An Optimization Based Framework for Human Pose Estimation in Monocular Videos

Priyanshu Agarwal[†], Suren Kumar[†], Julian Ryde[‡], Jason J. Corso[‡], and Venkat N. Krovi[†]

Mechanical and Aerospace Engineering Department[†]
Computer Science and Engineering Department[‡]
University at Buffalo, Buffalo, NY, USA
{priyansh, surenkum, jryde, jcorso, vkrovi}@buffalo.edu

Abstract. Human pose estimation using monocular vision is a challenging problem in computer vision. Past work has focused on developing efficient inference algorithms and probabilistic prior models based on captured kinematic/dynamic measurements. However, such algorithms face challenges in generalization beyond the learned dataset.

In this work, we propose a model-based generative approach for estimating the human pose solely from uncalibrated monocular video in unconstrained environments without any prior learning on motion capture/image annotation data. We propose a novel Product of Heading Experts (PoHE) based generalized heading estimation framework by probabilistically-merging heading outputs (probabilistic/ non-probabilistic) from time varying number of estimators to bootstrap a synergistically integrated probabilistic-deterministic sequential optimization framework for robustly estimating human pose. Novel pixel-distance based performance measures are developed to penalize false human detections and ensure identity-maintained human tracking. We tested our framework with varied inputs (silhouette and bounding boxes) to evaluate, compare and benchmark it against ground-truth data (collected using our human annotation tool) for 52 video vignettes in the publicly available DARPA Mind's Eye Year I dataset [1]. Results show robust pose estimates on this challenging dataset of highly diverse activities.

1 Introduction

Estimating and tracking 3D pose of humans in unrestricted environments using monocular vision poses several technical challenges due to high-dimensionality of human pose, self-occlusion, unconstrained motions, variability in human motion and appearance, observation ambiguities (left/right limb ambiguity), ambiguities due to camera viewpoint, motion blur and unconstrained lighting [1]. Efforts at addressing this challenging problem can be broadly classified into: (i) model-based approaches, and (ii) model-less approaches [2]. Sminchisescu [3] alternately categorizes the research into: (i) generative approaches and (ii) discriminative approaches. While generative approaches

[1] Available at: www.visint.org

are highly generalizable, the use of stochastic sampling methods to deal with the multimodal posterior/likelihood function increases their computational complexity. On the other hand, discriminative approaches are computationally tractable (for moderate sized training sets) but lack generalizability to unseen exemplars. However, there is always one or more fundamental assumptions involved that there is a priori knowledge about the physical properties (e.g. mass, inertia, limb lengths, ground plane and/or collision geometries), the activity in the scene, calibrated camera, imagery from multiple cameras (often in laboratory settings), availability of similar motion dataset [4,5,6].

No formal studies exist on which methods are employed by the human visual system for its marvelous visual perception. However, studies have constantly shown that humans use motion based cues (the instantaneous retinal optical flow) for instantaneous retino-centric heading (3D translation direction), eye-body rotation, and the relative depth of points in the world [7]. Humans appear to use motion based cues whenever motion is present in the scene and resort to visual cues (color, texture) when no/subtle motion is present in the scene. To the best of our knowledge, no prior work has used motion based cues for the task of explicitly estimating human heading direction.

Our work employs a model-based generative approach for the task of human pose estimation for general human movements in unrestricted environments. Unlike many previous approaches, our framework is fully automatic, without using camera calibration, prior motion (motion capture database), prior activity, appearance, body size information about the scene. Evaluations on a challenging dataset (DARPA Mind's Eye Year I) show the robustness of the presented framework.

Research Contributions

1. *Product of Heading Experts* - We model the heading estimation task independent of features/types of individual estimators using the proposed Product of Heading Experts (PoHE) based generalized heading estimation framework which probabilistically merges heading outputs from time varying number of estimators to produce robust heading estimates under varied conditions in unconstrained scenarios.
2. *Motion Cues Based Heading Estimation* - We propose a novel generative model for estimating heading direction of the subject in the video using motion-based cues thus, significantly reducing the pose search space.
3. *Decoupled Pose Estimation* - We propose a sequential optimization based framework optimizing the uncoupled pose states (camera/body location, body joint angles) separately using a combination of deterministic and probabilistic optimization approaches to leverage the advantages associated with each.
4. *Probabilistic-Deterministic Optimization Scheme* - We achieve faster convergence to the global minima by obtaining initial guesses using population based global optimization technique for deterministic convex optimization scheme.
5. *Identity Maintained Pose Evaluation Metric* - We introduce the notion of pose evaluation for videos with multiple humans by defining identity maintained pose evaluation metrics.

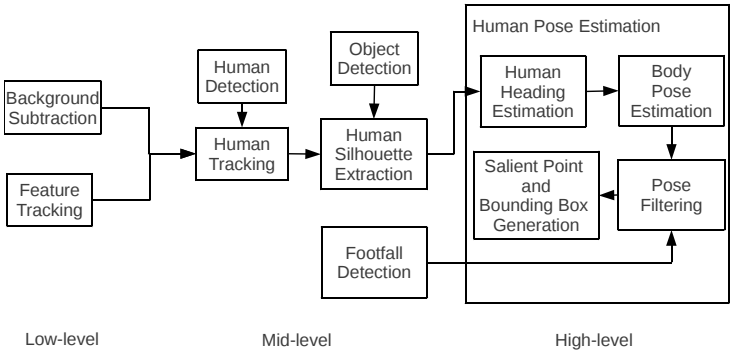


Fig. 1: System diagram for human pose estimation framework.

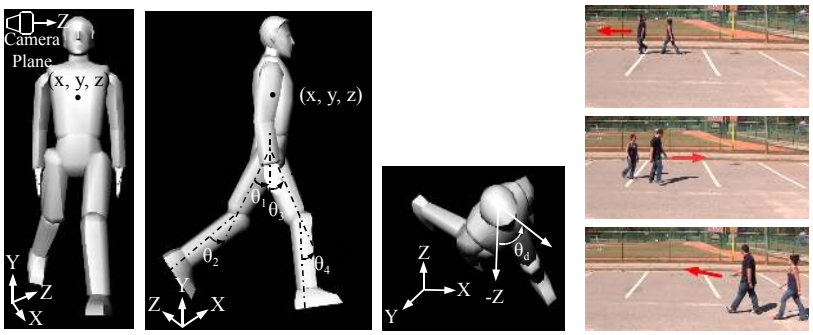
2 Optimization Based Pose Estimation

Fig. 1 provides an overview of the human pose estimation framework. We use background subtracted binary images [8] and point features at the low-level to detect/track humans/objects in the scene, extract human silhouettes at the mid-level leading to human heading and pose/salient-point estimation/filtering at the high-level. We consider 3 position variables and 5 angular variables to define the pose of a human (Figs. 2(a)-(c)).

2.1 Human Heading Estimation

Knowledge regarding the heading direction can significantly restrict the pose search space and can result in better pose estimates at lower computational costs. In the past, the task of heading estimation is not addressed separately from the actual body pose which significantly increases the complexity of the problem. Furthermore, heading is often modeled as a discrete variable using discriminative approaches with few possible values [9]. Fig. 2(d) illustrates a sequence of frames where invaluable human heading direction information can be inferred from following cues: (i) human silhouette centroid, (ii) human silhouette bounding box centroid, (iii) detected human bounding box centroid, (iv) area of human silhouette, (v) aspect ratio of bounding boxes, (vi) human silhouette/bounding boxes centroid velocity (x and y coordinates), (viii) regression/classification-based estimation of heading direction (Adaboost/Support Vector Machine), and/or (ix) optical flow.

Product of Heading Experts: We use a time evolving Product of Experts (PoE) [10] model to optimally fuse hypothesis from various heading estimators at each instant in time to propose a Product of Heading Experts (PoHE). We consider each estimator $T_1, T_2, ..., T_K$ as experts for predicting the heading direction. Product of experts model for heading ensures that the resulting model for heading is explained by all the experts. Let θ_k be the parameters associated with probability distribution of each expert ($= [\mu^k, \Sigma^k]^T$ in current case). Probability of any direction ϕ to be true heading of a human



(a) **(b)** **(c)** **(d)**

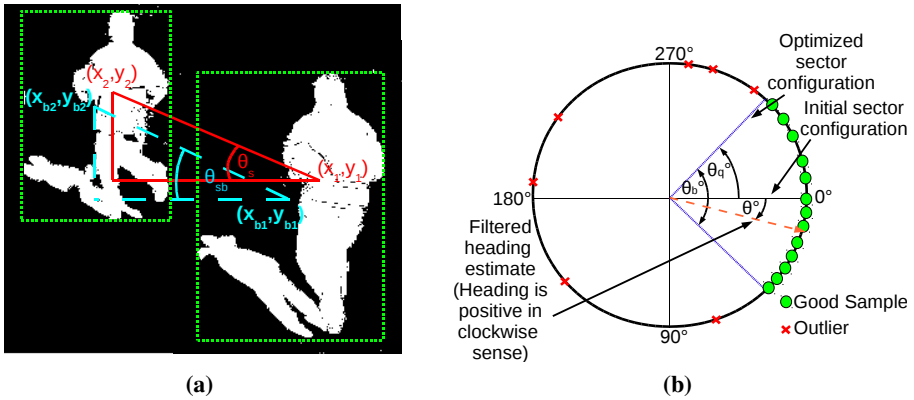
Fig. 2: Variables used in the model. (a) side view, (b) front view, and (c) top view of the human model (d) frames from a vignette in the DARPA corpus depicting that the motion cues provide significant information regarding the heading direction of a human. The red arrow portrays the direction of motion of the human in the respective frame.

as explained by all the expert estimators is given by Equation 1.

$$p(\phi|\theta_{T_1}, \theta_{T_2}, ..., \theta_{T_K}) = \frac{\prod_{k=1}^{K} p_k(\phi|\theta_k)}{\int \prod_{k=1}^{K} p_k(\phi|\theta_k) \mathrm{d}\phi} \tag{1}$$

This model results in robust estimation because it allows to incorporate (or leave out) arbitrary number of estimators, even those providing non-probabilistic output, which could also be incorporated using Equation 3.

A wealth of information about the heading direction of the human torso can be inferred solely from information regarding the human motion. In the current implementation, we focus on a PoHE based generative heading estimation method using (i) human silhouette centroid, and (ii) human silhouette bounding box centroid. Once a silhouette corresponding to a detected/tracked human is found in a frame, internal holes/gaps are filled [11] for subsequent use in the pose-estimation process. The silhouette centroid and the silhouette bounding box centroid are then evaluated for every valid frame and any gaps are filled using linear interpolation. We model the 3D heading direction as a continuous variable and approximate it as the 2D heading angle (which is the projected 3D heading angle) which works fairly well as will be evident in results. Fig. 3a depicts two human silhouettes from two different frames (N frames/δt time apart) in a video. The red triangle (solid line) connects the centroid of the two silhouettes ((x_1, y_1) to (x_2, y_2)) and the blue triangle (dashed line) connects the centroid of the two silhouette bounding boxes ((x_{b1}, y_{b1}) to (x_{b2}, y_{b2})). It can be seen that the true silhouette centroid and the silhouette bounding box centroid information are corrupted by the merging of the silhouette due to the shadow in the original human silhouette. In cases where partial silhouette information is obtained, the silhouette centroid tends to be biased towards



(a) **(b)**

Fig. 3: Human heading estimation modeling. (a) Silhouette and bounding box centroid modeling of human heading estimation, and (b) Outlier detection in angular data using optimization based sector positioning.

the region where the foreground pixels are concentrated. However, the bounding box centroid locates the centroid of the region irrespective of the foreground pixel density. By merging information from both the sources we tend to reduce the effect of noise in estimating heading direction. Equation 2 is used to evaluate an estimate of the heading direction given the centroid information for two frames.

$$\mu_k = tan^{-1}\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \tag{2}$$

We build a Gaussian distribution for each heading estimate considering the distribution mean to be situated at the corresponding estimated value and the variance to be equal to the variation in the value from its vector mean in a local temporal window. *Intuitively, we seek to weight the heading direction changes with uncertainty within each temporal window.* Please note that directional statistics [12] is required to deal with the heading angle data.

$$p(\phi_k(t)) = N(\mu_k(t), \sigma_k^2), \tag{3}$$

where, $\sigma_k^2(t) = \phi_k(t) - \bar{\phi}_k(t)$, $\bar{\phi}_k(t) = Arg(\bar{\rho}_k), k \in \{s, sb\}, \bar{\rho}_k = \frac{1}{N}\sum_{n=1}^{N} z_{kn}$, $z_{kn} = \cos\phi_k(t) + i\sin\phi_k(t)$

Outlier detection in angular data: The raw heading estimates obtained are noisy due to noise in silhouettes and so contain outliers which are eliminated using outlier detection. For outlier detection, we use an optimization based sector positioning technique in which the data lying within a sector is considered to be fit for evaluating the heading estimate within a local temporal window (Fig. 3b). The green circles on the main circle represents good samples and the red crosses represents the outliers. The blue sector represents the angular region (of angle $\theta_b = \pi/2$ degrees) samples in which are considered to be good and valid for heading estimation. Initially the sector is aligned with

the main quadrant ($\theta_q = 0$) and the sector positioning (θ_q) is determined by solving the optimization problem in (4) which maximizes the number of samples lying in the angular region:

$$\arg\max_{\theta_q}(\max \#\{\theta_q|\theta \in bin(k)\})\ k = 1...K, K = \frac{2\pi}{\theta_b} \tag{4}$$

where the symbol '#' stands for angular histogram. The optimization is carried out in local temporal sliding window to remove the outliers and Gaussian filtering is carried out on the filtered data considering the same temporal window. *Intuitively, we rely on the continuity of motion i.e. the human heading direction does not change within a fraction of a second.*

2.2 Optimization Based Body Position Estimation

We formulate the problem of determining the position of the body relative to the camera as two optimization subproblems.

Z Coordinate Estimation: The camera depth (z coordinate) estimation is based on the fact that an actual body with proportional dimensions and similar orientation in space will roughly occupy a similar area in an actual image as that of the model in the synthetic image. We set up an optimization problem based on the difference in the silhouette area in the original image and the model generated image, and minimize the square of this difference as in Fig. 4 (c_z is the z coordinate of the camera in the model coordinate system, A_o and A_m is the silhouette area in the original and model generated image, respectively.). We also specify an upper and lower bound on z coordinate such that the model generates a reasonable area in the synthetic image.

X,Y Coordinate Estimation: The estimation of the x, y coordinate is based on the fact that the centroid of the silhouette in the original image and the model generated image should roughly be the same for model with similar orientation. We setup another optimization problem in which square of the distance between the centroid of the original silhouette and the model generated silhouette is minimized constraining the (x,y) coordinates such that the model silhouette is within the synthetic image as in Fig. 4 ((c_x, c_y) is the (x, y) coordinate of the camera in the model coordinate system, (x_{co},y_{co}) and (x_{cm},y_{cm}) is the centroid of the silhouette in the original and model generated image, respectively).

2.3 Optimization Based Pose Estimation

For a given camera position, the difference between original and model generated images is minimum when the correct limb pose is achieved. The absolute subtracted image (of model generated and actual human silhouettes) measures the extent of mismatch and serves as the objective function (Fig. 4 where the subscript i indicates i^{th} joint in the human body model, I_a and I_m denotes the actual and model generated silhouette image, respectively). Limits on the human joint angles are imposed based on the biomechanical constraints set by the human body [13].

3 Optimization Approach

The probabilistic optimization techniques are good at identifying promising areas of the search space (exploration), but slow at fine-tuning the approximation to the minimum (exploitation) [14]. Thus, a much faster convergence to the local minima can be achieved if initial guesses are obtained using population based global optimization technique (Genetic Algorithm (GA) [15]) and then convergence to the global optima is accomplished using convex optimization techniques.

3.1 Convex Optimization

We use Augmented Lagrangian method (ALM) [16] for solving the optimization problem considering its advantage over penalty methods which are less robust due to sensitivity to penalty parameter chosen. In order to solve the ND unconstrained optimization subproblem, we use Powell's conjugate direction method [17] as it requires only the objective function value and is more robust to noise in function evaluation, which is often the case with image based objective functions. For 1D optimization subproblem we employ Golden section with Swann's bounding [18].

3.2 Optimization Framework

The optimization subproblems in Fig. 4 are highly coupled and cannot be solved independently. While a weighted/combined optimization problem may be posed, it suffers from multiple local minima as well as sensitivity to weightage of each objective. Hence, in lieu of this, we adopt a sequential optimization framework as shown in Fig. 4. Once, we have the heading estimates for each frame in the video, we first optimize for the camera parameters (relative location of body with respect to camera) and then for the pose assuming fixed geometries for the human body parts. In order to deal with the well-known problem of pose ambiguity due to symmetric nature of human body and keep the framework computationally feasible, we only resort to GA when either the difference between the joint angles for the the left and the right leg are below a certain threshold or the joint angle limits are exceeded, to obtain good initialization for pose. The presented framework is executed on each frame in the video to estimate two corresponding poses (left leg forward and right leg forward).

4 Experiments

We evaluated the proposed human pose estimation framework on 52 challenging video vignettes in the DARPA Mind's Eye Year I[1] dataset (resolution: 1280×720) of different activities (collide, enter, follow, flee, leave, run, jump, walk, approach, fall, pass, stop, replace, take, turn, throw, kick, go, hold, get) performed by multiple people interacting with other entities (humans/objects) in outdoor scenes.

[1] Available at https://sites.google.com/site/poseestimation/

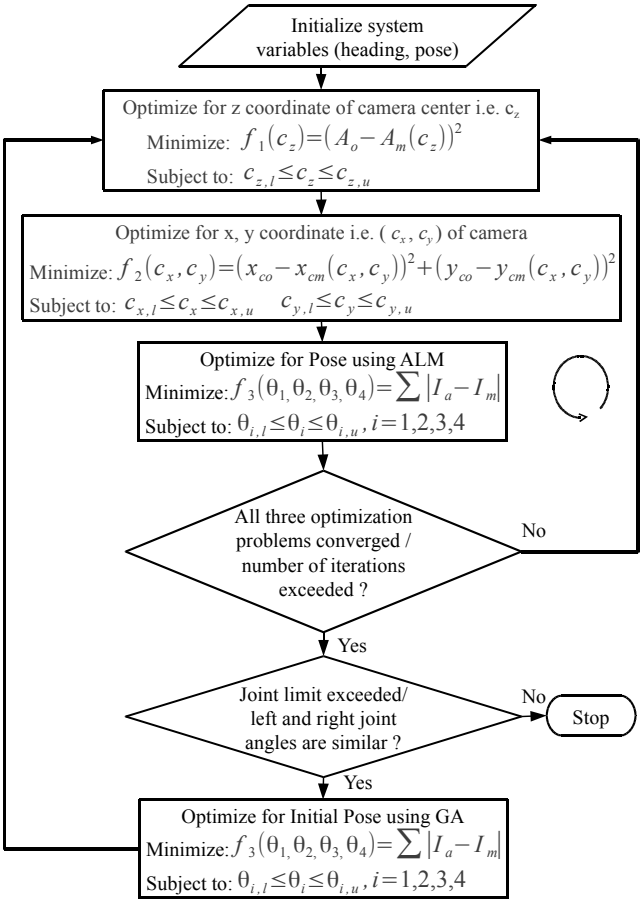


Fig. 4: Summary of optimization framework implemented for pose estimation on each frame.

Inputs: In order to thoroughly test the system performance we test the pose estimation framework on three different types of inputs: (i) Manually Labeled Silhouette (MLS) and Manually Labeled Human Bounding Boxes (MLHBB) for a selected set of the videos (6 in number) where it was possible to get good pose estimates as significant lower limb movement was involved. This trial was carried out to establish the benchmark against which to compare the performance of the algorithm with inputs of varying fidelity; (ii) Background Subtracted Silhouette (BSS), Detected Human Bounding Boxes (DHBB) [19], and Detected Object Bounding Boxes (DOBB) for the entire dataset to establish the system performance over a larger set and all algorithm generated inputs. We observed that the human detection results contains a lot of false positives along with ambiguity in entity identity while tracking; (iii) Background Subtracted Silhouette (BSS), Tracked Human Bounding Boxes (THBB) and Detected Object Bound-

ing Boxes (DOBB) for the entire corpus again to establish the system performance over a large set and more reliable human detections [20].

Pose Evaluation Metrics: Human annotation GUI[2] was developed in order to assess and quantify the performance of the pose estimation algorithm. 13 salient points on human body: head center, right shoulder, right elbow, right hand, left shoulder, left elbow, left hand, right hip, right knee, right foot (ankle), left hip, left knee, left foot (ankle) were manually marked for all videos in the corpus. We build upon the pose error metric proposed in [21] and define the following pose evaluation metrics for each vignette in the corpus: (a) Average error per frame as in (5), (b) Average error per marker per frame (D_{aepmpf}) (average of (5) for number of markers) , (c) Average error for different markers per frame as in (6).

$$D_{aepf}(X, \hat{X}) = \frac{1}{N}\left(\sum_{n=1}^{N}\sum_{m=1}^{M=13} ||x_m - \hat{x}_m||_2\right) \tag{5}$$

$$D_{aedmpf}(X, \hat{X}, m) = \frac{1}{N}\left(\sum_{n=1}^{N} ||x_m - \hat{x}_m||_1\right) \tag{6}$$

where N is the number of processed frames in the considered vignette. For vignettes with multiple humans, we first associate the estimated pose tracks with the ground truth pose tracks by using the nearest neighbor approach on the entire track, as in (7).

$$j_i = \arg\min_k \sum_{n=1}^{N}\sum_{m=1}^{M=13} ||x_{mk} - \hat{x}_{mi}||_1, E_x = \frac{1}{K}\left(\sum_{n=1}^{P} D_x\right), \text{x}\in\{\text{aepf, aepmpf, aedmpf}\} \tag{7}$$

where x_{mk}, $\hat{x}_{mi}$ are the coordinates of the m^{th} marker in the ground truth data of the k^{th} person and in the estimated data of the i^{th} person, respectively, j_i is the ground truth track associated with the i^{th} detected track, K is the number of humans present in the ground truth and P is the number of detected humans. The error over the entire corpus is the average error obtained considering all the vignettes in the corpus as in 7.

5 Results

Fig. 6 depict the stick figure and bounding boxes superimposed over the original video frame for vignettes corresponding to the verbs “pass”, “collide”, and “run” in the dataset, respectively. As can be seen the tracking is carried out while maintaining the identity of people in the video. Please note that the presented framework works well with different types of verbs[2] and does not make assumptions regarding the activity in the scene which is an unstated assumption in many state-of-the-art pose trackers.

Fig. 5 shows the error metric obtained for the two probable pose estimates using BBSS and THBB. Table. 1 shows a comparison of the pose evaluation metric for different inputs described in the Section 4. As expected, the average error per frame per

[2] Available at https://sites.google.com/site/poseestimation/

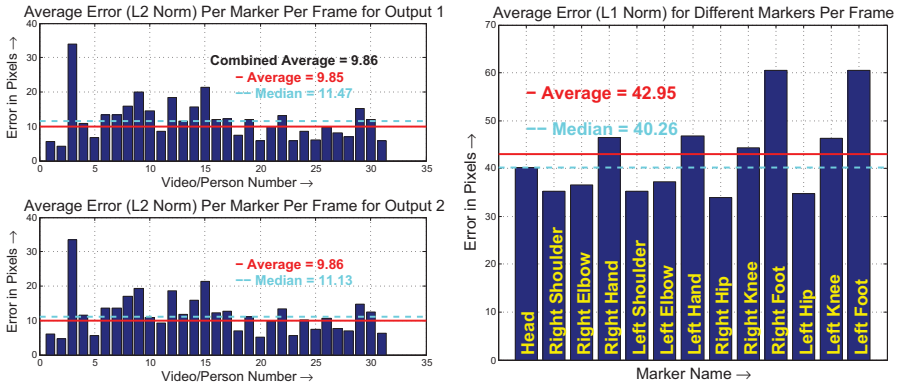


Fig. 5: Error metric on the two probable pose estimates using the BSS, THBB and DOBB. (a) Average error (L2 norm) per marker per frame, and (b) Average error distribution across markers (L1 norm) per frame.

marker increased from a value of 6 to 13 when BSS, DHBB, DOBB are provided as input as opposed to MLS, MLHBB. However, the error reduced from 13 to 10 when tracked human bounding box detections are used showing the performance of the pose estimation framework over the entire dataset. Please note that an average human head for the current dataset has a dimension of ≈50 pixels (ground truth), so an accuracy of around 10 pixels (L2 norm) and 40 pixels (L1 norm) is fairly good. Since, the current framework does not reliably distinguish between the left and the right leg the error corresponding to the foot and the knee markers is relatively high (Fig. 5b).

6 Discussion

In this work, we propose a Product of Heading Experts (PoHE) based generalized heading estimation framework bootstrapping an integrated probabilistic-deterministic optimization framework for human pose estimation in uncalibrated monocular videos. We benchmarked the standalone performance of the pose estimation framework against ground-truth data for the DARPA video corpus using the proposed pixel-distance based metrics emphasizing identity maintained human tracking and low false human detections. Results showed the robustness and performance of the proposed framework.

7 Acknowledgements

The authors gratefully acknowledge the support from Defense Advanced Research Projects Agency Mind's Eye Program (W911NF-10-2-0062).

Table 1: Error metric comparison for different inputs provided to the developed human pose estimation framework on 52 vignettes from DARPA ARL-RT1 dataset

Input/Error Metric	Average Error Per Frame (L2 Norm) (pixels)	Average Error Per Frame Per Marker (L2 Norm) (pixels)	Average Error for Different Markers Per Frame (L1 Norm) (pixels)
MLS + MHLBB (6 vignettes)	80	6	17
BSS + DHBB + DOBB (52 vignettes)	166	13	65
BSS + THBB + DOBB (52 vignettes)	128	10	43

References

1. Hen, Y.W., Paramesran, R.: Single camera 3d human pose estimation: A review of current techniques. In: International Conference for Technical Postgraduates. (2009) 1 –8
2. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding **108** (2007) 4–18
3. Sminchisescu, C.: 3d human motion analysis in monocular video: techniques and challenges. Computation Imaging and Vision **36** (2008) 185
4. Balan, A., Black, M.: An adaptive appearance model approach for model-based articulated object tracking. In: CVPR. Volume 1. (2006) 758–765
5. Sigal, L., Isard, M., Haussecker, H., Black, M.: Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. IJCV (2011) 1–34
6. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011) 1385–1392
7. Perrone, J., Zanker, J., Zeil, J.: A closer look at the visual input to self-motion estimation. Motion vision: Computational, neural, and ecological constraints (2001) 169–179
8. Ryde, J., Waghmare, S., Corso, J., Fu, Y.: ISTARE quaterly report: Signal unit. Technical report, SUNY Buffalo (2011)
9. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR. (2010) 623–630
10. Hinton, G.E.: Products of experts. In: ICANN. (1999) 1–6
11. Soille, P.: Morphological Image Analysis: Principles and Applications. Springer-Verlag, Berlin (2003)
12. Mardia, K., Jupp, P.: Directional statistics. John Wiley & Sons Inc (2000)
13. Anderson, F., Pandy, M.: Dynamic optimization of human walking. Journal of Biomechanical Engineering **123** (2001) 381
14. Costa, L., Santo, I., Denysiuk, R., MGP, E.: Hybridization of a Genetic Algorithm with a Pattern Search Augmented Lagrangian Method. In: International Conference on Engineering Optimization. (2010)
15. Goldberg, D.: Genetic algorithms in search, optimization, and machine learning. Addison-wesley (1989)
16. Schuldt, S.B.: A method of multipliers for mathematical programming problems with equality and inequality constraints. Journal of Optimization Theory and Applications **17** (1975) 155–161
17. Powell, M.J.D.: An efficient method for finding the minimum of a function of several variables without calculating derivatives. The Computer Journal **7** (1964) 155–162

18. Swann, W.H.: Report on the development of a new direct search method of optimization. Research Note (**64**)
19. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI (2009) 1627–1645
20. Kumar, S., Agarwal, P., Corso, J., Krovi, V.: ISTARE proxy evaluation report: Human tracking. Technical report, SUNY Buffalo (2011)
21. Sigal, L., Black, M.: Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. International Journal of Computer Vision **87** (2010) 4–27

Fig. 6: Raw pose estimation results for the verbs “pass”, “collide”, and “run” using the system generated inputs. N.B. Identity of the persons is maintained before and after collision for the verb “collide”. (Please view in color)