Finding Causal Interactions in Video Sequences

Mustafa Ayazoglu, Burak Yilmaz, Mario Sznaier, and Octavia Camps*
Dept. of Electrical and Computer Engineering
Northeastern University, Boston, MA 02115
{hacettepeli.muhendis,yilmazbur}@gmail.com, {msznaier,camps}@coe.neu.edu

Abstract

This paper considers the problem of detecting causal interactions in video clips. Specifically, the goal is to detect whether the actions of a given target can be explained in terms of the past actions of a collection 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. As shown in the paper, this leads to a block-sparsification problem that can be efficiently solved using a modified Group-Lasso type approach, capable of handling missing data and outliers (due for instance to occlusion and mis-identified correspondences). Moreover, this approach also identifies time instants where the interactions between agents change, thus providing event detection capabilities. These results are illustrated with several examples involving non–trivial interactions amongst several human subjects.

1. Introduction and Motivation

The problem of identifying causal interactions amongst targets in a video sequence has been the focus of considerable attention in the past few years. A large portion of the existing body of work in this field uses human annotated video to build a *storyline* that includes both recognizing the activities involved and the causal relationships between them (see for instance [10] and references therein). While these methods are powerful and work well when suitably annotated data is available, annotating video clips is expensive and parsing relevant actions requires domain knowledge which may not be readily available. Indeed, in many situations, unveiling potentially hidden causal relationships is a first step towards building such knowledge.

In this paper we consider the problem of identifying causal interactions amongst targets, not necessarily human, from unannotated video sequences and without prior domain knowledge. Our approach exploits the concept of "Granger Causality" [9], that formalizes the intuitive idea that if a time series $\{x(t)\}$ is causally related to a second one $\{y(t)\}$, then knowledge of the past values of $\{y\}_1^t$ should lead to a better prediction of future values of $\{x\}_t^{t+k}$. In [14], Prabhakar *et. al.* successfully used a frequency domain reformulation of this concept to uncover pairwise interactions in scenarios involving repeating events, such as social games. This technique was later extended in [17] to model causal correlations between human joints and applied to the problem of activity classification. However, since this approach is based upon estimating the cross-covariance density function between events, it cannot handle situations where these events are non repeating, are too rare to provide an accurate estimate, or where these estimates are biased by outliers or missing data. Further, estimating a pairwise measure of causal correlation requires a spectral factorization of the cross-covariance, followed by numerical integration and statistical thresholding, limiting the approach to moderately large problems.

To circumvent these problems, in this paper we propose an alternative approach based upon recasting the problem into that of identifying the topology of a sparse (directed) graph, where each node corresponds to the time traces of relevant features of a target, and each link corresponds to a regressor. The situation is illustrated in Fig. 1 using as an example the problem of finding causal relations amongst 4 tennis players, leading to a graph with 4 nodes, and potentially 12 (directed) links. Note that in general, the problem of identifying causal relationships is ill posed (unless one wants to identify the set of all individuals that could possibly have causal connections), due to the existence of secondary interactions. To illustrate this point, consider a very simplistic scenario with three actors A, B, and C, where A copies (with some delay) the actions of B, which in turn mimics C, also with some delay. In this situation, the actions of A can be explained in terms of either those of B delayed one time sample, or those of C delayed by two samples. Thus, an algorithm based upon a statistical analysis

*This work was supported by NSF grants IIS–0713003, IIS-1318145, and ECCS–0901433, AFOSR grant FA9559–12–1–0271, and the Alert DHS Center of Excellence under Award Number 2008-ST-061-ED0001.

would identify a causal connection between A and C, even though there is no direct link between them. Further, if the actions of C can be explained by some simple autoregressive model of the form:

$$C(t) = \sum a_i C(t-i)$$

then it follows that the actions of A can be explained by the same model, e.g.

$$A(t) = \sum a_i A(t-i)$$

Hence, multiple graphs topologies, some of which include self-loops, can explain the same set of time-series. On the other hand, note that in this situation, the sparsest graph (in the sense of having the fewest links) is the one that correctly captures the causality relations: the most direct cause of A is B and that of B is C, with C potentially being explained by a self-loop. To capture this feature and regularize the problem, in the sequel we will seek to find the sparsest graph, in the sense of having the least number of interconnections, that explains the observed data, reflecting the fact that, when alternative models are possible, often the most parsimonious is the correct one. Our main result shows that the problem of identifying sparse graph structures from observed noisy data can be reduced to a convex optimization problem (via the use of Group Lasso type arguments) that can be efficiently solved. The advantages of the proposed methods are:

- Its ability to handle complex scenarios involving non-repeating events, environmental changes, collections of targets that do not necessarily split into well defined groups, outliers and missing data.
- The ability to identify the sparsest interaction structure that explains the observed data (thus avoiding labeling as causal connections those indirect correlations mediated only by an intermediary), together with a sparse "indicator" function whose support set indicates time instants where the interactions between agents change.
- Since the approach is not based on semantic analysis, it can be applied to the motion of arbitrary targets, not necessarily humans (indeed, it applies to arbitrary time series including for instance economic or genetic data).
- From a computational standpoint, the resulting optimization problems have a specific form amenable to be solved by a class of iterative algorithms [5, 3], that require at each step only a combination of thresholding and least-squares approximations. These algorithms have been shown to substantially outperform conventional convex-optimization solvers both in terms of memory and computation time requirements.

The remainder of the paper is organized as follows. In section 2 we provide a formal reformulation of the problem of finding causal relationships between agents as a sparse graph identification problem. In section 3, we show that this problem can be efficiently solved using a re-weighted Group Lasso approach. Moreover, as shown there, the resulting problem can be solved one node at a time using first order methods, which allows for handling situations involving a large number of agents. Finally, the effectiveness of the proposed method is illustrated in section 4 using both simple scenarios (for which ground truth is readily available) and video clips of sports, involving complex, non-repeating interactions amongst many agents.

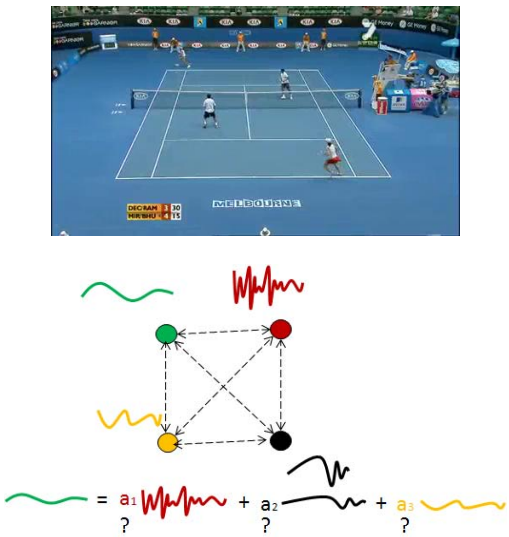

Figure 1. Finding causal interactions as a graph identification problem. Top: sample frame from a doubles tennis sequence. Bottom: Representation of this sequence as a graph, where each node represents the time series associated with the position of each player and the links are vector regressive models. Causal interactions exist when one of the time series can be explained as a combination of past values of the others.

2. Preliminaries

For ease of reference, in this section we summarize the notation used in the paper and give a formal definition of the problem under consideration.

2.1. Notation

$\sigma_i(\mathbf{M})$	i^{th} largest singular value of the matrix $\mathbf{M}$.
$\|\mathbf{M}\|_*$	nuclear norm: $\|\mathbf{M}\|_* \doteq \sum_i \sigma_i(\mathbf{M})$.
$\|\mathbf{M}\|_F$	Frobenious norm: $\|\mathbf{M}\|_F^2 \doteq \sum_{i,j} M_{ij}^2$
$\|\mathbf{M}\|_1$	ℓ_1 norm: $\|\mathbf{M}\|_1 \doteq \sum_{i,j} \|M_{ij}\|$.
$\|\mathbf{M}\|_o$	ℓ_o quasi-norm: $\|\mathbf{M}\|_o \doteq$ number of non-zero elements in $\mathbf{M}$.
$\circ$	Hadamard product of matrices: $(\mathbf{A} \circ \mathbf{B})_{i,j} = \mathbf{A}_{i,j}\mathbf{B}_{i,j}$.

2.2. Statement of the Problem

Next, we formalize the problem under consideration. Consider a scenario with P moving agents, and denote by $\tilde{\mathbf{Q}}_p(t)$ the 3D homogenous coordinates of the p^{th} individual at time t. Motivated by the idea of Granger Causality, we will say that the actions of this agent depend causally from those in a set $\mathcal{I}_p$ (which can possibly contain p itself), if $\tilde{\mathbf{Q}}_p(t)$ can be written as:

$$\tilde{\mathbf{Q}}_p(t) = \sum_{j\in\mathcal{I}_p}\sum_{n=0}^{N} a_{jp}(n)\tilde{\mathbf{Q}}_j(t-n) + \tilde{\eta}_p(t) + \tilde{\mathbf{u}}_p(t) \quad (1)$$

Here a_{jp} are unknown coefficients, and $\tilde{\eta}_p(t)$ and $\tilde{\mathbf{u}}_p(t)$ represent measurement noise and a piecewise constant signal that is intended to account for relatively rare events that cannot be explained by the (past) actions of other agents. Examples include interactions of an agent with the environment, for instance to avoid obstacles, or changes in the interactions between agents. Since these events are infrequent, we will model $\tilde{\mathbf{u}}$ as a signal that has (component-wise) a sparse derivative. Note in passing that since (1) involves homogeneous coordinates, the coefficients $a_{j,p}(.)$ satisfy the following constraint[1]

$$\sum_{j\in\mathcal{I}_p}\sum_{n=0}^{N} a_{jp}(n) = 1 \quad (2)$$

Our goal is to identify causal relationships using as data 2D measurements $\mathbf{q}_p(t)$ in F frames of the affine projections of the 3D coordinates $\tilde{\mathbf{Q}}_p(t)$ of the targets. Note that, under the affine camera assumption, the 2D coordinates are related exactly by the same regressor parameters [2]. Thus, (1) holds if and only if:

$$\mathbf{q}_p(t) = \sum_{j\in\mathcal{I}_p}\sum_{n=0}^{N} a_{jp}(n)\mathbf{q}_j(t-n) + \tilde{\mathbf{u}}_p(t) + \eta_p(t) \quad (3)$$

In this context, the problem can be precisely stated as: Given $\mathbf{q}_p(t)$ (in F number of frames) and some a-priori bound N on the order of the regressors (that is the "memory" of the interactions), find the sparsest set of equations of the form (3) that explains the data, that is:

$$\min_{a_{j,p},\eta_p,\mathbf{u}_p} \sum n_{I_p} \quad (4)$$

subject to (2) and:

$$\mathbf{q}_p(t) = \sum_{j\in\mathcal{I}_p}\sum_{n=0}^{N} a_{jp}(n)\mathbf{q}_j(t-n) + \mathbf{u}_p(t) + \eta_p(t),\ p = 1\ \ldots, P \text{ and } t = 1,..F \quad (5)$$

[1]This follows by considering the third coordinate in (1)

where $n_{\mathcal{I}_p}$ denotes the cardinality of the set $\mathcal{I}_p$. Rewriting (5) in matrix form yields:

$$[\mathbf{x}_p; \mathbf{y}_p] = [\mathbf{B}_p, \mathbf{I}][\mathbf{a}_p^T \mathbf{u}_{x_p}^T \mathbf{u}_{y_p}^T]^T + \eta_p \quad (6)$$

where

$$\begin{aligned}
\mathbf{q}_p(t) &= [\mathbf{x}_p(t)^T \mathbf{y}_p(t)^T]^T \\
\mathbf{u}_p(t) &= [\mathbf{u}_{x_p(t)}^T \mathbf{u}_{y_p(t)}^T]^T \\
\eta_p(t) &= [\eta_{xp}(t)^T \eta_{yp}(t)^T]^T \\
\mathbf{x}_p &= [x_p(F)x_p(F-1)...x_p(1)]^T \\
\mathbf{y}_p &= [y_p(F)y_p(F-1)...y_p(1)]^T \\
\mathbf{a}_p &= [\mathbf{a}_{1p}^T, \mathbf{a}_{2p}^T, ..., \mathbf{a}_{Pp}^T]^T \\
\mathbf{a}_{ip} &= [a_{ip}(0), a_{ip}(1), ..., a_{ip}(N)]^T \\
\mathbf{u}_{x_p} &= [u_{x_p(F)} u_{x_p(F-1)} ... u_{x_p(1)}]^T \\
\mathbf{u}_{y_p} &= [u_{y_p(F)} u_{y_p(F-1)} ... u_{y_p(1)}]^T \\
\mathbf{B}_p &= [\mathbf{X}_p; \mathbf{Y}_p] \\
\mathbf{X}_p &= [hankel(\mathbf{x}_1, N), ..., hankel(\mathbf{x}_P, N)] \\
\mathbf{Y}_p &= [hankel(\mathbf{y}_1, N), ..., hankel(\mathbf{y}_P, N)]
\end{aligned}$$

and where, for a sequence $z(t)$, hankel$(\mathbf{z}, N)$ denotes its associated Hankel matrix:

$$hankel(\mathbf{z}, N) = \begin{pmatrix} z(F) & z(F-1) & \cdots & z(F-N) \\ z(F-1) & z(F-2) & \cdots & z(F-N-1) \\ z(F-2) & z(F-3) & \cdots & z(F-N-2) \\ \vdots & \vdots & \cdots & \vdots \\ z(N+1) & z(N) & \cdots & z(1) \end{pmatrix}$$

It follows that a description of all the interactions amongst agents (that is the complete graph structure) is captured by a matrix equation of the form:

$$\mathbf{q} = [\mathbf{B}, \mathbf{I}][\mathbf{a}^T \mathbf{u}^T]^T + \eta \quad (7)$$

where

$$\begin{aligned}
\mathbf{q} &= \left[\mathbf{q_1}^T, \mathbf{q_2}^T, \mathbf{q_3}^T, \cdots, \mathbf{q_P}^T\right]^T \\
\mathbf{u} &= \left[\mathbf{u_1}^T, \mathbf{u_2}^T, \mathbf{u_3}^T, \cdots, \mathbf{u_P}^T\right]^T \\
\mathbf{a} &= \left[\mathbf{a_1}^T, \mathbf{a_2}^T, \mathbf{a_3}^T, \cdots, \mathbf{a_P}^T\right]^T \\
\eta &= \left[\eta_\mathbf{1}^T, \eta_\mathbf{2}^T, \eta_\mathbf{3}^T, \cdots, \eta_\mathbf{P}^T\right]^T
\end{aligned} \quad (8)$$

and

$$\mathbf{B} = \begin{bmatrix} \mathbf{B_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B_P} \end{bmatrix}$$

Thus, in this context, the problem of interest can be formalized as finding the *block*–sparsest solution to the set of linear equations (2) and (7).

The problem of identifying a graph structure subject to sparsity constraints, has been the subject of intense research in the past few years. For instance, [1] proposed a Lasso type algorithm to identify a sparse network where each link corresponds to a VAR process. The main idea underlying this method is to exploit the fact that penalizing the ℓ_1 norm of the vector of regression coefficients tends to produce sparse solutions. However, enforcing sparsity of the entire vector of regressor coefficients does not necessarily result in a sparse graph structure, since the resulting solution can consist of many links, each with a few coefficients. This difficulty can be circumvented by resorting to group Lasso type approaches [18], which seek to enforce *block sparsity* by using a combination of ℓ_1 and ℓ_2 norm constraints on the coefficients of the regressor. While this approach was shown to work well with artificial data in [11], exact recovery of the underlying network can be only guaranteed when the data satisfies suitable "incoherence" type conditions [4]. Finally, a different approach was pursued in [13], based on the use of a modified Orthogonal Least Squares algorithm, Cyclic Orthogonal Least Squares. However, this approach requires enforcing an *a-priori* limit on the number of links allowed to point to a single node, and such information may not be readily available, specially in cases where this number has high variability amongst nodes. To address these difficulties, in the next section we develop a convex optimization based approach to the problem of identifying sparse graph structures from observed noisy data. This method is closest in spirit to that in [11], in the sense that it is also based on a group Lasso type argument. The main differences consist in the ability to handle the unknown inputs $\tilde{\mathbf{u}}_p(t)$, needed to model exogenous disturbances affecting the agents, and in a reformulation of the problem, that allows for using a re-weighted iterative type algorithm, leading to substantially sparser solutions, even when the conditions in [4] fail.

3. Causality Identification Algorithm

In this section we present the main result of this paper, an algorithm to search for block-sparse solutions to (7). For each fixed p, the algorithm searches for sparse solutions to (6) by solving (iteratively) the following problem (suggested by the re-weighted heuristic proposed in [7])

$$\min_{\mathbf{a}_p,\mathbf{u}_{x_p},\mathbf{u}_{y_p}} \sum_{i=1}^{P} \mathbf{w}_j^a(\|\mathbf{a}_{ip}\|_2) + \lambda \left\|diag(\mathbf{w}^u)[\Delta\mathbf{u}_{x_p};\Delta\mathbf{u}_{y_p}]\right\|_1$$
$$\text{subject to: } \|\eta_{\mathbf{p}}\|_\infty \leq \epsilon,\ p=1,\ldots,P.$$
$$\sum_{i=1}^{P}\sum_{n=0}^{N} a_{ip}(n) = 1,\ p=1,\ldots,P. \tag{9}$$

where $[\Delta\mathbf{u}_{x_p};\Delta\mathbf{u}_{y_p}]$ represents the first order differences of the exogenous input vector $[\mathbf{u}_{x_p};\mathbf{u}_{y_p}]$, $\mathbf{W}^a$ and $\mathbf{W}^u$ are weighting matrices, and λ is a Lagrange multiplier that plays the role of a tuning parameter between graph sparsity and event sensitivity.

Intuitively, for a fixed set of weights $\mathbf{w}$, the algorithm attempts to find a block sparse solution to (6) and a set of sparse inputs $\Delta\mathbf{u}_{x_p};\Delta\mathbf{u}_{y_p}$, by exploiting the facts that minimizing $\sum_i \|\mathbf{a}_{ip}\|_2$ (the $\ell_{2,1}$ norm of the vector sequence $\{\mathbf{a}_{ip}\}$) tends to maximize block-sparsity [18], while minimizing the ℓ_1 norm maximizes sparsity [16]. Once these solutions are found, the weights $\mathbf{w}$ are adjusted to penalize those elements of the sequences with small values, so that in the next iteration solutions that set these elements to zero (hence further increasing sparsity) are favored. Note however, that proceeding in this way, requires solving at each iteration a problem with $n = P(Pn_r + F)$ variables, where P and F denote the number of agents and frames, respectively, and where n_r is a bound on the regressor order. On the other hand, it is easily seen that both the objective function and the constraints in (9) can be partitioned into P groups, with the p^{th} group involving *only* the variables related to the p^{th} node. It follows then that problem (9) can be solved by solving P smaller problems of the form:

$$\min_{\mathbf{a}_p,\mathbf{u}_{x_p},\mathbf{u}_{y_p}} \sum_{i=1}^{P} \mathbf{w}_j^a(\|\mathbf{a}_{ip}\|_2) + \lambda \left\|diag(\mathbf{w}^u)[\Delta\mathbf{u}_{x_p};\Delta\mathbf{u}_{y_p}]\right\|_1$$
$$\text{subject to: } \|\eta_{\mathbf{p}}\|_\infty \leq \epsilon \text{ and } \sum_{i=1}^{P}\sum_{n=0}^{N} a_{ip}(n) = 1 \tag{10}$$

leading to the algorithm given below:

Algorithm 1: REWEIGHTED CAUSALITY ALGORITHM

for each p
$\mathbf{w}^a = [1,1,...,1]$
$\mathbf{w}^{\mathbf{u}} = [1,1,...,1]$
$S > 1$ (self loop weight)
$\mathbf{s} = [1,1,...,S,...,1]$ (p'th element is S)
while not converged **do**
1. solve (9)
2. $\mathbf{w}_j^a = 1/(\|\mathbf{a}_{ip}\|_2 + \delta)$
3. $\mathbf{w}_j^a = \mathbf{w}_j^a \circ \mathbf{s}$ (Penalization self loops)
4. $\mathbf{w}^u = 1./(abs([\Delta\mathbf{u}_{x_p};\Delta\mathbf{u}_{y_p}]) + \delta)$
end while
5. At this point $a_{jp}(.)$, I_p and $\mathbf{u}_p(t)$ have been identified
end for

It is worth emphasizing that, since the computational complexity of standard interior point methods grows as n^3, solving these smaller P problems leads to roughly a $\mathcal{O}(P^2)$

reduction in computational time over solving a single, larger optimization. Thus, this approach can handle moderately large problems using standard, interior-point based, semi-definite optimization solvers. Larger problems can be accommodated by noting that the special form of the objective and constraints allow for using iterative Augmented Lagrangian Type Methods (ALM), based upon computing, at each step, the closed form solution to suitable intermediate optimization problems. While a complete derivation of such an algorithm is beyond the scope of this paper, using results from [12] it can be shown that each step requires only a combination of thresholding and least-squares approximations. Moreover, it can be shown that such an algorithm converges Q-superlinearly.

4. Handling Outliers and Missing Data

The algorithm outlined above assumes an ideal situation where the data matrix $\mathbf{B}$ is perfectly known. However, in practice many of its elements may be outliers (due to misidentified correspondences) or missing (due to occlusion). As we briefly show next, these situations can be efficiently handled by performing a structured robust PCA step [3] to obtain a "clean" data matrix, prior to applying Algorithm 1. From equation (6) it follows that, in the absence of exogenous inputs and noise:

$$\begin{bmatrix} \mathbf{x}_1 \dots \mathbf{x}_P \\ \mathbf{y}_1 \dots \mathbf{y}_P \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \dots \mathbf{X}_P \\ \mathbf{Y}_1 \dots \mathbf{Y}_P \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \dots \mathbf{a}_P \end{bmatrix} \tag{11}$$

Since $\mathbf{x}_i \in \{\text{col}(\mathbf{X}_j)\}$ and $\mathbf{y}_i \in \{\text{col}(\mathbf{Y}_j\})$, it follows that the sets $\{\text{col}(\mathbf{X}_i)\}$ and $\{\text{col}(\mathbf{Y}_i)\}$ are self-expressive, or, equivalently, the matrices $\mathcal{X} \doteq \begin{bmatrix} \mathbf{X}_1 \dots \mathbf{X}_N \end{bmatrix}$ and $\mathcal{Y} \doteq \begin{bmatrix} \mathbf{Y}_1 \dots \mathbf{Y}_N \end{bmatrix}$ are rank deficient. Consider now the case where some elements x_i, y_i of $\mathcal{X}$ and $\mathcal{Y}$ are missing. From the self-expressive property of $\{\text{col}(\mathbf{X}_i)\}$ and $\{\text{col}(\mathbf{Y}_i)\}$ it follows that these missing elements are given by:

$$x_i = \underset{x}{argmin}\, \mathbf{rank}(\mathcal{X}),\ y_i = \underset{y}{argmin}\, \mathbf{rank}(\mathcal{Y}) \tag{12}$$

Similarly, in the presence of outliers, $\mathcal{X}, \mathcal{Y}$ can be decomposed into the sum of a low rank matrix (the clean data) and a sparse one (the outliers) by solving a problem of the form

$$\min \mathbf{rank} \begin{pmatrix} \mathcal{X}_\mathbf{o} \\ \mathcal{Y}_\mathbf{o} \end{pmatrix} + \lambda \left\| \begin{matrix} \mathbf{E}_X \\ \mathbf{E}_Y \end{matrix} \right\|_o \text{ s. t.: } \begin{bmatrix} \mathcal{X}_\mathbf{o} \\ \mathcal{Y}_\mathbf{o} \end{bmatrix} + \begin{bmatrix} \mathbf{E}_X \\ \mathbf{E}_Y \end{bmatrix} = \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \end{bmatrix}$$

From the reasoning above it follows that in the presence of noise and exogenous outputs, the clean data record can be recovered from the corrupted, partial measurements by solving the following optimization problem:

$$\begin{aligned} &\min \left\| \begin{matrix} \mathcal{X}_\mathbf{o} \\ \mathcal{Y}_\mathbf{o} \end{matrix} \right\|_* + \lambda_1 \left\| \begin{matrix} \mathbf{M}_\mathbf{X} \circ \mathbf{E}_X \\ \mathbf{M}_\mathbf{Y} \circ \mathbf{E}_Y \end{matrix} \right\|_1 + \lambda_2 \left\| \begin{matrix} \mathbf{M}_\mathbf{X} \circ \mathbf{\Delta U}_X \\ \mathbf{M}_\mathbf{Y} \circ \mathbf{\Delta U}_Y \end{matrix} \right\|_1 \\ &+\lambda_3 \left\| \begin{matrix} \mathbf{M}_\mathbf{X} \circ \mathbf{\Xi}_\mathbf{X} \\ \mathbf{M}_\mathbf{Y} \circ \mathbf{\Xi}_Y \end{matrix} \right\|_F \\ &\text{subject to:} \\ &\begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \end{bmatrix} = \begin{bmatrix} \mathcal{X}_\mathbf{o} \\ \mathcal{Y}_\mathbf{o} \end{bmatrix} + \begin{bmatrix} \mathbf{E}_X \\ \mathbf{E}_Y \end{bmatrix} + \begin{bmatrix} \mathbf{U}_X \\ \mathbf{U}_Y \end{bmatrix} + \begin{bmatrix} \mathbf{\Xi}_X \\ \mathbf{\Xi}_Y \end{bmatrix} \end{aligned} \tag{13}$$

where we have used the standard convex relaxations of rank and cardinality[2]. Here $\mathbf{\Xi}$ and $\mathbf{U}$ denote noise and piecewise constant exogenous matrices, $\mathbf{\Delta U}$ denotes the matrix obtained by taking the difference between consecutive elements in $\mathbf{U}$, and $\mathbf{M}_X$ ($\mathbf{M}_Y$) is a "mask" matrix, with $m_{i,j} = 0$ if the element (i, j) in $\mathcal{X}$ ($\mathcal{Y}$) is missing, $m_{i,j} = 1$ otherwise, used to avoid penalizing elements in $\mathbf{E}, \mathbf{\Xi}, \mathbf{U}$ corresponding to missing data. Problem (13) is a structured robust PCA problem (due to the Hankel structure of $\mathcal{X}, \mathcal{Y}$) that can be efficiently solved using the first order method proposed in [3], slightly modified to handle the terms containing $\mathbf{\Delta U}$.

5. Experimental Results

In this section we illustrate the effectiveness of the proposed approach using several video clips (provided as supplemental material). The results of the experiments are displayed using graphs embedded on the video frames: An arrow indicates causal correlation between agents, with the point of the arrow indicating the agent whose actions are affected by the agent at its tail. The internal parameters of the algorithm were experimentally tuned, leading to the values $\epsilon = 0.1$, $\lambda = 0.05$, self loop weights $S = 10$. The algorithm is fairly insensitive to the value of the regularization parameters λ and S, which could be adjusted up or down by an order of magnitude without affecting the structure of the resulting graph. Finally, we used regressor order N=2 for the first three examples and N=4 for the last one, a choice that is consistent with the frame rate and the complexity of the actions taking place in each clip.

5.1. Clips from the UT-Interaction Data Set

We considered two video clips from the UT Human Interaction Data Set [15] (sequences 6 and 16). Figures 2 and 5 compare the results obtained applying the proposed algorithm versus Group Lasso (GL) [11] and Group Lasso combined with the reweighted heuristic described in (9) (GLRW). In all cases, the inputs to the algorithm were the (approximate) coordinates of the heads of each of the agents, normalized to the interval $[-1, 1]$, artificially corrupted with 10% outliers. Notably, the proposed algorithm

[2] As shown in [6, 8] under suitable conditions these relaxations recover the exact minimum rank solution.

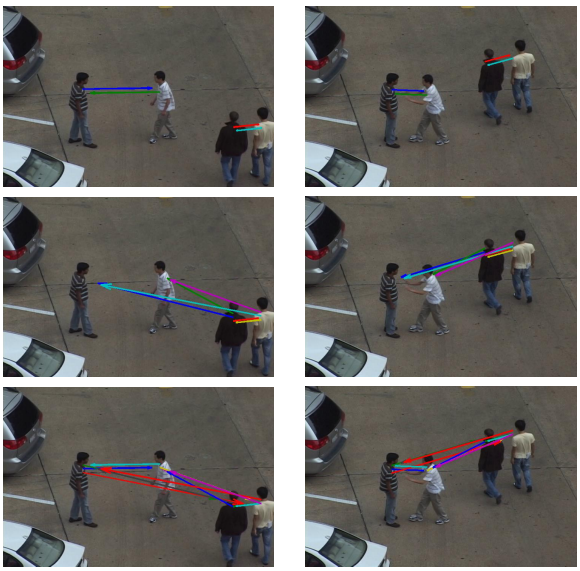

Figure 2. Sample frames from the UT sequence 6 with the identified causal connections superimposed. Top: Proposed Method. Center: Reweighted Group Lasso. Bottom: Group Lasso. Only the proposed method identifies the correct connections.

was able to correctly identify the correlations between the agents from this very limited amount of information, while the others failed to do so. Note in passing that in both cases none of the algorithms were directly applicable, due to some of the individuals leaving the field of view or being occluded. As illustrated in Fig. 3, the missing data was recovered by solving an RPCA problem prior to applying Algorithm 1. Finally, Fig. 4 sheds more insight on the key role played by the sparse signal u. As shown there, changes in u correspond exactly to time instants when the behavior of the corresponding agent deviates from the general pattern followed during most of the clip.

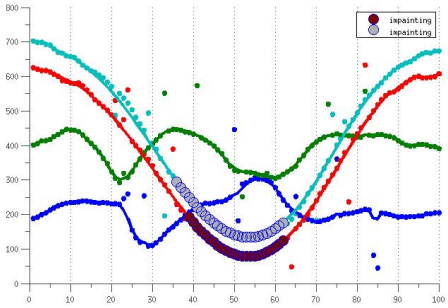

Figure 3. Time traces of the individual heads in the UT sequence 6, artificially corrupted with 10 % outliers. The outliers were removed and the missing data due to targets leaving the field of view was estimated solving a modified RPCA problem.

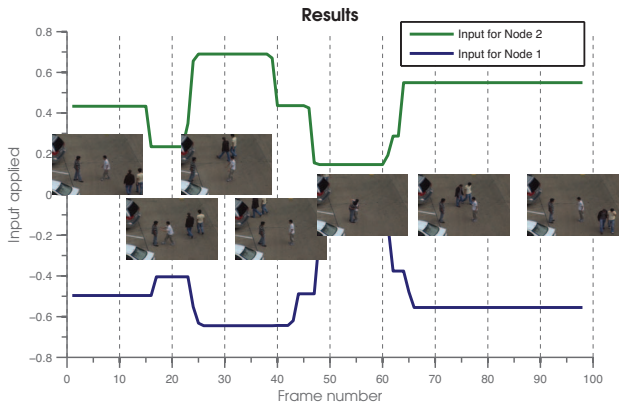

Figure 4. Sample (derivative sparse) exogenous signals in the UT sequence 6. The changes correspond to the instants when the second person starts moving towards the first, who remains stationary, and when the two persons merge in an embrace.

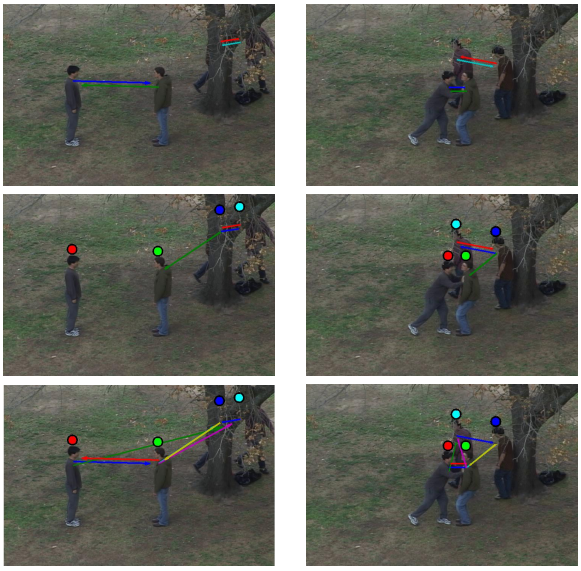

Figure 5. Sample frames from the UT sequence 16. Top: Correct correlations identified by the Proposed Method. Center and Bottom: Reweighted Group Lasso and Group Lasso (circles indicate self-loops).

5.2. Doubles Tennis Experiment

This experiment considers a non-staged real-life scenario. The data consists of 230 frames of a video clip from the Australian Open Tennis Doubles Final games. The goal here is to identify causal relationships between the different players using time traces of the respective centroid positions. Note that in this case the ground truth is not available. Nevertheless, since players from the same team usually look at their opponents and react to their motions, we expect a strong causality connection between members of

opposite teams. This intuition is matched by the correlations unveiled by the algorithm, shown in Fig. 6. The identified sparse input corresponding to the vertical direction is shown in Fig. 7 (similar results for the horizontal component are omitted due to space reasons.)

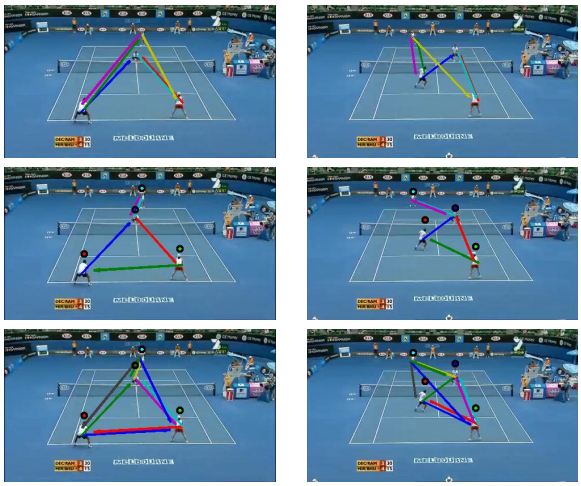

Figure 6. Sample frames from the tennis sequence. Top: The proposed method correctly identifies interactions between opposite team members. Center: Reweighted Group Lasso misses the interaction between the two rear-most individuals of opposite teams, generating self loops instead (denoted by the disks). Bottom: Group Lasso yields an almost complete graph.

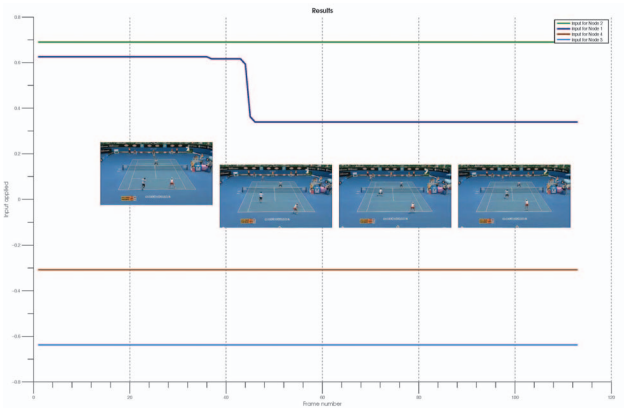

Figure 7. Exogenous signal corresponding to the vertical axis for the tennis sequence. The change in one component corresponds to the instant when the leftmost player in the bottom team moves from the line towards the net, remaining closer to it from then on.

5.3. Basketball Game Experiment

This experiment considers the interactions amongst players in a basketball game. As in the case of the tennis players, since the data comes from a real life scenario, the ground truth is not available. However, contrary to the tennis game, this scenario involves complex interactions amongst many players, and causality is hard to discern by inspection. Nevertheless, the results shown in Fig. 8, obtained using the position of the centroids as inputs to our algorithm, match our intuition. Firstly, one would expect a strong cause/effect connection between the actions of the player with the ball and the two defending opponents facing him. These connections (denoted by the yellow arrows) were indeed successfully identified by the algorithm. The next set of causal correlations is represented by the (blue, light green) and (black, white) arrow pairs showing the defending and the opponent players on the far side of the field and under the hoop. An important, counterintuitive, connection identified by the algorithm is represented by the magenta arrows between the right winger of the white team with two of his teammates: the one holding the ball and the one running behind all players. While at first sight this connection is not as obvious as the others, it becomes apparent towards the end of the sequence, when the right winger player is signaling with a raised arm. Notably, our algorithm was able to unveil this signaling without the need to perform a semantic analysis (a very difficult task here, since this signaling is apparent only in the last few frames). Rather, it used the fact that the causal correlation was encapsulated in the dynamics of the relative motions of these players.

6. Conclusions

In this paper we propose a new method for detecting causal interactions between agents using video data. The main idea is to recast this problem into a blind directed graph topology identification, where each node corresponds to the observed motion of a given target, each link indicates the presence of a causal correlation and the unknown inputs account for changes in the interaction patterns. In turn, this problem can be reduced to that of finding block-sparse solutions to a set of linear equations, which can be efficiently accomplished using an iterative re-weighted Group-Lasso approach. The ability of the algorithm to correctly identify causal correlations, even in cases where portions of the data record are missing or corrupted by outliers, and the key role played by the unknown exogenous input were illustrated with several examples involving non–trivial interactions amongst several human subjects. Remarkably, the proposed algorithm was able to identify both the correct interactions and the time instants when interactions amongst agents changed, based on minimal motion information: in all cases we used just a single time trace per person. This success indicates that in many scenarios, the dynamic information contained in the motion pattern of a single feature associated with a target is rich enough to enable identifying complex interaction patterns, without the need to track multiple features, perform a semantic analysis or use additional domain knowledge.

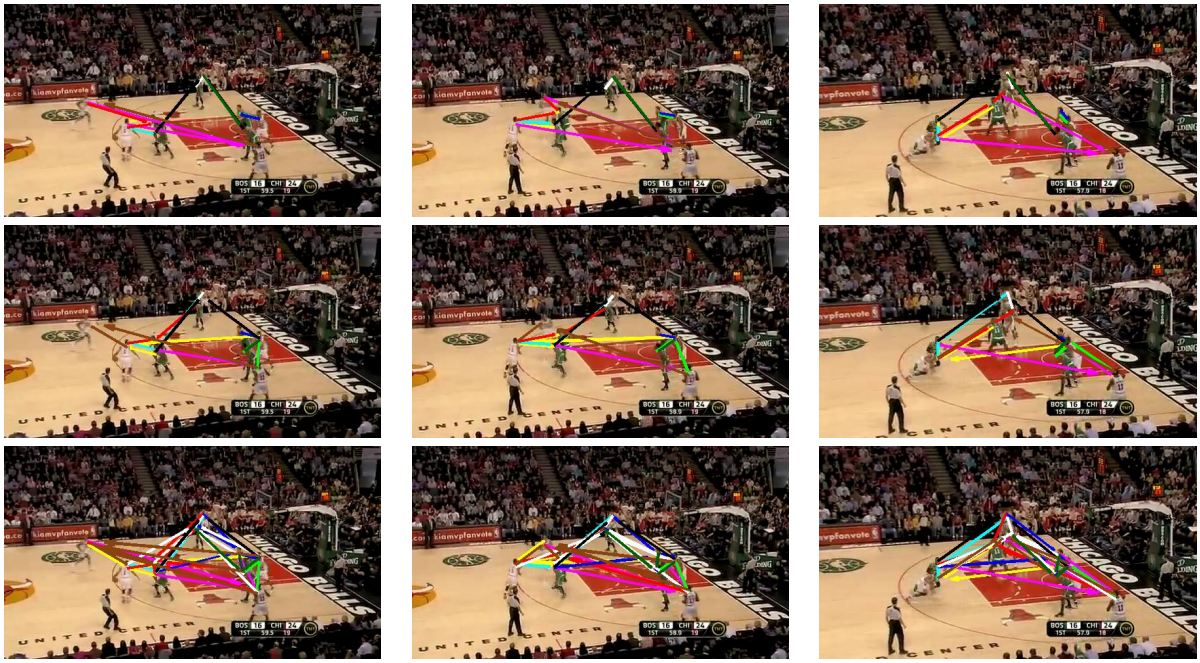

Figure 8. Sample frames from a Basketball game. Top: proposed method. Center: Reweighted Group Lasso misses the interaction between the signaling player and his teammates. Bottom: Group Lasso yields an almost complete graph.

References

[1] A. Arnold, Y. Liu, and N. Abe. Estimating brain functional connectivity with sparse multivariate autoregression. In *Proc. of the* 13^{th} *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 66–75, 2007. 4

[2] M. Ayazoglu, B. Li, C. Dicle, M. Sznaier, and O. Camps. Dynamic subspace-based coordinated multicamera tracking. In *2011 IEEE ICCV*, pages 2462–2469, 2011. 3

[3] M. Ayazoglu, M. Sznaier, and O. Camps. Fast algorithms for structured robust principal component analysis. In *2012 IEEE CVPR*, pages 1704–1711, June 2012. 2, 5

[4] A. Bolstad, B. Van Veen, and R. Nowak. Causal network inference via group sparse regularization. *IEEE Transactions on Signal Processing*, 59(6):2628–2641, 2011. 4

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011. 2

[6] E. Candes, X. Li, Y. Ma, and J.Wright. Robust principal component analysis? *J. ACM*, (3), 2011. 5

[7] E. J. Candes, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, December 2008. 4

[8] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *Siam J. Optim.*, (2):572–596, 2011. 5

[9] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, pages 424–438l, 1969. 1

[10] A. Gupta, P. Srinivasan, J. Shi, and L. Davis. Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos. In *2009 IEEE CVPR*, pages 2012–2019, 2009. 1

[11] S. Haufe, G. Nolte, K. R. Muller, and N. Kramer. Sparse causal discovery in multivariate time series. In *Neural Information Processing Systems*, 2009. 4, 5

[12] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 1663–670, 2010. 5

[13] D. Materassi, G. Innocenti, and L. Giarre. Reduced complexity models in identification of dynamical networks: Links with sparsification problems. In *48th IEEE Conference on Decision and Control*, pages 4796–4801, 2009. 4

[14] K. Prabhakar, S. Oh, P. Wang, G. Abowd, and J. Rehg. Temporal causality for the analysis of visual events. In *IEEE Conf Comp. Vision and Pattern Recog. (CVPR).*, pages 1967–1974, 2010. 1

[15] M. S. Ryoo and J. K. Aggarwal. UT Interaction Dataset, ICPR contest on Semantic Description of Human Activities. http://cvrc.ece.utexas.edu/SDHA2010/Human Interaction.html, 2010. 5

[16] J. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006. 4

[17] S. Yi and V. Pavlovic. Sparse granger causality graphs for human action classification. In *2012 ICPR*, pages 3374–3377. 1

[18] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006. 4