F3-F: A System Theoretic Approach to Robust Detection Of Potential Threats From Video

Abstract— This research aims at a substantial enhancement of the ability to conduct autonomous, video based, persistent intelligent surveillance, and reconnaissance and threat assessment in highly uncertain, adversarial scenarios such as urban environments. The main idea is the use of operator theoretic and convex analysis methods to recast several key sub-problems arising in this context -tracking, dynamic appearance, and activity recognition -into a finite dimensional convex optimization that can be efficiently solved.

I. PARTICIPANTS

<table>
<thead>
<tr>
<th>Faculty/Staff</th>
<th>Name</th>
<th>Title</th>
<th>Institution</th>
<th>Email</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Octavia Camps</td>
<td>Professor</td>
<td>NEU</td>
<td><a href="mailto:camps@coe.neu.edu">camps@coe.neu.edu</a></td>
<td>617-373-4663</td>
</tr>
<tr>
<td></td>
<td>Mario Sznaier</td>
<td>Professor</td>
<td>NEU</td>
<td><a href="mailto:mksznaier@coe.neu.edu">mksznaier@coe.neu.edu</a></td>
<td>617-373-5364</td>
</tr>
<tr>
<td></td>
<td>Gilead Tadmor</td>
<td>Professor</td>
<td>NEU</td>
<td><a href="mailto:tadmor@coe.neu.edu">tadmor@coe.neu.edu</a></td>
<td>617-373-5277</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Students</th>
<th>Name</th>
<th>Degree Pursued</th>
<th>Institution</th>
<th>Email</th>
<th>Intended Year of Graduation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Teresa Mao</td>
<td>MSEE</td>
<td>NEU</td>
<td><a href="mailto:Mao.t@husky.neu.edu">Mao.t@husky.neu.edu</a></td>
<td>2011</td>
</tr>
<tr>
<td></td>
<td>Caglayan Dicle</td>
<td>Ph. D.</td>
<td>NEU</td>
<td><a href="mailto:cdicle@gmail.com">cdicle@gmail.com</a></td>
<td>2013</td>
</tr>
<tr>
<td></td>
<td>Oliver Lehmann</td>
<td>Ph. D.</td>
<td>NEU</td>
<td><a href="mailto:Lehmann.o@husky.neu.edu">Lehmann.o@husky.neu.edu</a></td>
<td>2014</td>
</tr>
<tr>
<td></td>
<td>Mustafa Ayazoglu</td>
<td>PhD</td>
<td>NEU</td>
<td><a href="mailto:hacettepeli.muhendis@gmail.com">hacettepeli.muhendis@gmail.com</a></td>
<td>2012</td>
</tr>
<tr>
<td></td>
<td>Binlong Li</td>
<td>PhD</td>
<td>NEU</td>
<td><a href="mailto:li.b@neu.edu">li.b@neu.edu</a></td>
<td>2012</td>
</tr>
</tbody>
</table>

II. PROJECT OVERVIEW AND SIGNIFICANCE

This research effort aims at a substantial enhancement of our ability to exploit surveillance camera networks to predict and isolate threats from explosive devices in heavily crowded public spaces, and to guide complementary detection modalities, subsequent to a threat alert. At its core is a novel approach, stressing dynamic models as a key enabler for automatic, real time interpretation of what is currently an overwhelming profusion of video data streams. It includes both theory development in an emerging new field, and an investigation of implementation issues. As part of the ALERT center of excellence, it will synergistically link with several concurrent efforts at NU, and with NU’s partners, both in academia and industrial R&D.

Video-based methods have an enormous potential for providing advance 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 to a very large extent 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, the critical impediment to fully realize this potential is the absence of reliable technology for ro-
bust, 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 vs. dangerous behaviors, make automatic threat detection extremely challenging. Indeed, state-of-the-art turnkey software, such as that in use by complementary projects at NU, heavily relies on human operators, which, in turn, severely limits the scope of its use.

The research is 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 information content on targeted behavior in dynamic models. Drawing on solid theoretical foundations, robust system identification and adaptation methods, along with model (in)validation tools, will yield quantifiable characterization of threats and benign behaviors, provable uncertainty bounds, and alternatives for viable explanations of observed activities. The resulting systems will integrate real time data from multiple sources over dynamic networks, cover large areas, extract meaningful behavioral information on a large number of individuals and objects, and strike 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.

III. RESEARCH AND EDUCATION ACTIVITY

A. State-of-the-Art and Technical Approach

Conceptual foundation: Dynamic models as an information encoding paradigm for data streams.

The assumption that encoded information is manifested by (low dimensional) cross-correlations, is standard, including extensions to spatio-temporal settings. Our basic premise, as simple as it is powerful, is that a leap forward is enabled by the encapsulation of temporal (and some spatial) correlations in dynamic models. It amounts to a reasonable “localization” hypothesis for temporal correlations, and is a given in mechanical and biological motion (including human). Embedding problems in the conceptual world of dynamical systems makes available an extremely versatile and powerful knowledge base and ensemble of methods, including analysis and design tools and quantifiable comparisons. As a simple illustration, note that while a comparison of two dynamical systems may be as simple as solving a low dimensional linear matrix inequality (LMI), the comparison of high-dimensional data streams necessitates temporal overlap and requires a computationally expensive synchronization search.

Technical foundation: Operator theoretic embedding.

Having translated problems arising in the context of activity analysis and threat detection into systems theoretic terms, the proposed approach focal point is the realization that a shared underlying property is the ability to reformulate them as constrained operator interpolation problems, efficiently solvable with commercially available tools. While tools developed to date might possibly not yet resolve all open issues for each of the sub-problems, the proposed framework provides the foundation for building an unified set of tools to address robustness and performance issues across a wide range of technical problems addressed in this project as briefly summarized below.

I. Robust Operator Based Tracking. The ability to persistently track and disambiguate is a key enabler for both activity recognition, as discussed below, and suspect identification at a distance. However, this process is far from trivial in urban environments: barriers include missing data due to occlusion; target appearance changes, not always visible and compounded by the (potential) existence of multiple targets with similar appearance; and the high dimensionality intrinsic to video streams. The key idea to overcome these barriers is to model the temporal evolution of relevant video image features as the output of a dynamical system, driven by a stochastic input. Tracking targets in a sequence of frames is then stated as an operator theoretic robust identification/model (in)validation problem using a priori information derived e.g. from physical con-
straints, long term crowd motion models (see below), etc. As recently shown by the Co–PIs in the simpler case of linear target dynamics and few features, combining the identified dynamics with a particle filter leads to both considerable robustness improvement and substantial computational complexity reduction, compared with prevalent methods. Realistic situations require addressing the issues of high dimensionality of the data, appearance changes and nonlinear, possibly time varying, dynamics. We plan to address these using manifold discovery methods, such as Locally Linear Embeddings (LLE), to map image features to points on low dimensional manifolds where dynamics are locally linear time invariant, effectively decoupling appearance from intrinsic dynamics. Generalized interpolation theory reduces the identification of local linear dynamics, along the manifold, to a convex Linear Matrix Inequalities (LMIs) optimization problem. The transition between local models can be detected by an LMI model invalidation feasibility step. Figure 1 illustrates this approach, showing close agreement between the actual time traces over the manifold representing a walking sequence, and the positions predicted using a locally linear, dynamic model, identified from training data, and successful tracking in the presence of occlusion.

II. The multiple camera case. Here the goal is to exploit data from multiple cameras to cooperatively enhance tracking capabilities, including (i) disambiguation between similar targets, (ii) sustain occlusion of some participating cameras, and (iii) exploit multiple perspectives for enhanced target appearance and behavior analysis. Inputs from several (roughly) registered cameras are incorporated as sources for cumulative geometric constraints that, once again, translate into additional convex constraints for associated identification/invalidation problems. The entailed increased computational complexity will be alleviated using data streams projections onto the low dimensional manifolds, associated with each viewpoint as discussed in I. above, rather than raw video data. The manifolds will be dynamically registered, using robust identification to extract dynamical systems (hence input-output operators) that mediate between trajectories across different manifolds. These operators are the key enablers to the sought enhanced capabilities: Coordinating data from participating cameras will reduce uncertainty through the intersection of uncertainty sets, will associate appearance descriptors of a single object from several perspectives, and will predict the location of a target, momentarily occluded from one camera, using inputs from other cameras.

III. Robust Activity Recognition. Activities are modeled as stationary stochastic processes. Expert rules and adaptive learning of prototype normal and threatening human and object (e.g., abandoned luggage or automobiles) activities will be translated into a model database through training data. Solving a sequence of model invalidation problems, the database will be interrogated, to find the closest match. We have recently tested this approach in a simplified, low noise version of a human gait recognition problem where it achieved a success rate of 97% in real images. Realistic, noisy data, leads to a non–convex, NP hard problem. We plan to address this by using a moments-based approach that provides (in)validation certificates in polynomial time. As noted earlier, a complete a priori cataloguing of benign behavior is not feasible, and real-time crowd

![Figure 1: (a) 3 dimensional representation of a walking sequence by Locally Linear Embedding (LLE). (b) Dynamics Identification by Caratheodory-Fejer (CF) interpolation and prediction for next 38 frames. (c) Handling occlusion by a background object.](image-url)
activity model extraction will be used – along with successive model invalidation – to detect abrupt changes of behavior. Sites of such changes will become the subject of heightened scrutiny, such as a turn to individual tracking and motion analysis, accompanied by reallocation of observation cameras and computational resources.

IV. Multiple Targets. It is vital to have the ability of tracking and disambiguating multiple moving objects (as e.g., in crowd activity modeling and analysis). Thus, it is of interest to identify (i) the number of objects, (ii) their individual dynamics and, (iii) assign points in the image to each object. Existing approaches employ factorizations of matrices formed by time traces of positions in the image of tracked features. These methods lead to elegant, computationally efficient solutions, yet they are fragile, sensitive to noise and missing correspondences, and have difficulties disambiguating objects that share some motion components such as subgroups of moving people in a crowded area. The proposed approach to alleviate these difficulties is to base segmentation on the consistent identification of distinct dynamic models for distinct objects. This amounts, once again, to an operator identification problem, involving structural rank constraints. These conditions precisely encapsulate the physical spatio–temporal constraints that recover the geometry of the scene, exploiting objects dynamics to accurately track them over time, even in the presence of occasional occlusion.

V. Crowd motion modeling. Considering densely crowded areas as natural terrorist targets, the combination of sheer numbers, partial occlusion, and the proximity between people, might make tracking and analyzing the motion of each individual in some large inspected area (e.g., the already sitting spectators in a sport arena) prohibitively difficult. Our team’s approach is therefore to pass to mass flow models at the macroscopic levels. A succession of increased resolution refinements from mass through group to the individual levels of interrogation will be triggered by anomaly detection at suspect areas, before a human observer is alerted. Granular fluid models provide the framework for mass flow models. Such models have evolved from observations regarding crowd behavior, dating to the early 1970’s. Appeals to the conceptual and technical framework of fluid dynamics and statistical mechanics have grown rapidly over the past decade and are now part and parcel of crowd motion analysis and planning, including response programs, evacuation plans and the architectural design of high traffic public areas.

B. Major Contributions

In this section we briefly review the main contributions made to the ALERT project by our team during 2009 - 2011. A more complete description of these results is available in the papers listed at the end of this report. Copies of these papers, as well as several demos are available at the Robust Systems Laboratory web page: http://robustsystems.ece.neu.edu/

(a) Identification of Wiener systems. Tracking complex targets such as humans that can undergo substantial appearance changes requires sophisticated models that capture the time evolutions of target templates and shape -- i.e. time evolution of shape moments, contour descriptors, etc. However, moving beyond just a few simple descriptors such as color histograms or the width and height of the target, requires addressing the issues of high computational costs, due to the poor scaling properties of LMI based identification algorithms, and nonlinear appearance changes. We plan to address this challenge through the use of nonlinear dimensionality reduction techniques to map the data to a lower dimensional manifold where the identification/tracking is performed. This idea is appealing since it leads to a separation type principle, decoupling the motion dynamics of the target and its appearance changes: motion of the points in the manifold correspond to the intrinsic dynamics of the target, while the mapping from the manifold to descriptors accounts for appearance changes. Further, since this last mapping does not have dynamics, it can be modeled as a static nonlinearity, leading naturally to the problem of identification of Wiener systems. As part of this project, we have established that this problem is NP-hard [1]. Thus, in order to obtain computationally tractable solutions, amenable to real time implementations, we have developed risk-adjusted relaxations, whose complexity only grows polynomially with the size of the data [2]. We are also currently exploring a deterministic alternative,
based on recent results on polynomial optimization using the method of moments [3].

(b) Identification of switched systems. Cases involving a transition between different models, e.g. different motion modalities or substantially different appearance, can be modeled as a mode-transition in a piecewise affine switched system. While identification of these systems has been the subject of intense research in the past few years, a comprehensive framework is still lacking. Indeed, in the case of noisy measurements, existing methods lead to computationally hard problems with poor scaling properties. We have obtained preliminary results indicating that recasting the problem into a sparsification form can circumvent these difficulties and exploiting recent results from convex optimization on obtaining convex envelopes[4, 5, 6, 7, 8]. In addition, we have developed efficient methods for detecting mode switches, based on estimating the rank of suitable Hankel matrices [9, 10, 11].

(c) Constrained interpolation of noisy data. In most surveillance scenarios only partial data is available, due for instance to occlusion, transmission blackouts or limited sensing capabilities. In these situations, it is of interest to estimate the missing data, for instance in order to perform data association (e.g. in the context of tracklet stitching) or to uncover correlations mediated by the missing elements. We have shown that this interpolation can be reduced to a rank minimization problem, which in turn (due to its Hankel structure) can be efficiently solved using convex relaxations [9, 10]. Figure 2 shows an example where tracklets resulting from occlusion in a video of people walking on a parking lot (provided by Siemens Corporate Research) are stitched together to preserve the identity of the target in spite of the occlusion.

(d) Fast event detection in video sequences. We have developed an efficient algorithm based upon recasting the problem as finding modes-switches in a piecewise affine system and using the theoretical tools developed in item (a) above. Figure 3 shows the effectiveness of this approach in detecting the transition between two different motion modalities: walking, and stopping to remove some clothing in a video capturing the suspect of a bombing attack attempted in New York City in 2007. Further, combination with the data interpolation tools of (b), allows for uncovering events occurring while the target is occluded, as illustrated in Figure 4 [9, 10].

(e) Contra-flow in the exit detection. In this application, the goal is to flag an event when an individual...
enters a sterile area in an airport terminal through an exit lane bypassing security checks and to tag the individual for future tracking. While it is possible to detect such an event by tracking everybody and flagging targets using the method described above in (d), such an approach would unnecessarily severely tax computing resources since in this case the event of interest is well specified. Thus, instead, we have designed an algorithm to simply detect areas where localized (both spatially and temporarily) motion is against a pre-defined direction. We have recently applied this approach to videos provided by TSA and have obtained very promising results. Figure 5 shows an example of a frame where an individual going against the flow is detected. We are currently in the process of deploying the algorithm at Cleveland Airport for real time, on-site testing.

(f) Fluid models for event detection in crowds. Our roadmap to alleviate the prohibitive complexity of anomaly detection in dense crowds includes developing (i) a flow-based modeling paradigm, suitable for real time identification in large crowds, (ii) fitting model reduction methods, (iii) translating anomaly detection to parameter estimation in reduced order models, and (iv) couple the mass flow and discrete modeling levels. Each of these tasks involves both conceptual issues and intricate technicalities, implied by the inherent complexity of distributed parameter models. Previous efforts focused on the development of viable global-modes (Galerkin) model reduction methods, as essential enablers, and as means to narrow down feasible flow-based models that could be employed. Here we leveraged technical overlaps with other fluid flow applications to address the critical issues of unsteady base flows, moving boundaries, deformable modal expansions, and the evolution of statistical (“thermodynamic”) flow properties [12, 13, 14, 15, 16, 17, 18].

We are undertaking the development of a mass-flow modeling paradigm suitable for real-time analysis of dense crowd behavior. Work over the last 6 month postulates traffic of N-populations, each characterized by a preferred destination (e.g., train platforms, station exits, etc.). A compressible, multi-phase variant of the Navier-Stokes equations was chosen as a target framework, with the goal to capitalize on simplifications enabled by the continuity assumption (as compared with existing particle-based simulation models, inspired by statistical mechanics and far too complex for the task). Necessary points of departure from classical fluid dynamics include: (i) Pressure singularity at a finite, density capacity, (ii) potential-gradients and “ground friction”, realized by a linear filter term, are used to jointly represent preferred routes and velocities of sub-populations, (iii) stochastic forcing represents natural and density-induced variations in velocity and orientations, and (iv) the mixing of sub-population, down to the microscopic level, to remove the complexity of inter-species bubbles and boundaries. Our starting point for the momentum equation of the ith population is:

\[
\partial_t (\rho_i \mathbf{u}_i) = \frac{\rho_i}{\sum_j \rho_j} \left\{ \mu \Delta \mathbf{u}_i - \nabla_x \left( \frac{\overline{\rho_{\max}}}{\rho_{\max} - \rho} \right) \right\} + \kappa_i (\mathbf{u}_i - \overline{\mathbf{u}_i}) + \eta_i, \quad i = 1, \ldots, N.
\]

Here \( x, \mathbf{u}, \rho \) and \( \eta \) indicate the location in space, velocity, mass density, and stochastic force, an over-line indicates the multi-population average, and the pressure is substituted by the saturated density ratio. We have
implemented this model using OpenFOAM®, an open source computational fluid dynamics toolbox. This effort begun with the construction, structural refinement of a working simulation model, and continues with the verification of the model’s capacity to reproduce generic crowd behavior patterns. Figure 6 illustrates the spontaneous generation of a two-lanes flow when two populations move through a corridor in opposite directions.

(g) Robust estimation under \( l_\infty \) bounded disturbances. Traditional noise models often do not capture key features of the problems of interest here. As a simple example, noise in images should be bounded. While in principle this feature can be captured using truncated distributions, the resulting problems are computationally hard. In addition, the resulting filters are fragile to errors in estimating the parameters of the distribution. To circumvent this difficulty we are developing a new framework for robust estimation in the presence of unknown-but-bounded noise. Preliminary results indicate that using a concept similar to superstability leads to robust filters that can be synthesized by simply solving a linear programming problem [19]. A salient feature of this framework is that it explicitly allows for trading off filter complexity against worst-case estimation error.

(h) Activity Recognition. Current approaches to modeling and recognizing actions of single actors [20] rely on local features extracted at the frame level and lack of strong relations among features across frames or assume a dynamical model, which is often too simplistic, that must be estimated from extensive experimental data, often corrupted by noise. In this research, we propose a time-series approach for activity recognition that, in contrast with previous approaches, requires neither assuming nor identifying a dynamical model [21]. Instead, we simply hypothesize that the temporal data is the output trajectory of an underlying, unknown linear (possibly slowly varying) dynamical system. In this context, different realizations of the same activity correspond to trajectories of the same system in response to different initial conditions. Exploiting the fact, derived from realization theory, that these trajectories are constrained to evolve in the same subspace spanned by their corresponding Hankel matrices (directly determined from the experimental data), allows for measuring the similarity between activities by simply computing the angle between the associated subspaces. To further increase robustness to noise, we propose to use a support vector machine (SVM) in conjunction with a discriminative canonical correlation [22] transformation to simultaneously decrease the inter-class and increase the intra-class distances. The proposed approach was tested with the KTH database [23] which consists of six types of human activities (walking, running, boxing, hand waving, hand clapping,
and jogging) performed by 25 subjects in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothing, and indoors. The proposed approach achieved an overall accuracy of 93.68%, which is higher than the current state of the art. Figure 7 shows a diagram illustrating all the required steps of the proposed method and a comparison of the achieved performance against competing methods.

(i) Cross-view Activity Recognition. The problem of recognizing an activity from a viewpoint different to the one used during training is considerably less studied. Some approaches rely on geometric constraints [24], body joints detection and tracking [25, 26], and 3D models [27, 28, 29, 30]. More recent approaches transfer features across views [31, 32] or use self-similarities as quasi-view invariant features [33, 34]. However, the performances for these approaches are still far below the performances achieved for single view activity recognition. In [35] we introduced a new type of feature, the “Hankelet” that captures dynamic properties of short tracklets. Hankelets are defined as the Hankel matrices associated with short tracklets as the ones illustrated in Figure 8. While Hankelets do not carry any spatial information, they bring invariant properties to changes in viewpoint that allow for robust cross-view activity recognition, i.e. when actions are recognized using a classifier trained on data from a different viewpoint. Our experiments on the IXMAS dataset show that using Hankelets improves the state of the art performance by over 20% as shown in Table 1.

(j) Contextually Abnormal Activity Sequence Recognition. We have modeled activities as second order stationary processes reducing the recognition problem to a model (in)validation one. Specifically, given a sequence of frames of an unknown activity, recognition can be achieved by interrogating models in a database as to whether this sequence is compatible with each model and its associated uncertainty description. A difficulty here is that a single activity can consist of the concatenation of several sub-activities of various lengths. For instance a “normal” activity could consist of walking for two minutes, standing for one, and then resuming walking. On the other hand, while bending over by itself could be benign, a combination walking,
stopping, bending over and resuming walking could indicate an abnormal activity where an explosive threat is left behind. We proposed to detect contextually abnormal activity sequences, by recasting the problem into that of (in)validating the output of a switched system, where each mode corresponds to a given, known to be benign activity. In this context, any sequence of activities that cannot be shown to have been generated by an admissible switching sequence of this system (e.g. an overall benign concatenation of benign sub-activities) triggers an increase level of scrutiny. While the resulting problem is in principle NP-hard, our results [37] indicate that tractable convex relaxations can be obtained by appealing to recent results in the fields of polynomial optimization and semi-algebraic geometry. An example of these results is shown in Figure 9.

(k) Recovering 3-D geometry from 2-D video. Structure from motion techniques allow the recovery of 3D geometry from video data. However, existing approaches can only recover structure up to a (frame-dependent) perspective transformation. Thus, they are of limited use when trying to detect suspicious geometries. In contrast, we introduced a new approach that implicitly exploits the temporal ordering of the frames [38, 39]. This new method leads to a provably correct algorithm to find Euclidean structure (up to a single scaling factor) without the need to alternate between projective depth and motion estimation, estimate the Fundamental matrices or assume a camera motion model. Finally, the proposed approach does not require an accurate calibration of the camera. The accuracy of the algorithm is illustrated with examples using synthetic and real data in Figures 10 and 11, respectively.

(l) Data Integration from Multiple Cameras: Tracking a target using multiple cameras can increase robustness.

Figure 9: Detecting contextually abnormal activity sequences as a model (in)validation problem. Top: Walk-Wait-Walk (normal, not invalidated). Middle: Running (contextually abnormal, invalidated). Bottom: Walk-Jump (contextually abnormal, invalidated).

Figure 10. (a) Frames 1, 5 and 10 of a teapot sequence. (b)-(d) 3D structure recovered using our approach and two competing methods, respectively. Note that the proposed approach does not distort the teapot.
against occlusion and clutter since, even if the target appears largely occluded to some sensors, the system can recover by using the others. Furthermore, examining data from spatially distributed cameras can reveal activity patterns not apparent to single or closely clustered sensors. However, in order for a multi-camera tracking system to take full advantage of the additional information available from its multiple sensors, it must maintain consistent identity labels of the targets across views and recover their 3D trajectories. Previous approaches to the "correspondence across views" problem include matching features, using camera calibration information or making assumptions about planar surfaces. However, it can be difficult to find matching features across significantly different views, camera calibration information is not always available and planar world hypothesis can be too restrictive.

We propose a new approach to the problem of finding correspondences across views that does not require feature matching, camera calibration or planar assumptions [40]. The key insight of the proposed method is the fact that, under mild conditions, the 2D trajectories of the target in the image planes of each of the cameras are constrained to evolve in the same subspace. This observation allows for identifying, at each time instant, a single (piecewise) linear model that explains all the available 2D measurements. In turn, this model can be used in the context of a modified particle filter to predict future target locations. In the case where the target is occluded to some of the cameras, the missing measurements can be estimated using the facts that they must lay both in the subspace spanned by previous measurements and satisfy epipolar constraints. Hence, by exploiting both dynamical and geometrical constraints the proposed method can robustly handle substantial occlusion, without the need for performing 3D reconstruction, calibrated cameras or constraints on sensor separation. The performance of the proposed tracker is illustrated in Figure 12 (a) with a challenging example involving two cameras with very different viewpoints tracking a pedestrian in a train terminal under substantial occlusion in one of the views. In contrast, combining only geometrical information from the two cameras without exploiting dynamic invariance across views leads the tracker to get distracted by the occluders as shown in Figure 12 (b).

(m) Dynamics-based dimensionality reduction. The problem of finding low complexity representations of high-dimensionality data is ubiquitous in computer vision and pattern recognition. Applications include, among others, appearance-based tracking and activity recognition from video data. A common feature of dimensionality reduction methods is that, while they exploit the local spatial topology of the data, they stop short of taking full advantage of the available dynamical information, encapsulated in its temporal ordering. Thus, the resulting representations are not necessarily the most compact ones. In addition, neglecting
dynamical information can lead to embeddings that are fragile in the presence of outliers or missing data. In [41] we propose a new manifold embedding algorithm (LDE) for dynamical sequences that exploits both spatial and temporal information in order to obtain the simplest possible dynamical representation of the data. Briefly, the main idea is to model the data as the output of an underlying Wiener system of the form shown in Figure 13 consisting of the cascade of a piece-wise linear autoregressive model (generating the manifold trajectories) and a possibly time-varying nonlinearity that maps the low dimensional manifold representation back to the original data space. As illustrated in the paper, constraining target manifolds to those spanned by feasible trajectories of the Wiener system endows the proposed method with robustness against outliers and allows for handling missing data. Further, in this context the complexity of the representation is given by a combination of the order (e.g. the number of coefficients) of the regressor and the (local) dimension of the embedding manifold. By appealing to concepts from realization theory, the problem of minimizing this complexity can be reduced to a rank minimization form and efficiently solved using recently proposed convex relaxations. Figure 14 shows an example illustrating the resiliency of the proposed method to outliers. The figure shows the result of applying LDE to a video sequence from the KTH database, where the location of the actor in the 40th frame is shifted with respect to the location of the actor in other frames due to unintentional camera jitter. As shown in Figure 14, the outlier has minimal influence on the structure of the obtained manifold.

(n) Fast Algorithms for Structured Robust Principal Component Analysis. A large number of problems arising in computer vision can be reduced to the problem of minimizing the nuclear norm of a matrix, subject to additional structural and sparsity constraints on its elements. Examples of relevant ap-
Applications include, among others, robust tracking in the presence of outliers, manifold embedding, event detection, inpainting and tracklet matching across occlusion. In principle, these problems can be reduced to a convex semi-definite optimization form and solved using interior point methods. However, the poor scaling properties of these methods limit the use of this approach to relatively small sized problems. In [42] we show that structured nuclear norm minimization problems can be efficiently solved by using an iterative Augmented Lagrangian Type (ALM) method that only requires performing at each iteration a combination of matrix thresholding and matrix inversion steps. The proposed algorithm results in a substantial reduction of computational time and memory requirements when compared against interior-point methods, opening up the possibility of solving realistic, large sized problems. Figure 15 shows the results of applying the proposed algorithm to remove outliers from trajectories that are 250 frames long, manually corrupted with outliers added at random locations with probability 0.2. As illustrated there, our algorithm (SRPCA) was able to recover the original trajectories in about 25 seconds. On the other hand, this example could not be solved using a standard SDP solver in a computer with 24GB of RAM due to insufficient memory.

IV. FUTURE PLANS

This section provides technical detail of the proposed research plan for next year.

**Sustained Distributed Tracking under Abrupt Appearance and Motion Changes in Large Public Spaces:** A challenge in tracking complex targets such as humans stems from the fact that these targets can undergo simultaneous, abrupt changes in both appearance and motion modalities (for example, a walking human that discards an outer piece of clothing and starts running). Furthermore, monitoring large public spaces requires the use of multiple cameras that often have little or no overlapping field of view. This brings additional challenges since the targets can have very different appearance from different viewpoints and it is hard to predict when they will be visible at a given camera field of view. During 2012, we plan to address these issues through a combination of Receding Horizon, Incremental Locally Linear Embeddings (ILLE) techniques, and context support features. This combination is expected to provide adaptation to both time varying dynamics and abrupt appearance changes. Finally, in addition to enabling sustained tracking under these conditions, we expect that this approach will allow for efficiently detecting appearance changes by simply computing the distance between successive embeddings.

**Dynamics-based Causal Interactions Detection:** The goal is to detect whether the actions of a given target can be explained in terms of the past actions of other agents. We propose to solve this problem by recasting it into a directed graph topology identification, where each node corresponds to the observed motion of a given target, and each link indicates the presence of a causal correlation. Formulating the problem in this way...
framework leads to a block sparsification problem that we plan to solve using a Group-Lasso type approach combined with a re-weighted heuristic.

**Multiple Actors Activity Recognition:** We are currently working on extending our work on activity recognition to scenarios where multiple people interact with each other in cluttered video scenes. This is a very challenging task due to the complexity of the scenes, inherent occlusions due to the interactions between the actors, and because in some cases actors may act in a coordinated manner but at a distance from each other. We plan to address these challenges by using localized spatio-temporal features such as histogram and average of gradients (HOG/AOG) tracked over time and by exploiting their temporal correlation through the use of their underlying dynamics. Preliminary results testing the proposed approach using data from the very challenging TV interaction database [43] are very promising. On a set of 100 short clips from TV sitcoms with two types of human interactions (hand shaking and high-five), the proposed method achieved an accuracy of 68%, which is significantly better than the state of the art (54.45%).

**Crowd Motion Modeling:** We are currently conducting parametric studies, aimed to explore the validity envelope, need for refinement and parameter ranges of our stipulated mass-flow model structure. These studies include increasingly complex geometries of inter-connecting walkways, larger numbers of sub-populations, density variations and stochastic stops and turns. In parallel, we plan to shortly begin to develop video-based parameter estimation, and the real-time adaptation and tracking of such parameters, in Lagrangian (group-centered) and Eulerian (location centered) frameworks. Using the values and dynamics of model parameters and lumped state properties as anomaly indicators, the immediate purpose is to determine typical length and time scales for generic variations in targeted test cases: Dense sidewalks, street crossings, train stations, etc. Model reduction methods we previously developed for single species flows will then be adapted to multi-phase crowd flows, with the goal of using their states as lumped indicators of the distributed state.

V. **LEVERAGING OF RESOURCES**

O. Camps led the Imaging component of VOTERS: Versatile Onboard Traffic Embedded Roaming Sensors, a $9 Million grant awarded to Northeastern and partners by NIST.

The goal of this project is to develop a systemic approach to the problem of monitoring and maintenance of civil infrastructure of roads and interstate highways. The imaging component helps to register measurements from heterogeneous sensors through video of urban areas captured from cameras mounted on moving vehicles. An important component of this module is the ability to robustly detect and track features in the presence of clutter and occlusion.


VI. **DOCUMENTATION**

A. **Publications**


3. L. Mirkin and G. Tadmor, “Imposing FIR Structure on H_2 Preview Tracking and Smoothing Solutions,”


B. Technology Transfer

- During 2010, Octavia Camps worked with Siemens Corporate Research in the project “Intelligent Pedestrian Surveillance Platform Program”, an extension to the Suicide Bomber Detection Program Technology, funded by DHS S&T Explosives Division. As part of this project we started the transition of algorithms for tracklet stitching and contextually abnormal activity recognition developed at NEU to Siemens Corporate Research. Currently, we are continuing this effort, working with video data collected by Siemens at National Guard Camp Edwards (MA).

- We initiated collaborations with Balfour Technologies to incorporate robust tracking algorithms to their fourDscape browser. The goal is to provide persistent tracking of targets from live smart sensors integrated with user data generated, to provide automated situational awareness for incident commanders, first responders, and law enforcement.


- We initiated collaborations with TSA at Cleveland Airport (OH) and Mass Transport Authority at Logan Airport (MA) to design, deploy and test video analytics algorithms for airport security. As part of this effort we have recorded video sequences at these airports to test contra-flow in the exit detection and distributed visual tracking. We are currently in the process of deploying algorithms for contra-flow detection at Cleveland Airport.

C. Seminars, Workshops and Short Course.

1. Tutorial session on Computer Vision and Control, 2009 IFAC Symposium on Robust Control Design, Mario Sznaier and Octavia Camps.


10. “Dynamics based Information,” Short course, Univ. degli Studi di Calabria, Rende, May 24-28, 2010


VII. REFERENCES


based convex optimization , 4686-4691.


els for High Dimensional Time Series.


