h(x) = c^2{x_2}^2$ from Assumption 5, and $G = [G_1, G_2, G_3]$, and where

$$A = \begin{bmatrix} -\frac{\alpha_1}{m_1} & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & -\frac{\alpha_2}{m_2} \end{bmatrix}, \quad (25a) \qquad B = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{m_2} \end{bmatrix}, \quad (25b) \qquad C = \begin{bmatrix} 0 & c & 0 \end{bmatrix}, \quad (25c)$$

we can replace the Hamilton-Jacobi (HJ) equation (10a) and (10b) over domain $D \subset \mathsf{R}^{n\times 1}$ with

$$0 = A^{|}P + PA - PBB^{|}P + C^{|}C + G^{|}G, \quad (26a)$$
$$0 = N(A^{|}P + PA)N, \qquad N = I - C^{|}(CC^{|})^{-1}C. \quad (26b)$$

From the kernel condition (26b), we have

$$P = \begin{bmatrix} p_{11} & \frac{\alpha_1}{m_1}p_{11} & -\frac{\frac{\alpha_1}{m_1}p_{11} + \frac{\alpha_2}{m_2}p_{33}}{\frac{\alpha_1}{m_1} + \frac{\alpha_2}{m_2}} \\ \frac{\alpha_1}{m_1}p_{11} & p_{22} & -\frac{\alpha_2}{m_2}p_{33} \\ -\frac{\frac{\alpha_1}{m_1}p_{11} + \frac{\alpha_2}{m_2}p_{33}}{\frac{\alpha_1}{m_1} + \frac{\alpha_2}{m_2}} & -\frac{\alpha_2}{m_2}p_{33} & p_{33} \end{bmatrix}. \quad (27)$$

From the algebraic Riccati equation (26a), we obtain the following for p_{11}, p_{22}, p_{33} and G:

$$p_{11} = \frac{|c|\alpha_2 + |c|^2\frac{m_2}{\alpha_2} - |c|^2\frac{\frac{m_1}{\alpha_1}\frac{m_2}{\alpha_2}}{\frac{m_1}{\alpha_1} + \frac{m_1}{\alpha_1}}}{|c|\frac{\alpha_1}{\alpha_1 m_2 + \alpha_2 m_1} + \frac{{\alpha_1}^2}{{m_1}^2}}, \quad (28a) \qquad G_1 = -\frac{\alpha_1 m_2 p_{11} + \alpha_2 m_1 p_{33}}{\alpha_1 {m_2}^2 + \alpha_2 m_1 m_2}, \quad (28b)$$

$$p_{22} = |c|\alpha_2 + |c|^2\frac{m_2}{\alpha_2}, \quad (28c) \qquad G_2 = 0, \quad (28d)$$

$$p_{33} = |c|\frac{{m_2}^2}{\alpha_2}, \quad (28e) \qquad G_3 = |c|\frac{m_2}{\alpha_2}. \quad (28f)$$

Substituting (28a), (28c) and (28e) in (27), it follows that the principal minors of (27) are nonnegative; hence $P \geq \mathbf{0}$. ■

For specific numerical values of α_1, α_2, m_1 and m_2, a numerical procedure shown in Kučera and Souza (1995) and Gadewadikar et al. (2006) can be used to numerically solve (26).

Appendix B. Simulation Results

Due to space limitations, these results are provided in the technical report Abu-Khalaf et al. (2021).

References

Murad Abu-Khalaf, Sertac Karaman, and Daniela Rus. Feedback from pixels: Output regulation via learning-based scene view synthesis. *arXiv preprint arXiv:2103.10888*, 2021.

A. Amini, W. Schwarting, G. Rosman, B. Araki, S. Karaman, and D. Rus. Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 568–575, Oct 2018. doi: 10.1109/IROS.2018.8594386.

A. Amini, I. Gilitschenski, J. Phillips, J. Moseyko, R. Banerjee, S. Karaman, and D. Rus. Learning robust control policies for end-to-end autonomous driving from data-driven simulation. *IEEE Robotics and Automation Letters*, 5(2):1143–1150, 2020. doi: 10.1109/LRA.2020.2966414.

A. Astolfi and P. Colaneri. A hamilton-jacobi setup for the static output feedback stabilization of nonlinear systems. *IEEE Transactions on Automatic Control*, 47(12):2038–2041, Dec 2002. ISSN 2334-3303. doi: 10.1109/TAC.2002.805680.

Alessandro Astolfi and Patrizio Colaneri. Static output feedback stabilization: from linear to nonlinear and back. In Alberto Isidori, Françoise Lamnabhi-Lagarrigue, and Witold Respondek, editors, *Nonlinear control in the Year 2000*, pages 49–71, London, 2001. Springer London. ISBN 978-1-84628-568-4.

Ershad Banijamali, Rui Shu, Mohammad Ghavamzadeh, Hung Bui, and Ali Ghodsi. Robust locally-linear controllable embedding. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1751–1759, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/banijamali18a.html.

Q. Bateux, E. Marchand, J. Leitner, F. Chaumette, and P. Corke. Training deep neural networks for visual servoing. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3307–3314, 2018. doi: 10.1109/ICRA.2018.8461068.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016. URL http://arxiv.org/abs/1604.07316.

Bryan Chen, Alexander Sax, Gene Lewis, Iro Armeni, Silvio Savarese, Amir Zamir, Jitendra Malik, and Lerrel Pinto. Robust policies via mid-level visual representations: An experimental study in manipulation and navigation. *4th Conference on Robot Learning (CoRL 2020)*, 2020.

C. Collewet and E. Marchand. Photometric visual servoing. *IEEE Transactions on Robotics*, 27(4): 828–834, 2011. doi: 10.1109/TRO.2011.2112593.

Sarah Dean and Benjamin Recht. Certainty equivalent perception-based control. *arXiv preprint arXiv:2008.12332*, 2020.

Sarah Dean, Nikolai Matni, Benjamin Recht, and Vickie Ye. Robust guarantees for perception-based control. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 350–360, The Cloud, 10–11 Jun 2020a. PMLR. URL http://proceedings.mlr.press/v120/dean20a.html.

Sarah Dean, Andrew J Taylor, Ryan K Cosner, Benjamin Recht, and Aaron D Ames. Guaranteeing safety of learned perception modules via measurement-robust control barrier functions. *arXiv preprint arXiv:2010.16001*, 2020b.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

Jyotirmay Gadewadikar, Frank L. Lewis, and Murad Abu-Khalaf. Necessary and sufficient conditions for H-Infinity static output-feedback control. *Journal of Guidance, Control, and Dynamics*, 29(4): 915–920, 2006. doi: 10.2514/1.16794. URL https://doi.org/10.2514/1.16794.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/hafner19a.html.

N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese. GONet: A semi-supervised deep learning approach for traversability estimation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3044–3051, Oct 2018. doi: 10.1109/IROS.2018.8594031.

N. Hirose, A. Sadeghian, F. Xia, R. Martín-Martín, and S. Savarese. VUNet: Dynamic scene view synthesis for traversability estimation using an rgb camera. *IEEE Robotics and Automation Letters*, 4(2):2062–2069, April 2019a. ISSN 2377-3774. doi: 10.1109/LRA.2019.2894869.

N. Hirose, F. Xia, R. Martín-Martín, A. Sadeghian, and S. Savarese. Deep visual mpc-policy learning for navigation. *IEEE Robotics and Automation Letters*, 4(4):3184–3191, Oct 2019b. ISSN 2377-3774. doi: 10.1109/LRA.2019.2925731.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2017–2025. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and

fast inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2946–2954. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/7d6044e95a16761171b130dcb476a43e-Paper.pdf.

Hassan K Khalil. *Nonlinear Systems*. Prentice-Hall, 2002.

V. Kučera and C.E. De Souza. A necessary and sufficient condition for output feedback stabilizability. *Automatica*, 31(9):1357 – 1359, 1995. ISSN 0005-1098. doi: https://doi.org/10.1016/0005-1098(95)00048-2. URL http://www.sciencedirect.com/science/article/pii/0005109895000482.

W. Levine and M. Athans. On the optimal error regulation of a string of moving vehicles. *IEEE Transactions on Automatic Control*, 11(3):355–361, 1966. doi: 10.1109/TAC.1966.1098376.

K. Nagai and S. Sakai. A visual feedback design on matrix space for a liquid sloshing experiment. In *The SICE Annual Conference 2013*, pages 2088–2093, 2013.

E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–711, 2017. doi: 10.1109/CVPR.2017.82.

S. Sakai and M. Ando. On the visual systems control on matrix space. In *53rd IEEE Conference on Decision and Control*, pages 2173–2178, 2014. doi: 10.1109/CDC.2014.7039720.

S. Sakai and M. Sato. Visual systems control on polynomial space and its application to sloshing problems. *IEEE Transactions on Control Systems Technology*, 22(6):2176–2187, 2014. doi: 10.1109/TCST.2014.2309615.

Alexander Sax, Bradley Emi, Amir R. Zamir, Leonidas Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. *arXiv preprint arXiv:1812.11971*, 2019.

A. Saxena, H. Pandya, G. Kumar, A. Gaud, and K. M. Krishna. Exploring convolutional networks for end-to-end visual servoing. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3817–3823, 2017. doi: 10.1109/ICRA.2017.7989442.

H. J. Terry Suh and Russ Tedrake. The surprising effectiveness of linear models for visual foresight in object pile manipulation. In *The 14th International Workshop on the Algorithmic Foundations of Robotics*, 2020.

Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 322–337, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In

C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2746–2754. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. *Science Advances*, 6(10), 2020. doi: 10.1126/sciadv.aax5979. URL https://advances.sciencemag.org/content/6/10/eaax5979.

Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew Johnson, and Sergey Levine. SOLAR: Deep structured representations for model-based reinforcement learning. volume 97 of *Proceedings of Machine Learning Research*, pages 7444–7453, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/zhang19m.html.

Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 286–301, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.