Tracking-by-selection Pose Estimation in Videos

A Thesis Presented
by

Jing Xiao

to

The Department of Electrical and Computer Engineering

in partial fulfillment of the requirements
for the degree of

Master of Science

in

Electrical and Computer Engineering

Northeastern University
Boston, Massachusetts

April 2016

To my parents.

Contents

List of Figures

List of Tables

List of Acronyms

LTI Linear Time Invariant. An LTI system, in a simplified sense, will exhibit two behaviors: 1) time invariance, and 2) additive Superposition.

BFC Breadth First Cut. An algorithm introduced in 3.4 in the thesis.

CVX CVX is a Matlab-based modeling system for convex optimization problem.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Octavia Camps for the continuous support of my study and research, for her patience, motivation, enthusiasm, and knowledge. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance and steering when the front is a dead end. It was her outstanding teaching of Image Processing and Pattern Recognition and Advanced Computer Vision that let me decide to convert to computer vision area.

Besides my advisor, I am deeply grateful to the rest of my thesis committee: Prof. Mario Sznaier, who brings rigorous and critical comments as well as cheerful atmosphere at our regular lab meeting on every Thursday. And Prof. Masoud Salehi, for his encouragement and support. His clear and pinpoint explanations made certain difficult materials on the Digital Communications course much easier, and I can not forget the thrill of understanding these new materials in a way that makes perfect sense to me.

It has been a pleasure to be a member of the Robust Systems Lab. I am also indebted to the lab members, Mengran Gou, Wenqian Liu, Yin Wang, Xikang Zhang, Angels Rates-Borras, Bengisu Ozbay, Sadjad Asghari-Esfeden, Mohammad Reza, Zulqarnain Khan, Yuexi Zhang, Ichiro Maruta and others with whom I have interacted during my graduate studies, for their support and all the fun and cakes we have had in the last year. Particularly, I would like to acknowledge Caglayan Dicle and Yongfang Cheng, for many stimulating discussions that helped me understand my research better, for showing me how to start at first place and for their valuable advise. And I am also grateful to the following former members in the Robust Systems Lab, Fei Xiong and Binlong Li. Although I have never met them, the books and materials left by them are of great benefit to me, and their deeds told by current lab-mates constantly inspired me during my studies.

Many friends have influenced me through these years. Their support and care helped me overcome the difficulties in a foreign county and stay focused on my graduate study. I greatly appreciate their friendliness and their ability of making the life a delight in Boston. Their names are mentioned for avoiding the consequences of their threats against not thanking them in the acknowledgment: Tancheng Zhuang, Yichen Lin, Siran Zheng, Xiaofeng Su, Ye Liu, Simon Zou, Tianzi Liu, Jingyun Hu and all others.

Most importantly, none of these would have been possible without the love and patience of my parents, Haijiang Xiao and Mingying Liu. I would like to express my deepest gratitude to them, for their support and understanding of me studying abroad and for their efforts in giving me a carefree and joyful period of time in my life. 谢谢你们。

Abstract of the Thesis

Tracking-by-selection Pose Estimation in Videos

by

Jing Xiao

Master of Science in Electrical and Computer Engineering

Northeastern University, April 2016

Dr. Octavia I. Camps, Adviser

In this thesis, we propose a novel, effective, yet simple tracking-by-selection algorithm for human pose estimation in videos. The problem is solved by two steps, firstly, top N pose candidates are generated from each frame of the video; secondly, frame to frame temporal smoothness between poses across different frames are guaranteed by selecting the trajectory with the least Nuclear norm of its Hankel matrix among all the possible combinations. In the end, we also discuss the necessity of cleaning the selected trajectory by rank minimization to remove the effects of noise and outliers. Our dynamic based approach not only exhibits the ability to select the smooth trajectory from N-best detections accurately in the MoCap dataset, but also finds its value in video based tracking-by-selection human pose estimation framework.

Chapter 1

Introduction

1.1 Background

Pose estimation is an important task in vision based human activity recognition applications. A human body can be represented as an articulated system of rigid segments connected by joints, and human motion can be considered as a continuous evolution of the spatial configuration of these rigid segments [2]. Even though a huge amount of effort has been putting into estimating the articulated pose of a human from an image, one could still get very bad estimations where occlusions or overlapping are present. And some other failure reasons such as double counting occurs when the same region of the image is used to explain more than one body parts. [1] illustrated a typical example where any reasonable detector model might give those pose estimations high score. For the sake of understanding, the example from their work is shown in Figure 1.1. Thus it makes the pose estimation problem troublesome from one single image; in the example, how could a valid algorithm decide which estimation is more accurate? However, we argue that the correct pose should be extracted from temporal information. Namely if the accurate poses of former or next frames could be incorporated, the detector might be able to reason a correct estimation at this current frame. Based on this simple intuition, this thesis presents a novel, effective yet simple method that achieves excellence performance in synthetic experiments and produces state-of-the-art results for some of the tasks of estimating human poses in video sequences and generating skeleton representations accordingly.

Obviously, still image based pose detection method can be applied to every single frame in a continuous video to get a rough pose estimation, and then a further smoothing across adjacent frames could be used to make this sequence of pose estimations consistent and accurate across time instances. More details are discussed in related works.

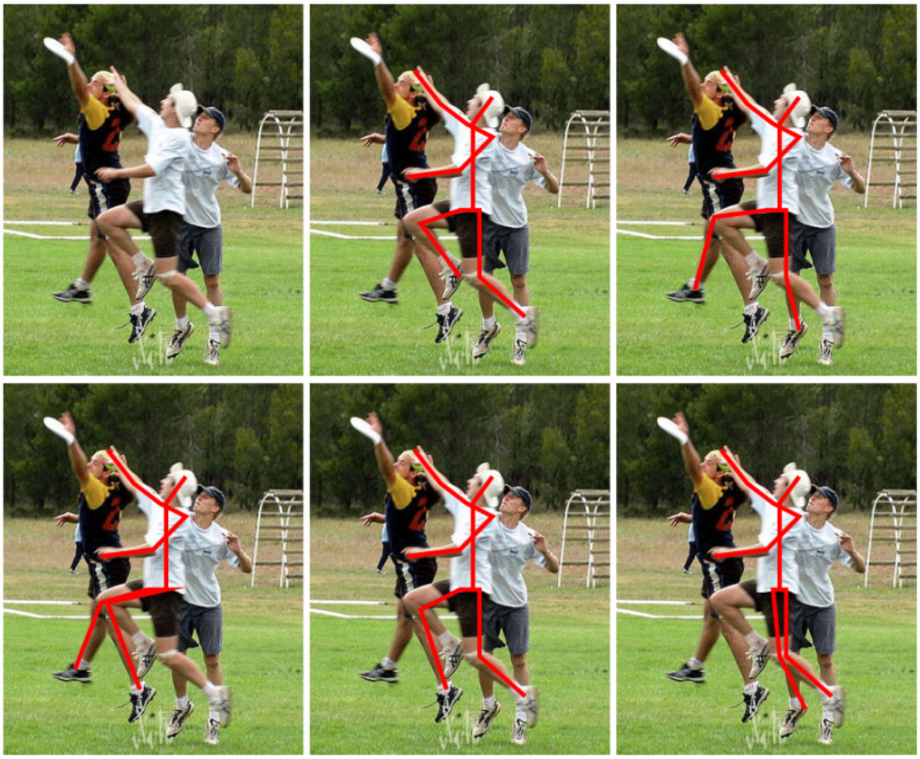


Figure 1.1: To localize articulated objects in cluttered scenes, one will need to reason about multiple pose hypotheses. In the image above, an accurate detection in the **top middle** is shown along with other hypotheses that may achieve high score in a given reasonable detector. Image from [1].

1.2 Problem Statement

Human pose estimation in videos is intrinsically a very challenging problem due to the large diversity of possible human configurations, non-rigidity of the human body, different illuminations or viewpoints, cluttered background and self occlusions etc. One commonly considered approach is the tracking-by-selection framework [1, 3], which is similar to tracking-by-detection and breaks down the whole problem into two stages: in the first step, a set of candidate poses are generated from each frame of a video; in the second step, frame-frame temporal smoothness metrics among poses or different body parts are introduced, and by minimizing an optimization formulation the best suitable set of pose estimations for the video is obtained.

1.3 Related Works

Related works are categorized according to the tracking-by-selection framework. The first group of recent works is about estimating poses in static images, because it turns out that such techniques will serve initializing video-based articulated trackers. The second group of recent works is on how to consider the useful spatial and temporal constraints between body parts across different frames. And the third group provides the intuition that the joint evolutions of human activities are smooth by nature and can be modeled as a low order auto-regressive process.

Static images: Felzenszwalb and Huttenloche in [4] provided an efficient framework for part-based modeling and recognition of objects. Their idea is to represent an object (or a human face or body) by several parts arranged in a deformable configuration, together with geometric constraints on pairs of parts, often visualized as springs. These pictorial structure models allow for qualitative descriptions of visual appearance and immediately become the influential and dominant approach for human pose estimation. Inspired by their work, many human pose estimation methods have been proposed. In [5] Yang and Ramanan proposed a mixture of small, non-oriented parts model that jointly captures spatial relations between part locations and co-occurrence relations between part mixtures, augmenting standard pictorial structure models that encode just spatial relations. Ferrari *et al.* [6] proposed an approach that progressively reduces the search space for different body parts, to greatly improve the chances that pose estimation will succeed. Moreover, they also proposed an integrated spatio-temporal model covering multiple frames to refine pose estimation from individual frames.

Videos: Ramakrishna *et al.* [7] presented an occlusion aware algorithm for tracking human pose in an image sequence, that addresses the problem of double counting, besides, they considered the continuity constraints among three frames, previous frame, current frame and next frame. Park and Ramanan [1] proposed an N-best algorithm that gives top N candidate pose estimations for a single input image, and then they used this algorithm in video sequences, and stitched them together with dynamic programming on a trellis graph, which only takes into account pairwise (previous frame and current frame) penalty. Specifically, they simply used the negative of the total squared pixel difference between each joint in pose at time t-1 and time t. Tokola *et al.* [3] proposed the tracking-by-selection framework, in which each body part is tracked separately and parts are then combined at a latter stage. Cherian *et al.* [8] used the tracking-by-selection two stage scheme as well, and the temporal edges consist of every pair of frames at time t-1 and time t, besides they mainly focused on pose estimation on upper body. Zhang and Shah [9] conceived a sophisticated temporal edge including optical flow predicted location distance and the Chi-square distance of HOG features. All of the above methods are insightful, however, none of them exploited the temporal information deep enough, i.e., [1] treated the problem as a memory one system, and this problem improved a bit in [3, 8, 9], which treated as a memory two system in a broad sense.

Dynamics: Linear dynamic systems have been recently used in a wide range of computer vision applications, and its validity has been tested in [10, 11, 12]. They proposed Hankelets (the Hankel matrix of a short tracklet) as a new representation for activities. The method proposed in this thesis uses Hankel matrix of the human joints to capture the spatial and temporal dynamics and aims to incorporate much deeper temporal relations across time frames.

1.4 Thesis Organization

The following provides a brief explanation to each chapter:

Chapter 2, explains the background of dynamic systems, auto regression, Hankel matrix and Nuclear norm.

Chapter 3, presents a Nuclear norm of Hankel matrix metric for tracking-by-selection video based pose estimation framework, and also introduces the BFC search algorithm.

Chapter 4, demonstrates the effectiveness of the proposed dynamic-based selection algorithm for human pose estimation in MoCap dataset and in real videos, and the reasons of failure cases are discussed.

Chapter 5, draws the conclusion for the entire thesis and briefly discusses the future works.

Chapter 2

Background

2.1 Dynamical Systems

A dynamical system is a system in which a function describes the time dependence of a point in a geometrical space. Recently, dynamical systems play an important role in a wide range of practical computer vision applications, such as object tracking, human activity recognition and dynamic texture recognition. The main advantage of dynamics is that it captures the essence of the temporal evolution of the data in a compact way that it is suitable for both analysis and synthesis. Given a temporal sequence of a measurement vector $\mathbf{y}_k \in \mathbb{R}^n$, the goal is to model its temporal evolution as a function of a relatively low dimensional state vector $\mathbf{x}_k \in \mathbb{R}^d$ that changes over time. The simplest dynamical model is the linear time invariant (LTI) system as following:

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{w}_k \tag{2.1a}$$

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1}, \qquad \mathbf{x}_0 \text{ given} \tag{2.1b}$$

where both equations are linear, the matrices $\mathbf{C}$ and $\mathbf{A}$ are time invariant constants, and where $\mathbf{w}_k \sim N(0, Q)$ is uncorrelated zero mean Gaussian measurement noise. $\mathbf{x}_k \in \mathbb{R}^d$ is the d-dimensional hidden state of the LTI system, while $\mathbf{y}_k \in \mathbb{R}^n$ is the n-dimensional measurement. The dimension of the stare vector $\mathbf{x}_k$, d, is the order (memory) of the system and is a standard measure of the complexity of the system [10].

It should be noted that there is a non-negligible limitation for the LTI model in the practical computer vision applications. One must assume or estimate the dimensions and values of the matrices $\mathbf{A}$ and $\mathbf{C}$ and the initial vector $\mathbf{x}_0$. Furthermore, given a finite number of measurements of $\mathbf{y}_k$, the set of triples $(\mathbf{A}, \mathbf{C}, \mathbf{x}_0)$ that could have generated this data is not unique. This is related to the concepts of consistency set and diameter of information. As a result, any attempt to jointly identify the dynamics $(\mathbf{A}, \mathbf{C})$ and the initial condition $\mathbf{x}_0$ leads to computationally challenging non convex problems. To bypass these identification difficulties, in this thesis we will build the Hankel matrices to capture the dynamic information of the systems rather than working with the model representation itself (2.1).

2.2 Hankel Matrices

Given a sequence of output measurements from the system (2.1), $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, ...$, its associated block Hankel matrix $\mathbf{H}_{\mathbf{y}}^{s,r}$ is:

$$\mathbf{H}_{\mathbf{y}}^{s,r} = \begin{bmatrix} \mathbf{y}_0 & \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_r \\ \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 & \cdots & \mathbf{y}_{r+1} \\ \vdots & \vdots & \ddots & \vdots & \\ \mathbf{y}_s & \mathbf{y}_{s+1} & \mathbf{y}_{s+2} & \cdots & \mathbf{y}_{r+s} \end{bmatrix} \tag{2.2}$$

Note the columns of the Hankel matrix correspond to overlapping subsequences of data, shifted by one, and that the block anti-diagonals of the matrix are constant as visualized in Figure 2.1.

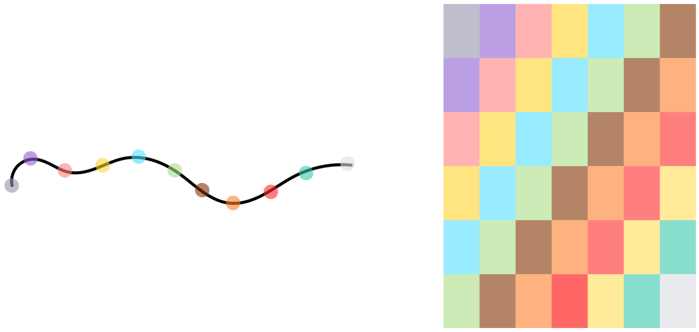


Figure 2.1: The line in the left represents a trajectory with 11 data points, and each point is a measurement vector; the matrix in the right shows how to stack the measurement vectors of the data points from overlapping subsequencies into a Hankel matrix that has repeating block-antidiagonals. Each color block could be a multidimensional vector.

As explained in [13], this special structure of Hankel matrix is what encapsulates the dynamic information of the system. In particular, a well known result from realizaion theory is that, under mild conditions, the rank of the Hankel matrix is the order n of the system $\text{rank}(\mathbf{H}_{\mathbf{y}}^{s,r}) = n$ provided that $r, s \geq n$. Furthermore, by writing $\mathbf{y}_k$ using an n^{th} order auto regressive model of the form:

$$\mathbf{y}_k = \sum_{i=1}^{n} a_i \mathbf{y}_{k-i} \tag{2.3}$$

$$= [\mathbf{y}_{k-1}\ \mathbf{y}_{k-2}\ \ldots\ \mathbf{y}_{k-n}]\mathbf{a} \tag{2.4}$$

where $\mathbf{a} = [a_1\ a_2\ ...\ a_n]^T$ are a set of coefficients, and by setting $r = n$ in (2.2), it is easy to see that the last column of the Hankel matrix is a linear combination of the previous ones and that the coefficients of this combination, $\mathbf{a}$, are exactly the coefficients of the auto regressor. That means, the coefficents in (2.4) also satisfies the following equation:

$$\mathbf{H}_{\mathbf{y}}^{s,n}\ [\ \mathbf{a}^T\ \ -1\ ]^T = 0 \tag{2.5}$$

Then the coefficents $\mathbf{a}$ can be obtained by computing sigular value decomposition (SVD) on Hankel matrix $\mathbf{H}_{\mathbf{y}}^{s,n}$:

$$\mathbf{H}_{\mathbf{y}}^{s,n}\ \mathbf{v} = 0 \cdot \mathbf{v} \tag{2.6}$$

where $\mathbf{v} = [\ \mathbf{a}^T\ \ -1\ ]^T$. Therefore $\mathbf{v}$ is the eigenvector corresponding to the zero eigenvalue of the matrix.

A useful property of Hankel matrix worth mentioning is dynamic subspace invariance to affine transformations. The columns of two Hankel matrices corresponding to a trajectory and its affine transformation, span the same linear subspace which is orthogonal to the auto regressor vector of the trajectories $\mathbf{v} = [\ \mathbf{a}^T\ \ -1\ ]^T \in \mathbb{R}^{n+1}$. This property can be easily shown by writing $\mathbf{Y}_k = \sum a_i \mathbf{Y}_{k-i}$ and using the fact that affine transformations $\mathbf{y}_k = \Pi \mathbf{Y}_k$ are linear [10]. Then,

$$\begin{aligned}\mathbf{y}_k &= \Pi \mathbf{Y}_k = \Pi \sum_{i=1}^{n} a_i \mathbf{Y}_{k-i} \\ &= \sum_{i=1}^{n} a_i \Pi \mathbf{Y}_{k-i} = \sum_{i=1}^{n} a_i \mathbf{y}_{k-i} \end{aligned} \tag{2.7}$$

and hence the two Hankel matrices generated by $[\mathbf{y}_0, \mathbf{y}_1, ...]$ and $[\mathbf{Y}_0, \mathbf{Y}_1, ...]$ share the same auto regressor, and the same system complexity.

2.3 Rank and Nuclear Norm

From section 2.2, we know the fact that the order of the system is the rank of its associated Hankel matrix, however only the rank of a clean matrix can be calculated, and the rank of any noisy matrix is hard to estimate. In general, the rank of a matrix can be estimated by singular value decomposition (SVD). Namely, $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ is an $s \times s$ unitary matrix, $\mathbf{\Sigma}$ is a $s \times r$ rectangular diagonal matrix with singular values on its diagonal, and $\mathbf{V}$ is an $r \times r$ unitary matrix. Then we estimate the dominant singular values by principal component analysis to measure the rank. We increase the value of n and sum the normalized singular values until the following inequality is satisfied:

$$\frac{\sum_{i=1}^{n \leq min(s,r)} \mathbf{\Sigma}_{i,i}}{\sum_{j=1}^{min(s,r)} \mathbf{\Sigma}_{j,j}} \geq t \tag{2.8}$$

where t is the threshold between 0 and 1, typical values are $t = 0.95, 0.98, 0.99$.

However, calculating rank according to (2.8) is time consuming. An alternative measure that closely relates to rank is Nuclear norm, who often substitute the positions of rank in rank minimization problems. Nuclear norm is equivalent to the l_1-norm of the vector of its eigenvalues. Thus, when minimizing the Nuclear norm of a matrix we are injecting sparsity to the vector of eigenvalues. Essentially, this sparsity means the rank of the original matrix is reduced.

The Nuclear norm of a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, denoted by $||\mathbf{X}||_*$, is defined as the sum of all its singular values, i.e.,

$$||\mathbf{X}||_* = \sum_{i=1}^{r} \sigma_i(\mathbf{X}) \tag{2.9}$$

where $\sigma_i(\mathbf{X})$ means the i-th largest singular value of $\mathbf{X}$ and r is simply the rank. The Frobenius norm is defined as the square root of the sum of the squares of its elements, and it also equals to the Euclidean norm of the vector of singular values, i.e.,

$$||\mathbf{X}||_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{r} \sigma_i(\mathbf{X})^2 \right)^{\frac{1}{2}} \tag{2.10}$$

We also introduce the operator norm of $\mathbf{X}$, which is equal to its largest singular value, i.e., $||\mathbf{X}|| = \sigma_1(\mathbf{X})$. These three norms have the following relations which hold for any matrix $\mathbf{X}$ of rank at most r:

$$||\mathbf{X}|| \leq ||\mathbf{X}||_F \leq ||\mathbf{X}||_* \leq \sqrt{r}||\mathbf{X}||_F \leq r||\mathbf{X}|| \tag{2.11}$$

By the chain of (2.11), we know that $r \geq ||\mathbf{X}||_*/||\mathbf{X}||$, and it shows in [14] that this is the tightest convex lower bound for the rank of matrix $\mathbf{X}$.

Chapter 3

Selection by Nuclear Norm of Hankel Matrices

Human pose estimation is crucial for many computer vision applications, and in recent years, a significant effort has been devoted to estimating human poses in single images [5, 15, 16]. However, human pose estimation in videos is a relatively new and challenging problem [9]. The focus of this thesis is to estimate poses from videos, and in particular based on detection results from single images, we want to improve the results through imposing temporal constraints.

3.1 Framework

Due to the innate complexity of video data, the problem formulations of video based pose estimation models are very complex (usually NP-hard). One commonly considered approach is the tracking-by-selection framework [1, 3], which breaks down the whole problem into two stages: in the first step, a set of (for example, N) candidate poses are generated from each frame of a video; in the second step, frame-frame temporal smoothness metrics among poses or different body parts are introduced, and by minimizing an optimization formulation the best suitable set of pose estimations for the video is obtained.

Based on the tracking-by-selection framework, the problem boils down to find a proper metric that captures the temporal nature in videos. For the purpose of illustration, we introduce the general notions here. since most of the works formulate the video based human pose estimation problem into a graph-based optimization framework. Assume there is a set of entities $\mathcal{E} = \{e^i|_{i=1}^{M}\}$ where each entity can only be in one of the many states $\mathcal{S} = \{s^k|_{k=1}^{N}\}$, with the unary scoring functions $\{\Phi(e^i, s^k)|e^i \in \mathcal{E}, s^k \in \mathcal{S}\}$, which gives the likelihood that an entity e_i in state s^k. And there is a binary compatibility function for each pair of entities $\{\Psi(e^i, e^j, s^k, s^l)|e^i, e^j \in \mathcal{E}, s^k, s^l \in \mathcal{S}\}$, which represents the compatibility of entity e^i in state s^k and entity e^j in state s^l. The goal is to determine the best states for each entity such that all of them have high unary scores and they are also compatible with each other. The hypothesis graph $G = (V, E)$ represents the relationship of a set of which are represented by entity nodes $\{v^i|_{i=1}^{|V|}\}$, and the relationships between pairs of hypothesis are represented in edges E. See an example in Figure 3.1

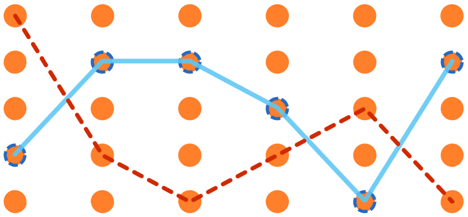


Figure 3.1: This is a typical graph model for a 6-frame image sequences, and at each frame, we have 5 candidates to choose from. Ground truth trajectory in blue and an arbitrary trajectory is the dotted red line. A good graph model considers the blue trajectory better than the red dotted one.

3.2 Previous Works

To our best knowledge, almost all the previous works formalize the problem into the following optimization problem – the model selects the ”best” poses $(P_1^{n(1)}$,$P_2^{n(2)}$,...,$P_T^{n(T)})$ for T frames by minimizing the cost function $Cost$:

$$n(1:T) = \arg \min_{(n(1),...,n(T))} Cost(P_1^{n(1)}, P_2^{n(2)}, ..., P_T^{n(T)}) \tag{3.1}$$

$$Cost = \sum_{i=1}^{T} \Phi(P_i^{n(i)}) + \sum_{i=1}^{T-1} \Psi(P_i^{n(i)}, P_{i+1}^{n(i+1)}) \tag{3.2}$$

where $\Phi(P_i^{n(i)})$ is a unary term that measures the likelihood of the pose and $\Psi(P_i^{n(i)}, P_{i+1}^{n(i+1)})$ is a pairwise term that measures consistency of the joints in consecutive frames. For example, in [1], they simply use the negative of the total squared pixel difference between each joint in pose at time $t-1$ and pose at time t. And in a one-step better approach [8, 9, 17, 18, 19], they used second order difference (in other words, they assume every joint moves at its own constant velocity), for example, they use the difference between optical flow predicted location and the detection location, besides, some of them also use appearance constraints, such as the shape of each body part should not change rapidly between adjacent video frames [9, 19], and the color of body part should be stable in successive video frames [18, 19]. They all only introduce temporal links between every pair of current frame and its previous frame for imposing temporal consistency, ending up with a trellis graph, which could be solved by dynamic programming used in viterbi decoding. Indeed, they craft the graph so that global optimal solution is achievable, but only very shallow temporal information is exploited by this energy function of memory one (or two because of optical flow) and by no means should one consider this trellis graph as the most suitable model for reality. However, by introducing Nuclear norm to Hankel matrix, we are able to exploit much more temporal information and thus temporal consistency is guaranteed easily. We show detail in next section.

3.3 Proposed Method

This section describes the proposed selection method. We want to obtain a set of skeletons that describe the action of a human in a video. Use off-the-shelf algorithm [1] we are able to generate several different pose estimations at each frame, but due to the imperfections of the algorithm, these pose estimations may not overlap with the real pose in this frame, however they may have different level of accuracy, therefore the problem is to devise a metric to stitch the best pose estimations among the estimations at each frame such that (a): pose configurations are consistent across adjacent frames and (b): pose configurations at each frame should be as accurate as possible. To be sure, (a) and (b) are typically interweaving with each other. Accurate pose configuration at every frame ensures temporal consistency across frames. In our method, we are trying to guarantee temporal consistency among adjacent frames so as to boost detection accuracy.

Two reasonable assumptions are needed for our method to work perfectly. Firstly, natural human activities are inherently characterizable in low order dynamic systems. This has been proven a safe supposition in a number of applications [11, 10]. Secondly, stitching noisy observations of each frame tends to have a higher order dynamics. This is because noise follows a certain probability distribution around the true value, and since observations are independent between frames, a low order noise trajectory with length n would happen with a probability that is the product of n probabilities, which is usually close to zero for large n. Having this two assumptions, we could claim that among all the trajectories the one with the lowest dynamic order is bound to be the true human activity trajectory.

As mentioned earlier, the problem can be abstracted as following. Assume we generate

N human pose candidates at every frame individually in a video of T frames, see Fig. 3.1 for an illustration. A pose estimation $P \in \mathbb{R}^{2k}$, which denotes the 2-D coordinates of k joints, namely $P_{\cdot}^{(\cdot)} = \begin{bmatrix} x_1 & y_1 & \dots & x_k & y_k \end{bmatrix}^T$. Let $P_t^{n(t)}$ be the $n(t)$-th pose estimation in the t-th frame, where $n(t) \in \{1, 2, \dots, N\}$, and we name it selection trajectory because its element denotes which candidate is selected at this particular frame, and $t \in \{1, 2, \dots, T\}$. Given a sequence of pose estimations up to time T, $P_1^{n(1)}$,$P_2^{n(2)}$,...,$P_T^{n(T)}$, its associated Hankel matrix $\mathbf{H}_{1:T}^{n(1:T)} \in \mathbb{R}^{2km\times(T-m+1)}$ as:

$$\mathbf{H}_{1:T}^{n(1:T)} = \begin{bmatrix} P_1^{n(1)} & P_2^{n(2)} & \dots & P_{T-m+1}^{n(T-m+1)} \\ P_2^{n(2)} & P_3^{n(3)} & \dots & P_{T-m+2}^{n(T-m+2)} \\ \vdots & \vdots & \ddots & \vdots \\ P_m^{n(m)} & P_{m+1}^{n(m+1)} & \dots & P_T^{n(T)} \end{bmatrix} \tag{3.3}$$

Note that the columns of the Hankel matrix correspond to overlapping sub-sequences of the data, shifted by one. As shown in section 2.2, Hankel matrix carries useful invariant properties. Specifically, if m and T are selected in a way that $r = \text{rank}(\mathbf{H}_{1:T}^{n(1:T)}) < \min\{m, T-m+1\}$, then r measures the complexity of the underlying dynamics. We can easily formalize the selection part as below:

$$\begin{aligned} \min \quad & \text{rank}(\mathbf{H}_{1:T}^{n(1:T)}) \\ s.t. \quad & n(1:T) \in \{1, 2, \dots, N\} \\ & \mathbf{H}_{1:T}^{n(1:T)} = \text{hankel}(P_1^{n(1)}, ..., P_T^{n(T)}) \end{aligned} \tag{3.4}$$

A major challenge in computing equation (3.4) is that one has to estimate the rank of noisy structured Hankel matrices. The simplest way is to count the singular values until the sum exceeds a threshold as in inequality (2.8), but this will not work as expected, and we will show in later experiments. An alternative to address these issues is to solve a similar convex relaxtion using the method in [20]:

$$\begin{aligned} \min \quad & ||\mathbf{A}||_* + \lambda||\mathbf{E}||_F \\ s.t. \quad & \mathbf{H}_{1:T}^{n(1:T)} = \mathbf{A} + \mathbf{E} \in \mathcal{S}_{\mathcal{H}} \\ & \text{Hankel constraints on } \mathbf{A} \text{ and } \mathbf{E} \end{aligned} \tag{3.5}$$

where $\mathcal{S}_{\mathcal{H}}$ represents the set of Hankel matrices it can choose from and is defined as in equation (3.4), and the Hankel structural matrices $\mathbf{A}$ and $\mathbf{E}$ are the denoised Hankel matrix and noise Hankel matrix respectively. As pointed in section 2.3 and [14, 12, 21, 22], Nuclear norm is a tight convex lower bound to rank, and we are assuming that by minimizing the lower bound, the original matrix tends to have lower rank, thus the skeleton sequence tends to be more smooth. However, given an arbitrary vector $n(1{:}T)$, we can compose a Hankel metrix $\mathbf{H}_{1:T}^{n(1:T)}$ accordingly. Obviously, the trajectory $n(1{:}T)$ has T (the number of frames in the video) elements and each element is a choice from N (the number of pose estimations at each frame) candidates, which shows that $n(1{:}T)$ has N^T (the size of the sets $\mathcal{S}_{\mathcal{H}}$) that many different selections and the same amount of Hankel matrices associated with them. Each selection trajectory $n(1{:}T)$ should map a distinct penalty score for this selection. And for each possible trajectory, we need to optimize equation (3.5) to obtain the penalty score. This procedure requires a number of SVDs at each iteration, and this is unbearably time consuming and not practical for the time being. We change objective function in equation (3.5) into:

$$\begin{aligned} \min \quad & ||\mathbf{H}_{1:T}^{n(1:T)}||_* \\ s.t. \quad & \mathbf{H}_{1:T}^{n(1:T)} \in \mathcal{S}_{\mathcal{H}} \end{aligned} \tag{3.6}$$

where $\mathcal{S}_{\mathcal{H}}$ follows the same definition in (3.4). Now we only need a single SVD for each trajectory, so it scales properly with the size of the solution set. One thing needs more explanation is that why is optimizing (3.6) a good approximation of (3.5). We will try to answer this in the following part.

An interesting linear algebra problem is to find how the eigen-values $\lambda_1(\mathbf{A})$, ..., $\lambda_n(\mathbf{A})$ and $\lambda_1(\mathbf{B})$, ..., $\lambda_n(\mathbf{B})$ of two $n \times n$ Hermitian matrices $\mathbf{A}, \mathbf{B}$ constrains the eigen-values $\lambda_1(\mathbf{A}+\mathbf{B})$,..., $\lambda_n(\mathbf{A}+\mathbf{B})$ of their sum, all the eigen-values are in decreasing (non-increasing) order. Now, the eigen-values are not linear functions of the matrix, and no simple relation is apparent, except one, the trace of $\mathbf{A}$, denoted by tr$(\mathbf{A})$ is the sum of the diagonal entries of $\mathbf{A}$ and also the sum of engenvalues of $\mathbf{A}$, following the linearity of trace:

$$\text{tr}(\mathbf{A}+\mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B}) \tag{3.7}$$

when expressed in terms of eigenvalues, gives the trace constraint:

$$\begin{aligned} &\lambda_1(\mathbf{A}+\mathbf{B}) + \ldots + \lambda_n(\mathbf{A}+\mathbf{B}) \\ &= \lambda_1(\mathbf{A}) + \ldots + \lambda_n(\mathbf{A}) + \lambda_1(\mathbf{B}) + \ldots + \lambda_n(\mathbf{B}) \end{aligned} \tag{3.8}$$

And some other not so obvious inequalities, such as,

$$\lambda_1(\mathbf{A}+\mathbf{B}) \leq \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B}) \tag{3.9}$$

$$\lambda_n(\mathbf{A}+\mathbf{B}) \geq \lambda_n(\mathbf{A}) + \lambda_n(\mathbf{B}) \tag{3.10}$$

We know that every Hermitian operator $\mathbf{A}$ can be diagonalized in some orthonormal basis, $\mathbf{A} = \sum \lambda_j(\mathbf{A}) u_j u_j^T$. Using this, it is easy to see that the set $\{\langle x, \mathbf{A}x \rangle : ||x|| = 1\}$ is equal to the interval $[\lambda_n(\mathbf{A}), \lambda_1(\mathbf{A})]$, and this implies us inequality (3.9), and equality occurs exactly when the same vector is a principal eigen-vector for both matrices, similar reason for inequality (3.10).

The complete answer to this problem is a tricky one and far beyond the scope of this thesis, requiring a strangely recursive description (once known as Horn's conjecture, which is now solved), and connected to a large number of other fields of mathematics, such as geometric invariant theory, intersection theory, and the combinatorics of a certain gadget known as a ”honeycomb” [23].

In our application, one of the matrices is ”small” in some sense, say $\mathbf{B}$ is the noise matrix added to the clean data matrix $\mathbf{A}$, so that $\mathbf{A} + \mathbf{B}$ is a perturbation of $\mathbf{A}$. In this case, one does not need the full strength of the above theory, and instead rely on the simple Weyl's inequalities [24]:

$$\lambda_{i+j-1}(\mathbf{A} + \mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}) \tag{3.11}$$

valid whenever $i, j \geq 1$ and $i + j - 1 \leq n$. We simply choose $j = 1$, and now we have:

$$\lambda_i(\mathbf{A} + \mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B}) \tag{3.12}$$

valid for $1 \leq i \leq n$. And notice that the eigen-values of $-\mathbf{B}$ are the same as the negatives of the eigenvalues of $\mathbf{B}$, but taking negatives reverses order, namely:

$$\lambda_j(-\mathbf{B}) = -\lambda_{n-j+1}(\mathbf{B}) \tag{3.13}$$

valid for $1 \leq j \leq n$. Combining inequality (3.12) and equation (3.13), we end up with:

$$\lambda_i(\mathbf{A}) + \lambda_n(\mathbf{B}) \leq \lambda_i(\mathbf{A} + \mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B}) \tag{3.14}$$

for $1 \leq i \leq n$. Note inequalities (3.14) is stronger than equation (3.3). One consequence of these inequalities is that the singular values of a Hermitian matrix is stable with respect to small perturbations [24].

However, we are dealing with Hankel matrices, which are not necessarily Hermitian matrices. We know the singular values of a $n \times m$ matrix $\mathbf{H}$ are more or less the eigenvalues of the $(n + m) \times (n + m)$ matrix $[\,\mathbf{0} \quad \mathbf{H};\ \mathbf{H}^T \quad \mathbf{0}\,]$. Using this, one can deduce inequalities for the singular values from that of the Hermitian matrices problem. Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ be given, and let $q = \min\{m, n\}$. The following inequalities hold for the decreasingly (non-increasingly) ordered singular values of $\mathbf{A}, \mathbf{B}$ and $\mathbf{A}+\mathbf{B}$ [25]:

$$\sigma_{i+j-1}(\mathbf{A}+\mathbf{B}) \leq \sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B}) \tag{3.15}$$

$$|\sigma_i(\mathbf{A}+\mathbf{B}) - \sigma_i(\mathbf{A})| \leq \sigma_1(\mathbf{B}) \tag{3.16}$$

$$\sum_{i=1}^{k} \sigma_i(\mathbf{A}+\mathbf{B}) \leq \sum_{i=1}^{k} \sigma_i(\mathbf{A}) + \sum_{i=1}^{k} \sigma_i(\mathbf{B}) \tag{3.17}$$

valid when $1 \leq i, j, k \leq q$ and $i+j-1 \leq q$. Please note that except for the largest singular value (when $k=1$ in inequality (3.17)), individual singular values need not obey the triangle inequality (only the sum obeys). Therefore in our case, let $\mathbf{B} = \mathbf{E}$ as the small additive noise in Hankel structure, let $\mathbf{A}$ be the Hankel matrix formed by clean trajectory, and $\mathbf{H} = \mathbf{A}+\mathbf{E}$, the corrupted measurement trajectory. From the above discussion evidently,

$$||\mathbf{A}||_* - q||\mathbf{E}|| \leq ||\mathbf{A}||_* - ||\mathbf{E}||_* \leq ||\mathbf{A}+\mathbf{E}||_* \leq ||\mathbf{A}||_* + ||\mathbf{E}||_* \leq ||\mathbf{A}||_* + q||\mathbf{E}|| \tag{3.18}$$

where $q = \text{min}\{m, n\}$. Using the chain of (2.11), we know,

$$||\mathbf{A}||_* - \sqrt{q}||\mathbf{E}||_F \leq ||\mathbf{A}+\mathbf{E}||_* \leq ||\mathbf{A}||_* + \sqrt{q}||\mathbf{E}||_F \tag{3.19}$$

Thus, the optimization in (3.6) minimizes the upper bound for $||\mathbf{A}||_* - \sqrt{q}||\mathbf{E}||_F$ and the lower bound for $||\mathbf{A}||_* + \sqrt{q}||\mathbf{E}||_F$, which is close to $||\mathbf{A}||_* + \lambda||\mathbf{E}||_F$ in some sense. Therefore we have the same conclusion in Hermitian matrix, that Nuclear norm is stable with respect to small perturbations. This explains the validity of using Nuclear norm of noisy data.

3.4 BFC Search Algorithm

Now, we address problem that the solution space exponentially increases with respect to candidates number N at each frame and the length of the video T. We use a Breadth First Cut (BFC) search algorithm, at first we keep the all the possible trajectories until frame T_a, this is for accumulating input data for meaningful Hankel matrix. For now we have N^{T_a} candidates, and we expand the combination at frame T_a+1, so we end up with N^{T_a+1} candidate trajectories, who are ordered according to the Nuclear norm of their associated Hankel matrix. Next, only top Q promising trajectories are left, and the other $N^{T_a+1} - Q$ candidate trajectories are pruned for the sake of memory as well as time cost. Then, we again expand the trajectory combinations, by one frame, to frame T_a+1, the total trajectories are $Q \times N$ this time, and again they are ordered by the value of their Nuclear norm, next top Q trajectories are kept, and the rest are pruned, and so forth. In the end, we have Q trajectories, and the first one is the desired solution.

BFC search algorithm is not the one returns the global optimal solution, but it works perfectly with the structure of Hankel matrix. We conducted several experiments on examining its ability to hit the global optima. In the experiment setting, we choose several activity sequences from MHAD dataset, which is introduced in the next section, each activity sequence was truncated at length of 11 frames, and 4 other noisy candidates are generated at each frame. Therefore, in these experiments, the solution space is of size $5^{11} = 48,828,125$. We run the BFC search algorithm with $Q = 10,000$, and it turns out the kept top 10,000 trajectories from BFC search algorithm overlap completely in the same order as the top 10,000 trajectories from exhaustive breath first search results. This shows that the Nuclear norm for the Hankel matrices built from trajectories are order-preserving from frame to frame and BFC search algorithm fits Nuclear norm metric well. Outline of the whole algorithm is in below.

Algorithm 1: BFC search on Nuclear norm of Hankel matrices

```
Data: Detection data of length $T$ frames, $N$ hypotheses at each frame.
Result: A trajectory consists of a sequence of best hypothesis at each frame.
initialization: $i = 0, T_a, Q$;
while $i \leq T$ do
    $i = i + 1$;
    if $i < T_a$ then
        build all the Hankel matrices associated with the $N^i$ trajectories;
        sort the $N^i$ trajectories according to the Nuclear norm of their Hankel
        matrices;
    else
        build the Hankel matrices associated with all the kept ($N^{T_a}$ or $QN$)
        trajectories;
        sort all the trajectories according to the Nuclear norm of their Hankel
        matrices;
        keep the top $Q$ trajectories and prune the rest;
    end
end
return the first trajectory.
```

Chapter 4

Application in MoCap Data and Videos

4.1 Pose Estimation in MoCap Data

4.1.1 MHAD Dataset

The motion capture dataset we choose is Berkeley Multimodal Human Action Database (MHAD), it contains 11 actions performed by 12 subjects for 5 repetitions, yielding around 660 (659, to be precise) action sequences [26]. All the 659 actions in MHAD dataset are used in the experiment results. Two actions in MHAD dataset, a jumping jack action (the 10-th sequence), and a bending action (the 180-th sequence), are used here for illustrative purpose. The former action is with movement in both upper and lower extremities and the latter action performed with still lower extremities and high dynamics in upper extremities. The frequency of the motion capture sampling is extremely high; every second it samples 480 times, so the difference between adjacent frames are extremely small given these activities are performed by human beings. Therefore we resampled the motion capture data points at a frequency of around 5 Hz. The activity sequences could be treated as clean data because the resolution of joint locations is less than 1 mm according to [26] and the Nuclear norm of the activity trajectory is very close to the Nuclear norm of ideal trajectory under such small noise.

4.1.2 Experiment Results

Firstly, we corrupt the clean sequence with different kinds of noise, such as additive noise, as well as some structured noise to obtain distracting candidates at each frame. In here, we generate 10 hypothesis at each frame, only one of them is the clean, and we keep top 100 promising trajectories

Table 4.1: Experiments Results on MHAD dataset, MaxNoise = 1. Small noise ($0.5\times$ MaxNoise) corrupted clean trajectory (ground truth), additive noise with MaxNoise added to the ground truth togenerate other 9 candidates.

	6 joints		**25 joints**	
Method	**Average Accuracy**	**Error**	**Average Accuracy**	**Error**
1-st order Difference	0.3339	29.8506	0.8234	24.5283
2-nd order Difference	0.3461	29.1815	0.8649	19.2938
SVD	0.1907	32.5022	0.4332	59.0333
Hankel+Nuclear, 1 R	**0.5269**	**22.9796**	**0.9429**	**10.2245**
Hankel+Nuclear, 2 R	0.4867	23.9951	0.9098	15.5085
Hankel+Nuclear, 3 R	0.4379	25.5970	0.8012	28.6562
Hankel+Nuclear, 4 R	0.4110	26.6804	0.7387	33.7511
Hankel+Nuclear, 5 R	0.3945	27.2430	0.6967	36.9594

in the BFC search algorithm. And then we select a trajectory with the smallest Nuclear norm based on the premise that natural action sequences performed by human should feature low order dynamics (comparing to random noise) therefore small Nuclear norm.

Full results are shown in Table 4.1. In the table, the 1-st Difference and 2-nd Difference method are baselines. In 1-st Difference, we define the pairwise term in (3.2) as $\Psi(P_i^{n(i)}, P_{i+1}^{n(i+1)}) = ||P_i^{n(i)} - P_{i+1}^{n(i+1)}||_2$ as in [1], and for 2-nd Difference, we difine the pairwise term in (3.2) as $\Psi(P_i^{n(i)}, P_{i+1}^{n(i+1)}) = ||P_{i-1}^{n(i-1)} - 2P_i^{n(i)} + P_{i+1}^{n(i+1)}||_2$ simulating the ideas in [8, 9, 17, 18, 19]. The SVD method is simply counting the number of singular values until their sum exceeds a threshold. This method does not work well, mainly because the score it assigns to each trajectory is an integer, which is not discriminative enough for ordering. Finally Hankel+Nuclear norm is the method we proposed, the ”1 R” means we form the Hankel matrices in one row, basically, they are the trajectories themselves without repeating. And ”5 R” means we form the Hankel matrices with 5 rows; some candidates appear in the structure as many as 5 times. The reason ”1 R” works better than ”5 R” is because spatial smoothness is enough to explain the data. The original skeleton poses consist of 35 joints, we remove some unnecessary joints to form the 25 joints skeleton poses, and we also reduce them to 6 joints to see if a small set of joints catches the dynamics of the activity. The Average Accuracy measures the frequencies of the algorithm selecting the ground truth poses, while Error measures the Euclidean distance between the selected trajectory to the ground truth trajectory. These two measurements are closely related as shown in the table. One thing worth mentioning is that, even though the 1 row form Hankel matrix method achieves best accuracy and lowest error overall, it does not implies it achieves best results for every activity. In fact, these results are outcomes of simple activities with low dynamics.

4.1.3 Illustrative Examples and Discussion

We change the experiment setting a bit in these examples, such as, the frame rate is now 30 Hz instead of 5 Hz. In Figure 4.1, the results of the jumping jack action (the 10-th sequence, re-sampled to 30 Hz, 25 joints) are shown. And Figure 4.2 is the visualization. Please note our algorithm picks exactly the ground truth pose configuration in every frame of both sequences, in other words, the accuracy of the selection given by our algorithm is 100%. However, the 1-st Difference approach and 2-nd Difference approach perform fairly well in this experiment settings as well, therefore we re-sampled the original sequence to 6 Hz to see the limits of these methods.

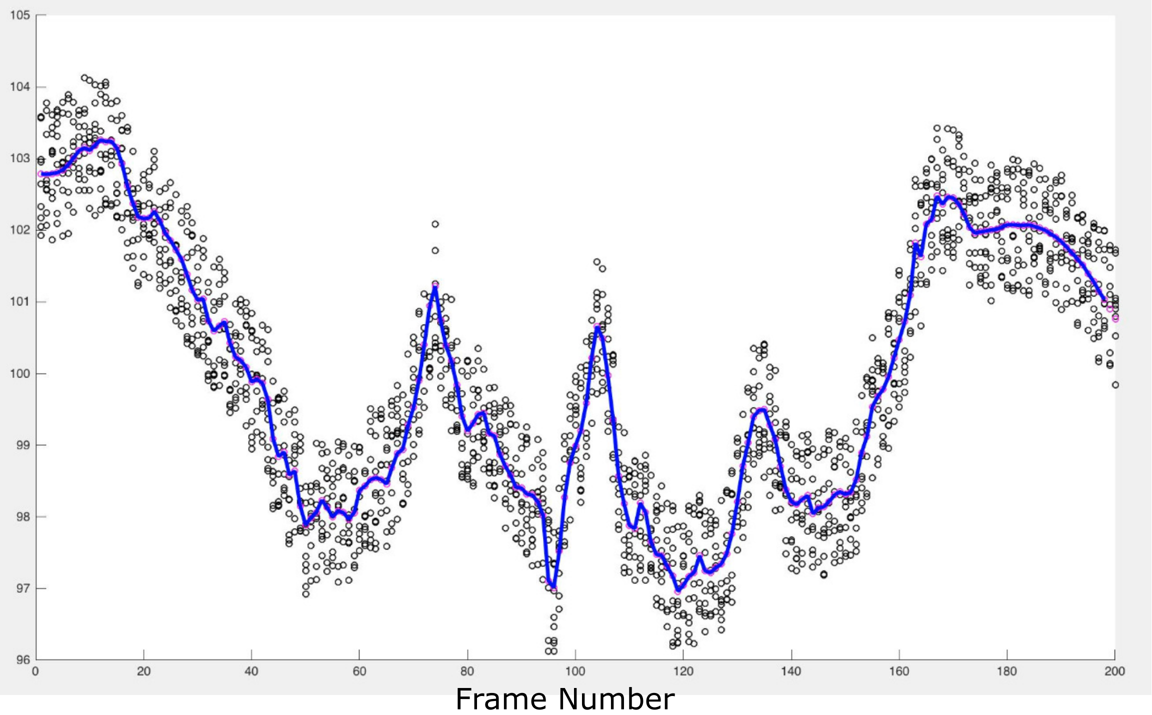


Figure 4.1: The x axis is the frame number, while the y axis indicates the illustrative positions of the 10 candidate poses (actually, the circle stands for the position of the first coordinate of the heap joint in that pose skeleton). Ground truth trajectory are the circles in purple and our BFC search algorithm finds all the ground truth point in every frame correctly.

During our experiments, we find that if the dynamic of the structured noise is in lower order

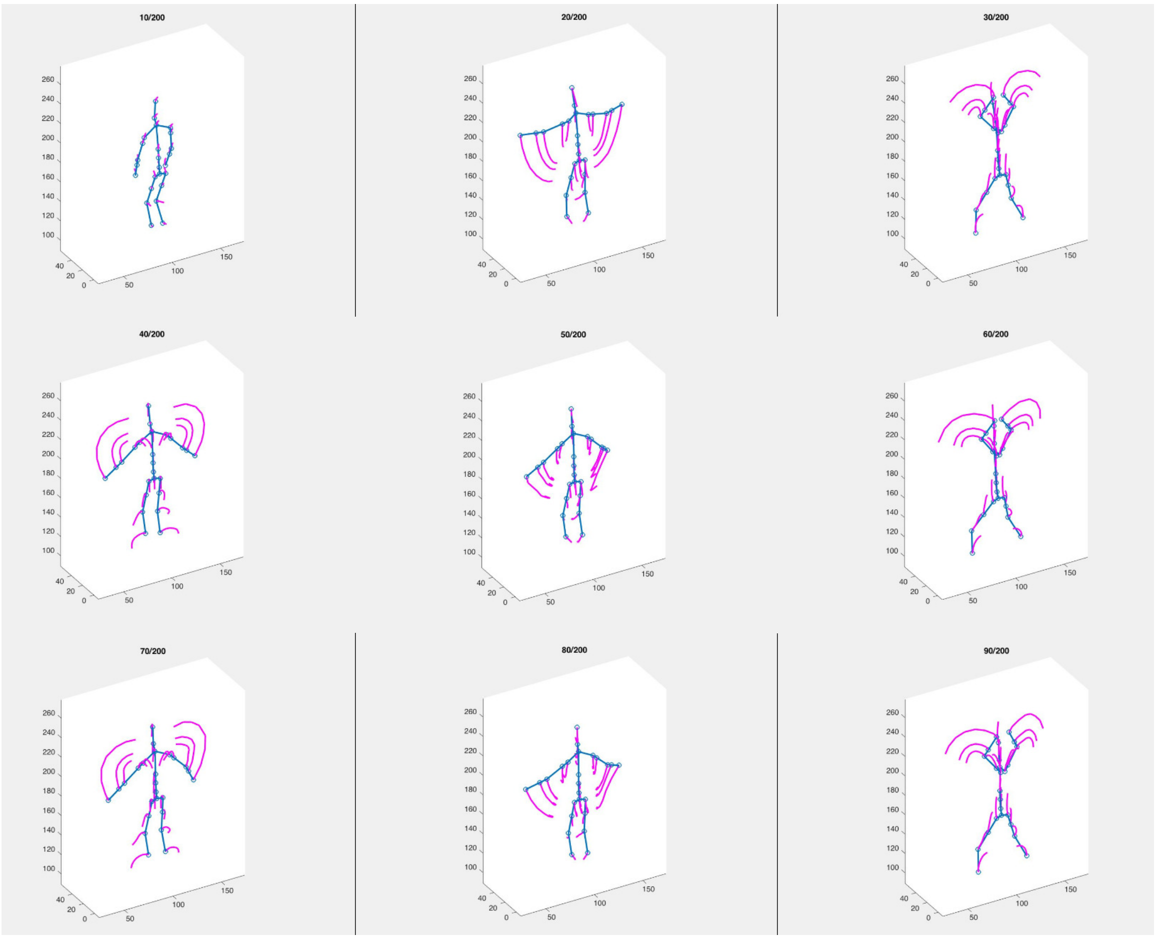


Figure 4.2: The visualization of the selected trajectory in Figure 4.1 is shown. Notice that the purple trajectories are smooth.

than the ground truth action, our algorithm picks the lower order structured noise Figure 4.3, which means that our algorithm works as expected exactly, and in reality it is a fairly weak assumption that noise should have higher order dynamics than activity sequences, in other words we are assuming that noise should be random and independent across adjacent frames instead of acting in a structured way. As time goes by and more data comes in, our algorithm demonstrates its ability to correct previous inferior selections, which is a desirable advantage.

An interesting phenomenon is that there are two different implementations of 1-st order Difference method, 1) as pointed out in [1, 9], since it is a trellis graph, it can be solved by a viterbi decoding dynamic programming algorithm, and the global optima is returned by the algorithm; 2) the other implementation is using our BFC search framework. Note this algorithm does not guarantee the global optima of the graph definition. However, the latter BFC search framework achieves 100% accuracy while the global optima solution via dynamic programming fails on more than half of the frames. This experiment shows even though global optimal solution is desirable, but it may work worse when the modeling is inappropriate than an approximate solution given by a suitable modeling. Please see Figure 4.4. The visualization is in Figure 4.5, and only the 1-st order Difference results are shown, because the other two, Hankel+Nuclear norm and 1-st Difference via BFC search algorithm are basically the same as in Figure 4.2, thus they are neglected for space.

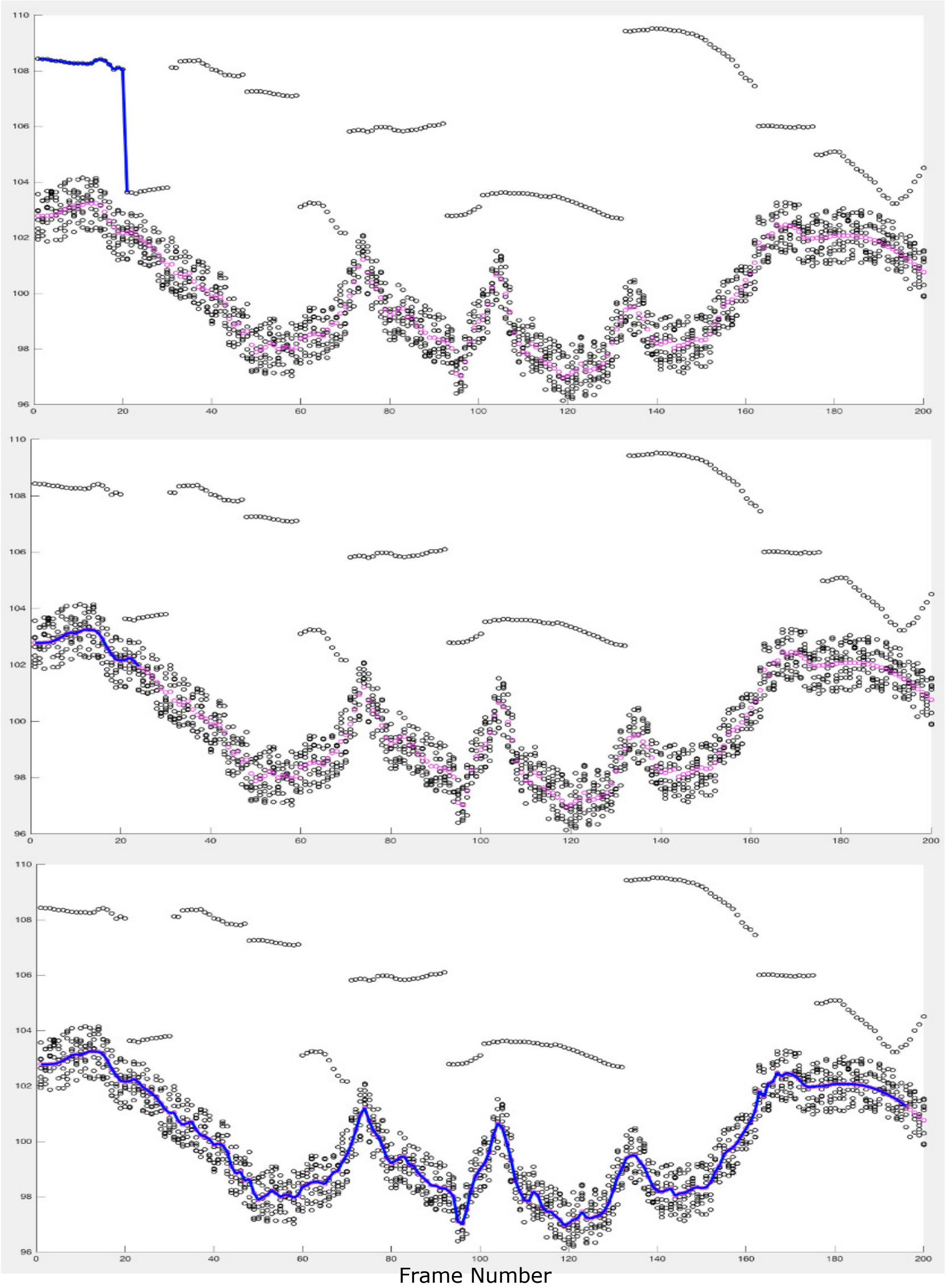


Figure 4.3: The x axis is the frame number, while the y axis indicates the illustrative positions of the 10 candidate poses. Ground truth trajectory are the circles in purple and our BFC search algorithm finds all the ground truth point in every frame correctly. Please notice that in the first few frames, due to the presence of structured noise, our algorithm picks the lower order structured noise rather than the slightly higher order ground truth points. As time goes by, our algorithm demonstrates its ability to correct previous false selections and achieves 100% accuracy, which is a desirable advantage.

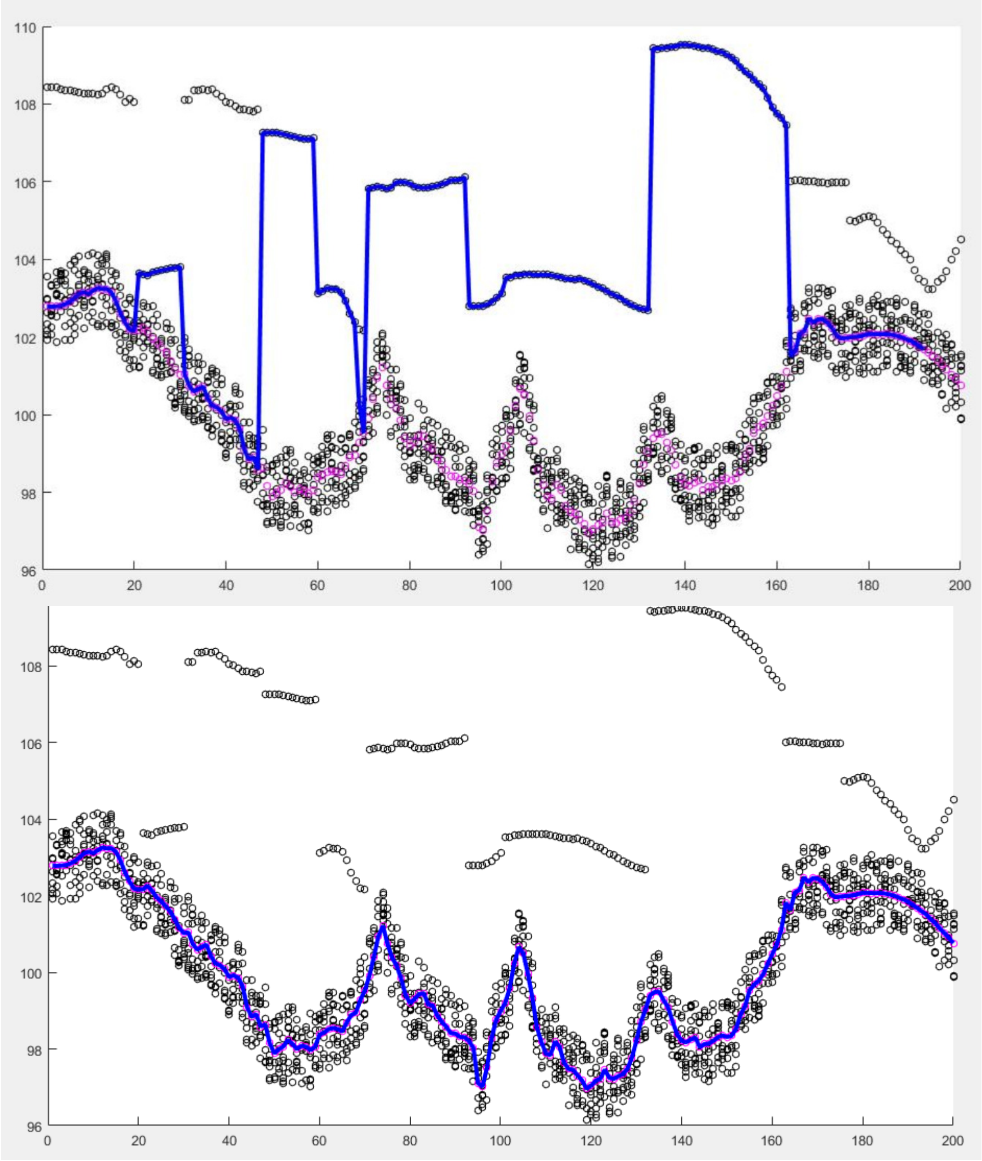
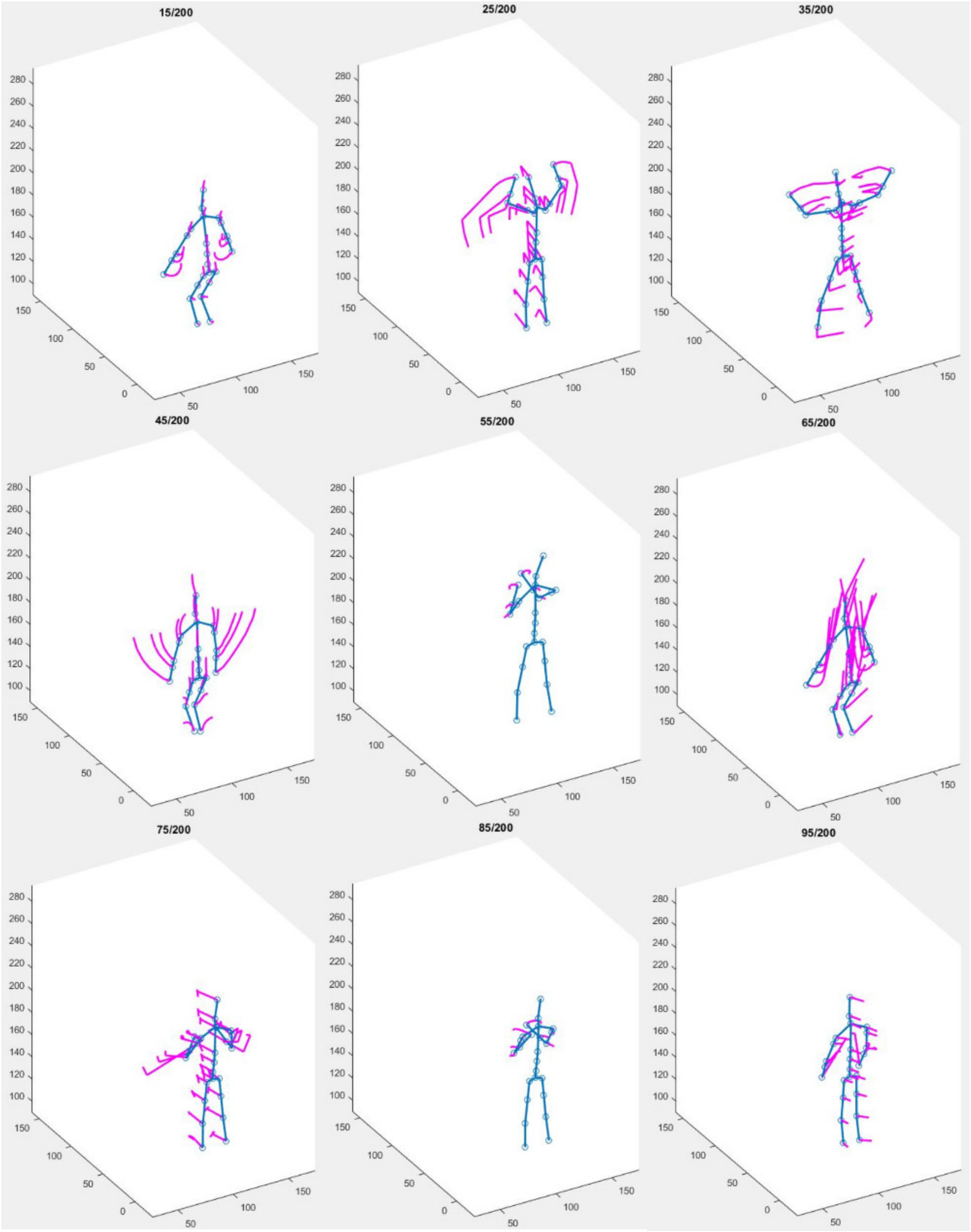


Figure 4.4: The x axis is the frame number, while the y axis indicates the illustrative positions of the 10 candidate poses. Ground truth trajectory are the circles in purple. The **upper** and **bottom** images show the selection results of global solution via dynamic programming and approximate solution via BFC search framework respectively. According to the definition of the graph, the selection trajectory in **upper** costs 5.2024×10^3, while the selection trajectory in **bottom** costs 6.2642×10^3. Therefore, dynamic programming indeed gives the global optimal solution in the sense of the definition of the graph model, which may not be suitable for the problem in reality.



15/200 25/200 35/200

45/200 55/200 65/200

75/200 85/200 95/200

Figure 4.5: The visualization of the selected trajectory of **upper** in Figure 4.4 is shown. Notice that there are some sharp changes on the purple trajectories in the images of **North**, **Northeast** and **Southwest**. Also some distractions from other activities are selected in the **Center** and **South** images.

4.1.4 Dealing with Outliers

One common problem often encountered in reality is the presence of outliers. Our BFC search algorithm does not consider the case when outliers are present, it forces to pick one candidate at each frame without thinking whether it is an outlier or not. Because of these outliers, the trajectory we select is still noisy, a simple approach to address this is to clean the trajectory using a similar optimization equation in (3.5) via the method in [20], we repeat here for reference:

$$\begin{aligned} \min \quad & ||\mathbf{H}||_* + \lambda ||h - d||_1 \\ \text{s.t.} \quad & \mathbf{H} = \text{hankel}(h) \end{aligned} \tag{4.1}$$

where d is the trajectory returned from BFC search algorithm, h is the cleaned trajectory and $\mathbf{H}$ is the Hankel matrix associated with it, λ is a parameter balancing the importance of smoothness or fidelity to the original data. The experiment results are shown in Figure 4.6. In the presence of a group of consecutive outliers (more than 10 frames around 40-th frame), the algorithm forces to pick one candidate from the hypothesis set regardless of how far it is from the reasonable range. The post smoothing processing pulls back the trajectory around the outliers at 40-th frame, but also it alters the positions of right selections at around the 77-th frame, 95-th frame and 145-th frame.

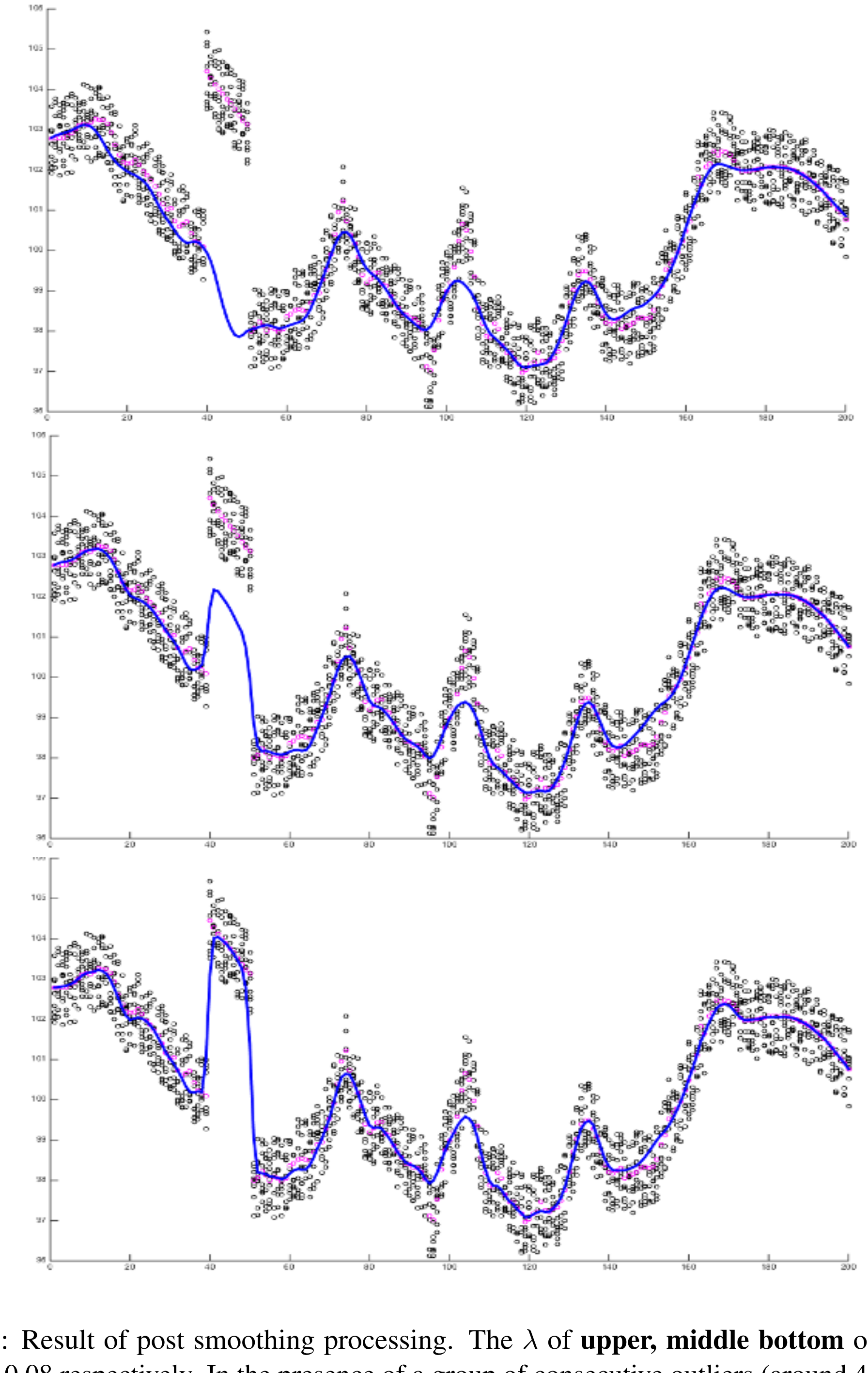


Figure 4.6: Result of post smoothing processing. The λ of **upper, middle bottom** ones is 0.07, 0.075, and 0.08 respectively. In the presence of a group of consecutive outliers (around 40-th frame), the algorithm forces to pick one candidate from the hypothesis set regardless of how far it is from the reasonable range. The post smoothing processing pulls back the trajectory around the outliers at 40-th frame, but also it alters the positions of right selections at around the 77-th frame, 95-th frame and 145-th frame.

4.2 Video Experiments

In previous section, we demonstrated the effectiveness of proposed method on synthetic experiments, and now we will use the proposed method on real videos. Here we introduce the main dataset, **N-Best Dataset:** this dataset was collected by the authors of [1], which consists of 4 challenging videos with annotations available. Following previous works, we also test our method on sequences *walkstraight* and *baseball* for a fair comparison with [7, 9].

4.2.1 Evaluation metrics

Following [1, 5], we quantitatively evaluate the performance of our approach by Percentage of Correct Parts (PCP), which is a standard criteria first introduced in [6]: a body part returned by the algorithm is considered correct if its segment endpoints lie within 50% (or any other threshold) of the length of the ground-truth segment from their annotated location. In [8], the focus was on detecting the best symmetric parts, and they report the maximum accuracy of the best localized symmetric part. This protocol has an obvious shortcommming: if a pose estimator mis-detects one hand completely and detect the other hand with full precision, then the metric will report a 100% accuracy, which can lead to a poor assessment. For our evaluations, we report the mean accuracy of symmetric keypoints. Apart from this, however, there are at least two different interpretations about PCP definition because of ambiguous wording according to [5]. The original paper [6] did not mention averaging over endpoints before comparing with the threshold, thus some latter works [27, 28] consider a body part correct only if *both* of its endpoints are closer to their ground-truth locations than a threshold. After some time, the authors of [6] clearified this on their project websites. However, most of the previous works did not mention these two types of interpretation on their results. And we use the strict interpretation in our result. Even with the strict PCP, our method still performs better than the state of the art [7, 9] on *walkstraight* video.

PCP, however, is sensitive to the amount of foreshortening of a limb, it can be too strict or too loose in some cases [5]. Therefore we also use another metric Keypoint Localization Error (KLE) as in [8, 7, 9], which measures the average Euclidean distance from the ground truth to the estimated keypoints, normalized by the size of the head in this ground truth frame to correct scale changes. KLE is an error metric, thus the smaller the better the results are.

Table 4.2: Results on *walkstraight* and *baseball* videos. Our method is evaluated using strict PCP, while other methods do not mention strict or standard PCP in their works. And our results are reported separately rather than the average of the two videos as other methods. PCP is an accuracy measure with maximum at 1. KLE is an error measure, the less the better.

	Method	**Head**	**Torso**	**U.Legs**	**L.Legs**	**U.Arms**	**L.Arms**	**Avg.**
PCP	[7]	1.0	0.69	0.91	0.89	0.85	0.42	0.80
	[8]	1.0	1.0	0.91	0.90	0.69	0.39	0.82
	Nbest [1]	1.0	0.61	0.86	0.84	0.66	0.41	0.73
	Zhang [9]	1.0	1.0	0.92	0.94	0.93	0.65	0.91
	Ours, walk	**1.0**	**1.0**	**0.97**	0.92	**0.95**	**0.70**	**0.92**
	Ours, baseball	**1.0**	**1.0**	0.81	0.67	0.67	0.26	0.68
KLE	[7]	0.53	0.88	0.67	1.01	1.70	2.68	1.25
	[8]	0.15	0.23	0.31	0.37	0.46	1.18	0.45
	Nbest [1]	0.54	0.74	0.80	1.39	2.39	4.08	1.66
	Zhang [9]	0.15	0.17	0.24	0.37	0.30	0.60	0.31
	Ours, walk	0.16	**0.14**	**0.21**	**0.28**	**0.29**	**0.59**	**0.28**
	Ours, baseball	0.16	0.18	0.43	0.71	0.43	1.12	0.57

4.2.2 Results and Discussions

We compare our proposed method with four state-of-the-art video based human pose estimation methods on *walkstraight* video: N-best method [1], Symmetric Tracking method [7], Mixing Body-part method [8], and the method by Zhang [9]. Since [8] was designed for upper-body pose estimation, Zhang re-implemented its algorithm and reported the results in [9], and we use his results for consistency. Table 4.2 shows the results. ”U.Legs”, ”L.Legs”, ”U.Arms” and ”L.Arms” are short for upper legs, lower legs, upper arms and lower arms respectively. ”Avg.” stands for the average performance of all the previous body parts. All the results of the body parts are evaluated using mean accuracy and strict PCP.

Limitations: The proposed method relies on the N-best method [1]; therefore, if N-best method cannot generate any correct hypothesis for a specific frame, it is not possible to obtain improved results by the proposed method. And according to our experience, for some difficult videos, such as *baseball* in N-best dataset, N-best algorithm fails to return correct poses frequently. Table 4.3 and Table 4.4 show the detection results from N-best algorithm [1]; ”Fr.#” is short for frame numbers, while ”C.C. #” column is the correct candidate indexes among top 200 candidates

Table 4.3: Detection results for *baseball* video 1. Fr. # and C.C. # are the frame numbers in the video and the correct candidate indexes among top 200 detections at each frame. ”None” means all candidates are considered outliers.

Fr. #	**C.C. #**	**Fr. #**	**C.C. #**	**Fr. #**	**C.C. #**	**Fr. #**	**C.C. #**	**Fr. #**	**C.C. #**
1	1	**16**	16	**31**	2	**46**	None	**61**	None
2	2	**17**	4	**32**	4	**47**	None	**62**	None
3	5	**18**	2	**33**	23	**48**	None	**63**	None
4	2	**19**	128	**34**	21	**49**	None	**64**	None
5	8	**20**	9	**35**	12	**50**	99	**65**	None
6	3	**21**	10	**36**	22	**51**	68	**66**	None
7	2	**22**	10	**37**	16	**52**	20	**67**	None
8	12	**23**	11	**38**	68	**53**	20	**68**	None
9	12	**24**	7	**39**	173	**54**	None	**69**	None
10	5	**25**	16	**40**	57	**55**	55	**70**	None
11	9	**26**	7	**41**	24	**56**	47	**71**	None
12	12	**27**	1	**42**	3	**57**	149	**72**	None
13	23	**28**	22	**43**	None	**58**	6	**73**	None
14	87	**29**	1	**44**	None	**59**	54	**74**	None
15	128	**30**	15	**45**	None	**60**	None	**75**	None

returned by the algorithm, and ”None” means we cannot find any pose estimation overlapping with the human in current frame, basically, all the estimations are different from the ground truth in at least one arm or one leg. These ”None” frames are outliers generated by N-best algorithm. Thus, our algorithm cannot give satisfying selection in this *baseball* video after a post smoothing processing, see Table 4.2.

Table 4.4: Detection results for *baseball* video 2. Fr. # and C.C. # are the frame numbers in the video and the correct candidate indexes among top 200 detections at each frame. ”None” means all candidates are considered outliers.

Fr. #	**C.C.#**	**Fr. #**	**C.C.#**	**Fr. #**	**C.C.#**	**Fr. #**	**C.C.#**	**Fr. #**	**C.C.#**
76	None	**91**	137	**106**	None	**121**	3	**136**	79
77	None	**92**	2	**107**	None	**122**	1	**137**	1
78	None	**93**	9	**108**	None	**123**	1	**138**	8
79	None	**94**	1	**109**	54	**124**	44	**139**	60
80	None	**95**	1	**110**	None	**125**	15	**140**	39
81	None	**96**	3	**111**	None	**126**	21	**141**	None
82	None	**97**	7	**112**	None	**127**	None	**142**	None
83	None	**98**	1	**113**	28	**128**	1	**143**	4
84	None	**99**	1	**114**	9	**129**	7	**144**	15
85	None	**100**	16	**115**	41	**130**	12	**145**	None
86	None	**101**	11	**116**	59	**131**	15	**146**	None
87	40	**102**	8	**117**	9	**132**	182	**147**	None
88	32	**103**	3	**118**	184	**133**	4	**148**	None
89	1	**104**	5	**119**	5	**134**	7	**149**	2
90	22	**105**	13	**120**	15	**135**	1		

Chapter 5

Conclusion and Future Work

In this thesis, we propose a simple, yet effective tracking-by-selection algorithm for human pose estimation in videos. The problem is solved by two steps, firstly, top N candidates are generated from each frame of the video; secondly, frame to frame temporal smoothness between poses across different frames are guaranteed by selecting the trajectory with the least Nuclear norm of its Hankel matrix among all the possible combinations. In the end, it is necessary to clean the selected trajectory by a rank minimization algorithm to remove the effects of noise and outliers. Our dynamic based approach not only exhibits the ability to select the smooth trajectory from N-best detections accurately in the MoCap dataset, but also finds its value in video based tracking-by-selection human pose estimation framework.

However, our algorithm forces to select wrong path when all the N-best candidates are all outliers. In fact, this disadvantage happens in all of the tracking-by-selection approaches [1, 9, 8, 7, 3]. We have two solutions to this problem: the first, is to generate more promising candidate pose combinations that reduces the frequencies of outliers. This method was used in [9, 8]. They basically break down the whole skeleton into several body parts, and track different body parts separately to build M trajectories for each body part, and at latter stage one of the M trajectories is selected and used as the body part to assemble the whole skeleton trajectory. The second is to detect whether there are reasonable candidates at current frame, if they are not all outliers, then the selection continues as usual, if they are all outliers, then none of them is supposed to be selected, and we predict a reasonable pose estimation based on observed data at current frame, and continue to next frame. We have done some preliminary works on the prediction part. We formulate it as a matrix completion problem, so we want to find a skeleton pose that does not increase the complexity of the Hankel matrix built by previous poses and the desired current pose. Inspired by [22, 29], we use the CVX toolbox [30] to solve the prediction by optimizing the weighted Nuclear norm of the Hankel matrix iteratively, and it works but extremely slow.

Bibliography

[1] D. Park and D. Ramanan, “N-best maximal decoders for part models,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov 2011, pp. 2627–2634.

[2] V. M. Zatsiorsky, *Kinetics of human motion*. Human Kinetics, 2002.

[3] R. Tokola, W. Choi, and S. Savarese, “Breaking the chain: Liberation from the temporal markov assumption for tracking human poses,” in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[4] P. F. Felzenszwalb and D. P. Huttenlocher, “Pictorial structures for object recognition,” *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000042934.15159.49

[5] Y. Yang and D. Ramanan, “Articulated human detection with flexible mixtures of parts,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2878–2890, Dec 2013.

[6] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, “Progressive search space reduction for human pose estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2008.

[7] V. Ramakrishna, Y. Sheikh, and T. Kanade, “Tracking human pose by tracking symmetric parts,” in *CVPR*, 2013.

[8] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, “Mixing body-part sequences for human pose estimation,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[9] D. Zhang and M. Shah, “Human pose estimation in videos,” in *Proceedings of International Conference on Computer Vision*, 2015.

[10] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 1362–1369. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2012.6247822

[11] F. Xiong, O. I. Camps, and M. Sznaier, "Low order dynamics embedding for high dimensional time series," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 2011, pp. 2368–2374. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2011.6126519

[12] C. Dicle, M. Sznaier, and O. Camps, "The way they move: Tracking targets with similar appearance," in *ICCV*, 2013.

[13] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier, "Activity recognition using dynamic subspace angles," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 3193–3200.

[14] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[15] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," *International Journal of Computer Vision*, vol. 99, no. 2, pp. 190–214, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11263-012-0524-9

[16] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 1014–1021.

[17] M. W. Lee and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 27–38, Jan 2009.

[18] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 915–922.

[19] Y. Bo and H. Jiang, "Scale and rotation invariant approach to tracking human body part regions in videos," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 1041–1047.

[20] M. Ayazoglu, M. Sznaier, and O. I. Camps, "Fast algorithms for structured robust principal component analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1704–1711.

[21] T. Ding, M. Sznaier, and O. I. Camps, "Fast track matching and event detection," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[22] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2862–2869.

[23] A. Knutson and T. Tao, "Honeycombs and sums of hermitian matrices," *Notices Amer. Math. Soc*, vol. 48, no. 2, 2001.

[24] T. Tao, *Topics in random matrix theory*. American Mathematical Soc., 2012, vol. 132.

[25] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991, cambridge Books Online. [Online]. Available: http://dx.doi.org/10.1017/CBO9780511840371

[26] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, Jan 2013, pp. 53–60.

[27] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proceedings of the British Machine Vision Conference*, 2010, doi:10.5244/C.24.12.

[28] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[29] K. Mohan and M. Fazel, "Reweighted nuclear norm minimization with application to system identification," in *American Control Conference (ACC), 2010*, June 2010, pp. 2953–2959.

[30] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” http://cvxr.com/cvx, Mar. 2014.