Pose Merging in Collaborative Workspace

NCS Lab

December 3, 2021

1 Coordinate Transformations

3D positions and transformations exist within coordinate systems called spaces. World space is the coordinate system for the entire scene. Its origin is at the center of the scene. Object space is the coordinate system from an object's point of view. The origin of object space is at the object's predefined centroid. Camera space is the coordinate system from the camera's point of view. The origin of the camera space is on the chip plane, right in the center of the pixel matrix. The camera's z axis is normal to its chip and pointing opposite to the direction in which the object is visible.

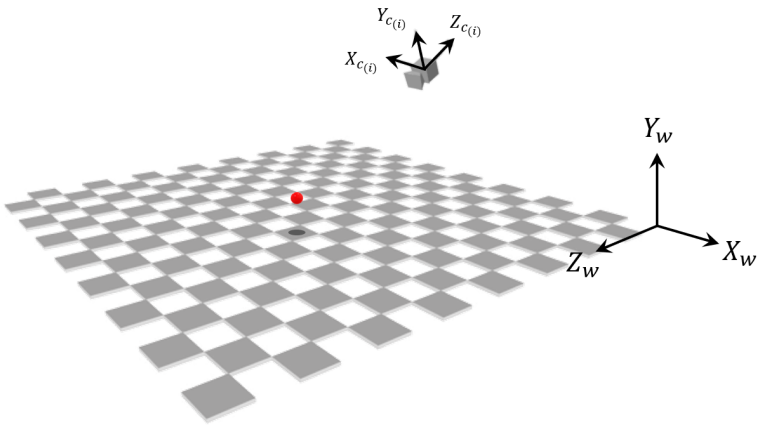


Figure 1: Camera Space

To convert observations made by a camera i with pose $[c_{x_{(i)}}, c_{y_{(i)}}, c_{z_{(i)}}, \alpha_{(i)}, \beta_{(i)}, \gamma_{(i)}]$ (where $\alpha_{(i)}, \beta_{(i)}, \gamma_{(i)}$ are the euler angles for pitch, yaw and roll) a transformation matrix $M_{c2w_{(i)}}$ is required. Such matrix is defined as follows:

$$M_{c2w_{(i)}} = T_{xyz_{(i)}} \cdot R_{z_{(i)}} \cdot R_{y_{(i)}} \cdot R_{x_{(i)}} = \begin{bmatrix} m11_{(i)} & m12_{(i)} & m13_{(i)} & m14_{(i)} \\ m21_{(i)} & m22_{(i)} & m23_{(i)} & m24_{(i)} \\ m31_{(i)} & m32_{(i)} & m33_{(i)} & m34_{(i)} \\ m41_{(i)} & m42_{(i)} & m43_{(i)} & m44_{(i)} \end{bmatrix} \tag{1}$$

where:

$$R_{x_{(i)}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(\alpha_{(i)}) & -sin(\alpha_{(i)}) & 0 \\ 0 & sin(\alpha_{(i)}) & cos(\alpha_{(i)}) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_{y_{(i)}} = \begin{bmatrix} cos(\beta_{(i)}) & 0 & sin(\beta_{(i)}) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(\beta_{(i)}) & 0 & cos(\beta_{(i)}) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_{z_{(i)}} = \begin{bmatrix} cos(\gamma_{(i)}) & -sin(\gamma_{(i)}) & 0 & 0 \\ sin(\gamma_{(i)}) & cos(\gamma_{(i)}) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$T_{xyz_{(i)}} = \begin{bmatrix} 1 & 0 & 0 & c_{x_{(i)}} \\ 0 & 1 & 0 & c_{y_{(i)}} \\ 0 & 0 & 1 & c_{z_{(i)}} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

An object with centroid at location $P_{c_{(i)}} = [x_{c_{(i)}}, y_{c_{(i)}}, z_{c_{(i)}}, 1]$, as observed by such camera, will actually be located, in the world coordinate frame, at $P_{w_{(i)}} = [x_{w_{(i)}}, y_{w_{(i)}}, z_{w_{(i)}}, 1] = M_{c2w_{(i)}} \cdot P_{c_{(i)}}$ which can be expressed with the following equations:

$$x_{w_{(i)}} = m_{11_{(i)}} \cdot x_{c_{(i)}} + m_{12_{(i)}} \cdot y_{c_{(i)}} + m_{13_{(i)}} \cdot z_{c_{(i)}} + m_{14_{(i)}} \tag{2}$$

$$y_{w_{(i)}} = m_{21_{(i)}} \cdot x_{c_{(i)}} + m_{22_{(i)}} \cdot y_{c_{(i)}} + m_{23_{(i)}} \cdot z_{c_{(i)}} + m_{24_{(i)}} \tag{3}$$

$$z_{w_{(i)}} = m_{31_{(i)}} \cdot x_{c_{(i)}} + m_{32_{(i)}} \cdot y_{c_{(i)}} + m_{33_{(i)}} \cdot z_{c_{(i)}} + m_{34_{(i)}} \tag{4}$$

2 Gaussian Distributions

The normal probability distribution function (PDF) for a continuous random variable x is defined by a Gaussian with mean μ and standard deviation σ:

$$PDF(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Which, can also be written as follows:

$$x \sim \mathcal{N}[\mu_x, \sigma_x^2]$$

The PDF of a linear function of continuous random variables is also a Gaussian. For example, when:

$$x \sim \mathcal{N}[\mu_x, \sigma_x^2] \quad and \quad y \sim \mathcal{N}[\mu_y, \sigma_y^2] \quad and \quad z \sim \mathcal{N}[\mu_z, \sigma_z^2]$$

if $f(x, y, z) = a \cdot x + b \cdot y + c \cdot z + d$ then:

$$f(x, y, z) \sim \mathcal{N}[a \cdot \mu_x + b \cdot \mu_y + c \cdot \mu_z + d, (a \cdot \sigma_x)^2 + (b \cdot \sigma_y)^2 + (c \cdot \sigma_z)^2]$$

which basically means that:

$$\mu_{f(x,y,x)} = a \cdot \mu_x + b \cdot \mu_y + c \cdot \mu_z + d \tag{5}$$

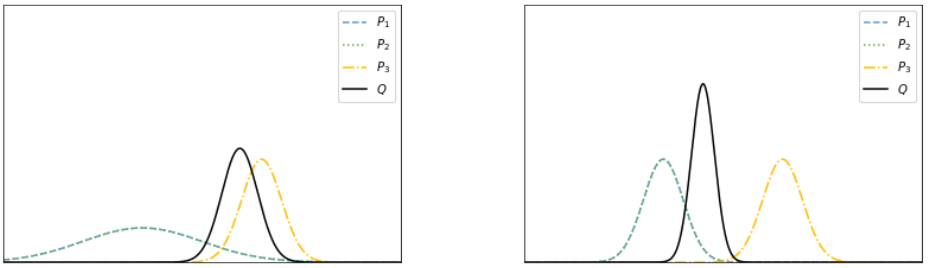
$$\sigma^2_{f(x,y,z)} = (a \cdot \sigma_x)^2 + (b \cdot \sigma_y)^2 + (c \cdot \sigma_z)^2 \tag{6}$$

2.1 Conflation

The conflation of a finite number of probability distributions is a consolidation of those distributions into a single probability distribution $Q = \&(P_1, ..., P_n)$. The conflation is the distribution determined by the normalized product of the probability density or probability mass functions. When $P_1, ..., P_n$ are Gaussian, Q is Gaussian with mean and standard deviation given by Equations 7 and 8. Examples of conflation are shown in Figure 2: case a) shows how the expected value of the consolidated distribution tends towards the expected value of the base distribution with lower standard deviation and case b) shows how 'majority wins'. At first glance it may seem counter-intuitive that the conflation of three relatively broad distributions can be a much narrower one. However, if the sources of the PDFs (i.e. the measurments) are assumed to be equally valid, then with relatively high probability the true value should lie in the overlap region between the distributions.

$$\mu_Q = \frac{\sigma_2 \cdot \sigma_3 \cdot \mu_1 + \sigma_3 \cdot \sigma_1 \cdot \mu_2 + \sigma_1 \cdot \sigma_2 \cdot \mu_3}{\sigma_1 \cdot \sigma_2 + \sigma_2 \cdot \sigma_3 + \sigma_3 \cdot \sigma_1} \tag{7}$$

$$\sigma_Q = \sqrt{\frac{\sigma_1 + \sigma_2 + \sigma_3}{\sigma_1 \cdot \sigma_2 + \sigma_2 \cdot \sigma_3 + \sigma_3 \cdot \sigma_1}} \tag{8}$$



(a) $\sigma_1 = \sigma_2 > \sigma_3$ (b) $\sigma_1 = \sigma_2 = \sigma_3$

Figure 2: Conflation of P_1, P_2 and P_3; $\mu_1 = \mu_2 = -\mu_3$.

3 Pose merging

Figure 3 shows the workflow needed in a setup where three cameras (in blue, green and yellow) are used to get an estimation of the coordinates at which a particular object (in red) is located. The workflow, which employs the concepts introduced in sections §1 and §2, can be briefly described as follows: each camera observes the workspace where an object is found, these data (events or RGB images) are fed to an instance of a pre-trained neural network (SNN or not) so predictions of object's location can be made in camera space; those predictions come with some uncertainty and it is assumed that the random process at their core can be represented by a normal distribution (i.e. 9 Gaussian PDFs, 3 per camera: x, y and z); normal distributions are transformed from camera space to world space and conflated to consolidate a unique estimate for object location (i.e. 3 Gaussian PDFs: x, y and z). Sub-sections §3.1 and §3.2 will focus on explaining the pose merging workflow.

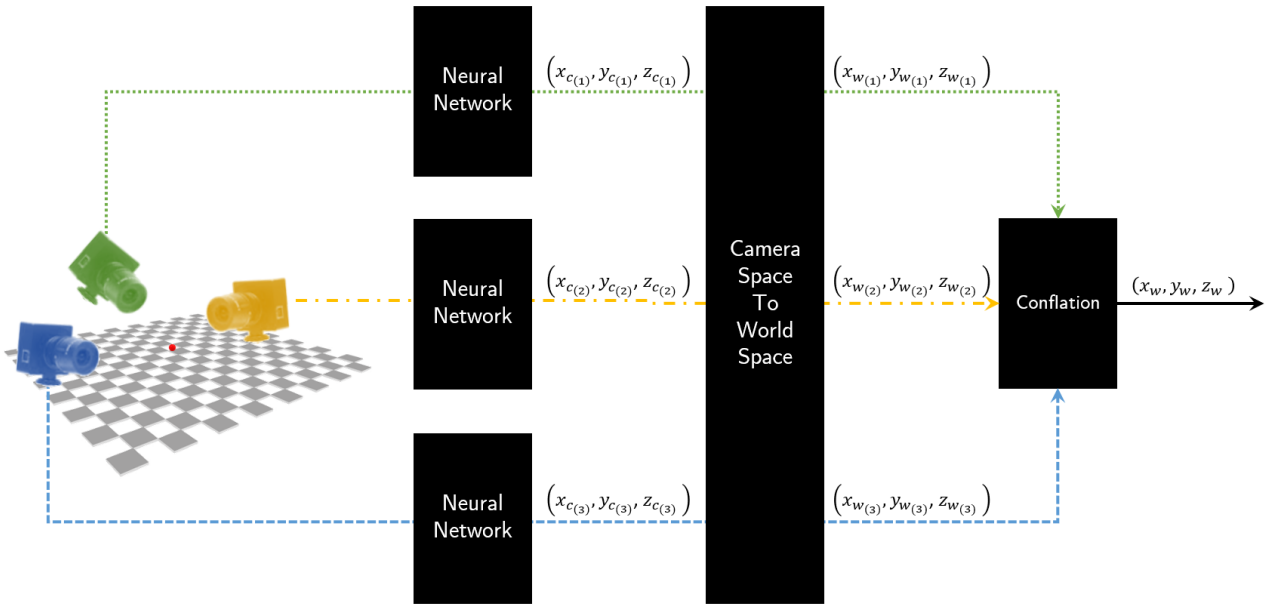


Figure 3: From camera raw data to pose estimation using SNNs, coordinate frame transformation and conflation of normal distributions; $x_{w_{(i)}}$, $y_{w_{(i)}}$ and $x_{z_{(i)}}$ are given by equations 2, 3 and 4.

3.1 Principles

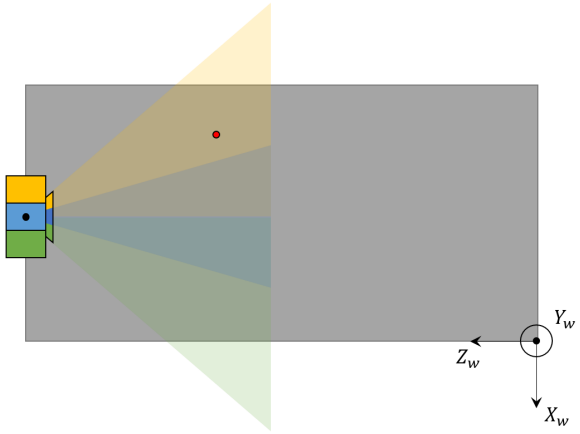
Generally the uncertainty associated to observations made in camera space by a single camera along the axis x/y can't be smaller than half a pixel length/height. This uncertainty can grow due to lens distortion, specially for outer pixels, i.e. pixels closer to the edges of the pixel matrix. As for the z axis, in absence of fancy/smart software, a camera can't estimate depth, which leads to an uncertainty tending to infinity. This, of course, can be counteracted by different methods. Neural networks (spiking or not) are one of such methods used to translate 2D projections of 3D environments into 3D coordinates of objects immersed in such environments. Those networks are trained using large datasets in which ground truth, i.e. actual pose of object and camera, is known and where images taken from different perspectives allow to create latent representations of the world incorporating relevant *implicit* information such as object size or distance object-to-camera. These datasets are hard to label in real life, for this reason it is of utmost importance to 1) train models with simulated data and 2) transfer trained models from virtual to real environments. In both steps, sources of error can degrade accuracy, but it is on the latter where, in spite of how good the model validation with simulated data can be, this degradation can be more noticeable and that is the reason why it is still the general case that the uncertainty for observations along the z axis in camera space is considerably higher than the uncertainties along axes x and y. Now, when it comes to world space, the situation changes since, depending on the camera pose, the uncertainties along the world's axes will be determined by a composition of the uncertainties along the camera's axes. In the particular case of a setup consisting of 3 cameras, each camera observes the workspace and provides locations of objects relative to its pose. As seen in section 1, it is possible to estimate the equivalent location of such objects in world space. Since each camera will provide a set of Cartesian coordinates in 3D, 9 values are expected. Now, ideally $x_{w_{(1)}} = x_{w_{(2)}} = x_{w_{(3)}}$, $y_{w_{(1)}} = y_{w_{(2)}} = y_{w_{(3)}}$ and $z_{w_{(1)}} = z_{w_{(2)}} = z_{w_{(3)}}$. However, that will be rarely the case because each camera will introduce some uncertainty to its observations along the three axes due to at least three factors mentioned above. This disagreement can be solved by consolidating probability distributions by means of conflation, as explained in section 2.1

3.2 Scenarios

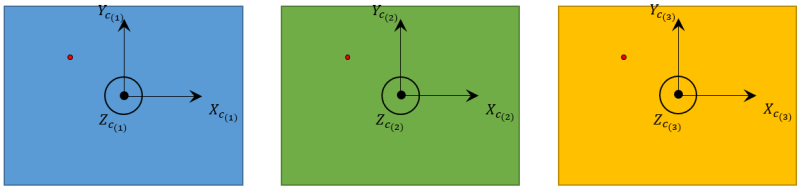
The scenarios in Figures 4, 5, 6 and 7 illustrate how observation uncertainties along axes in camera space influence uncertainties along axes in world space based on camera poses. Each scenario consists of a setup where three cameras observe a workspace (in grey) where an object (in red) is located. Each figure is divided in fours subfigures, subfigure *a*) shows a top view of the setup; subfigure *b*) shows views of the cameras' pixel matrices and subfigure *c*) and *d*) for show respectively for camera space and world space the density of probability (centered around the predicted object position) for each camera along each axis and the corresponding conflation of Gaussians. It must be said that Figures 4c, 5c, 6c and 7c contain exactly the same information because the nature of the cameras does not change from one scenario to the next and the plots are centered around the Gaussian's expected value.

3.2.1 Scenario 1

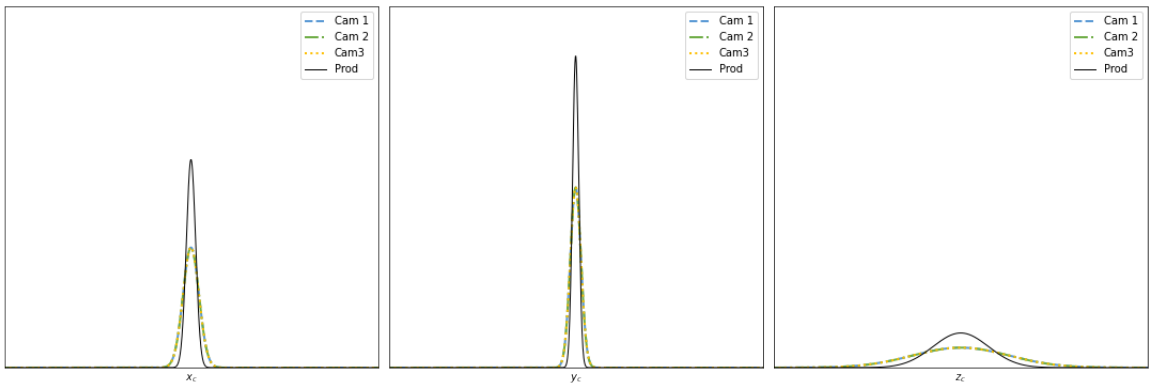
In this scenario, which is not plausible in reality, three cameras are located at exactly the same coordinates. The camera space and the world space are aligned; i.e. $X_w = X_c$, $Y_w = Y_c$ and $Z_w = Z_c$. All the cameras are located at the same height. The uncertainty along Z_w is higher the ones along X_w and Y_w in the same way the uncertainty along Z_c is higher the ones along X_c and Y_c.



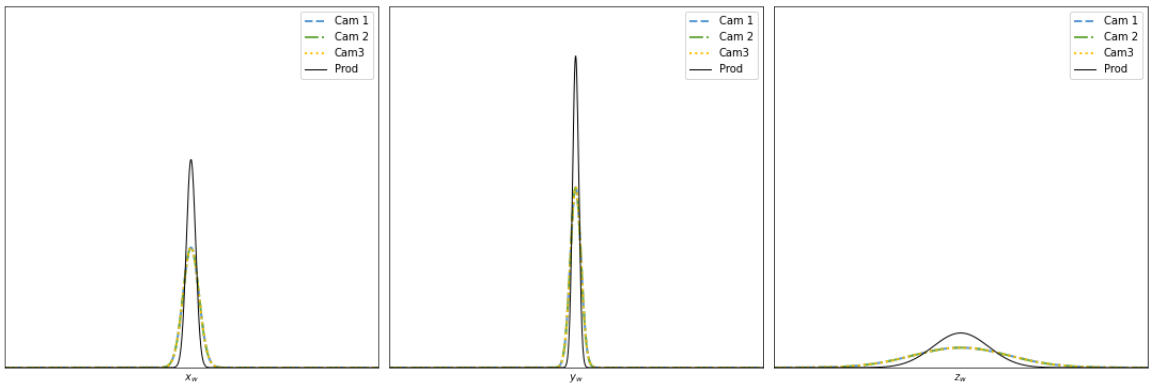
(a) Top view of the setup



(b) Chip view (camera space) for cameras 1, 2 and 3



(c) Density of probability for $x_{c_{(i)}}$, $y_{c_{(i)}}$ and $z_{c_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

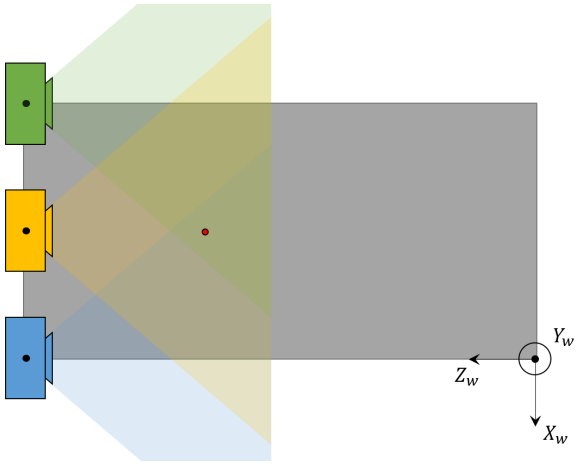


(d) Density of probability for $x_{w_{(i)}}$, $y_{w_{(i)}}$ and $z_{w_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

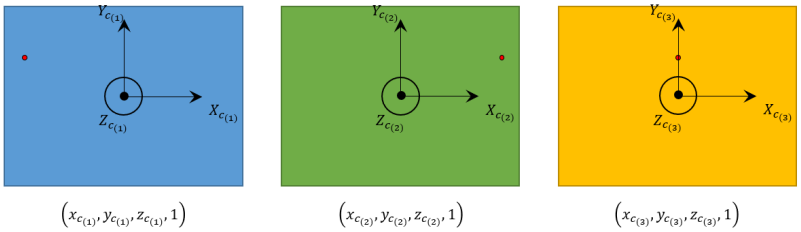
Figure 4: Scenario 1

3.2.2 Scenario 2

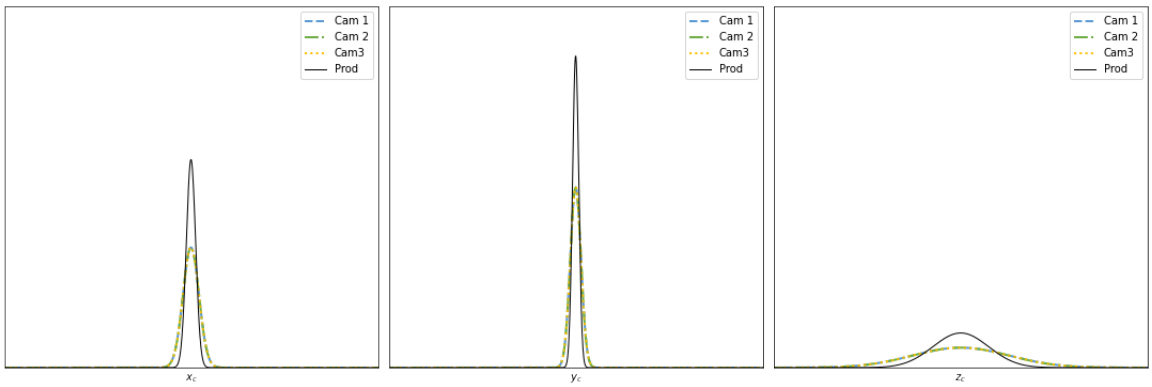
In this scenario, three cameras are located at exactly the same z_w coordinate. The camera space and the world space are aligned; i.e. $X_w = X_c$, $Y_w = Y_c$ and $Z_w = Z_c$. All the cameras are located at the same height (y_w). However, $x_{w_1} \neq x_{w_2} \neq x_{w_3}$. The uncertainties don't change, compared to scenario 1 since camera space and world space are still aligned.



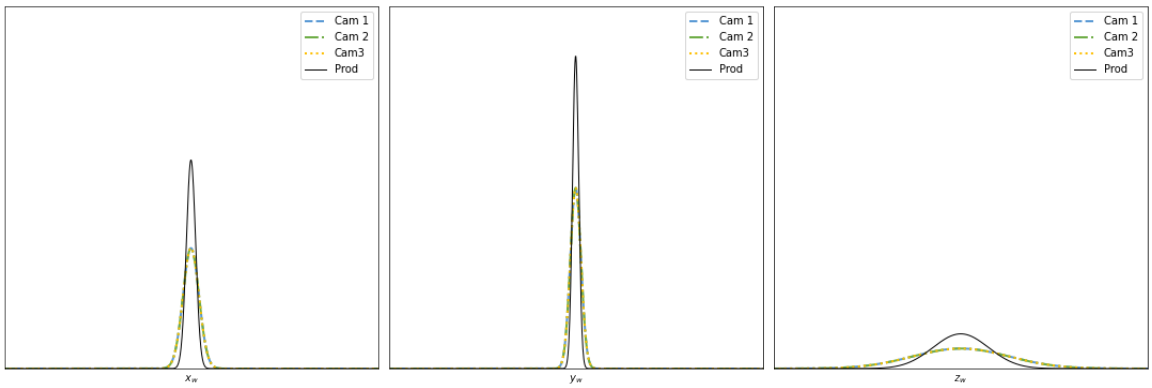
(a) Top view of the setup



(b) Chip view (camera space) for cameras 1, 2 and 3



(c) Density of probability for $x_{c_{(i)}}$, $y_{c_{(i)}}$ and $z_{c_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

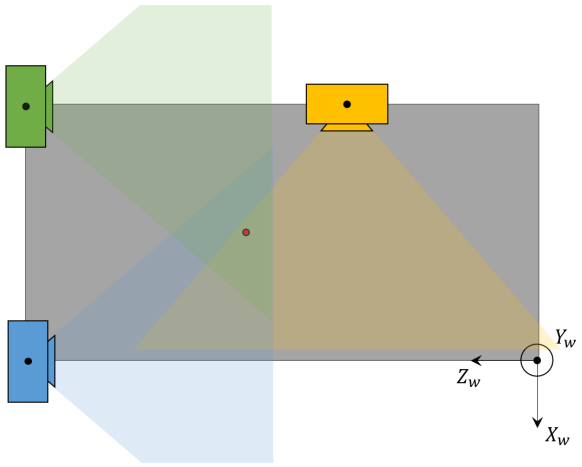


(d) Density of probability for $x_{w_{(i)}}$, $y_{w_{(i)}}$ and $z_{w_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

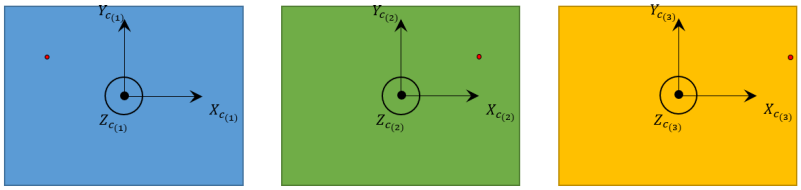
Figure 5: Scenario 2

3.2.3 Scenario 3

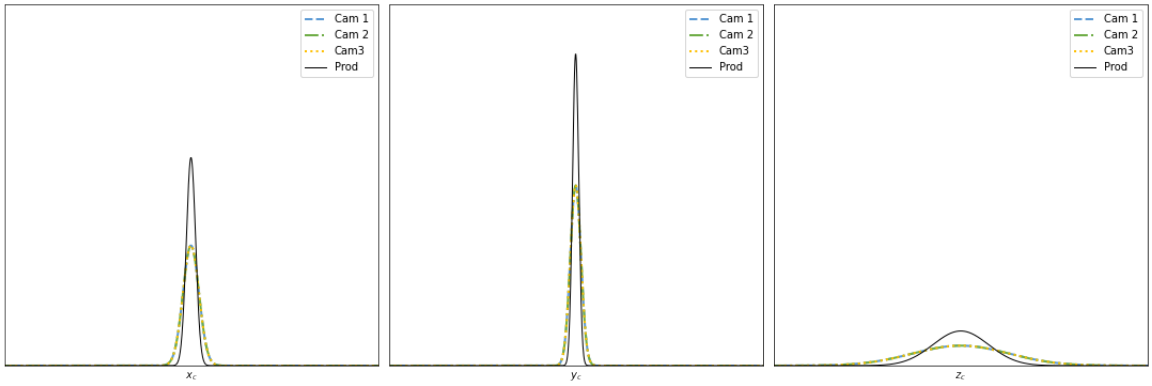
In this scenario, cameras 1 and 2 are in the same location they were in scenario 2. Moreover, $y_{w_1} = y_{w_2} = y_{w_3}$). However, $x_{w_3} \neq x_{w_{1,2}}$ and $z_{w_3} \neq z_{w_{1,2}}$. Camera 3 has been rotated $-90°$ around the Y_w axis. The uncertainties don't change for cameras 1 and 2, compared to scenarios 1 and 2. However, that's not the case for camera 3 whose PDF now exhibits a lower/wider shape along x_w and a higher/narrower shape along z_w.



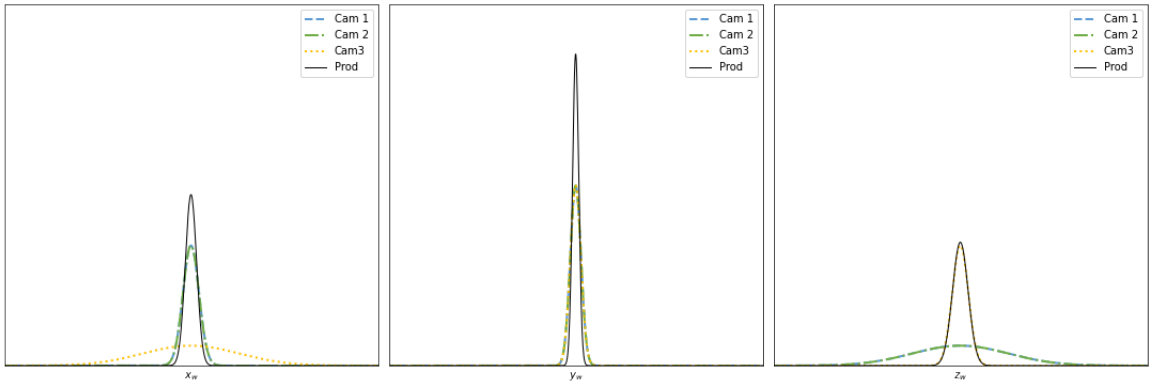
(a) Top view of the setup



(b) Chip view (camera space) for cameras 1, 2 and 3



(c) Density of probability for $x_{c_{(i)}}$, $y_{c_{(i)}}$ and $z_{c_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

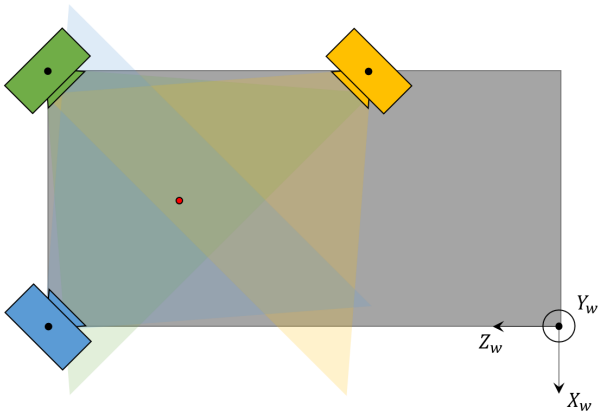


(d) Density of probability for $x_{w_{(i)}}$, $y_{w_{(i)}}$ and $z_{w_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

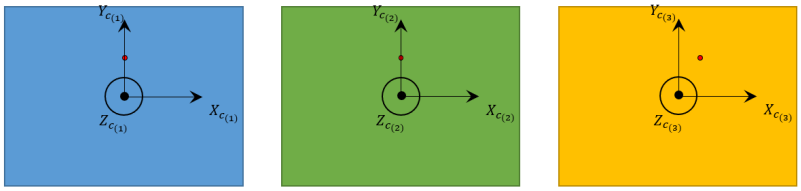
Figure 6: Scenario 3

3.2.4 Scenario 4

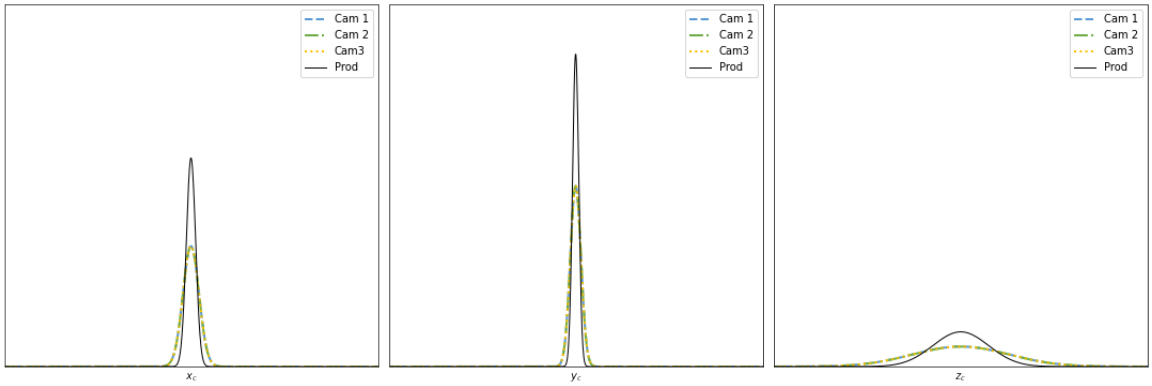
In this scenario, all the cameras are located at the same height (y_w). Cameras 1, 2 and 3 have respectively been rotated around the Y_w axis so $\beta_{(1)} = 45°$, $\beta_{(2)} = -45°$ and $\beta_{(3)} = -135°$. The cameras are not rotated around X_w nor Z_w: $\alpha_{(1)} = \alpha_{(2)} = \alpha_{(3)} = 0°$ and $\gamma_{(1)} = \gamma_{(2)} = \gamma_{(3)} = 0°$. The uncertainties change for all the cameras along axes x and z.



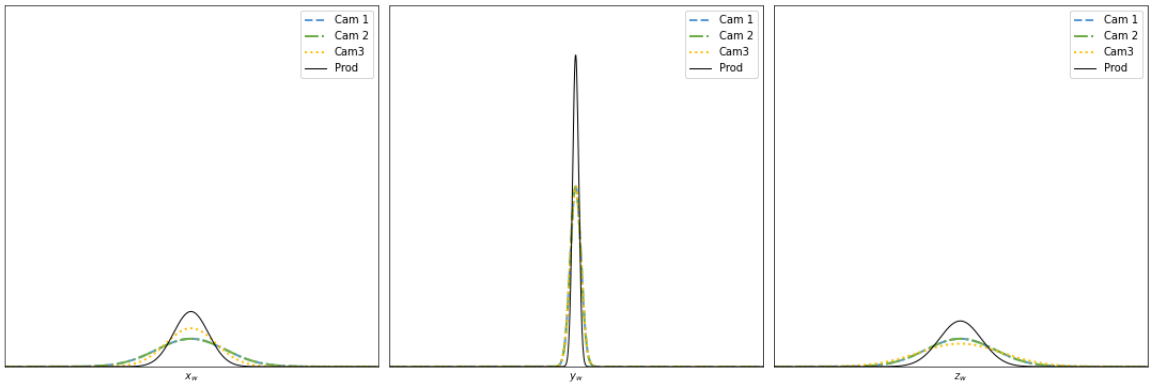
(a) Top view of the setup



(b) Chip view (camera space) for cameras 1, 2 and 3



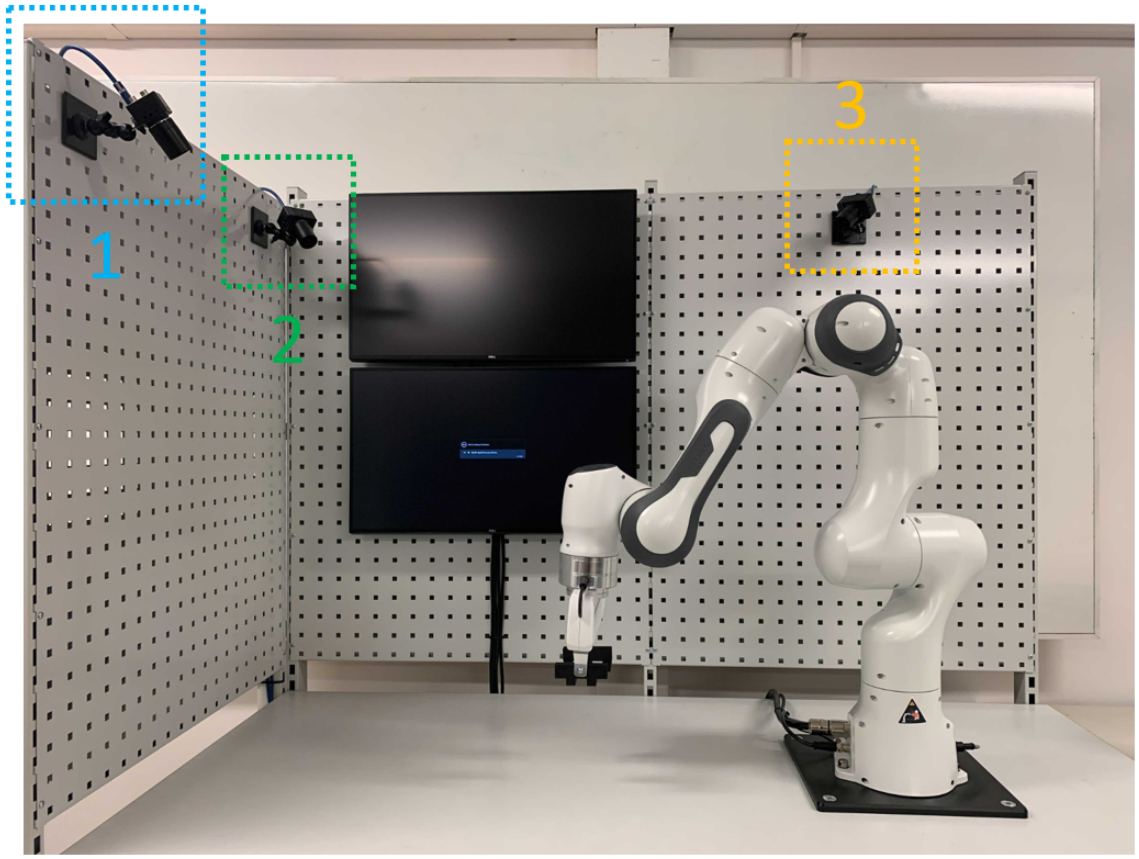
(c) Density of probability for $x_{c_{(i)}}$, $y_{c_{(i)}}$ and $z_{c_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).



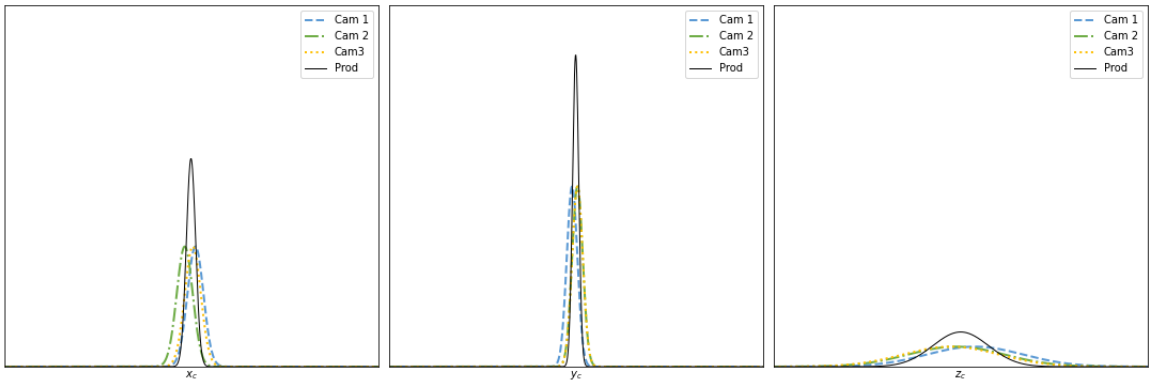
(d) Density of probability for $x_{w_{(i)}}$, $y_{w_{(i)}}$ and $z_{w_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

Figure 7: Scenario 4

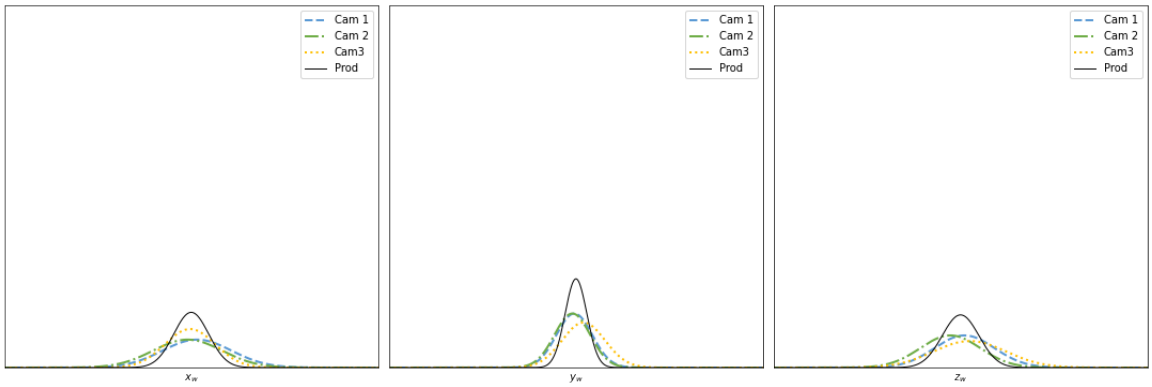
3.3 Real Setup



(a) Front view of the setup



(b) Density of probability for $x_{c_{(i)}}$, $y_{c_{(i)}}$ and $z_{c_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).



(c) Density of probability for $x_{w_{(i)}}$, $y_{w_{(i)}}$ and $z_{w_{(i)}}$ corresponding to cameras 1, 2 and 3 ($i \in \{1, 2, 3\}$).

Figure 8: Real Setup

4 Summary

The following equations summarize the pose merging workflow. These equations are implemented by function *get_world_gaussian* from coremerge.py in the repository PoseMerge where the jupyter notebook MergeTester.ipynb shows the scenarios presented in section 3.2.

4.1 Camera Space

$$x_{c_{(i)}} \sim \mathcal{N}[\mu_{x_{c_{(i)}}}, \sigma^2_{x_{c_{(i)}}}] \quad ; \quad y_{c_{(i)}} \sim \mathcal{N}[\mu_{y_{c_{(i)}}}, \sigma^2_{y_{c_{(i)}}}] \quad ; \quad z_{c_{(i)}} \sim \mathcal{N}[\mu_{z_{c_{(i)}}}, \sigma^2_{z_{c_{(i)}}}]$$

4.2 Camera To World

$$[x_{w_{(i)}}, y_{w_{(i)}}, z_{w_{(i)}}, 1] = M_{c2w_{(i)}} \cdot [x_{c_{(i)}}, y_{c_{(i)}}, z_{c_{(i)}}, 1]$$

where:

$$M_{c2w_{(i)}} = R_{x_{(i)}} \cdot R_{y_{(i)}} \cdot R_{z_{(i)}} \cdot T_{xyz_{(i)}} = \begin{bmatrix} m_{11_{(i)}} & m_{12_{(i)}} & m_{13_{(i)}} & m_{14_{(i)}} \\ m_{21_{(i)}} & m_{22_{(i)}} & m_{23_{(i)}} & m_{24_{(i)}} \\ m_{31_{(i)}} & m_{32_{(i)}} & m_{33_{(i)}} & m_{34_{(i)}} \\ m_{41_{(i)}} & m_{42_{(i)}} & m_{43_{(i)}} & m_{44_{(i)}} \end{bmatrix}$$

so:

$$x_{w_{(i)}} = m_{11_{(i)}} \cdot x_{c_{(i)}} + m_{12_{(i)}} \cdot y_{c_{(i)}} + m_{13_{(i)}} \cdot z_{c_{(i)}} + m_{14_{(i)}}$$

$$y_{w_{(i)}} = m_{21_{(i)}} \cdot x_{c_{(i)}} + m_{22_{(i)}} \cdot y_{c_{(i)}} + m_{23_{(i)}} \cdot z_{c_{(i)}} + m_{24_{(i)}}$$

$$z_{w_{(i)}} = m_{31_{(i)}} \cdot x_{c_{(i)}} + m_{32_{(i)}} \cdot y_{c_{(i)}} + m_{33_{(i)}} \cdot z_{c_{(i)}} + m_{34_{(i)}}$$

4.3 World Space

$$x_{w_{(i)}} \sim \mathcal{N}[\mu_{x_{w_{(i)}}}, \sigma^2_{x_{w_{(i)}}}] \quad ; \quad y_{w_{(i)}} \sim \mathcal{N}[\mu_{y_{w_{(i)}}}, \sigma^2_{y_{w_{(i)}}}] \quad ; \quad z_{w_{(i)}} \sim \mathcal{N}[\mu_{z_{w_{(i)}}}, \sigma^2_{z_{w_{(i)}}}]$$

where:

$$\mu_{x_{w_{(i)}}} = m_{11_{(i)}} \cdot \mu_{x_{c_{(i)}}} + m_{12_{(i)}} \cdot \mu_{y_{c_{(i)}}} + m_{13_{(i)}} \cdot \mu_{z_{c_{(i)}}} + m_{14_{(i)}}$$

$$\sigma^2_{x_{w_{(i)}}} = (m_{11_{(i)}} \cdot \sigma_{x_{c_{(i)}}})^2 + (m_{12_{(i)}} \cdot \sigma_{y_{c_{(i)}}})^2 + (m_{13_{(i)}} \cdot \sigma_{z_{c_{(i)}}})^2$$

$$\mu_{y_{w_{(i)}}} = m_{21_{(i)}} \cdot \mu_{x_{c_{(i)}}} + m_{22_{(i)}} \cdot \mu_{y_{c_{(i)}}} + m_{23_{(i)}} \cdot \mu_{z_{c_{(i)}}} + m_{24_{(i)}}$$

$$\sigma^2_{y_{w_{(i)}}} = (m_{21_{(i)}} \cdot \sigma_{x_{c_{(i)}}})^2 + (m_{22_{(i)}} \cdot \sigma_{y_{c_{(i)}}})^2 + (m_{23_{(i)}} \cdot \sigma_{z_{c_{(i)}}})^2$$

$$\mu_{z_{w_{(i)}}} = m_{31_{(i)}} \cdot \mu_{x_{c_{(i)}}} + m_{32_{(i)}} \cdot \mu_{y_{c_{(i)}}} + m_{33_{(i)}} \cdot \mu_{z_{c_{(i)}}} + m_{34_{(i)}}$$

$$\sigma^2_{z_{w_{(i)}}} = (m_{31_{(i)}} \cdot \sigma_{x_{c_{(i)}}})^2 + (m_{32_{(i)}} \cdot \sigma_{y_{c_{(i)}}})^2 + (m_{33_{(i)}} \cdot \sigma_{z_{c_{(i)}}})^2$$

4.4 Data Consolidation

$$\mu_{x_w} = \frac{\sigma_{x_{w_{(2)}}} \cdot \sigma_{x_{w_{(3)}}} \cdot \mu_{x_{w_{(1)}}} + \sigma_{x_{w_{(3)}}} \cdot \sigma_{x_{w_{(1)}}} \cdot \mu_{x_{w_{(2)}}} + \sigma_{x_{w_{(1)}}} \cdot \sigma_{x_{w_{(2)}}} \cdot \mu_{x_{w_{(3)}}}}{\sigma_{x_{w_{(1)}}} \cdot \sigma_{x_{w_{(2)}}} + \sigma_{x_{w_{(2)}}} \cdot \sigma_{x_{w_{(3)}}} + \sigma_{x_{w_{(3)}}} \cdot \sigma_{x_{w_{(1)}}}}$$

$$\mu_{y_w} = \frac{\sigma_{y_{w_{(2)}}} \cdot \sigma_{y_{w_{(3)}}} \cdot \mu_{y_{w_{(1)}}} + \sigma_{y_{w_{(3)}}} \cdot \sigma_{y_{w_{(1)}}} \cdot \mu_{y_{w_{(2)}}} + \sigma_{y_{w_{(1)}}} \cdot \sigma_{y_{w_{(2)}}} \cdot \mu_{y_{w_{(3)}}}}{\sigma_{y_{w_{(1)}}} \cdot \sigma_{y_{w_{(2)}}} + \sigma_{y_{w_{(2)}}} \cdot \sigma_{y_{w_{(3)}}} + \sigma_{y_{w_{(3)}}} \cdot \sigma_{y_{w_{(1)}}}}$$

$$\mu_{z_w} = \frac{\sigma_{z_{w_{(2)}}} \cdot \sigma_{z_{w_{(3)}}} \cdot \mu_{z_{w_{(1)}}} + \sigma_{z_{w_{(3)}}} \cdot \sigma_{z_{w_{(1)}}} \cdot \mu_{z_{w_{(2)}}} + \sigma_{z_{w_{(1)}}} \cdot \sigma_{z_{w_{(2)}}} \cdot \mu_{z_{w_{(3)}}}}{\sigma_{z_{w_{(1)}}} \cdot \sigma_{z_{w_{(2)}}} + \sigma_{z_{w_{(2)}}} \cdot \sigma_{z_{w_{(3)}}} + \sigma_{z_{w_{(3)}}} \cdot \sigma_{z_{w_{(1)}}}}$$