CASCADED REGRESSION FOR 3D POSE ESTIMATION FOR MOUSE IN FISHEYE LENS DISTORTED MONOCULAR IMAGES

Ghadi Salem[1]*, Jonathan Krynitsky*[1]*, Monson Hayes*[2]*, Thomas Pohida*[1] *and Xavier Burgos-Artizzu*[3]

[1] National Institutes of Health, Bethesda, Maryland 20892
[2] George Mason University, Fairfax, VA 22030
[3] Transmural Biotech, Barcelona, Spain

ABSTRACT

We address the challenging problem of estimating 3D position of laboratory mouse key-points from monocular images acquired through a fisheye lens positioned very close to the mouse 'home-cage'. This video acquisition system optical configuration is used when designing practical compact systems for large scale automated monitoring of mice in animal facility ventilated racks. The space constraints in the ventilated racks necessitate the use of fisheye lenses to ensures a more complete view of the home-cage. We extend a cascaded pose regression (CPR) algorithm that has proven to be successful for 2D pose estimation and introduce novel enhancements to the algorithm. Our 3D CPR algorithm reliably generates pose estimates starting from a rough initial pose estimate.

1. INTRODUCTION AND BACKGROUND

Research institutions rely heavily on mice for biomedical and basic research. Animal facilities utilize ventilated cage racks to house large number of mice. Researchers and animal facility staff are interested in using automated systems to quantify activity and behavior of the mice in their home-cages over long durations (e.g., multiple circadian cycles). Salem et. al. [1] reported a compact system designed specifically for use in the cage racks. The compact design is due, in large part, to the use of fisheye lenses positioned close (i.e., $< 5mm$) to the cage walls. As a result of this optical configuration, the appearance of the mouse in video is highly dependent on the mouse position within the cage. Pose estimation in this setting is further complicated by the mouse body high deformability and lack of visible articulation. Automated video-based analysis of mice activity has resulted in several custom implementations [2, 3] offering varied capabilities to answer emerging questions for researchers [4, 5]. Existing systems, with the exception of [1], utilize standard lenses with the camera positioned on the side of the cage or above the cage at a distance sufficient to ensure the field-of-view encompasses the cage volume. These monitoring setups allow for straightforward 2D planar parameterization of mouse pose, but are not well-suited for scalable use in animal vivaria [1]. Defining an image domain 2D mouse pose such as an ellipse [6, 7], oriented ellipse [8, 9, 10], deformable contour template [11], or rigid-body physics based model [12], might be uninformative in a fisheye lens based compact system as similar physical poses would have greatly varying image appearances and correspondingly varying pose parameters. In this work, we define and estimate pose in 3D space. Our pose estimation method is a novel extension of an established algorithm called Cascaded Pose Regression (CPR) [8]. CPR trains a sequence of weak regressors to iteratively improve the pose estimate relying on pose-indexed features, i.e., features that depend on the current estimate of the pose. Following the notation of the work upon which we are expanding, a cascaded pose regressor R is composed of T weak regressors, i.e., $R = (R^1, \cdots, R^T)$. Each regressor R^t produces a pose perturbation θ_δ^t of the previous pose estimate θ^{t-1} towards the 'true' pose. At the core of CPR is that the features h^t fed into R^t to produce θ_δ^t are *pose-indexed*, i.e., dependent on the previous pose estimate θ^{t-1}. Hence h^t is a function of θ^{t-1} and the image I. The pose estimate at stage t in the cascade, θ_t, is computed as

$$\theta^t = \theta^{t-1} \circ \theta_\delta^t, \qquad \theta_\delta^t = R^t(h^t(\theta^{t-1}, I)) \tag{1}$$

where the operator $\circ$ is defined over pose space Θ, such that $(\Theta, \circ)$ form a group with a properly defined inverse operator. One condition for convergence to true pose is that the pose-indexed features h are *weakly invariant*. Namely, the features depend on the difference between pose estimate and true pose. Mathematically stated, if true pose is θ and the current pose estimate is θ^{t-1}, then $h^t(\theta^{t-1}, I_\theta) = h^t(\vartheta^{t-1}, I_\vartheta)$, where $\vartheta^{t-1} = \theta^{t-1} \circ \theta_\epsilon$ is the pose estimate for a different true pose $\vartheta = \theta \circ \theta_\epsilon$. Cao et al. [13] introduced a variation on CPR that included many enhancements, most notably elimination of a parameterized model and implementing boosted regressors at each stage in the cascade.Another variation, [14], introduced features that are more invariant to scale and rotation. In both [13, 14] the desired output pose is defined in 2D image coordinates. The output of each stage in the cascade is defined relative to 2D normalized target bounding box. A simple geometric transformation projects the normalized pose onto the

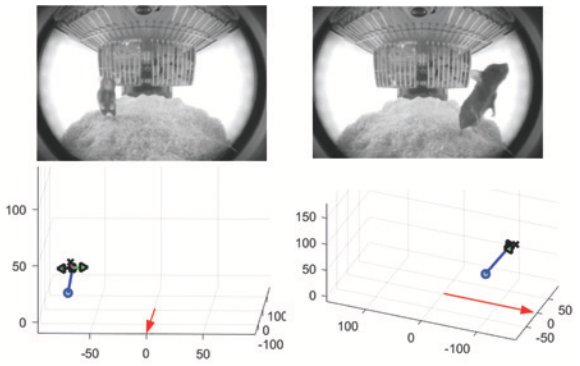

Fig. 1. Two examples of the fisheye lens distorted images acquired by the system described in [1] along with the corresponding 3D pose estimates for the mouse. The red arrow in the 3D plots points to the front of the cage.

image coordinates. In our problem, the desired output pose is defined in 3D physical coordinates. The relation between image coordinates and physical coordinates is ambiguous due to absence of scaled orthography. In what follows we present an extension of CPR to 3D enabling pose estimation in the challenging optical configuration of the acquisition system. We use the cascade structure employed in [13] to directly estimate 3D pose without a parameterized model. We introduce robust features that mitigate the violation of weak-invariance caused by the recording environment.

2. METHOD

The goal is to estimate ϕ, the 3D physical coordinates of four mouse key points in a Cartesian coordinate system with origin defined (arbitrarily) at the center of the cage floor. Estimating ϕ is accomplished from a monocular image acquired by a fisheye lens positioned horizontally (i.e., not overhead) within $5mm$ of the cage wall. Mathematically stated,

$$\phi = [\phi_i], \qquad \phi_i = [x_i, y_i, z_i], \qquad i \in \{1, \cdots, 4\} \tag{2}$$

where the sequence of i corresponds to tail, nose, left-ear, and right-ear key points in order. Since CPR is a supervised learning framework, the algorithm requires availability of ground truth values for the 3D coordinates of the key points. To generate the ground truth set of key point coordinates, an overhead camera fitted with a fisheye lens was mounted at each end of the cage. The camera is strictly used for the purpose of generating the training set as the pose estimation described in this work is achieved from the monocular image acquired through the horizontal camera only. All cameras were calibrated, and acquisition was synchronized. Annotators marked the four key points of interest in the horizontal and overhead images. In cases where the mouse was symmetrical along the spine, only the more visible of the two ears is annotated and the position of the other was recovered using geometrical constraints. The image annotations were converted to physical 3D coordinates, i.e. ϕ's, via the camera calibration mappings. The training set is comprised of approximately 50,000 ground truth annotations.

2.1. Foreground detection

In [8, 13, 14] a trained object detector supplies a bounding box for the target in the image, e.g., face detector. We capitalize on the constrained recording environment in our setup to build a classification model to identify mouse pixels in the image. Generating a segmentation map is more informative than supplying a detection window and is exploited in other parts of the algorithm. A set of 250 images each with corresponding manually generated segmentation mask is used to train a decision forest classifier to predict a per-pixel binary foreground/background label. To derive discriminative features from the image, linear and non-linear transformations are applied to generate additional information channels which are pixel-to-pixel registered to the image as was described in [15]. Namely, we compute per pixel image gradients as well as 4-bin HOG, each bin of which serves as a separate information channel. In training, each pixel ground-truth label (i.e., foreground or background) is paired with a feature vector formed by selection of features from all available information channels. We also take advantage of the fixed camera positioning by utilizing the pixel image position as a feature. Hence, the feature vector is augmented with the labelled pixel (x, y) location in the image. At run-time, the feature channels for the whole image are computed. Each pixel's feature vector is constructed in the same method used during training. The trained classifier is evaluated for the given feature vector and a segmentation label prediction is output.

2.2. Pose definition

In [14, 13], the desired pose ϕ is the image coordinates of the face landmarks. The regression cascade, however, operates on θ^t in equation (1) defined as 2D offsets relative to a normalized detection window. Representing pose relative to the detection window achieves scale and translation invariance. The output of the cascade, θ^T, is projected back onto the detection window by a simple translation and scaling to obtain ϕ. In our case, ϕ is defined in (2). We herein present our definition of a translation-tolerant θ^t. Conversion between θ^t and ϕ is taken directly from the definition. Our definition of θ^t establishes a 3D reference point for the tail position by exploiting the stationary camera placement and defining an initial image-to-physical coordinate mapping $\mathcal{M}(\sigma)$ based on the camera calibration. The mapping $\mathcal{M}(\sigma)$ assigns to each image pixel σ a 3D position $(\mathcal{M}_x(\sigma), \mathcal{M}_y(\sigma), \mathcal{M}_z(\sigma))$ in the same Cartesian coordinates system on which ϕ in (2) is defined. Since each pixel in the image corresponds to a line in 3D space, the initial assignment for $\mathcal{M}$ involves selecting a

point along the line by applying an arbitrary constraint. For example, for all pixels with 3D lines intersecting the cage-floor, we choose the initial mapping by constraining z to lie on the cage-floor plane.In θ^t, the tail is expressed relative to $\mathcal{M}(\sigma_a)$, where σ_a is the foreground ellipse fit major axis endpoint closest to the lower left corner of the image. Overall,

$$\theta^t = [\phi_1 - \mathcal{M}(\sigma_a), \phi_i - \phi_1], \qquad i \in \{2, 3, 4\} \tag{3}$$

where, $\forall i$, ϕ_i's are the key points coordinates as per (2).

2.3. Initializing the cascade

In [14, 13], θ^0 is computed offline as the mean positions of the landmarks in the normalized detection window. In our case, defining θ^0 as the mean of θ_i in the training set would be uninformative and often place the initial estimate too far for convergence. Instead, we generate an initial ϕ^0 (as opposed to θ^0) based on the binary silhouette and the mapping $\mathcal{M}$. θ^0 is then computed from ϕ^0 as per equation (3). We now describe how ϕ^0 is computed. Let e be the ellipse fit of the detected foreground binary silhouette. Furthermore, let σ_a, σ_b be be the end points of e's major axis such that σ_a is the end point closest to the lower left corner of the image. ϕ_1^0, the tail coordinates, is taken to be $\mathcal{M}(\sigma_a)$, whereas ϕ_2^0, the nose coordinates, is taken to be $\mathcal{M}(\sigma_b)$. Since the head of the mouse is a rigid structure, we rely on the known fixed distances between nose and ears to compute an initial assignment for ears, i.e. $\{\phi_3^0, \phi_4^0\}$. Rigid 3D similarity transforms compute ϕ_3^0, ϕ_4^0 with the assumption that they are coplanar with the tail ϕ_1^0 and nose ϕ_2^0. Furthermore, the plane on which $\phi_1^0, \phi_2^0, \phi_3^0, \phi_4^0$ are computed to lie is assumed to have no roll relative to the XY plane in 3D Cartesian coordinates, rather only a pitch and a yaw determined by the relative position of nose ϕ_2^0 to tail ϕ_1^0.

2.4. 3D Pose-indexed features

As stated earlier, the key to the success of CPR and its variants is the utilization of pose-indexed features. Our feature selection method is similar to that described in [14] but extended to 3D with additional enhancements. The features are chosen to lie in the space along the line connecting two landmarks, however not strictly on the line. Instead, we allow the feature positions to deviate by up to a pre-specified offset from the line. Furthermore, features are extracted along all possible pairings of the key points as opposed to a random subset of the pairings as in [14]. For each pairing, the 3D feature positions are defined relative to a normalized segment (i.e., of length = 1). To get the corresponding pixel locations, the 3D similarity transformation matrix between the normalized segment endpoints and the paired keypoints in the current pose ϕ^t is computed. The transformation matrix is applied to the normalized feature positions to map the positions to the same 3D coordinate system on which the keypoints are defined. The camera calibration mappings are then utilized to obtain the image coordinates to which the 3D feature positions project.

2.5. Weak-invariance assumption

One requirement for convergence, as stated in [8], is weak invariance of features. The assumption does not hold in our setup since the mouse appearance in the image is highly dependent on the mouse position relative to the camera. The two images in Figure 1 show the mouse in similar posture (i.e., position of key points relative to tail) but at different tail position within the cage. The extent of appearance dependence on position is quite evident. To combat the observed violation of features weak-invariance, we include the detected binary silhouette statistics (which are correlated, albeit ambiguously, with position) as features alongside the pairwise intensity difference image features described in section 2.4. Namely, we use area, orientation, major and minor axis lengths, major-to-minor axis length ratio, ellipse fit axes end points, and bounding box corners as features. In order to integrate usage of binary silhouette features with image intensity values pairwise comparisons within the same ferns framework described in [14], we perform two operations. First we normalize each binary silhouette feature by subtracting the mean of the feature encountered in the training set and divide by a multiple of the standard deviation of the feature. Second, to maintain a meaningful pairwise difference, whenever a normalized binary silhouette feature is selected in a fern, it is paired up with the corresponding mean (i.e., 0) as opposed to any other feature.

3. RESULTS

The algorithm was applied to two separate video sequences that were excluded from training. The sequences have a total of $\sim 1,100$ annotated frames with the θ's reconstructed from the image annotations regarded as ground truth. To assess the quality of pose estimates, we define a distance measure between two poses θ^1 and θ^2 similar to that proposed in [8], namely $d(\theta^1, \theta^2) = \sqrt{\frac{1}{12}\sum_{i=1}^{12} \frac{1}{\sigma_i^2}(\theta^1(i) - \theta^2(i))^2}$ where $i \in \{1, \ldots, 12\}$ to account for all three parameters (i.e., coordinates) in each of the 4 key points. The σ_i^2's refer to the variance in each parameter between two human annotators. To compute σ_i^2's, approximately 6,000 frames were redundantly annotated by two annotators. The corresponding $\phi(i)$'s were reconstructed as described in section 2. The Euclidean distance between $\phi(i)$'s was computed, i.e., $d_i = \|\phi^1(i) - \phi^2(i)\|$ where the superscript denotes the annotator. σ_i^2 is computed as the variance of d_i for the whole set of redundantly annotated frames. The utilized distance measure weighs the error contribution of each key point by the observed variance in the redundant annotations, hence equalizing the error from each key point and providing a comparison between our method and human annotators. Following [8], we define a normalized distance threshold d_{thr} for a successful estimate. d_{thr} is set to be such that the normalized distance for 99% of the redundantly annotated frames fall below d_{thr}.

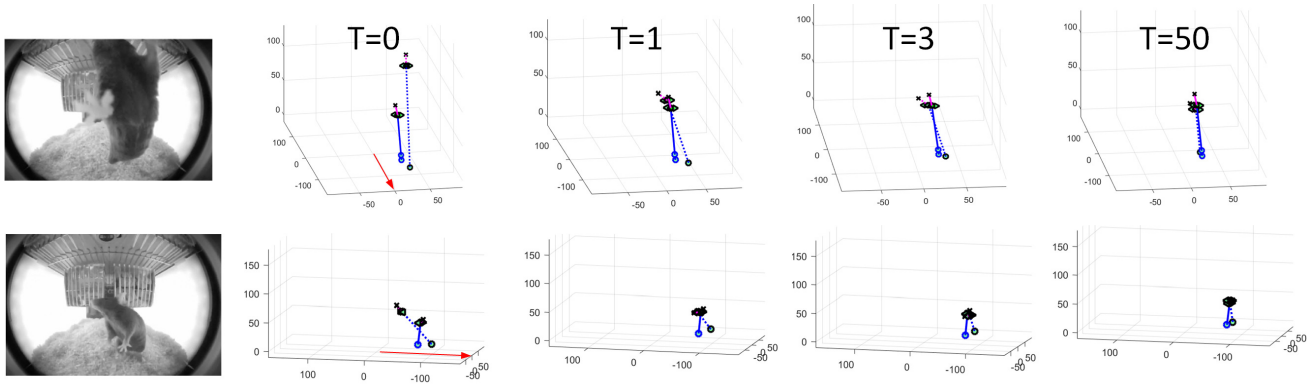

Fig. 2. Ground truth (solid stick-figure) compared to pose estimate (dashed stick-figure) at different stages in the regression cascade for two example images. The red arrow in the 3D plots for $T = 0$ points to the front of the cage.

Table 1. Performance as a function of number of stages

Stages (T)	% Failure	Mean d for success
1	73.0	2.92
5	64.0	2.66
15	58.3	2.56
25	54.8	2.49
50	52.2	2.42
100	49.8	2.34

So if the estimation output has a normalized distance (i.e., that computed via the aforementioned distance measure) exceeding d_{thr}, then it is considered a failed estimate. The performance was evaluated under different parameter settings. Figure 2 pictorially shows example results at different stages in the cascade for the optimal model generated with the following settings: T=50, feature position selection with offsets, and inclusion of binary silhouette features. The coarse-to-fine manner in which CPR works is evident as the first stage (i.e., $T = 1$) results in the largest pose correction whereas subsequent stages further fine-tune the estimates. Figure 3 shows the normalized distance distribution plots for all 1,100 frames under different parameter settings. The plots include the distribution for the initial estimate θ^0 used to seed the cascade, the optimal model (marked OPTIMAL), a model without binary silhouette features (marked w/o BSF), and lastly one where feature positions are strictly chosen on the line connecting two landmarks (called NO OFFSET). Table 1 shows the performance as a function of number of cascade stages.

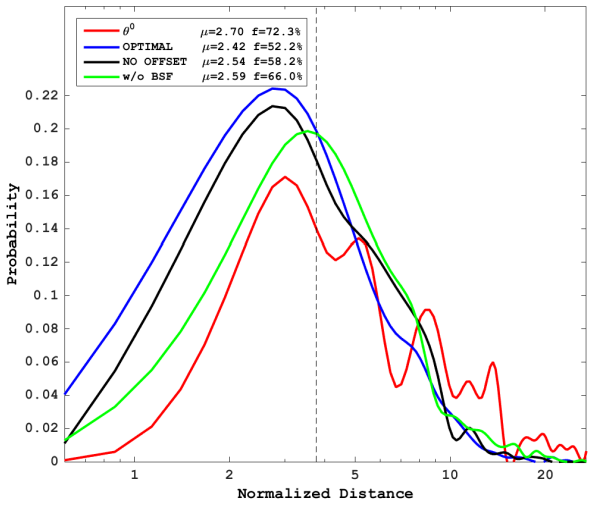

Fig. 3. Normalized distance distributions for initial pose estimate and several parameter settings. The dashed vertical line marks d_{thr}, the success/failure threshold.

4. DISCUSSION

We have successfully implemented a cascaded pose-indexed regression algorithm for estimating 3D pose of a mouse in monocular fisheye lens distorted images. Our algorithm not only extends CPR to 3D but also introduces key enhancements in feature choice that considerably increase accuracy. While a failure rate of 52% for our optimal model may seem high, two points are worth noting. First, in the seemingly easier task of 2D mouse pose estimation the failure rates reported in [8, 9] exceeded 30%. Given the greater dimensionality and complexity of the 3D estimation problem, confounded by the challenging imaging distortion, the 3D algorithm is expected to provide decreased performance relative to the original 2D case. Second, the distance measure used to set the failure metric does not account for the fact that estimates are computed from a single 2D image (i.e., horizontal view) whereas the ground-truth poses to which they are compared were reconstructed from annotations of a pair of images (i.e., horizontal and top views). The dataset is online: (scorhe.nih.gov).

5. REFERENCES

[1] Ghadi H Salem, John U Dennis, Jonathan Krynitsky, Marcial Garmendia-Cedillos, Kanchan Swaroop, James D Malley, Sinisa Pajevic, Liron Abuhatzira, Michael Bustin, Jean-Pierre Gillet, et al., "Scorhe: A novel and practical approach to video monitoring of laboratory mice housed in vivarium cage racks," *Behavior research methods*, vol. 47, no. 1, pp. 235–250, 2015.

[2] Anthony I. Dell, John A. Bender, Kristin Branson, Iain D. Couzin, Gonzalo G. de Polavieja, Lucas P.J.J. Noldus, Alfonso Perez-Escudero, Pietro Perona, Andrew D. Straw, Martin Wikelski, and Ulrich Brose, "Automated image-based tracking and its application in ecology," *Trends in Ecology and Evolution*, vol. 29, no. 7, pp. 417 – 428, 2014.

[3] Xavier P Burgos-Artizzu, Piotr Dollár, Dayu Lin, David J Anderson, and Pietro Perona, "Social behavior recognition in continuous video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1322–1329.

[4] Monya Baker, "Animal models: inside the minds of mice and men," *Nature*, vol. 475, no. 7354, pp. 123–128, 2011.

[5] Andreas T Schaefer and Adam Claridge-Chang, "The surveillance state of behavioral automation," *Current opinion in neurobiology*, vol. 22, no. 1, pp. 170–176, 2012.

[6] Hemerson Pistori, Valguima Victoria Viana Aguiar Odakura, João Bosco Oliveira Monteiro, Wesley Nunes Gonçalves, Antonia Railda Roel, Jonathan de Andrade Silva, and Bruno Brandoli Machado, "Mice and larvae tracking using a particle filter with an auto-adjustable observation model," *Pattern Recognition Letters*, vol. 31, no. 4, pp. 337–346, 2010.

[7] Kristin Branson, Vincent Rabaud, and Serge J Belongie, "Three brown mice: See how they run," in *VS-PETS Workshop at ICCV*, 2003.

[8] P. Dollar, P. Welinder, and P. Perona, "Cascaded pose regression," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 1078–1085.

[9] Xavier P Burgos-Artizzu, David C Hall, Pietro Perona, and Piotr Dollár, "Merging pose estimates across space and time," in *BMVC*, 2013.

[10] Weizhe Hong, Ann Kennedy, Xavier P. Burgos-Artizzu, Moriel Zelikowsky, Santiago G. Navonne, Pietro Perona, and David J. Anderson, "Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning," *Proceedings of the National Academy of Sciences*, vol. 112, no. 38, pp. E5351–E5360, 2015.

[11] K. Branson and S. Belongie, "Tracking multiple mouse contours (without too many samples)," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, June 2005, vol. 1, pp. 1039–1046 vol. 1.

[12] Fabrice de Chaumont, Renata Dos-Santos Coura, Pierre Serreau, Arnaud Cressant, Jonathan Chabout, Sylvie Granon, and Jean-Christophe Olivo-Marin, "Computerized video analysis of social interactions in mice," *Nature Methods*, vol. 9, no. 4, pp. 410–417, 2012.

[13] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[14] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollar, "Robust face landmark estimation under occlusion," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[15] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *BMVC*, 2009.