DEEP LEARNING AND ANTHROPOMETRIC PLANE BASED WORKFLOW MONITORING BY DETECTING AND TRACKING WORKERS

N. A. Gard[1]*, J. Chen[2], P. Tang[2], A. Yilmaz[1]

[1] Photogrammetric Computer Vision Laboratory,
Dept. of Civil, Environmental, and Geodetic Engineering, Columbus, OH - (ajamgard.1, yilmaz.15)@osu.edu
[2] School of Sustainable Engineering and the Built Environment, Tempe, AZ - (jchen311, tangpingbo)@asu.edu

Commission I, WG I/5

KEY WORDS: Detection, Tracking, Anthropometric Measures, Critical Path

ABSTRACT:

The worker productivity, a critical variable in project management, significantly affects the progress of a project. The key to measuring productivity is analysis of activities, which provides necessary information by identifying how workers spend their time at certain areas in the site. In this work, we propose a novel joint image-trajectory space for automatic detection and tracking of workers using a single fixed camera. A two-branch convolutional neural network detects workers and their body joints. Instead of tracking the body joints in the image space, we transform detected joints onto virtual parallel planes called "Anthropometric Planes". The detected joints are, then, tracked using a Kalman Filter on these planes which are created based on anthropometric measures of an average American male. Finally, an uncertainty measure is introduced to reduce the number of identity changes and to handle missing joints. The experiments conducted on an image sequence captured in a nuclear plant shows promising detection and tracking results.

1. INTRODUCTION

In the United States, many nuclear power plants (NPPs) were built more than 40 years, and they require regular refueling and maintenance, which are called outages. NPP outages are challenging because they require tracking and coordinating thousands of activities in a short span of time, usually between twenty to thirty days. Moreover, any delays in the NPP outage processes will cause significant economic losses. If an NPP is shut down for one more day because of a delay from an outage, the deficiency of power will lead to up to two million dollars losses for the energy company. NPP outages require a significant supplemental workforce that consists of thousands of contract workers, which increases the complexity of communication and information flows. Other challenges, including scheduling, work group coordination, nuclear safety concerns arising from different system configurations, and resource allocation issues, can create delays and schedule overruns, driving up outage costs. These features of NPP outages call for a real-time, robust, effective workflow progress monitoring to identify and resolve delays or critical path changes.

For timely and effective outage coordination at an NPP, it is required to have efficient and effective monitoring and control of non-wrench time activities (e.g., obtaining parts, tools or instructions, the travel associated with tasks) and tasks that are near the critical path(s). Duration variations and non-wrench time associated with tasks near critical paths could cause critical path changes and unexpected delays. The first step for achieving such monitoring and control of non-wrench-time and near-critical-path activities is to automatically and precisely detect and track workers during each activity to estimate future non-wrench time and task variations, which will help with effective scheduling and decision making. In this research, we proposed the automatic computer vision-based workflow monitoring methodology and carried out the following analysis of documentation and video data collected during two outages (April 2017, and October 2017).

To validate the proposed methodology, we collected 24-hour video data in the Radiation Protection Island (RPI) the spring outage in Palo Verde Nuclear Power Plant in Arizona. This proposed methodology tested on collected data shows promising results and will help prevent the delays in advance and quickly identify and diagnose the deviations between as-planned and as-is workflows for decision support.

2. RELATED WORK

Workflow monitoring is a major aspect in determining whether a project can be finished on time and on budget (Cheng et al., 2013, Ghanem and AbdelRazig, 2006, Girardeau-Montauta et al., 2005). Many researchers have explored to develop an effective and timely method to manage workers activity and thereby to improve the productivity. (Cheng et al., 2013) used the data fusion of spatio-temporal and workers posture data to monitor workers activity. (Ghanem and AbdelRazig, 2006) used Radio Frequency Identification (RFID) system to collect the trajectories of workers to monitoring the work done on construction sites. However, tag-based human tracking technologies are not suitable for NPP outages because NPP has restrictions on the devices that can be installed on the site and trackable tasks may cause confidential issues (Zhang et al., 2017). Visual tracking, on the other hand, is inexpensive solution which can be adapted to the confidential needs of a project.

With the appearance of deep convolutional neural networks (CNNs) many obstacles in the field of computer vision have been successfully overcome. Especially, significant improvement in prediction and estimation of human poses. (He et al., 2017) proposed a top-down method to first locate a bounding box around humans, and then estimate their body joints. (Cao

*Corresponding author

et al., 2017) instead proposed a bottom-up approach, where first body joints are detected, and then human pose is estimated after grouping joints into individual skeletons. One perk of using a bottom-up over top-down approach is when partial occlusion happens, in order to detect a human, most of his body should be visible, while a partial skeleton can be made using only visible joints.

Human tracking is still an active area of research as there are challenging problems in real-world cases that need to be solved. The formulation of tracking as a bipartite matching problem (Pirsiavash et al., 2011) and solving it using the Hungarian algorithm (Kuhn, 1955) has been a common practice among the tracking community. However, it always has been prone to frequent change of identity and losing track of people due to occlusion. More recent methods use graphical model to predict the joints locations over time (Insafutdinov et al., 2017, Iqbal et al., 2016). The complexity of these models is a serious drawback despite their promising performance.

3. METHODOLOGY

We propose to use an joint image-trajectory space that creates a new space to efficiently and accurately tracks workers across time. We use a 2D human pose predictor (Cao et al., 2017) which takes as input an online video stream and predicts the poses of all people in the video. To track the instances in time, first, we transform detections to a new trajectory space which consists of a set of parallel planes. Then, we perform tracking in the new space. We show how our proposed space overcome challenges that most tackers face in the regular image space.

3.1 Human Joint Detection

The first space of our joint image-trajectory space is the image space, $\mathcal{I}$, where detection occurs. Although our approach can build upon any frame-based pose estimation system, we use the top-down 2D human pose estimator (Cao et al., 2017) due to its robust and near real-time detection performance. A person is represented with a skeleton and the joints are labeled accordingly. A two-branch network (Fig. 1) takes an image as input, and through a refining process, it detects the body joints and their connecting limbs along with their orientations. A graph matching algorithm is responsible for mixing and matching the joints of a person. Given the orientation and the limbs as the edge weights of the $k-$partite graph, and the labeled joints as the vertices of the graph, the matcher finds the joints that belong to a person. However, since the detection randomly chooses an id for people in the video per frame, keeping track of the assigned id, when a person first appears in the scene, remains as a challenge. Furthermore, missing joints due to partial or complete occlusion or even simply failing to detect a worker aggravates the situation.

The output of this stage, which is the group of the labeled joints for a person, is the input to our joint trajectory space, where tracking occurs.

3.2 Anthropometric Planes

Tracking different body joints on a single camera image is prone to inconsistent displacements. A consistent tracking algorithm must be able to track a worker regardless of his position in an environment. Consider the case when a worker approaches a single fixed camera. As he gets closer to the camera, his displacement in

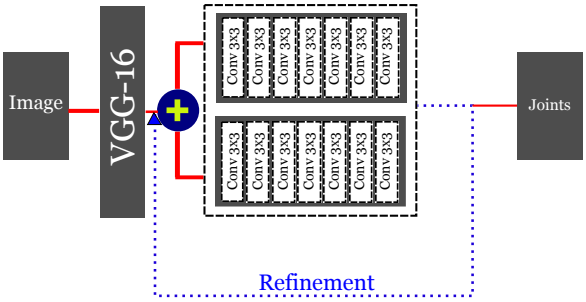

Figure 1. Joint Detection Architecture: Images are fed to VGG16, and generated feature maps are fed to a two branch network. Branch 1 (top) finds the confidence map for a labeling a joint. Branch 2 (bottom) is in charge of estimating the orientation of the limb between two detected joints

the image space becomes larger and larger. In other words, his velocity changes although in the object space he has a constant velocity. Now, consider another worker who moves away from the same camera. The worker's displacement becomes smaller and smaller resulting in a lower velocity in the image space. There could be other workers walking across the room, running, standing still, etc. These issues created by loss of depth due to the fact that we have only a single fixed camera, makes any tracking algorithm unreliable. On top of loss of depth and workers walking patterns, the pattern of each joint across time varies from one joint to another. We will study these patterns in the experiment section.

To overcome these issues, we propose to transform detected joints from the camera image to a set of virtual planes parallel to the floor.

The creation of anthropometric planes is inspired by the work of (Lai and Yilmaz, 2008) where they eliminate the use of camera calibration for shape reconstruction and instead adopt the silhouette images. The basic idea is to utilize homography transform to generate virtual planes at the level of all the body joints, parallel to a reference plane.

Let a set of points, $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, n \geq 4$, be located on a reference plane, π, defined in the object space $\mathcal{O}$. Define a transform, $T(\mathbb{X}, z)$, which (de-)escalates $\mathbb{X}$ to a new set of points, $\mathbb{X}_z$, by $z \in \mathbb{R}$ in the direction of π's normal. $\mathbb{X}_z = \{\mathbf{x}_1^{(z)}, \mathbf{x}_2^{(z)}, \ldots, \mathbf{x}_n^{(z)}\}$ are in the new plane, π_z which is parallel to π. Now, consider the set of lines, $\mathbb{L}$, passing through all the pairs, $(\mathbf{x}_i, \mathbf{x}_i^{(z)}), i \in \{1, 2, \ldots, n\}$. From the definition, one can see that $\mathbf{l}_i$'s are parallel and they intersect in infinity.

The two point sets $\mathbb{X}'$ and $\mathbb{X}'_z$ are the projection of the two point sets $\mathbb{X}$ and $\mathbb{X}_z$ from the object space $\mathcal{O}$ to the image space $\mathcal{I}$. It can be shown that the set of vanishing lines, $\mathbb{L}_v$ are the lines passing through $\mathbb{X}'$ and $\mathbb{X}'_z$, which intersect at the vanishing point, $\mathbf{v}_z$ (Fig 2).

The relation between $\mathbb{X}'$ and $\mathbb{X}'_z$ is, then, established as follows:

$$\begin{aligned} \lambda_i \mathbf{x}_i'^{(z)} = \mathbf{P}\mathbf{x}_i' &= \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \mathbf{p}_4 \end{bmatrix} \begin{bmatrix} X_i' \\ Y_i' \\ Z_i' \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_4 \end{bmatrix} \begin{bmatrix} X_i' \\ Y_i' \\ 1 \end{bmatrix} + \mathbf{p}_3 Z \end{aligned} \tag{1}$$

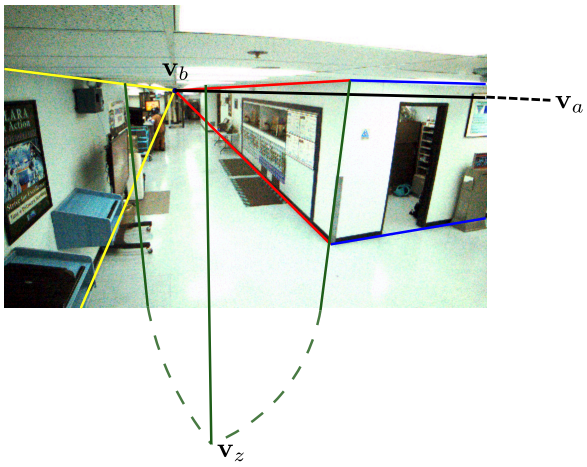

Figure 2. Vanishing Lines and Points: $\mathbf{v}_a$ and $\mathbf{v}_b$ are the vanishing points in the horizontal direction. $\mathbf{v}_z$ is the vanishing point in the vertical direction

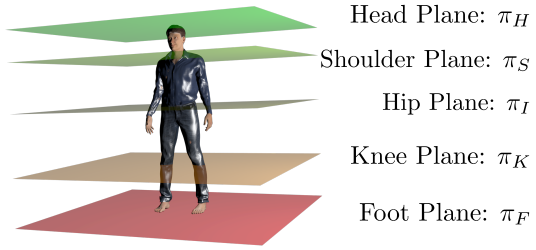

Figure 3. Anthropometric Planes for Human: body joints are tracked on their corresponding planes

where λ = scale factor
P = projection matrix.

Substituting $\mathbf{v}_z$ in equation (1) as $\mathbf{p}_3$ yields:

$$\lambda_i \mathbf{x}_i'^{(z)} = s_i \mathbf{x}_i' + \mathbf{v}_z Z, \tag{2}$$

where $\lambda_i = \sum s_i + Z$.

Given two points $\mathbf{x}_1$ and $\mathbf{x}_2$ the scale factor s_i can be calculated as follows:

$$s_1 \mathbf{x}_1' - s_2 \mathbf{x}_2' = (s_1 - s_2)\mathbf{v} \tag{3}$$

where **v** = the vanishing point in the direction of $\overrightarrow{\mathbf{x}_1\mathbf{x}_2}$.

Fig. (3) illustrates anthropometric planes for head, shoulder, hip, knees, and feet created for an average American male. Depending on the application, one can produce more planes intersecting with other joints. Fig. (4) shows projected anthropometric planes onto the image space.

3.3 Multi-People Multi-Joint Tracking

In this section, we break down a general tracking scheme and define necessary terms that help formulating it.

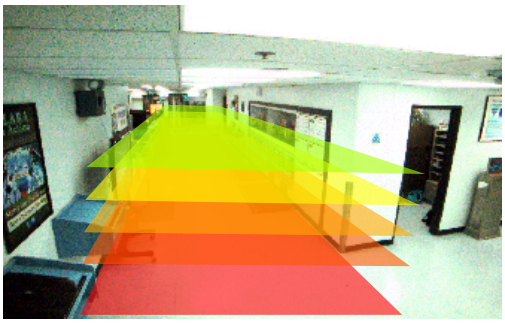

Figure 4. Anthropometric Planes: A new trajectory space for tracking joints of multiple people

Object State The object comprises of 15 body joints, for which the state is defined as its location if the joint is visible or labeled as occluded if the joint is not visible. Since, the joints are being detected and labeled in the detection phase, it is no longer needed to use spatial similarities such as overlap between bounding boxes or nearest neighbors. Instead, we use the Hungarian algorithm to solve the assignment problem.

Object Appearance At each frame, the object is represented as the mean value of all the observed or predicted locations of joints, and an uncertainty region, which is defined by the standard deviation of all the locations of the joints for one person.

Object Trajectory The history of the object written by its state and appearance in the image sequence encapsulates the trajectory of the object. The trajectory is readily available by connecting the mean locations in the previous frames.

Object Tracking Given body joint predictions grouped in the image space for the latest frame, the main task is to correctly find a person who corresponds to the same person in the previous frame. Our strategy is to construct anthropometric planes at the level of an average American male anthropometric measures. The object trajectory for each joint will be transformed to the corresponding plane. These anthropometric planes, in fact, create a new space in which one can perform all the previous tracking methods. For this work, we focus only on the Kalman Filter (Kalman, 1960). Therefore, our **Model Update** strategy is defined using the Kalman Filter.

3.4 Uncertain Joints

Anthropometric measures used in this paper are based on average anatomy measures. If a person is taller or shorter than the average, the anthropometric planes do not exactly intersect with the joints. This will cause uncertainty in projecting joints from the image space onto the joint space. For example, consider the situation in Fig. 5. Given are the four points locating under, on, and above an anthropometric plane. We can see that unless a joint is exactly on the plane, its projection will have an uncertainty corresponding to how far the joint is from the plane. Fig. 6 demonstrates a stack of anthropometric planes with joints and their projections on the joint space.

The two main reasons that projected joints won’t coincide exactly the same pixel are: 1) The uncertainty associating with difference

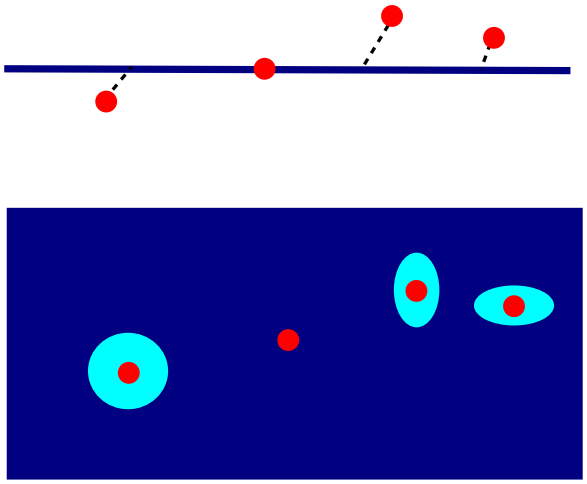

Figure 5. Projection uncertainty arisen from the fact that anthropometric planes are based on average measures

in anthropometric planes and actual person’s height, 2) Although neck and head are on the same line splitting the body into half, left and right shoulders are located on each side of this line. Therefore, even for joints locating exactly on an anthropometric plane, projections won’t coincide. 3) Detection results are not reliable.

We formulate this uncertainty with the following strategy:

Let μ_x the mean value and σ_x the variance of a stochastic variable $\mathbf{x}$ describing the projected joints on the joint space for one worker.

If the relative precision, σ_x/μ_x, of a collection of joints is high enough, propagation of uncertainty can be approximated by simple variance propagation. If the uncertain random vector $\{\boldsymbol{\mu}_x, \Sigma_{xx}\}$ is transformed by function $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, we obtain an uncertain vector $\{\boldsymbol{\mu}_y, \Sigma_{yy}\}$ with mean and variance

$$\boldsymbol{\mu}_y = \boldsymbol{f}(\boldsymbol{\mu}_x) \text{ and } \quad \Sigma_{yy} = J_{yx}\Sigma_{xx}J_{yx}^T \tag{4}$$

where J = Jacobian evaluated at $x = \mu_x$.

Here, we can interpret the distance between joints, as the inverse of the covariance matrix. The tool to measure the distance between joints with uncertainty is the Mahalanobis distance. For two points $\mathbf{x}_1$ and $\mathbf{x}_2$ with covariance matrices $\Sigma_{x_1x_1}$ and $\Sigma_{x_2x_2}$, the Mahalanobis distance is

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T(\Sigma_{x_1x_1} + \Sigma_{x_2x_2})^{-1}(\mathbf{x}_1 - \mathbf{x}_2)} \tag{5}$$

4. EXPERIMENTS

This section introduces the performance of the human detection and tracking results. We installed a camera in the Radiation Protection Island (RPI) at the time of the spring outage in Palo Verde Nuclear Power Plant and collected 24-hour video data on Apr. 16th, 2017. We use this video to investigate the capabilities of the human detection and tracking algorithm. First, we only track workers ankles and compare it with the results on the joint space. Then, we study the walking pattern of a human and examine how using different joints is beneficial for tracking in both image and joint space.

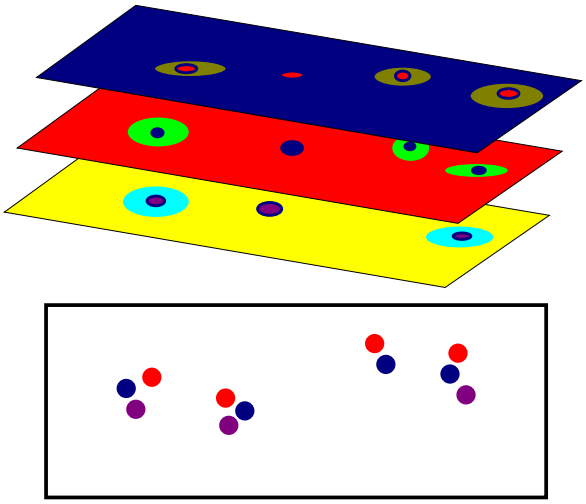

Figure 6. Top view of the projection of all the joints onto the joint space.

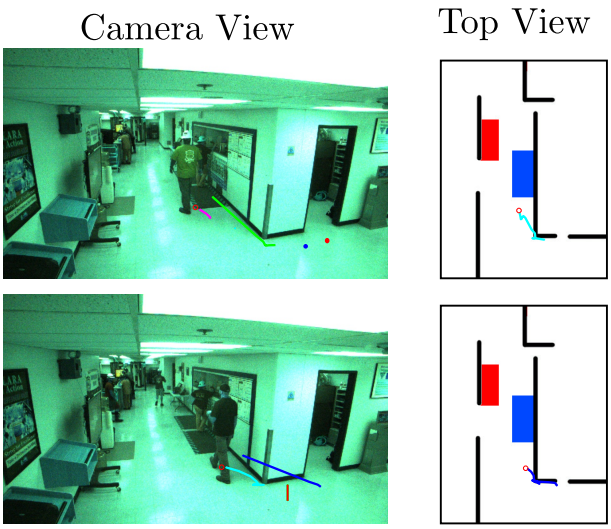

Figure 7. Only Ankle: Tracking in the image space (left images) is very unstable, while tracking on the anthropometric planes (right images) does not change the identity of the worker.

Ankle

For this experiment, we use only the detected ankles of workers and track them once in the image space, and simultaneously in the joint space. A constant velocity Kalman Filter with 4 inputs $\begin{bmatrix}\mathbf{x}\\ \dot{\mathbf{x}}\end{bmatrix}$ is used for tracking in both spaces. Fig. (7) demonstrates how unstable tracking in the image space is. On the other hand, since transforming the trajectory of ankles onto the joint space is more linear, the same tracking method with the same parameters significantly performs better. In the image space, we observe *4* and *3* identity changes in the top and bottom images, respectively. While no change of identity occurs in the joint space.

Upper Body Joints

Tracking ankles due to high non-linearity in their trajectory pattern across time, encounters with many change of identities. More importantly, ankles are more prone to occlusion as all people walk on the same floor but because they have different heights, it is easier to distinguish their upper bodies joints such as neck, head and shoulders (Fig. 8).

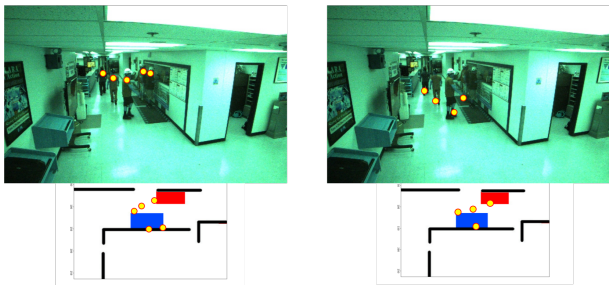

Figure 8. Detections on anthropometric planes: All workers heads are detected (left) while not all ankles are detected (right)

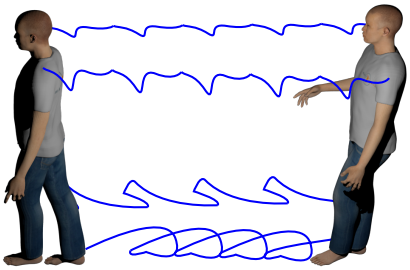

Figure 9. Walking Pattern: Foot and knee trajectory has a highly non-linear pattern. Head and shoulder trajectories appear to be more linear

Fig. 9 depicts the pattern of normal walking. As we can see, ankles (also feet) and knees are highly non-linear. Consider the situation when the left leg is moving. The next step requires the left leg to be still and the right leg to move. With the same idea, knees are also experiencing the same pause-go phases. Thus, our tracking method must be able to capture this type of movements. It is noteworthy that, although pause-go type of movements share the same nature, they have different patterns.

On the other hand, upper body joints are always on the go or have shorter pause intervals. As long as the person is moving, Head, and neck are moving with the person. Shoulders will also be on the go but with a negligible lag. Overall, we recommend to use the upper body joints.

In this experiment, for the image space, we use the head points as they have a more linear trajectory compared to other joints, and for the joint space, we use the average location of neck, head, left and right shoulders depending on which joints are available. If the distance between joints is larger than 10 pixels, we ignore those joints and only rely on other available joints as detection results may not always be reliable. With this strategy, not only the we handle the frequent change of identity, but also partial occlusion (Fig. 10).

5. CONCLUSION

We have presented a framework for detecting and tracking workers using a single fixed camera. Our approach combines a state-of-the-art human pose estimation methods with novel joint trajectory space. Transforming joints from the image space to the joint space significantly improve tracking performance that even a simple tracking algorithm such as Kalman Filter along with

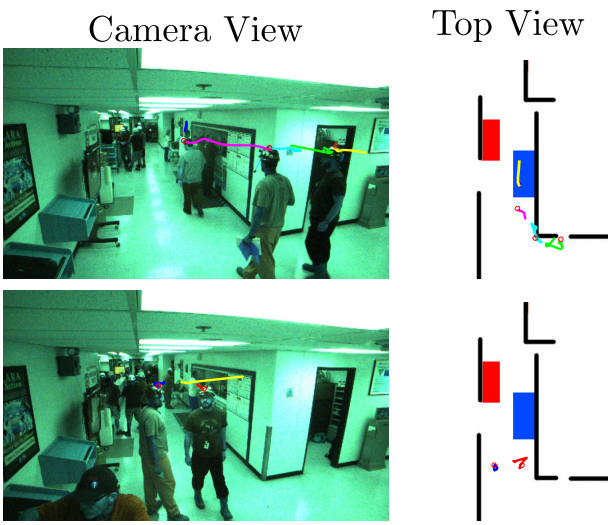

Figure 10. Tracking head in image space vs. tracking all points in joint space

a Hungarian algorithm is sufficient. Our strategy to use Mahalanobis distance as a tool to handle uncertainty helps handle situations when some of the joints are missing. The results obtained from applying our framework to a 24-hour dataset collected in a nuclear power plant, shows promising results where the number identity changes was reduced compare to the baseline. We also, studied the walking pattern of a human and suggested to use upper body joints due to their linear temporal pattern, if possible.

REFERENCES

Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: *CVPR*.

Cheng, T., Teizer, J., Migliaccio, G. C. and Gatti, U. C., 2013. Automated task-level activity analysis through fusion of real time location sensors and worker’s thoracic posture data. *Automation in Construction* 29, pp. 24–39.

Ghanem, A. and AbdelRazig, Y., 2006. A framework for real-time construction project progress tracking. In: *Earth & Space 2006: Engineering, Construction, and Operations in Challenging Environment*, pp. 1–8.

Girardeau-Montauta, D., Rouxa, M., Marcb, R. and Thibaultb, G., 2005. Change detection on points cloud data acquired with a ground laser scanner. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*.

He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. In: *Computer Vision (ICCV), 2017 IEEE International Conference on*, IEEE, pp. 2980–2988.

Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B. and Schiele, B., 2017. Arttrack: Articulated multi-person tracking in the wild.

Iqbal, U., Milan, A. and Gall, J., 2016. Posetrack: Joint multi-person pose estimation and tracking. *arXiv preprint arXiv:1611.07727*.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* 82(1), pp. 35–45.

Kuhn, H. W., 1955. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)* 2(1-2), pp. 83–97.

Lai, P.-L. and Yilmaz, A., 2008. Efficient object shape recovery via slicing planes. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, IEEE, pp. 1–6.

Pirsiavash, H., Ramanan, D. and Fowlkes, C. C., 2011. Globally-optimal greedy algorithms for tracking a variable number of objects. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, pp. 1201–1208.

Zhang, C., Tang, P., Cooke, N., Buchanan, V., Yilmaz, A., Germain, S. W. S., Boring, R. L., Akca-Hobbins, S. and Gupta, A., 2017. Human-centered automation for resilient nuclear power plant outage control. *Automation in Construction* 82, pp. 179–192.