R4-A.1: Dynamics-Based Video Analytics

I. PARTICIPANTS INVOLVED FROM JULY 1, 2019 TO JUNE 30, 2020

Faculty/Staff			
Name	**Title**	**Institution**	**Email**
Octavia Camps	Co-PI	Northeastern University	o.camps@northeastern.edu
Mario Sznaier	Co-PI	Northeastern University	m.sznaier@northeastern.edu
Graduate, Undergraduate and REU Students			
Name	**Degree Pursued**	**Institution**	**Month/Year of Graduation**
Sadjad Esfeden Asghari	PhD	Northeastern University	12/2021
Wenqian Liu	PhD	Northeastern University	5/2021
Dan Luo	PhD	Northeastern University	12/2023
Bengizu Ozbay	PhD	Northeastern University	12/2021
Dong Yin	PhD	Northeastern University	12/2022
Yuexi Zhang	PhD	Northeastern University	12/2021
Armand Comas	MS, PhD	Northeastern University	6/2019 (MS), 6/2024 (PhD)
Timothy Rupprecht	MS	Northeastern University	8/2020
Can Uner	MS	Northeastern University	12/2019

II. PROJECT DESCRIPTION

A. Project Overview

Video-based methods can provide advanced warning of terrorist activities and threats. In addition, they can assist and substantially enhance localized, complementary sensors that are more restricted in range, such as radar, infrared, and chemical detectors. Moreover, since the supporting hardware is relatively inexpensive and largely already deployed (stationary and mobile networked cameras, including camera cell phones, capable of broadcasting and sharing live video feeds), the additional investment required is minimal.

Arguably, a critical impediment to fully realizing this potential was the absence of reliable technology for robust, real-time interpretation of the abundant, multi-camera video data. The dynamic and stochastic nature of this data, compounded with its high dimensionality, and the difficulty to characterize distinguishing features of benign versus dangerous behaviors, makes automatic threat detection extremely challenging. Indeed, state-of-the-art turnkey software relies heavily on human operators, which in turn severely limits the scope of its use.

This research effort was motivated by an emerging opportunity to address these challenges, exploiting advances at the confluence of robust dynamical systems, computer vision, and machine learning. A fundamental feature and key advantage of the envisioned methods is the encapsulation of metadata on targeted behavior using dynamics-based and statistical-based invariants. Drawing on solid theoretical

foundations, robust system identification and adaptation methods, along with model (in)validation and artificial intelligence tools, we designed algorithms for quantifiable characterization of threats and benign behaviors, provable uncertainty bounds, and alternatives for viable explanations of observed activities.

Specifically, this research sought to predict and isolate threats in crowded public spaces—such as sports arenas, airports, bus terminals—and vulnerable urban spaces, as illustrated in Figure 1.

Figure 1: This research sought to predict and isolate threats in crowded public spaces, such as sport arenas and transport terminals, and vulnerable urban spaces.

Toward this goal, we developed algorithms to:

- answer the "who, what, where, and why" questions from video data;
- identify security breaches at portals;
- track movements across distributed camera networks;
- detect suspicious, potentially threatening activities; and
- flag objects left behind.

The resulting systems integrate real-time data from multiple sources over dynamic networks, covering large areas, extracting meaningful behavioral information on a large number of individuals and objects, and striking a difficult compromise between the inherent conservatism demanded from threat detection and the need to avoid a high false-alarm ratio, which heightens vulnerability by straining resources.

The impact of successful video analytics such as the ones developed in this project are very relevant to the Department of Homeland Security (DHS). Our goal was to provide tools to automatically process vast amounts of visual data, most of which is not relevant, and to localize, both in space and time, critical actionable information that is needed to ensure safety in large public spaces.

B. State of the Art and Technical Approach

Recent advances in the accuracy and efficiency of object detectors [1, 2], particularly pedestrian detectors, have inspired and fueled multi-target tracking approaches for detection. These techniques proceed by detecting the targets frame by frame, using a high quality object detector, and then associating these detections by using online or offline trackers [3-5]. Often, these associations are based on appearance and location similarity; however, these approaches fail when the appearance of the targets is discriminative and the targets display simple motion patterns. While there are trackers that rely less on appearance [6-10], they often require the tuning of a large number of parameters and the expertise to adapt the algorithms to these more challenging scenarios. Alternatively, Ding et al. [11] showed that it is possible to use dynamics

to compare tracks and disambiguate between targets without assuming a motion model a priori; however, the computational and memory complexity of this approach has limited its application to short trajectories of a few targets.

Multiple cameras are used to cover wide areas and provide different viewpoints of targets. Maintaining consistent identity labels across cameras is a difficult problem since the appearance of the targets can be quite different when seen from different angles. Previous approaches to this problem include matching features such as color and apparent height [12-14, 15], using 3D information from camera calibration [13, 16-20], using the epipolar constraint [21-23], modeling the relationship between the appearance of a target in different views through a linear time invariant system [24], or computing homographies between views [25-29]. When the cameras do not have overlapping fields of view, targets must be re-identified (re-IDed) across cameras. A good overview of existing re-ID methods can be found in [30-34] and the references therein.

The three most important aspects in re-ID are the features used, the matching procedure, and the performance evaluation. Most re-ID approaches use appearance-based features that are viewpoint quasi-invariant [35-40], such as color and texture descriptors; however, the number of features used varies greatly across approaches, making it difficult to compare their impact on performance. Using standard metrics such as Euclidean distance to match images based on these types of features results in poor performance due to the large variations in pose, illumination, and limited training data. Thus, recent approaches [34, 41-44] design classifiers to learn specialized metrics that enforce features from the same individual to be closer than features from different individuals. Yet, state-of-the-art performance remains low, slightly above 30% for the best match. Performance is often reported on standard datasets, and while they are challenging, they bring in different biases. Moreover, the number of datasets and the experimental evaluation protocols used also vary greatly across approaches, making it difficult to compare them.

Video frame prediction is an active research topic [45-54]. Most approaches use convolutional networks, such as 3D convolutional networks [55] or generative adversarial networks (GANs) [56] to synthetize future frames. Many techniques work directly on pixel values [52, 57-60] while others [61, 62] use/predict optical flow as well. However, the performance of convolutional schemes is limited by short-range dependencies, and they often experience blurriness in the predicted frames.

Human pose estimation [63-67], which seeks to estimate the locations of human body joints, has many practical applications such as smart video surveillance[68, 69], human computer interaction [70], and virtual reality (VR) / augmented reality (AR) [71]. The most general pose estimation pipeline extracts features from the input and then uses a classification/regression model to predict the location of the joints. Recently, Bertasius et al. [72] introduced a Pose Warper capable of using a few manually annotated frames to propagate pose information across the complete video. However, it relies on annotations of every k^{th} frame, and thus it fails to fully exploit the dynamic correlation between them.

Dynamics, and more precisely dynamic invariants, can be used to extract critical information from data streams. Robust identification of piecewise affine dynamic systems has been the subject of recent intensive research, leading to a number of techniques meant to identify subsystem dynamics and switching surfaces [73]. A common feature is the computational complexity entailed in dealing with noisy measurements. In this case, algebraic procedures [74] lead to nonconvex optimization problems, while optimization methods lead to mixed integer/linear programming [75]. Similarly, methods relying on probabilistic priors [76] also lead to computationally complex combinatorial problems. An alternative approach is provided by clustering-based methods [77, 78]. Since these methods rely on local identification, they require “fair sampling” of each cluster, which places constraints on the data that can be used. More recently, the PIs of

this project have developed new sparsification-based techniques for identification of switched affine models that allow for several types of noise [79-81].

Finding dynamic invariants from corrupted data often requires the ability to solve optimization problems. Semidefinite programs seek to minimize a linear function subject to affine matrix equality and positive semidefinite constraints. These problems are convex (albeit nonsmooth) and thus tractable. Indeed, recent research efforts have led to numerous algorithms (for instance interior point algorithms) with polynomial complexity; excellent surveys are given in [82] and [83]. Of particular interest to this project are semidefinite programs resulting from the relaxation of constrained rank minimization problems [84, 85]. It has been recently shown [86] that in these cases, gradient-based methods outperform interior point ones. Polynomial optimization problems are highly nonconvex; however, one can find convex liftings leading to standard semidefinite programs. Two (related) approaches are usually used: the "sums of squares" approach [87], which provides convex certificates for positivity of a polynomial over a semi-algebraic set, and its dual approach, referred to as the "moments" approach [88]. Here, sufficient and asymptotically necessary conditions for a sequence to be a moment sequence of some Borel measure are used to convexify the problem [89].

C. Major Contributions

C.1. Year 7

C.1.a. Dynamics-Based Video Prediction

Humans and animals rely on making accurate predictions in order to survive in a dynamic world, as illustrated in Figure 2. Accurate predictions of the location of objects in the environment and their motion is of vital importance for autonomous navigation, estimating human intention, and controlling robots. Motivated by this need, there has been significant interest in the task of video prediction, where the goal is to synthesize future frames from previous ones.

Figure 2: Good timing is everything. In order to survive as they move in a dynamic world, humans and animals rely on having accurate predictions of where things are going to be.

During Year 6, we introduced a novel architecture, DYAN, to predict future frames from a given short video clip of a scene [90]. DYAN uses a dictionary made of a set of dynamics-based atoms to identify a dynamic model for the scene optical flow, which is used to generate future frames. This approach produces high-quality, realistic frames, but when looked at carefully, one can observe that the predictions exhibit significant lagging when compared against the ground-truth frames. We observed that one cause for this lag is that DYAN uses a Eulerian point of view: it makes predictions at a pixel location based on the previous values at this same location. This can lead to incorrect predictions when objects move in the scene

and partially occlude each other or the background. To address this shortcoming, in Year 6 we preprocessed the input flows using a recursive warping. The modified architecture, W-DYAN, tries to approximate a Lagrangian point of view, where predictions are made by following points in the 3D scene, using the optical flow as a surrogate for tracking.

In Year 7, we further improved the DYAN and W-DYAN architectures by introducing a new module to reduce prediction lag caused by abrupt changes of input dynamics and to allow the networks to process arbitrary long sequences.

Abrupt changes in dynamics happen at occlusion boundaries (Eulerian point of view) and when objects change their motion patterns (Eulerian and Lagrangian points of view). To detect these changes, the proposed module uses a recursive Kalman filter on the latent dynamic encoding of the input data. Since the filter is recursive, it allows the network to process the data as it becomes available in an online fashion. Given a new frame, the filter updates its generative model and the error covariance of its predictions. The filter compares the new measurement with its prediction from previous data and determines if it is within the current estimated error covariance. If it does, the new measurement is used to refine the current model and covariance. If it doesn't, a dynamic change is detected, and a new model is initialized. A diagram of the new architecture, K-DYAN (KW-DYAN), combining this module with DYAN (W-DYAN) is shown in Figure 3.

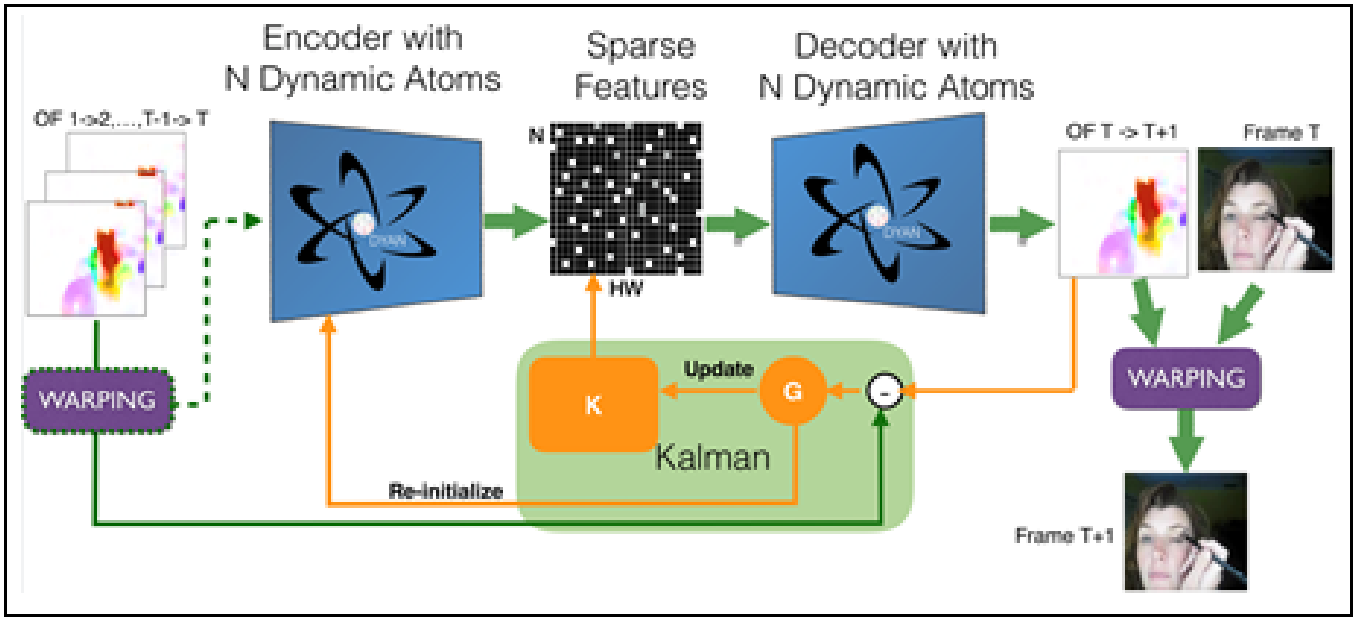

Figure 3: KW-DYAN's architecture. The latent space is filtered to estimate the error covariance, which is used by gate G to detect changes in the input dynamics, decide when to forget old inputs, and reset the system identification. The optional (dashed boundary) warping module at the input aligns the input optical flows to reduce the number of changes in dynamics in the input. All recurrent connections are shown in orange.

Figure 4 shows a qualitative example illustrating the benefits of using Kalman filtering with DYAN (K-DYAN) and warped WDYAN (KW-DYAN). As the figure shows, the module is able to significantly reduce the lag while still producing sharp images.

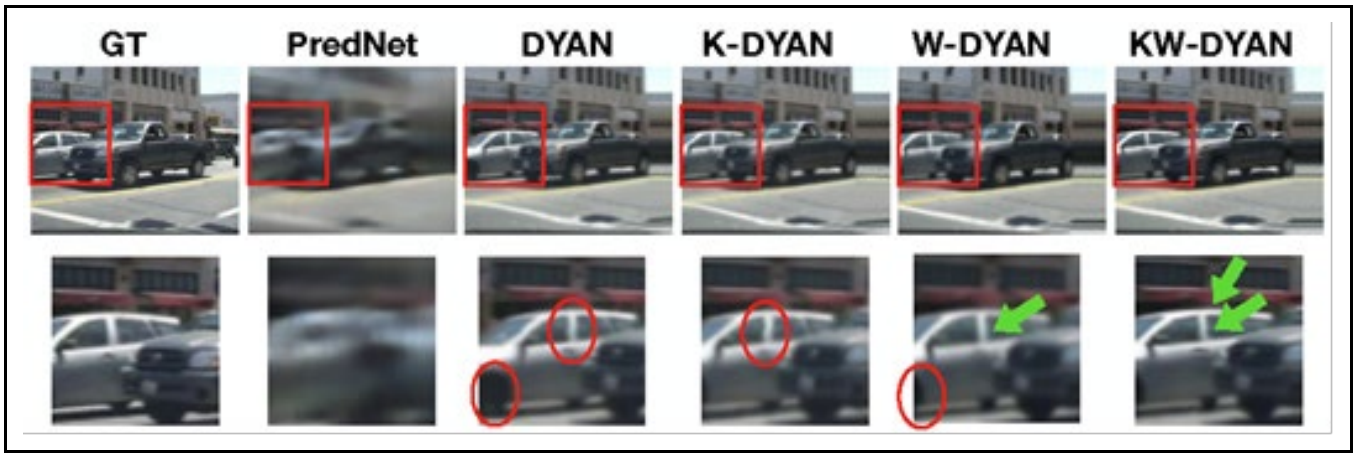

Figure 4: Qualitative example predicting the fifth future frame for the sequence set10V011 from the Caltech dataset. Top row shows the ground truth and predicted frames. Bottom row shows details inside the red box. Circled in red: incorrect position of the front tire and orientation of the window frame of the car. Green arrows: sharper roof line, correct orientation of the window frame.

We evaluated KW-DYAN and its variants and compared them against the state-of-the-art in Table 1 and Table 2, using the Caltech and UCF101 datasets, respectively. For next frame prediction, the networks predict the next optical flow and warp the last given frame. Comparisons were done with commonly used numerical measurements—mean peak-signal–to–noise ratio (PSNR) [58], mean square error (MSE), and structural similarity index measure (SSIM) [91]—to evaluate performance at the pixel level. Additionally, we report learned perceptual image patch similarity (LPIPS) distance [92], since it has been shown to be a good perceptual metric, and mean of the maximum optical flow (MMF) metric, proposed by us, to measure prediction lag. For the Human 3.6 M dataset, we evaluated long-term prediction performance using the mean of Euler angle error. Quantitatively, the higher PSNR/SSIM and the lower the MSE/LPIPS/MMF, the better the performance. As seen in the tables, the new architectures decreased lagging while still performing as good or better in the other metrics.

Method	Parameters	MSE x 10^3	SSIM	LPIPS x 10^{-2}	MMF
CopyLast	-	2.2	0.91	1.98	2.82
BeyondMSE [58]	8.9 million	3.26	0.88	-	-
PredNet [93]	6.9 million	-	0.91	7.47	-
ContextVP [49]	8.6 million	1.94	0.92	6.03	-
DualMoGan [94]	11.3 million	2.41	0.89	-	-
SDCNet [62]	-	1.62	0.92	-	-
CtrlGen [95]	-	-	0.90	6.38	-
FGVP [96]	-	-	0.92	5.04	-
DYAN	80	0.87	0.95	2.2	1.93
K-DYAN	82	0.69	0.96	2.1	1.75
W-DYAN	80	0.70	0.96	2.3	1.62
KW-DYAN	82	0.74	0.95	1.8	1.55

Table 1: Quantitative results for predictions on Caltech dataset, compared to best available open source methods.

Method	Parameters	PSNR	SSIM	LPIPS x 10^{-2}	MMF
CopyLast	-	28.6	0.89	3.8	6.25
BeyondMSE[58]	8.9 million	30.11	0.88	-	-
ContextVP[49]	8.6 million	34.9	0.92	-	-
DVF[61]	8.9 million	32.86	0.93	-	-
DYAN	80	34.26	0.95	**3.8**	5.59
K-DYAN	82	**34.90**	0.96	4.1	**4.95**
W-DYAN	80	32.07	**0.97**	4.4	5.75
KW-DYAN	82	33.59	0.95	4.1	5.87

Table 2: Quantitative results for predictions on UCF101 dataset, compared to best available open source methods. Red indicates the best scores.

C.1.b. Efficient Human Pose Estimation

In [97] we proposed an efficient pose estimation pipeline based on two observations: all frames are not equally informative, and the dynamics of the body joints can be modeled using simple dynamics. The new pipeline, shown in Figure 5, uses a light-weighted key frame proposal network (K-FPN) to select a small number of frames to apply a pose estimation model. One of the main contributions of our approach is a new loss function based on the recovery error in the latent feature space for unsupervised training of this network. The second module of the pipeline is an efficient human pose interpolation module (HPIM), which uses a dynamics-based dictionary to obtain the pose in the remaining frames.

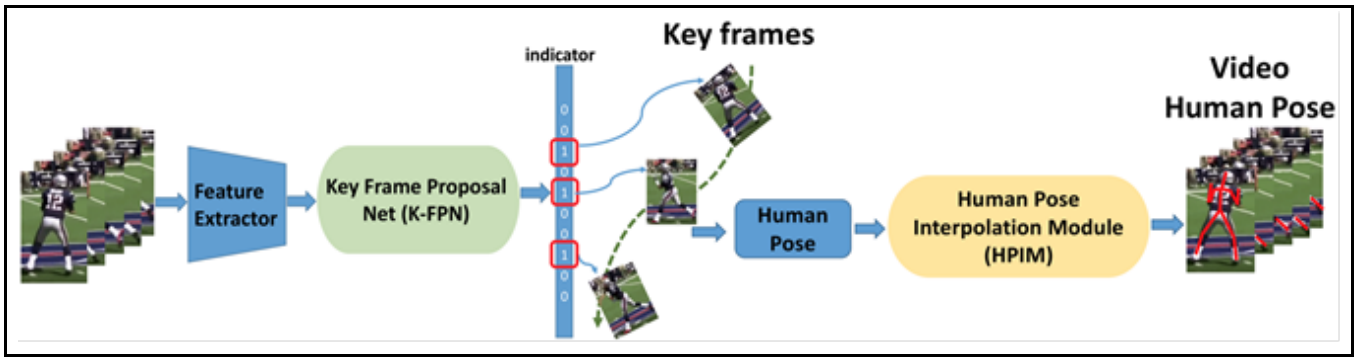

Figure 5: Proposed pipeline for video human pose detection. The K-FPN net, which is trained unsupervised, selects a set of key frames. The HPIM, trained to learn human pose dynamics, generates human poses for the entire sequence from the poses in the selected key frames.

Figure 6 shows two sample outputs of our pipeline, where the poses shown in purple were interpolated from the automatically selected red key frames. The advantages of the proposed approach are as follows:

- It uses a very light, unsupervised model to select "important" frames.
- It is highly efficient, since pose is estimated only at key frames.
- It is robust with respect to challenging conditions present in the non-key frames, such as occlusion, poor lighting conditions, and motion blur.
- It can be used to reduce annotation efforts for supervised approaches by selecting which frames should be manually annotated.

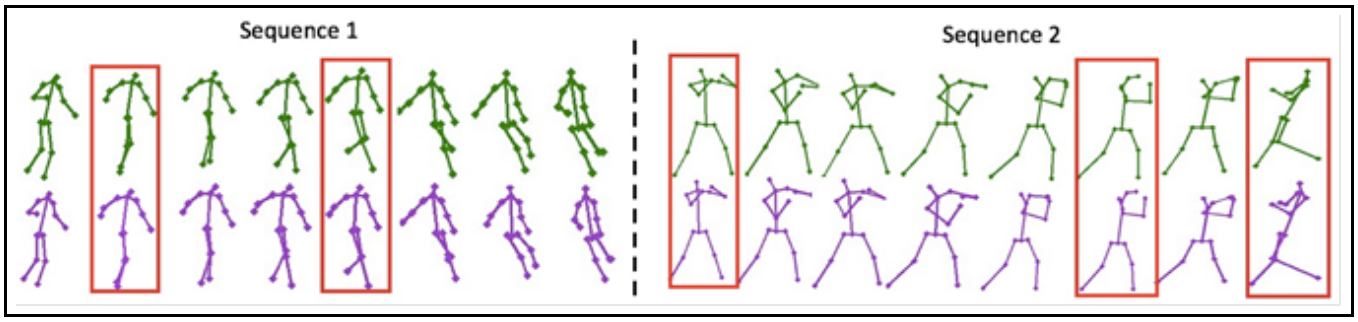

Figure 6: Two examples of the output of our pipeline, (top) ground truth and (bottom) poses recovered from the automatically selected key frames (red boxes).

The architecture of the frame selection network K-FPN is shown in Figure 7. It is trained completely unsupervised by minimizing the loss:

$$|[I - \rho^{-1}DD^{T}S]^{-1}Y|_F^2 + \lambda \sum_i s_i$$

where D is a dynamics-based DYAN dictionary, S is a diagonal matrix with diagonal elements s_i, which are the selection variables (1 for key frames, 0 otherwise), and Y is a tensor of image features of the input video. The first term of the loss penalizes reconstruction error of the input features from the features of the key frames while the second term penalizes the number of key frames.

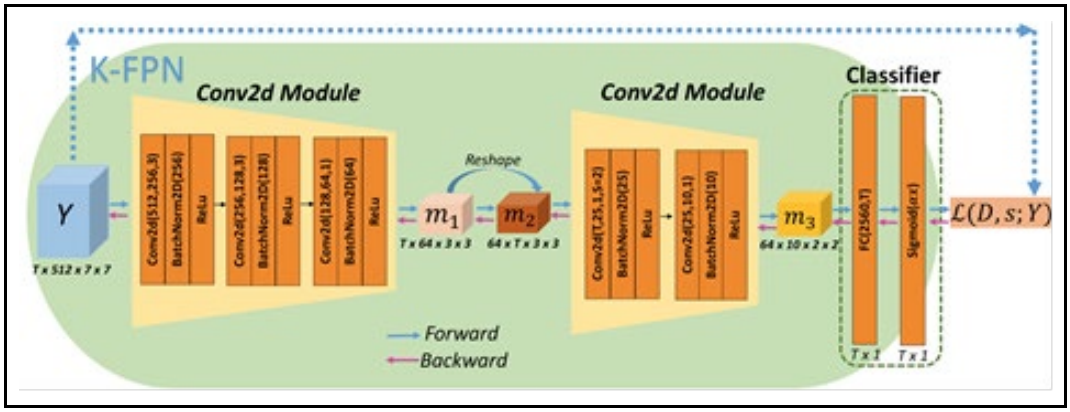

Figure 7: K-FPN architecture.

The network has two Conv2D modules followed by a fully connected and an adaptive Sigmoid layer, where the Sigmoid layer forces the output logits to be close to binary.

The HPIM efficiently interpolates the pose for the entire sequence, H, from the poses in the selected keyframes, H_r:

$$\boldsymbol{H} = \left(\boldsymbol{D}^{(h)}\boldsymbol{D}^{(h)^T}\right)\boldsymbol{P}_r^T\left[\boldsymbol{P}_r\left(\boldsymbol{D}^{(h)}\boldsymbol{D}^{(h)^T}\right)\boldsymbol{P}_r^T\right]^{-1}\boldsymbol{H}_r$$

where P_r is the selection matrix, $D^{(h)}$ is a dynamics-based DYAN dictionary for the poses, and $\left(D^{(h)}D^{(h)^T}\right)$ can be precomputed.

We evaluated the proposed approach on two widely used datasets: Penn Action (Table 3) and sub-JHMDB (Table 4). The input features *Y* were obtained using the ResNet family [98]. Our approach achieves the best performance and is 1.6 times faster (6.8 ms versus 11 ms) than the previous state-of-art [99] for the Penn Action dataset, using an average of 17.5 key frames. Moreover, if we use our lightest model (Resnet34), our approach is 2 times faster than [99] with a minor PCK degradation. For the sub-JHMDB dataset, we run more than 2 times faster than [100] without any degradation in accuracy.

Method	Time (ms)	Avg PCK	#Key Frames (avg,stdev)
Nie et al. [101]	-	48.0	N/A
Iqal et al. [102]	-	81.1	N/A
Gkioxari et al. [103]	-	91.9	N/A
Song et al. [104]	-	96.8	N/A
Luo et al. [105]	25.0	97.7	N/A
DKD (small CPM) [99]	12.0	96.8	N/A
Baseline [100]	11.3	97.4	N/A
DKD (ResNet50) [99]	11.0	97.8	N/A
Ours (ResNet50)	**6.8**	**98.0**	(17.5,4.9)
Ours (ResNet34)	**5.3**	**97.40**	(15.2,3.3)

Table 3: Performance evaluation on Penn Action dataset. Red indicates the best results.

Method	Time(ms)	Avg PCK	#Key Frames (avg,stdev)
Park et al.		52.5	N/A
Nie et al. [101]	-	55.7	N/A
Iqal et al. [102]	-	73.8	N/A
Song et al. [104]	-	92.1	N/A
Luo et al. [105]	24.0	93.6	N/A
DKD (ResNet50) [99]	-	94.0	N/A
Baseline [100]	10.0	94.4	N/A
Ours (ResNet50)	**7.0**	**94.7**	(17.8,1.4)
Ours (ResNet34)	**4.7**	**94.5**	(16.3,1.8)

Table 4: Performance evaluation on sub-JHDMB dataset. Red indicates the best results.

C.1.c. Explainable Variational Autoencoders and Anomaly Detection

Applications in safety-critical and consumer-focusing areas demand a clear understanding of the reasoning behind an algorithm's predictions, in addition certainly to robustness and performance guarantees. Consequently, there has been substantial recent interest in devising ways to understand and explain the underlying "why" driving the output of "what."

While progress in algorithmic generative modeling has been swift, explaining such generative algorithms is still a relatively unexplored field of study. There are certainly some ongoing efforts in using the concept of visual attention in generative models, but the focus of these methods is to use attention as an auxiliary information source for the particular task of interest, and not visually explain the generative model itself.

In [106], we take a step toward bridging this gap, proposing the first technique to visually explain variational autoencoders (VAE), by means of gradient-based attention. The intuition behind the proposed approach is that the latent space of a trained VAE captures key properties of the encoder, and thus explanations conditioned on the latent space will be informative about the downstream predictions.

More concretely, as illustrated in Figure 8, given a learned Gaussian distribution in the latent space, we use the re-parameterization trick to sample a latent code. Then, by backpropagating the activations in each dimension of the latent code to a convolutional layer in the model and aggregating all the resulting gradients, we generate the attention map.

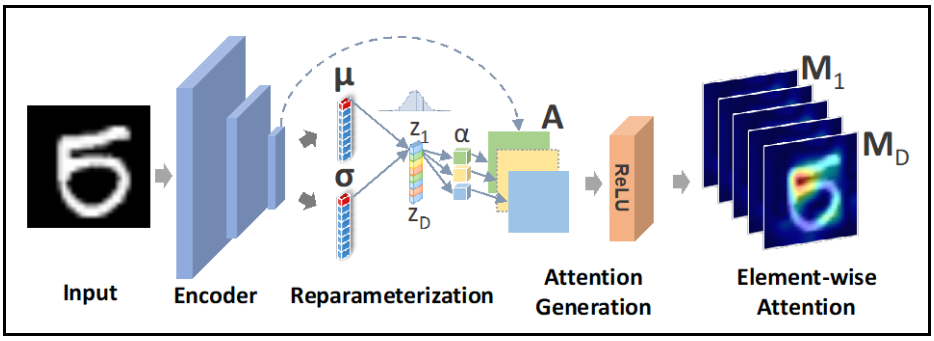

Figure 8: Element-wise attention generation with a VAE.

Let $q(z|x)$ be the posterior distribution inferred by the trained VAE for a sample x. For each element of latent vector z, we backpropagate gradients to the last convolutional feature map A, to obtain the attention map M^i corresponding to the element z_i:

$$M^i = ReLU\left(\sum_{k=1}^{n} \alpha_k A_k\right)$$

where the scalar α_k is given by:

$$\alpha_k = GAP\left(\frac{\partial z_i}{\partial A_k}\right) = \frac{1}{T}\sum_{p=1}^{h}\sum_{q=1}^{w}\frac{\partial z_i}{\partial A_k^{pq}}$$

and GAP is the global average pooling operator and A_k is the k^{th} feature channel of the feature map A.

Figure 9 shows an example of an attention map M. There, we can see that each component of the latent space has consistently high localization responses. All responses can be aggregated for an overall attention map using, for example, the mean of all the attention maps.

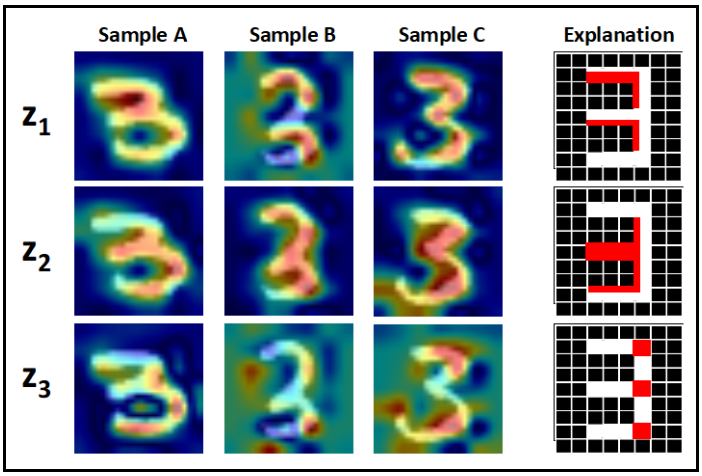

Figure 9: Each element in the latent space vector can be explained separately with the proposed attention map.

The attention maps obtained this way can also be used to localize anomaly regions, as illustrated in Figure 10 and Figure 11, given a one-class VAE trained on "normal" data (digit '1', for instance). When the VAE is given as input for an anomaly (i.e., digit "4"), the latent space for the given sample will be very different from the learned normal distribution. By simply computing the sum of all elements in the mean vector, we can obtain a score and backpropagate it to compute an anomaly attention map.

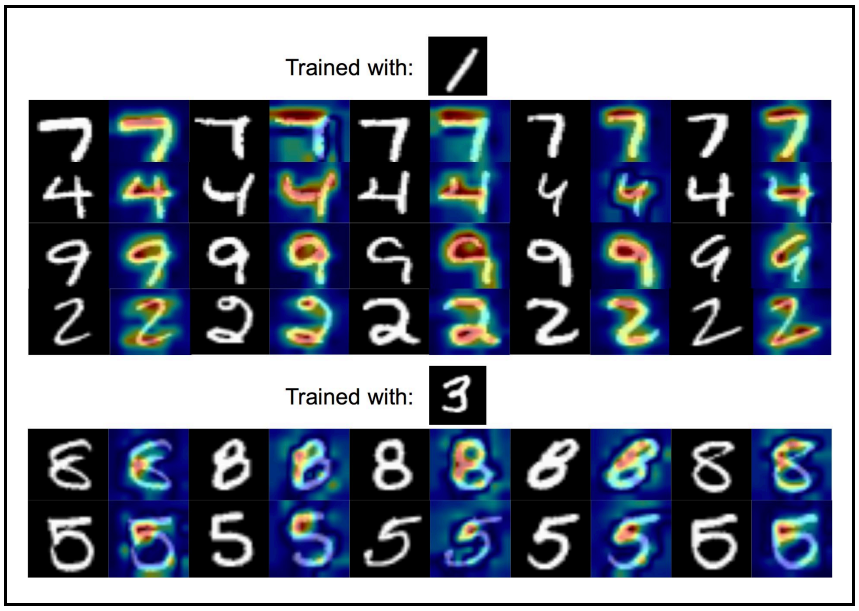

Figure 10: Anomaly localization for Modified National Institute of Standards and Technology database images.

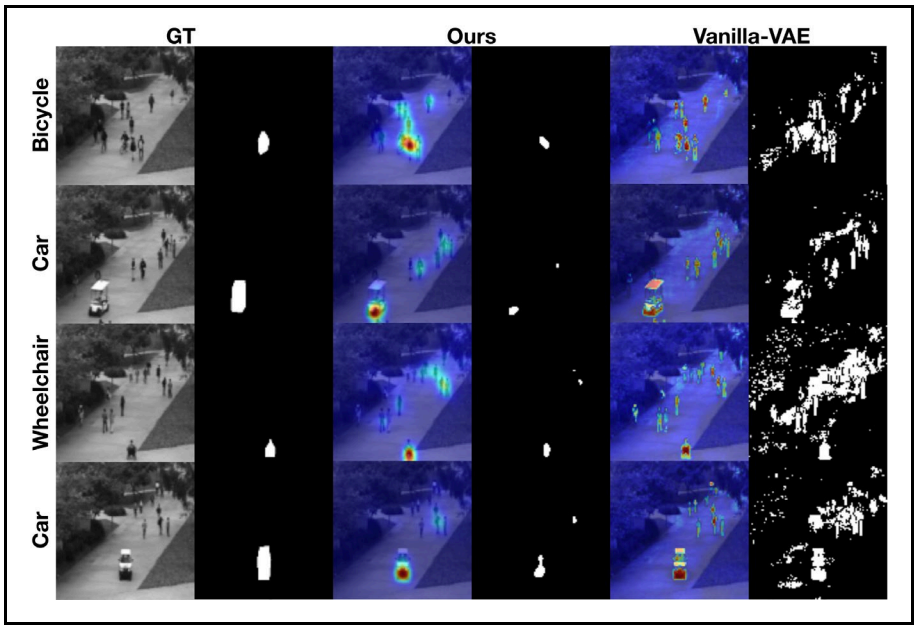

Figure 11: Anomaly detection using VAE attention maps; (left to right) original test image, ground truth masks, our attention maps, anomaly detection, and difference between input and VAE reconstruction. The anomalies in these examples are moving cars, bicycles, and wheelchair.

C.1.d. CLASP: Correlating Luggage and Specific Passengers

In addition to our core research work, we are working on a project with Rensselaer Polytechnic Institute (RPI) and Marquette University using ALERT's mock airport security checkpoint at the Kostas Research Institute. This supplement to ALERT's core cooperative agreement, Correlating Luggage and Specific Passengers (CLASP), allows us to generate large amounts of realistic data while facilitating ground truth annotation. We expect that this dataset will be the starting point for addressing many problems relevant to TSA.

During Year 7, we continued working on the CLASP task order project. In particular, we focused on the problem of activity recognition and passenger interactions. Toward this goal, we implemented several prototype learning neural networks that take combinations of RGB frames and detect actions such as putting down or picking up objects in or from bins on the conveyor belt and giving an object to another passenger.

Figure 12 shows a sample frame where the networks detect the action of putting objects in the conveyor belt bins. In our experiments, we tested using RGB alone, optical flow alone, and RGB and optical flow on a video with 67 bin transfers and 1 person-to-person transfer. RGB-alone had a precision of 0.88 and a recall of 0.90, while optical flow alone had 0.86 precision and 0.84 recall. Combining both inputs, precision improved to 0.91 and recall to 0.94.

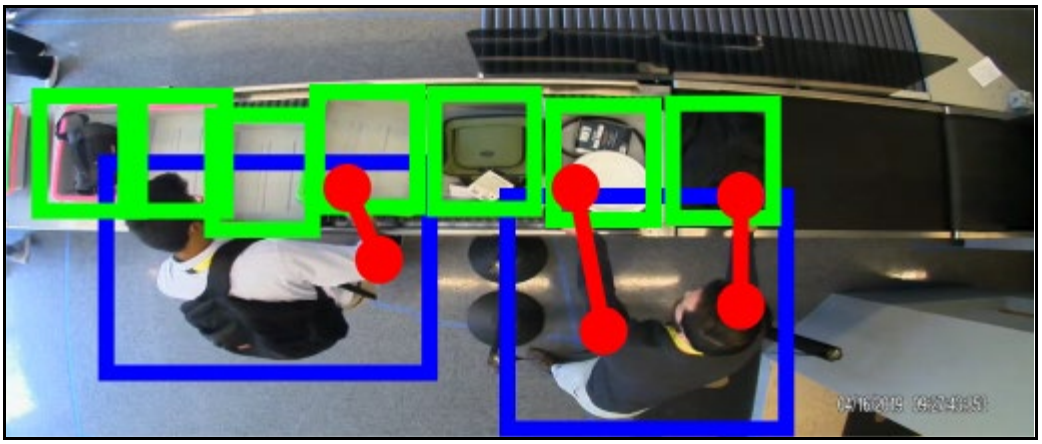

Figure 12: CLASP action detection (person to bin).

C.2. Years 1–6

The outcomes achieved in previous years (Years 1–6) are briefly summarized below.

- Tracking [107-109]:
  - We developed an algorithm that uses dynamics-based invariants to robustly track multiple targets with similar appearances. Our algorithm is faster and has better performance than the previous state-of-the-art algorithms. We have also developed a set of robust tools for tracking, including filtering and covariance propagation.
- Activity recognition [108, 110-119]:
  - We proposed an efficient algorithm to detect casual interactions by sparsifying a dynamics-based graph, where each node represents a time sequence associated with the location of an agent. This was applied to flag activity of people trying to breach security by moving into the secure area of the airport via the exit lane.
  - We developed a set of tools to compare and classify temporal sequences and applied them to the problem of activity recognition.
  - We developed a learning neural that takes an input human pose from an actor (e.g., joints or heat map of joints) and uses our DYAN encoder to capture dynamic information and classify activities.
  - We proposed a computationally efficient algorithm for the identification of error-in variables switched systems that can be used to segment activities from time traces of the position of a person's centroid in a video sequence.
- Human re-identification [120-123]:
  - In collaboration with Transportation Security Administration (TSA) and the Cleveland Hopkins International Airport (CLE), we collected and annotated a dataset for human re-identification at CLE. In collaboration with Great Cleveland Rapid Transit Authority (GCRTA) and the DHS Center of Excellence VACCINE (Visual Analytics for Command, Control, and Interoperability Environments), we collected and annotated a second dataset (where targets changed appearance) at bus terminals in Cleveland, OH.

- We benchmarked re-id algorithms and proposed a better kernel-based metric learning approach. We also addressed the problem of re-identifying targets in appearance-impaired scenarios, when targets have similar appearance or change appearance between views.

- Video prediction [90]:
  - We introduced a novel dynamics-based neural network, DYAN, that captures the dynamics of an input video sequence to predict future frames. The network is very compact, with only 80 parameters and achieved state-of-art performance.
  - We introduced a new metric to quantify prediction lagging.
  - We modified DYAN, using a recursive warping module to reduce temporal lagging in the predictions.
- Mathematical tools [108, 109, 111, 112, 114, 116, 117, 124-132]:
  - We developed theory-connecting machine learning and systems identification. For example, we developed the following tools:
    - A method for robustly estimating the fundamental matrix in stereo camera systems
    - A method for linear robust regression in the presence of gross outliers
    - A method for subspace clustering capable of incorporating prior information, which is suitable for motion and planar surfaces segmentation
    - An algorithm to chronologically sort crowd-sourced images in order to recover temporal information of an event
    - A robust algorithm for linear subspace clustering using a sum-of-squares approach
- Deep-model–based approaches [90, 133]:
  - We started incorporating our dynamics-based and statistical-based approaches into deep models. For example, we developed a deep architecture using moments embedding for fine-grain classification of objects that can only be distinguished by fine details. We also developed a deep architecture that incorporates dynamics-based layers for video encoding.
- Multi-camera motion segmentation [134]:
  - We developed an approach for motion segmentation of data collected with unsynchronized multiple cameras that combines shape and dynamical information but does not require spatiotemporal registration or shared features across video streams.
- CLASP [135]:
  - We implemented a set of algorithms for tasks related to CLASP, including passenger and bin detection, and upper body human pose estimation. The performance of the algorithms has been evaluated using data captured at ALERT's mock airport security checkpoint located at the NU Kostas Research Institute (KRI) in Burlington, MA.

D. Milestones

During Year 7, we continued working on the problems of activity segmentation, video prediction, and passenger-luggage association. We achieved the following milestones:

- Improved and tested dynamic-invariants–based deep architectures for video prediction to reduce prediction lag and process arbitrary length inputs.
- Designed and implemented an efficient deep pipeline to estimate human pose. The proposed approach is two times faster than the previous state of art.
- Designed and implemented a visual attention mechanism to explain variational autoencoders, which can be used for anomaly detection.
- Designed, implemented, and tested three preliminary architectures for action recognition for the CLASP project that can process RGB, optical flow, and a combination of the two to detect actions in real time.

E. Final Results as Project Completion (Year 7) / No-Cost Extension

Due to the current situation with COVID-19, we have not been able to collect (and label) additional data at the KRI for our CLASP research. This delay has impacted our ability to train and test our algorithms for activity detection and activity prediction. As a consequence, we have not been able to complete several of the anticipated milestones. Thus, we are planning to resume and continue this work through the no-cost extension period, ending in May 2021, to accomplish the following expected milestones:

- **MILESTONE 1**—Collect and annotate more data at the CLASP facility at KRI that captures a passenger's actions, from a side view.
  - NEXT STEPS: The collected data will be used to augment public available datasets and train a deep network for person and person-to-person activities for CLASP data. In particular, we will focus on interactions between passengers and transportation security officers during secondary screening and integrate human pose inputs.
- **MILESTONE 2**—Update the K-DYAN encoder to detect occlusions and fill gaps.
  - NEXT STEPS: We plan to extend the Kalman module to also run "backward in time" to provide further noise smoothing and to detect spatial occlusions and train a GAN to fill gaps.
- **MILESTONE 3**—We have implemented a prototype network to perform activity classification, which has been trained and tested using standard activity recognition datasets. However, the current network is not robust to partial detections and occlusions. We are making improvements to the current architecture by incorporating optical flow as part of the inputs.
  - NEXT STEPS: In addition to optical flow, we will incorporate human pose estimates as part of the input. Furthermore, we will use network distillation to eliminate the need to compute optical flow and pose during testing time in order to reduce computational complexity.
- **MILESTONE 4**—Fine-tune and test the new network performance using CLASP data.
  - NEXT STEPS: Once milestones 1 and 3 are completed, we will perform fine-tuning with CLASP data and will reevaluate the performance.
- **MILESTONE 5**—Extend the network to also perform action prediction. Train and test its performance using standard activity recognition datasets.
  - NEXT STEPS: This will start after milestone 4 is completed.
- **MILESTONE 6**—Fine-tune and test its performance using CLASP data.
  - NEXT STEPS: This will start after milestones 1 and 4 are completed.

III. RELEVANCE AND TRANSITION

A. Relevance of Research to the DHS Enterprise

This research addressed the challenge of processing vast amounts of video data in real-time to enhance security by: detecting dangerous situations as they evolve; providing supporting actionable information to mitigate damage; and aiding during forensic analysis of events.

Examples of benefits that a successful "who is doing what, where, and why" system could provide, include:

- Faster throughput in airport security lines without compromising security;
- Avoidance of airport terminal closure due to breach of security incidents (such as a person reaching the secure gates area through an exit, thus bypassing security);
- Quick identification of recurrent thieves in public transportation terminals; and
- Faster forensic analysis of security incidents.

All of these applications not only have a tangible effect in ensuring public safety, but also have clear economic benefits, such as reducing human resources needed at airport security checkpoints and reducing crime in bus terminals.

B. Status of Transition at Project End

The products of this research effort have direct application to the security and surveillance of large public spaces, such as airports, mass transport system terminals, sport venues, etc. In addition to directly supporting the homeland security enterprise's mission, systems endowed with activity analysis capabilities can assist law enforcement, allow elderly people to continue living independently, and help first responders and emergency workers prevent hazards from developing into full blown catastrophic situations. Finally, as part of this work, we continue collecting and labeling data, which will be distributed to the video analytics community to be used as benchmarks to aid the advancement of the state-of-the-art.

We engaged with potential customers by reaching out to DHS-related agencies such as TSA, and by presenting our work at professional and industrial meetings. Portions of this work have already been deployed and tested at CLE, where it was used by TSA officers to detect security threats caused by persons bypassing airport security at terminal exits. We believe that the system could be transferred to other airports in the near future. In addition, we are working on a project with RPI and Marquette University using ALERT's mock airport security checkpoint at KRI. This supplement to ALERT's core cooperative agreement, named Correlating Luggage and Specific Passengers (CLASP), allows us to generate large amounts of realistic data while facilitating ground truth annotation. We expect that this dataset will be the starting point for addressing many problems relevant to TSA. Finally, through the transition team at ALERT, we will also reach out to other DHS entities, such as the US Customs and Border Protection or the US Coast Guard, to explore transitioning our video-analytics-based threat detection and assessment tools to agency specific needs.

C. Transition Pathway and Future Opportunities

Our goal is to address the user needs for surveillance of large public spaces, such as airport terminals and bus stations. As part of this research, we are developing video analytics algorithms and implementing prototype systems, which are being tested using real-world data to show their feasibility.

D. Customer Connections

These are our customers from previous years:

- CLE airport commissioner, Mr. Fred Szabo
- CLE TSA, Mr. Michael Young (retired)
- GCRTA security chief, Mr. John Joyce
- DHS Science and Technology, Apex Screening at Speed Program manager, Mr. John Fortune

IV. PROJECT ACCOMPLISHMENTS AND DOCUMENTATION

A. Education and Workforce Development Activities

1. Course, Seminar, and/or Workshop Development
   a. Professor Octavia Camps taught regular and advanced courses in computer vision, where students worked on projects for object detection, tracking, and activity classification.
   b. Professor Mario Sznaier taught a course in control theory, where students applied concepts of system identification to design vision-based systems that can be used for surveillance.
2. Student Internship, Job, and/or Research Opportunities
   a. Wenqian Liu worked as a summer intern at Amazon.
   b. Armand Comas worked as an intern at MERL.

B. Peer Reviewed Journal Articles

1. Dai, T., & Sznaier, M. "A Semi-Algebraic Optimization Approach to Data-Driven Control of Continuous-Time Nonlinear Systems." *IEEE Control Systems Letters*, *5*(2), 18 June 2020, pp. 487–492. https://doi.org/10.1109/LCSYS.2020.3003505.

C. Peer Reviewed Conference Proceedings

1. Berberich, J., Sznaier, M., & Allgower, F. "Signal Estimation and System Identification with Nonlinear Dynamic Sensors." *IEEE Conference on Control Technology and Applications*, Hong Kong, China, 19–21 August 2019.
2. Taskazan, B., Miller, J., Inyang-Udoh, U., Camps, O., & Sznaier, M. "Domain Adaptation Based Fault Detection in Label Imbalanced Cyberphysical Systems." *IEEE Conference on Control Technology and Applications*, Hong Kong, China, 19–21 August 2019.
3. Dai, T., & Sznaier, M. "Worst-Case Optimal Data-Driven Estimators for Switched Discrete-Time Linear Systems." *IEEE Conference on Decision and Control,* Nice, France, 11–13 December 2019.
4. Miller, J., Zheng, Y., Roig-Solvas, B., Sznaier, M., & Papachristodoulou, A. "Chordal Decomposition in Rank Minimized Semidefinite Programs with Applications to Subspace Clustering." *IEEE Conference on Decision and Control*, Nice, France, 11–13 December 2019.
5. Ozbay, B., Camps, O., & Sznaier, M. "Efficient Identification of Error-in-Variables Switched Systems via a Sum-of-Squares Polynomial Based Subspace Clustering Method." *IEEE Conference on Decision and Control*, Nice, France, 11–13 December 2019.

6. Singh, R., & Sznaier, M. "A Convex Optimization Approach to Finding Low Rank Mixed Time/Frequency Domain Interpolants with Applications to Control Oriented Identification." *IEEE Conference on Decision and Control*, Nice, France, 11–13 December 2019.
7. Asghari-Esfeden, S., Sznaier, M., & Camps, O. "Dynamic Motion Representation for Human Action Recognition." *IEEE 2020 Winter Conference on Applications of Computer Vision*, Aspen, CO, 1–5 March 2020, pp. 557–566.
8. Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R., & Camps, O. "Towards Visually Explaining Variational Autoencoders." *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 16–18 June 2020, pp. 8642–8651.
9. Singh, R., & Sznaier, M. "A Loewner Matrix Based Convex Optimization Approach to Finding Low Rank Mixed Time/Frequency Domain Interpolants." *2020 American Control Conference*, Denver, CO, 1–3 July 2020.
10. Dai, T., Sznaier, M., & Roig-Solvas, B. "Data-Driven Quadratic Stabilization of Continuous LTI Systems." *2020 IFAC World Congress*, Berlin, Germany, 12–17 July 2020.
11. Miller, J., Zhang, Y., Sznaier, M., & Papachristodoulou, A. "Decomposed Structured Subsets for Semidefinite Optimization." *2020 IFAC World Congress*, Berlin, Germany, 12–17 July 2020.
12. Ozbay, B., Sznaier, M., & Camps, O. "An Algebraic Approach to Efficient Identification of a Class of Wiener Systems." *2020 IFAC World Congress*, Berlin, Germany, 12–17 July 2020.
13. Zhang, Y., Wang, Y., Camps, O., & Sznaier, M. "Key Frame Proposal Network for Efficient Pose Estimation in Videos." *European Conference on Computer Vision*, 23–28 August 2020.

Pending –

1. Comas Massague, A., Zhang, C., Feric, Z., Camps, O., & Yu, R. "Learning Disentangled Representations of Video with Missing Data." *Neural Information Processing Systems,* 5–12 December 2020, under review.
2. Liu, W., Comas Massague, A., Zhang, Y., Luo, D., Camps, O., & Szanier, M. "KW-DYAN: A Recurrent and Warping DYAN for Better Video Prediction." *Neural Information Processing Systems,* 5–12 December 2020, under review.
3. Miller, J., Wang, J., Sznaier, M., & Camps, O. "Model Fitting by Semialgebraic Clustering." *Neural Information Processing Systems,* 5–12 December 2020, under review.
4. Ozbay, B., Sznaier, M., & Camps, O. "SOS-Spaces: A Sum-of-Squares Polynomial Based Subspace Clustering Method." *Neural Information Processing Systems,* 5–12 December 2020, under review.
5. Sznaier, M. "A Convex Optimization Approach to Learning Koopman Operators." *Neural Information Processing Systems,* 5–12 December 2020, under review.
6. Chamanbaz, M., Sznaier, M., Lagoa, C.M., & Dabbene, F. "Probabilistic Discrete Time Robust H2 Controller Design." *59th IEEE Conference on Decision and Control*, Jeju Island, Korea, 14–18 December 2020, accepted.
7. Dai, T., & Sznaier, M. "A Semi-Algebraic Optimization Approach to Data-Driven Control of Continuous-Time Nonlinear Systems." *59th IEEE Conference on Decision and Control*, Jeju Island, Korea, 14–18 December 2020, accepted.

8. Miller, J., Singh, R., & Sznaier, M. "MIMO System Identification by Randomized Active-Set Methods." *59th IEEE Conference on Decision and Control*, Jeju Island, Korea, 14–18 December 2020, accepted.

D. Other Presentations

1. Seminars

   a. Camps, O. "Compact and Interpretable Dynamics-Based Video Representations." *2020 IEEE/CVF Area Chair Meeting, Conference on Computer Vision and Pattern Recognition*, San Diego, CA, February 2020.

E. Software Developed

1. We developed a suite of algorithms for fine grain classification, video prediction, motion segmentation, and outlier rejection. The code can be downloaded from http://robustsystems.coe.neu.edu.

V. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1: IEEE, pp. 886-893.

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008: IEEE, pp. 1-8.

[3] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011: IEEE, pp. 3457-3464.

[4] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009: IEEE, pp. 1200-1207.

[5] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008: IEEE, pp. 1-8.

[6] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012: IEEE, pp. 1926-1933.

[7] M. Betke, D. E. Hirsh, A. Bagchi, N. I. Hristov, N. C. Makris, and T. H. Kunz, "Tracking large variable numbers of objects in clutter," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007: IEEE, pp. 1-8.

[8] E. Brau, D. Dunatunga, K. Barnard, T. Tsukamoto, R. Palanivelu, and P. Lee, "A generative statistical model for tracking multiple smooth trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011: IEEE, pp. 1137-1144.

[9] R. T. Collins, "Multitarget data association with higher-order motion models," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012: IEEE, pp. 1744-1751.

[10] Z. Wu, T. H. Kunz, and M. Betke, "Efficient track linking methods for track graphs using network-flow and set-cover techniques," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011: IEEE, pp. 1185-1192.

[11] T. Ding, Mario Sznaier, and Octavia I. Camps, "Fast track matching and event detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, 2008.

[12] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 21, no. 11, pp. 1241-1247, 1999.

[13] T.-H. Chang and S. Gong, "Tracking multiple people with a multi-camera system," in *Multi-Object Tracking, 2001. Proceedings. 2001 IEEE Workshop on*, 2001: IEEE, pp. 19-26.

[14] D. Comaniciu, V. Ramesh, and F. Berton, "Adaptive resolution system and method for providing efficient low bit rate transmission of image data for distributed applications," ed: Google Patents, 2004.

[15] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. Van Gool, "Color-based object tracking in multi-camera environments," *Pattern Recognition,* pp. 591-599, 2003.

[16] J. Black and T. Ellis, "Multi camera image tracking," in *In International Workshop on Performance Evaluation of Tracking and Surveillance*, 2001: Citeseer.

[17] R. T. Collins, O. Amidi, and T. Kanade, "An active camera system for acquiring multi-view video," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, 2002, vol. 1: IEEE, pp. I-I.

[18] S. L. Dockstader and A. M. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proceedings of the IEEE,* vol. 89, no. 10, pp. 1441-1455, 2001.

[19] A. Mittal and L. S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *International Journal of Computer Vision,* vol. 51, no. 3, pp. 189-203, 2003.

[20] Z. Wu, N. I. Hristov, T. L. Hedrick, T. H. Kunz, and M. Betke, "Tracking a large number of objects from multiple views," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009: IEEE, pp. 1546-1553.

[21] A. Gaschler, D. Burschka, and G. Hager, "Epipolar-based stereo tracking without explicit 3d reconstruction," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010: IEEE, pp. 1755-1758.

[22] K. Ni and F. Dellaert, "Stereo tracking and three-point/one-point algorithms-a robust approach in visual odometry," in *Image Processing, 2006 IEEE International Conference on*, 2006: IEEE, pp. 2777-2780.

[23] D. Stoyanov, G. P. Mylonas, F. Deligianni, A. Darzi, and G. Z. Yang, "Soft-tissue motion tracking and structure estimation for robotic assisted MIS procedures," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2005: Springer, pp. 139-146.

[24] V. I. Morariu and O. I. Camps, "Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, vol. 1: IEEE, pp. 545-552.

[25] S. Calderara, R. Cucchiara, and A. Prati, "Bayesian-competitive consistent labeling for people surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, no. 2, 2008.

[26] Y. Caspi and M. Irani, "A step towards sequence-to-sequence alignment," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2000, vol. 2: IEEE, pp. 682-689.

[27] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, no. 3, pp. 505-519, 2009.

[28] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: Establishing a common coordinate frame," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 22, no. 8, pp. 758-767, 2000.

[29] S.-N. Lim and L. Davis, "An ease-of-use stereo-based particle filter for tracking under occlusion," *Human Motion–Understanding, Modeling, Capture and Animation,* pp. 225-239, 2007.

[30] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing,* vol. 32, no. 4, pp. 270-286, 2014.

[31] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher, "Appearance-based person reidentification in camera networks: problem overview and current approaches," *Journal of Ambient Intelligence and Humanized Computing,* vol. 2, no. 2, pp. 127-151, 2011.

[32] S. Gong, M. Cristani, S. Yan, C. C. Loy, and P. Re-Identification, "Springer Publishing Company," ed: Incorporated, 2014.

[33] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys (CSUR),* vol. 46, no. 2, p. 29, 2013.

[34] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE transactions on pattern analysis and machine intelligence,* vol. 35, no. 3, pp. 653-668, 2013.

[35] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, 2011: IEEE, pp. 179-184.

[36] M. Bauml and R. Stiefelhagen, "Evaluation of local features for person re-identification in image sequences," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, 2011: IEEE, pp. 291-296.

[37] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010: IEEE, pp. 2360-2367.

[38] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, vol. 2: IEEE, pp. 1528-1535.

[39] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *Computer Vision–ECCV 2008,* pp. 262-275, 2008.

[40] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007: IEEE, pp. 1-8.

[41] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610-3617.

[42] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012: IEEE, pp. 2666-2672.

[43] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318-3325.

[44] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, 2011: IEEE, pp. 649-656.

[45] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, "Predicting deeper into the future of semantic segmentation," in *of: ICCV 2017-International Conference on Computer Vision*, 2017, p. 10.

[46] V. Vukotić, S.-L. Pintea, C. Raymond, G. Gravier, and J. C. Van Gemert, "One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network," in *International Conference on Image Analysis and Processing*, 2017: Springer, pp. 140-151.

[47] J. Van Amersfoort, A. Kannan, M. A. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," *arXiv preprint arXiv:1701.08435,* 2017.

[48] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," *arXiv preprint arXiv:1710.11252,* 2017.

[49] W. Byeon, Q. Wang, R. Kumar Srivastava, and P. Koumoutsakos, "Contextvp: Fully context-aware video prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 753-769.

[50] F. Leibfried, N. Kushman, and K. Hofmann, "A deep learning approach for joint video frame and reward prediction in atari games," *arXiv preprint arXiv:1611.07078,* 2016.

[51] R. Mahjourian, M. Wicke, and A. Angelova, "Geometry-based next frame prediction from monocular video," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017: IEEE, pp. 1700-1707.

[52] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in neural information processing systems*, 2016, pp. 613-621.

[53] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," *arXiv preprint arXiv:1802.07687,* 2018.

[54] M. Oliu, J. Selva, and S. Escalera, "Folded recurrent neural networks for future video prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 716-731.

[55] L. B. D. Tran, R. Fergus, L. Torresani, and M. Palur, "Learning spatiotemporal features with 3d convolutional networks.," presented at the Int. Conf. on Computer Vision (ICCV), 2015.

[56] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.

[57] N. Kalchbrenner *et al.*, "Video pixel networks," in *International Conference on Machine Learning*, 2017, pp. 1771-1779.

[58] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440,* 2015.

[59] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," in *Advances in neural information processing systems*, 2015, pp. 2863-2871.

[60] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: a baseline for generative models of natural videos," *arXiv preprint arXiv:1412.6604,* 2014.

[61] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4463-4471.

[62] F. A. Reda *et al.*, "Sdc-net: Video prediction using spatially-displaced convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 718-733.

[63] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017: IEEE, pp. 468-475.

[64] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, 2016: Springer, pp. 483-499.

[65] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proceedings of the IEEE international conference on Computer Vision*, 2013, pp. 3487-3494.

[66] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653-1660.

[67] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724-4732.

[68] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing,* vol. 100, pp. 86-97, 2013.

[69] S. Park and M. M. Trivedi, "Understanding human interactions with track and body synergies (TBS) captured from multiple views," *Computer Vision and Image Understanding,* vol. 111, no. 1, pp. 2-20, 2008.

[70] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*, 2011: Ieee, pp. 1297-1304.

[71] H.-Y. Lin and T.-W. Chen, "Augmented reality with human body interaction based on monocular 3D pose estimation," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2010: Springer, pp. 321-331.

[72] G. Bertasius, C. Feichtenhofer, D. Tran, J. Shi, and L. Torresani, "Learning temporal pose estimation from sparsely-labeled videos," in *Advances in Neural Information Processing Systems*, 2019, pp. 3027-3038.

[73] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal, "Identification of hybrid systems a tutorial," *European journal of control,* vol. 13, no. 2-3, pp. 242-260, 2007.

[74] Y. Ma and R. Vidal, "Identification of deterministic switched ARX systems via identification of algebraic varieties," in *International Workshop on Hybrid Systems: Computation and Control*, 2005: Springer, pp. 449-465.

[75] J. Roll, A. Bemporad, and L. Ljung, "Identification of piecewise affine systems via mixed-integer programming," *Automatica,* vol. 40, no. 1, pp. 37-50, 2004.

[76] A. L. Juloski, S. Weiland, and W. Heemels, "A Bayesian approach to identification of hybrid systems," *IEEE Transactions on Automatic Control,* vol. 50, no. 10, pp. 1520-1533, 2005.

[77] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari, "A clustering technique for the identification of piecewise affine systems," in *International Workshop on Hybrid Systems: Computation and Control*, 2001: Springer, pp. 218-231.

[78] H. Nakada, K. Takaba, and T. Katayama, "Identification of piecewise affine systems based on statistical clustering technique," *Automatica,* vol. 41, no. 5, pp. 905-913, 2005.

[79] C. Feng, C. M. Lagoa, and M. Sznaier, "Hybrid system identification via sparse polynomial optimization," in *American Control Conference (ACC), 2010*, 2010: IEEE, pp. 160-165.

[80] N. Ozay, C. Lagoa, and M. Sznaier, "Robust identification of switched affine systems via moments-based convex optimization," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, 2009: IEEE, pp. 4686-4691.

[81] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps, "A sparsification approach to set membership identification of a class of affine hybrid systems," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, 2008: IEEE, pp. 123-130.

[82] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM review,* vol. 38, no. 1, pp. 49-95, 1996.

[83] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[84] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *American Control Conference, 2001. Proceedings of the 2001*, 2001, vol. 6: IEEE, pp. 4734-4739.

[85] M. S. Lobo, M. Fazel, and S. Boyd, "Portfolio optimization with linear and fixed transaction costs," *Annals of Operations Research,* vol. 152, no. 1, pp. 341-365, 2007.

[86] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in neural information processing systems*, 2009, pp. 2080-2088.

[87] D. Henrion and A. Garulli, *Positive polynomials in control*. Springer Science & Business Media, 2005.

[88] J. B. Lasserre, *Moments, positive polynomials and their applications*. World Scientific, 2009.

[89] J. B. Lasserre and M. Putinar, "Positivity and optimization for semi-algebraic functions," *SIAM Journal on Optimization,* vol. 20, no. 6, pp. 3364-3383, 2010.

[90] W. Liu, Abhishek Sharma, Octavia Camps, and Mario Sznaier, "DYAN: A Dynamical Atoms-Based Network for Video Prediction," presented at the European Conference on Computer Vision (ECCV), Munich, Germany, 2018.

[91] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing,* vol. 13, no. 4, pp. 600-612, 2004.

[92] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586-595.

[93] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," *arXiv preprint arXiv:1605.08104,* 2016.

[94] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion GAN for future-flow embedded video prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1744-1752.

[95] Z. Hao, X. Huang, and S. Belongie, "Controllable video generation with sparse trajectories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7854-7863.

[96] H. Gao, H. Xu, Q.-Z. Cai, R. Wang, F. Yu, and T. Darrell, "Disentangling propagation and generation for video prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9006-9015.

[97] Y. Zhang, Yin Wang, Octavia Camps, and Mario Sznaier, "Key Frame Proposal Network for Efficient Pose Estimation in Videos," in *European Conference on Computer Vision (ECCV)*, 2020.

[98] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[99] X. Nie, Y. Li, L. Luo, N. Zhang, and J. Feng, "Dynamic kernel distillation for efficient pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6942-6950.

[100] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466-481.

[101] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293-1301.

[102] U. Iqbal, M. Garbade, and J. Gall, "Pose for action-action for pose," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017: IEEE, pp. 438-445.

[103] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *European Conference on Computer Vision*, 2016: Springer, pp. 728-743.

[104] J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4220-4229.

[105] Y. Luo *et al.*, "Lstm pose machines," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5207-5215.

[106] W. Liu *et al.*, "Towards visually explaining variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8642-8651.

[107] C. Dicle, O. I. Camps, and M. Sznaier, "The way they move: Tracking multiple targets with similar appearance," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2304-2311.

[108] Y. Cheng, Y. Wang, M. Sznaier, and O. Camps, "Subspace clustering with priors via sparse quadratically constrained quadratic programming," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5204-5212.

[109] Y. Wang, M. Sznaier, O. Camps, and F. Pait, "Identification of a class of generalized autoregressive conditional heteroskedasticity (GARCH) models with applications to covariance propagation," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*, 2015: IEEE, pp. 795-800.

[110] M. Ayazoglu, B. Yilmaz, M. Sznaier, and O. Camps, "Finding causal interactions in video sequences," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3575-3582.

[111] F. Xiong, Y. Cheng, O. Camps, M. Sznaier, and C. Lagoa, "Hankel based maximum margin classifiers: A connection between machine learning and wiener systems identification," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, 2013: IEEE, pp. 6005-6010.

[112] L. Lo Presti, M. La Cascia, S. Sclaroff, and O. Camps, "Hankelet-based dynamical systems modeling for 3D action recognition," *Image and Vision Computing,* vol. 44, pp. 29-43, 2015.

[113] Y. Cheng, Y. Wang, and M. Sznaier, "A convex optimization approach to semi-supervised identification of switched ARX systems," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 2014: IEEE, pp. 2573-2578.

[114] N. Ozay, M. Sznaier, and C. Lagoa, "Convex certificates for model (in) validation of switched affine systems with unknown switches," *IEEE Transactions on Automatic Control,* vol. 59, no. 11, pp. 2921-2932, 2014.

[115] L. Lo Presti, M. La Cascia, S. Sclaroff, and O. Camps, "Gesture modeling by hanklet-based hidden markov model," in *Asian Conference on Computer Vision*, 2014: Springer, pp. 529-546.

[116] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps, "Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4498-4507.

[117] O. Camps, M. Sznaier, and X. Zhang, "Convex behavioral model (in) validation via Jensen-Bregman divergence minimization," in *American Control Conference (ACC), 2016*, 2016: IEEE, pp. 4575-4579.

[118] S. Asghari-Esfeden, M. Sznaier, and O. Camps, "Dynamic Motion Representation for Human Action Recognition," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 557-566.

[119] X. Zhang, M. Sznaier, and O. Camps, "Efficient Identification of Error-in Variables Switched Systems Based on Riemannian Distance-Like Functions," presented at the IEEE Conference on Decision and Control (CDC), 2018.

[120] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European conference on computer vision*, 2014: Springer, pp. 1-16.

[121] M. Gou, X. Zhang, A. Rates-Borras, S. Asghari-Esfeden, M. Sznaier, and O. Camps, "Person re-identification in appearance impaired scenarios," *arXiv preprint arXiv:1604.00367,* 2016.

[122] O. Camps *et al.*, "From the lab to the real world: Re-identification in an airport camera network," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 27, no. 3, pp. 540-553, 2017.

[123] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," *arXiv preprint arXiv:1605.09653,* 2016.

[124] Y. Cheng, J. A. Lopez, O. Camps, and M. Sznaier, "A convex optimization approach to robust fundamental matrix estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2170-2178.

[125] M. Sznaier, M. Ayazoglu, and T. Inanc, "Fast structured nuclear norm minimization with applications to set membership systems identification," *IEEE Transactions on Automatic Control,* vol. 59, no. 10, pp. 2837-2842, 2014.

[126] M. Sznaier and O. Camps, "Uncertainty and Robustness in Dynamic Vision," *Encyclopedia of Systems and Control,* pp. 1493-1499, 2015.

[127] Y. Wang, C. Dicle, M. Sznaier, and O. Camps, "Self scaled regularized robust regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3261-3269.

[128] B. Yılmaz and M. Sznaier, "Efficient identification of Wiener systems using a combination of atomic norm minimization and interval matrix properties," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*, 2015: IEEE, pp. 109-114.

[129] C. Dicle, B. Yilmaz, O. Camps, and M. Sznaier, "Solving Temporal Puzzles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5896-5905.

[130] Y. Wang, O. Camps, M. Sznaier, and B. R. Solvas, "Jensen Bregman LogDet Divergence Optimal Filtering in the Manifold of Positive Definite Matrices," in *European Conference on Computer Vision*, 2016: Springer, pp. 221-235.

[131] Y. Wang, M. Sznaier, and O. Camps, "A super-atomic norm minimization approach to identifying sparse dynamical graphical models," in *American Control Conference (ACC), 2016*, 2016: IEEE, pp. 1962-1967.

[132] M. Sznaier and O. Camps, "Sos-rsc: A sum-of-squares polynomial approach to robustifying subspace clustering algorithms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8033-8041.

[133] M. Gou, F. Xiong, O. Camps, and M. Sznaier, "MoNet: Moments Embedding Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3175-3183.

[134] X. Zhang, B. Ozbay, M. Sznaier, and O. Camps, "Dynamics Enhanced Multi-Camera Motion Segmentation from Unsynchronized Videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4668-4676.

[135] Y. Z. A. Islam, D. Yin, O. Camps, and R. J. Radke, "Correlating belongings with passengers in a simulated airport security checkpoint," presented at the 12th International Conference on Distributed Smart Cameras, 2018.