# TRAJECTORY TRIANGULATION: 3D MOTION RECONSTRUCTION WITH $\ell_1$ OPTIMIZATION

*Mingyu Chen, Ghassan AlRegib, and Biing-Hwang Juang*

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, Georgia 30332, U.S.A.
{mingyu, alregib, juang}@gatech.edu

## ABSTRACT

In this paper, we first explain the formulation of the trajectory triangulation: 3D reconstruction of a moving point from a series of 2D projections. The system has to be overconstrained to be solved by least squares techniques. We take advantage of the sparseness of real-world motions in the transformed domain, and borrow the concept of compressive sampling to reformulate the problem with $\ell_1$ optimization so that it is possible to reconstruct the trajectory even in an underconstrained system. Thus, fewer measurements are needed to reconstruct a 3D trajectory of even larger bandwidth coverage. We conduct experiments on both synthetic and real-world motion data to verify our proposed method, and compare the reconstruction results based on $\ell_1$ and $\ell_2$ optimization.

***Index Terms***— trajectory triangulation, motion tracking, compressive sampling, $\ell_1$ optimization

## 1. INTRODUCTION

The conventional optical motion tracking system relies on multiple synchronized cameras to simultaneously capture snapshots to triangulate the 3D positions of targets. In the context the motion trajectory is represented by sequential estimates of the position of the target. In this conventional scheme, the synchronization of cameras is critical, and the frame rate of the camera determines the sampling rate. Since the motion of a target may be quite general, a concern about the reconstruction of the motion trajectory and its accuracy thus arises, given that the camera's frame rate is a parameter preset for somewhat different needs. Furthermore, the requirement of calibrated synchronization among the cameras means a serious impediment to general deployment of a camera-based tracking system. We ask the following question: Can we relax the constraint on synchronization and reconstruct the trajectory from a sequence of monocular images from different viewing angles and positions? The answer is yes if we can have a priori knowledge or assumptions about how the object moves.

Avidan and Shashua proposed the idea of trajectory triangulation [1]. Given a series of images taken by a moving camera whose motion is general but known, they demonstrated that the trajectory can be reconstructed if the object moves along a line or a conic section. The motion constraint is relaxed to planes [2] and polynomial representations [3]. The pursuit of trajectory triangulation is similar to the problem of recovering a nonrigid structure from motion in computer vision. Instead of a shape space representation of the nonrigid structure, Akhter et al. [4] used Discrete Cosine Transform (DCT) basis functions to represent the time-varying structure in trajectory space. Park et al. [5] put together the idea of a DCT trajectory basis and the Direct Linear Transform algorithm [6] to propose a linear solution to reconstruct a moving point from a series of its image projections. They assume the trajectory can be well approximated by the DCT basis with relatively few low frequency components, and hence derive an overconstrained linear system with a unique least squares solution. They demonstrated that it is possible to achieve a precise 3D trajectory reconstruction using the DCT basis functions if the camera trajectory is random. An interesting real world example occurs when several photographers take asynchronous images of the same event from different locations, which can be interpreted as the random motion of the camera.

We are inspired by the idea proposed by Park et al. [5], and investigate the trajectory triangulation problem from the perspective of signal processing. In our setup, multiple cameras are mounted the same as in a conventional optical tracking system. The only difference is that we take one snapshot randomly or alternatively from the cameras instead of simultaneous measurements from all cameras at the instant of sampling. These cameras are not necessarily synchronized now. Among the real world motions of moving objects, we are most interested in tracking and reconstructing the human body motions. The 3D trajectory of the target is then represented with a linear combination of the DCT basis. We take advantage of the signal sparseness in the transformed domain, i.e., with non-zero DCT coefficients concentrated in the low frequency band, and borrow the concept of compressive sampling to tackle the trajectory triangulation problem with $\ell_1$ optimization. It is possible to reasonably reconstruct the motion from an underconstrained system. Therefore, we can cover the same bandwidth in the DCT domain with fewer measurements than the least squares approach.

In the following section, we analyze the motions of interests to verify the assumption of sparseness in the DCT basis representation. Section 3 formulates the trajectory triangulation problem in $\ell_1$ optimization. We present the experimental results on synthetic and real world data in section 4, and conclude this paper in Section 5.

## 2. MOTIONS OF INTERESTS

It is important to understand the characteristics of the signal being tracked so that we know how to represent and reconstruct it. The conventional 3D motion tracking system outputs the spatio-temporal signal of the target as a stream of 4-element samples: three for the spatial coordinates and one for the temporal information. We can define the trajectory as a composition of three functions of time in each 3D coordinates, $X(t) = [x(t), y(t), z(t)]^T$, and analyze them respectively. In general, all real world motions should be continuous in both position and speed over time, i.e., continuous in $X(t)$ and $X'(t)$ without singularities. Therefore, the DCT basis can be qual-

**Table 1**: Seq.1 is marker 21 in *Martial Art:Bassai*, seq.2 is marker 40 in *Breakdance:FancyFootWork*, and seq.3 is marker 40 in *General:RandomWalk*, where marker 21 is attached at the right wrist, and marker 40 is at right toe. The “-” sign indicates the percentage after rounding is 100%.

| seq. | len. | <5Hz |  | <10Hz |  | <15Hz |  |
|---|---|---|---|---|---|---|---|
|  |  | avg. | worst | avg. | worst | avg. | worst |
| 1 | 1 | 99.54 | 93.42 | 99.94 | 99.19 | 99.98 | 99.54 |
|  | 2 | 99.83 | 99.26 | 99.97 | 99.81 | 99.99 | 99.92 |
|  | 5 | 99.90 | 99.66 | 99.99 | 99.96 | - | 99.98 |
| 2 | 1 | 99.29 | 66.04 | 99.90 | 94.12 | 99.98 | 98.43 |
|  | 2 | 99.95 | 99.15 | 99.99 | 99.92 | - | 99.98 |
|  | 5 | 99.98 | 99.92 | - | 99.99 | - | - |
| 3 | 1 | 99.94 | 99.36 | 99.99 | 99.93 | - | 99.98 |
|  | 2 | 99.99 | 99.97 | - | - | - | - |
|  | 5 | - | - | - | - | - | - |

ified as a generic trajectory basis that compactly represents the real world motions.

In this paper, we focus on human motions because they are the most interesting subjects to track and lead to great variety of applications in entertainment, training and simulation, rehabilitation, and ergonomics. The motions of interests are further refined to the body motions, i.e., movements of torso, head, and limbs (minute motions such as finger movement or facial expression notwithstanding). We obtain the motion data from the CMU MoCap database: *mocap.cs.cmu.edu*, which are captured at 120 Hz with 41 markers placed over the actors body, and analyze the DCT basis representation of the motion trajectories.

First, we transform the motion $X(i) = [x(i), y(i), z(i)]^T$ to the DCT domain $F(k) = [f_x(k), f_y(k), f_z(k)]^T$ in the selected time frame, and analyze the distribution of energy over frequency. The trajectories of human motions are strongly time-variant, and generally have less correlation if we increase the time interval. No anatomically based motion modeling is assumed here. Thus, the length of the time frame is important. We perform the analysis with the time frame of 1, 2, and 5 seconds. For each frame length in one motion sequence, 100 frames are randomly selected to compute the average and worst (least concentrated) cases, and the trajectories within each frame are preprocessed to have zero mean to remove the DC bias. We intentionally select two extreme cases of the action sequences: the sequence contains intense or abrupt motions, and the one has slow and smooth motions. Among the 41 markers, we extract the trajectories from the markers placed at either wrist or ankle since they are more likely to contain complicated motions.

Table 1 shows a quantitative comparison of the energy distribution for different bandwidths and frame lengths. The results conclude that human body motions contains mostly low frequency components. In general, the band from 0 to 15 Hz covers 99.98% of energy, which means the trajectory can be well represented with the DCT basis under 15 Hz. The longer the time frame is, the more it concentrates in the low frequency band because the temporally local high frequency components weight even less in a longer frame. The selection of the frame length is twofold. A longer frame provides higher frequency resolution and denser energy distribution. On the other hand, the details (high frequency components) become less distinguishable. The number of variables and the size of the linear system increase proportional to the frame length as well. According to Table 1, the frame length of 1 second is a reasonable choice for our following problem setup.

## 3. FORMULATION OF THE TRAJECTORY TRIANGULATION PROBLEM

In the traditional triangulation, correspondences across multiple views allow us to triangulate the point in 3D space. Each 2D measurement provides us two equations, so we need at least two measurements from different views to form an overconstrained system to solve for a static 3D point. If the point is moving, the measurements from multiple views have to be taken at the same instant. The least squares solution to the overconstrained system minimizes the algebraic errors and is referred as the Direct Linear Transform algorithm [6].

Assume we use the same multi-camera setup, but the cameras are not necessarily synchronized now. We randomly pick up one among them to make a 2D measurement at each sampling instant, or shift the external control clock to make these cameras take snapshots interlacingly. Either case is the scenario of the trajectory triangulation problem, and we will show the formulation of this problem based on Park's work [5].

For the $i^{\text{th}}$ time sample with a given $i^{\text{th}}$ camera projection matrix $P_i$, a point in 3D, $\mathbf{X}_i$, is projected to a 2D point, $\mathbf{x}_i$. The projective transformation is defined as,

$$\begin{bmatrix}\mathbf{x}_i\\ 1\end{bmatrix} \equiv P_i \begin{bmatrix}\mathbf{X}_i\\ 1\end{bmatrix}, \text{ or } \begin{bmatrix}\mathbf{x}_i\\ 1\end{bmatrix}_\times P_i \begin{bmatrix}\mathbf{X}_i\\ 1\end{bmatrix} = 0, \qquad (1)$$

where $\equiv$ implies equality up to scale, and $[.]_\times$ is the skew symmetric matrix form of the cross product [6]. Equation (1) can be re-written as an inhomogeneous equation,

$$\begin{bmatrix}\mathbf{x}_i\\ 1\end{bmatrix}_\times \begin{bmatrix}p_1 & p_2 & p_3\end{bmatrix} \mathbf{X}_i = -\begin{bmatrix}\mathbf{x}_i\\ 1\end{bmatrix}_\times p_4, \text{ or } R_i\mathbf{X}_i = r_i, \qquad (2)$$

where $p_j$ is the $j^{\text{th}}$ column of $P_i$. Even though (2) gives us three equations, only two of them are linearly independent since the projection is up to scale. Thus, it is further refined as $Q_i\mathbf{X}_i = q_i$, where $Q_i$ and $q_i$ are the matrices made of the first two rows of $R_i$ and $r_i$ respectively. If we take $F$ samples in the time frame, the 3D point trajectory, $\mathbf{X}$, can be formulated as,

$$\begin{bmatrix}Q_1 & & \\ & \ddots & \\ & & Q_F\end{bmatrix} \begin{bmatrix}\mathbf{X}_1\\ \vdots\\ \mathbf{X}_F\end{bmatrix} = \begin{bmatrix}q_1\\ \vdots\\ q_F\end{bmatrix}, \text{ or } Q\mathbf{X} = q \qquad (3)$$

Obviously the problem is ill-posed with $3F$ unknowns and $2F$ constraints, i.e., $Q \in \mathbb{R}^{2F\times 3F}$. Given *a priori* knowledge that the trajectories are from real world motions studied in the previous section, we can utilize the sparseness of the signal in the DCT domain to help us solve this problem. The trajectory $\mathbf{X}$ can be approximated by a linear combination of trajectory bases, $\mathbf{X} \approx \Theta\beta$, where

$$\Theta = \begin{bmatrix}\theta_1^T & & \\ & \theta_1^T & \\ & & \theta_1^T\\ & \vdots & \\ \theta_F^T & & \\ & \theta_F^T & \\ & & \theta_F^T\end{bmatrix} \in \mathbb{R}^{3F\times 3K}, \qquad (4)$$

$\beta = [\beta_{x_1}\cdots\beta_{x_K}\beta_{y_1}\cdots\beta_{y_K}\beta_{z_1}\cdots\beta_{z_K}]^T \in \mathbb{R}^{3K}$, $\theta_i \in \mathbb{R}^K$ is formed with the first $K$ elements of the inverse DCT from coefficient space to original space at the sampling instant $i$, and $K$ can be interpreted as the number of the bases per coordinate. We then rewrite (3) as $Q\Theta\beta = q$. With the constraint $2F \geq 3K$, the system becomes overdetermined, which is solvable with least squares techniques. An overdetermined system can still be or close to singular, which is likely to happen when $2F \sim 3K$ and makes the least squares solution unstable. One possible workaround is to regularize the problem as,

$$\text{minimize } \|Q\Theta\beta - q\|_2 + \gamma_{\ell_2}\|\beta\|_2 \,, \tag{5}$$

where $\gamma_{\ell_2} > 0$ is a trade-off parameter. A question arises: Can super-resolved signals be reconstructed from fewer measurements?

As discussed in Section 2, the trajectory in the DCT domain concentrates in the low frequency band. Let $\beta_S$ denote the best $S$-sparse approximation of $\beta$, which is obtained by keeping the largest S entries of $[\beta_{x_1} \cdots \beta_{x_K}]^T$, $[\beta_{y_1} \cdots \beta_{y_K}]^T$, and $[\beta_{z_1} \cdots \beta_{z_K}]^T$ respectively and setting the rest zero. We re-examine $Q\Theta\beta_S = q$ with the knowledge of compressive sampling [7, 8, 9], where $Q$ is the measurement matrix, $\Theta$ is the representation orthobasis, $q$ is the measurement vector, and $\beta_S$ is the sparse signal we want to recover. The modified version of coherence between the measurement matrix $Q$ and the representation basis $\Theta$ can be defined as

$$\mu(Q, \Theta) = \sqrt{n}\max|Q_{\mathrm{N}}\Theta|, \tag{6}$$

where $Q_{\mathrm{N}}$ is obtained by normalizing each row of $Q$, and $\max|.|$ returns the element-wise maximum absolute value. In plain English, the coherence measures the largest correlation between any row of $Q_{\mathrm{N}}$ and column of $\Theta$. The smaller the coherence, the fewer samples are needed to reconstruct $\beta_S$. In [8], it states several ways to generate the measurement matrix $Q$ with low coherence so that the overall sensing matrix $A = Q\Theta$ satisfies the restricted isometry property (RIP) with overwhelming probability. Our measurement matrix $Q$ in (3) is generated very differently from those random based methods in [8], but fortunately the coherence in our case is comparatively low, e.g., $\mu \sim 2.0$ when $K = 120$ ($n = 360$) with full coverage of DCT spectrum. Therefore, the recovery via $\ell_1$ minimization should be reasonable even in an underconstrained system, and we formulate the trajectory triangulation problem as basis pursuit [10],

$$\text{minimize } \|Q\Theta\beta - q\|_2 + \gamma_{\ell_1}\|\beta\|_1 \,, \tag{7}$$

where $\gamma_{\ell_1} > 0$ is a parameter used to trade off the quality of the fit to the data and the sparsity of the coefficient vector. If we take account of the possible distortion caused by 2D measurements, camera matrices, or the DCT approximation, the $\ell_1$ optimization problem can be modified as,

$$\text{minimize } \|\beta\|_1 \text{ subject to } \|Q\Theta\beta - q\|_2 \leq \epsilon \tag{8}$$

## 4. EXPERIMENTS AND RESULTS

In this section, we evaluate the trajectory reconstruction on both synthetic and real-world motion data. For both cases, we simulate the camera setup to compute the camera projection matrices and the 2D measurements from 3D trajectories. The frame length is set to 1 second as discussed above.

### 4.1. Synthetic Motion Data

We are interested in the minimum number of measurements needed to solve the trajectory triangulation problem. There are various definitions of the lower bound of the measurements required for robust reconstruction depending on each instance [7][8], and our case doesn't fit in any of them. However, we can still approach the lower bound by simulations on synthetic motion data.

The synthetic trajectory contains 120 points that are temporally uniformly sampled in the time frame. To make it more realistic, the trajectory is synthesized from components of lower frequency band. We use $\beta_x$, $\beta_y$, and $\beta_z$ to represent the first $K$ of 120 DCT coefficients (in the order of increasing frequency) for each axis. The corresponding inverser DCT matrix can be written in matrix form,

$$D = \begin{bmatrix} d_1^T \\ \vdots \\ d_{120}^T \end{bmatrix} \in \mathbb{R}^{120\times K},$$

and the three time functions can be generated as $D\beta_x$, $D\beta_y$, and $D\beta_z$. We further sparsify them by setting only $S$ among the $K$ DCT coefficients non-zero. Both the indices and the magnitudes of the $S$ elements for each coordinate are randomized, and we can generate a $S$-sparse trajectory.

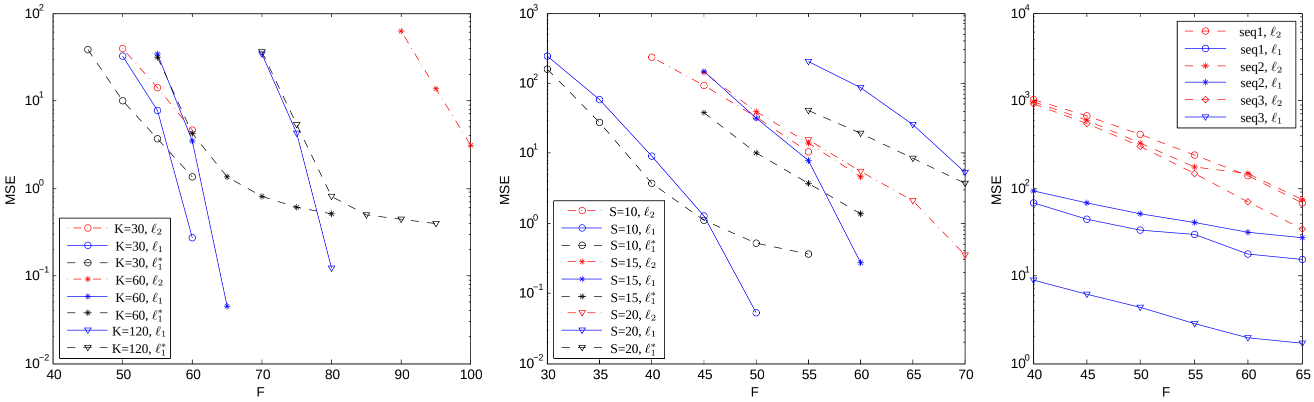
The next step is to create the 2D measurements of the synthetic trajectory and the corresponding $Q$ and $q$. In our virtual setup, there are four cameras with known camera projection matrices. We randomly select F points from the 120 samples. For each selected point $\mathbf{X}_i$, where $i \in [1\cdots F]$, we randomly choose one camera to measure its 2D projection and compute $Q_i$ and $q_i$. (3) puts $F$ measurements together to form the measurement matrix $Q \in \mathbb{R}^{2F\times 3F}$ and measurement vector $q \in \mathbb{R}^{2F}$. $\Theta$ is constructed as shown in (4) with $\theta_i = d_j$, where $j \in [1\cdots 120]$ is the temporal index of $\mathbf{X}_i$. We proceed to solve $Q\Theta\beta = q$ with formulations in (5), (7), and (8), with $\gamma_{\ell_2} = 0.001$, $\gamma_{\ell_1} = 0.01$, and $\epsilon = 0.01\|q\|_2$.

We conduct experiments on synthetic trajectories with different levels of sparseness $S$ and sizes of the DCT basis $K$ to see their impacts on the number of measurements needed to reconstruct the trajectory. Mean squared error (MSE) is used to quantify the accuracy of the reconstruction. We plot the MSE versus the number of measurements $F$ with different settings of $K$ and $S$ on Fig. 1(a) and Fig. 1(b). For each tuple of $S$, $K$, and $F$, we repeat the experiment 1000 times to compute its MSE.

In Fig. 1(a), we fix the sparseness $S$ and adjust the size of the DCT basis. Under the same $K$, $\ell_1$ optimization (7) or (8) needs less measurements to reconstruct the trajectory than $\ell_2$ optimization (5) at the same level of accuracy. The tendency is magnified after $K$ is doubled twice. $\ell_2$ optimization performs badly even when the system is overconstrained ($2F > 3K$). The $\ell_2$ MSE of $K$=120 blows up so that it is not shown. In Fig. 1(a), we vary the sparseness. The $\ell_2$ MSE is not affected by $S$ as we expect. The $\ell_1$ and $\ell_1^*$ MSE curves roughly shift to right by the amount of $3\Delta S$ as we increase $S$, and $S$=15 can be viewed as a threshold that $\ell_1$ optimization outperforms $\ell_2$ optimization. (8) performs slightly better than (7) when we have less measurements. By increasing $F$, $\ell_1$ MSE drops quickly ($< 10^{-4}$) while $\ell_1^*$ saturates due to the loose constraint of $\epsilon$.

### 4.2. Real-World Motion Data

The real-world motion trajectories tend to concentrate in the low frequency band, and they can be well represented with the DCT basis under 15 Hz. Assume the DCT domain for each coordinate has a temporal resolution of 120 points in the one-second time frame. We can construct $\Theta$ with $K = 30$ that spans the spectrum from 0 to 15 Hz. Larger $K$ is less meaningful here because the DCT coefficients above 15 Hz are negligible. From Table 1, it is reasonable to further approximate the signal as a $S$-sparse trajectory defined in previous section with $S$ in the range of 10 to 15.



(a) Synthetic trajectory: fixed sparseness ($S$=15)

(b) Synthetic trajectory: fixed size of DCT basis ($K$=30)

(c) Sequences in Table 1($K$=30)

**Fig. 1**: **MSE v.s. F:** Reconstruction results from (5), (7), and (8) are labeled as $\ell_2$, $\ell_1$, and $\ell_1^*$ respectively.

We use the same motion sequences and trackers in Table 1. For each sequence, we randomly extract a one-second frame that contains 120 sample points, and we can formulate the problem in the same manner as in the synthetic case. The trajectory of each frame is normalized to zero mean and identical variance so that errors can be compared across different frames. Given the selected sequence and marker, we repeat the experiment 1000 times for every $F$ in the range of interests and show the results in Fig. 1(c). Note that we do not show the curves of $\ell_1^*$ MSE because they are not stable and blow up at small $F$ for sequence 1 and 2. In all cases, $\ell_1$ outperforms $\ell_1^*$.

Even when the system is overconstrained ($F > 45$), $\ell_1$ optimization still produces better reconstruction results than $\ell_2$ optimization, and the gap between $\ell_1$ MSE and $\ell_2$ MSE increases as we reduce $F$. Fig. 1(b) shows that a sparser signal can be reconstructed with less measurements. Fig. 1(c) confirms that $\ell_1$ MSE directly relates to the sparseness of the signal as indicated in Table 1 that we can view sequence 3 as the sparsest signal and sequence 2 as the least sparse one.

## 5. CONCLUSION

In this paper, we first analyze the characteristics of human motion trajectories in the DCT domain and show that the human motions contain mainly low frequency components. The spatio-temporal signal of the trajectory can be well approximated by the DCT basis under 15 Hz. Within this bandwidth, the trajectory can be further considered as a sparse signal in the DCT domain with various degrees of sparseness. We investigate the coherence between the measurement matrix and representation matrix in the formulation of the trajectory triangulation, and the comparatively low coherence suggests that $\ell_1$ optimization is applicable to tackle this problem. Using this analysis, we demonstrate that $\ell_1$ optimization can help us reconstruct the motion trajectory especially when the system is underconstrained. Therefore, it is possible to reconstruct the motion signal of the same bandwidth with less measurements than the conventional optical tracking system. We also loose the constraint on camera synchronization. The snapshots can be taken at a constant rate or even arbitrarily during the time frame. If we know the sampling instant of each image, we can adjust the DCT basis functions according to the given time instants in the frame. In our experiments, the sampling instants are randomly selected from those of equal interval for the sake of convenience. Actually, our approach can be applied to not only the human motions but other trajectories as long as they are sparse and can be well approximated in the DCT domain. The future work is to develop the theoretic upper bound of measurements required for exact trajectory reconstruction.

## 6. REFERENCES

[1] S. Avidan and A. Shashua, “Trajectory triangulation: 3d reconstruction of moving points from a monocular image sequence,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 348–357, 2000.

[2] A. Shashua and L. Wolf, “Homography tensors: On algebraic entities that represent three views of static or moving planar points,” in *Proc. of the European Conf. on Computer Vision*, 2000, pp. 507–521.

[3] J. Y. Kaminski and M. Teicher, “A general framework for trajectory triangulation,” *J. Math. Imaging Vis.*, vol. 21, no. 1, pp. 27–41, 2004.

[4] I. Akhter, S. Khan, Y. Sheikh, and T. Kanade, “Nonrigid structure from motion in trajectory space,” in *Neural Information Processing Systems*, 2008.

[5] H. S. Park, T. Shiratori, I. Matthews, and Y. A. Sheikh, “3d reconstruction of a moving point from a series of 2d projections,” in *Proc. of the European Conf. on Computer Vision*, Sep 2010.

[6] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[7] E. Candés and M. B. Wakin, “An introduction to compressive sampling,” *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21 –30, Mar 2008.

[8] E. Candés, “Compressive sampling,” *Proc. of the International Congress of Mathematicians*, vol. 3, pp. 1433–1452, 2006.

[9] E. Candés and J. Romberg, “Sparsity and incoherence in compressive sampling,” *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.