# MODELING THE IMPACT OF KEYPOINT DETECTION ERRORS ON LOCAL DESCRIPTOR SIMILARITY

*André Araujo, Haricharan Lakshman, Roland Angst, Bernd Girod*

Department of Electrical Engineering, Stanford University, CA

## ABSTRACT

This paper presents a mathematical analysis of the impact of keypoint detection errors on the similarity of local image descriptors that are based on histogram of gradients. First, we derive a closed-form expression for the $L_p$ distance between two descriptors, for general translation, scale and orientation detection errors. Second, we introduce a detailed analysis for the special case where translation errors dominate, using the $L_2$ distance. We show that the individual components which form the squared $L_2$ distance can be approximated using Gamma distributions whose parameters are computed in closed-form by our model. We obtain approximate closed-form expressions for the expected squared $L_2$ distances when translation errors are fixed or uniformly distributed. Finally, these models are validated using image patches extracted from two standard image retrieval datasets, by comparing the predicted distributions to the ground-truth.

***Index Terms*—** local descriptors, keypoint detection, histogram of gradients

## 1. INTRODUCTION AND RELATED WORK

Gradient-based features have found broad applications in image processing and computer vision, such as motion tracking [1–3], image-based retrieval [4–7], action recognition [8], video copy detection [9], object detection [10, 11], image classification [12–14], among others. The most widely used algorithm is the Scale-Invariant Feature Transform (SIFT) [15]. With SIFT and its many proposed variations (e.g. [16–20]), keypoints are first detected in the image. Then, a descriptor is computed to encode information around the keypoint. Since each keypoint is associated with a predominant local scale and orientation, the descriptor can be computed in a canonical coordinate system to achieve invariance against scale changes and rotation. The descriptor is calculated by measuring gradient orientation histograms in different spatial bins, which are placed around the keypoint location.

The keypoint detection stage is often sensitive to imaging parameters such as changes in viewpoint or illumination. Studies on the impact of keypoints errors on descriptor similarity have been mainly empirical [19, 21]. In a comparison of several local descriptors, Mikolajczyk and Schmid [19] reported that SIFT and its variants are the top-performers when the overlap region between corresponding keypoints is small. In this work, we are interested in analytically modeling how a local image descriptor based on histogram of gradients is affected due to keypoint detection uncertainty. We believe that such models can find many applications within image processing and computer vision. For example, in image-based retrieval, they might be used to evaluate the robustness to keypoint detection errors of a particular descriptor design. For motion tracking, they could be used to understand how accurately a tracker needs to be for corresponding regions in consecutive frames to be similar enough in descriptor space. For image classification applications, where local descriptors are usually extracted over a dense grid, our models could help find the optimal grid spacing. While in this work we model descriptor distances due to keypoint uncertainty, recent works have focused on modeling image-level matching performance [22] or retrieval performance of global image descriptors [23].

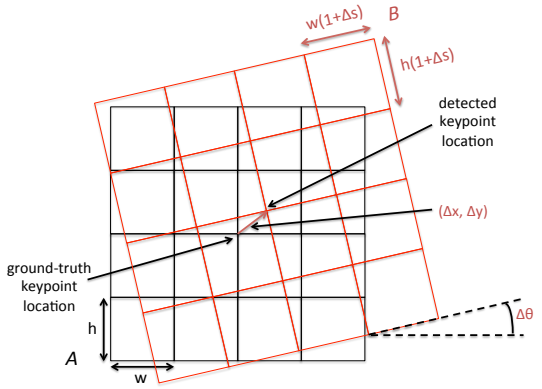


**Fig. 1**: Example of keypoint detection with errors. The detected keypoint depicts errors in location ($\Delta x$, $\Delta y$), scale ($\Delta s$) and orientation ($\Delta\theta$). The ground-truth patch $A$ and the detected patch $B$ are depicted using $4 \times 4$ spatial bins, as in SIFT [15]. In this paper, we are interested in modeling how the descriptor extracted from $B$ differs from that of $A$, as a consequence of the imprecise keypoint detection.

**Contributions.** To the best of our knowledge, this is the first work that models how local descriptor similarity is affected by keypoint detection uncertainty. Our main results are the following: (1) a closed-form expression that describes the $L_p$ descriptor distance for general errors in translation, scaling and orientation; (2) in the case where only translation errors are significant, we show that the individual components of the squared $L_2$ distance can be modeled using a Gamma distribution whose parameters are computed in closed-form by our model; (3) again in the case where only translation errors are significant, we obtain approximate closed-form expressions for the expected value of the squared $L_2$ distance. We validate the accuracy of our models using experimental data.

## 2. PROBLEM FORMULATION

Referring to Fig. 1, consider a ground truth patch $A$ and define $A_n$ as its $n$-th spatial bin. Consider a local image descriptor with $N$ spatial bins and $D$ gradient orientation bins. For example, for SIFT [15], $N = 16$ and $D = 8$. Denote the normalized histogram of gradient orientations of $A_n$ as $\mathbf{a}_n$, such that $\mathbf{a}_n = [a_n[1], a_n[2], \cdots, a_n[D]]$, and $a_n[d]$ is the proportion of area in $A_n$ where gradient orientations

are quantized to orientation $d$:

$$a_n[d] = \frac{|A_n^d|}{|A_n|} \tag{1}$$

where $|A_n|$ denotes the area of $A_n$, and $A_n^d$ is the region within $A_n$ where gradient orientations fall into bin $d$ ($|A_n^d| \le |A_n|$).

Define $f_A$ as the local feature descriptor for $A$, generated by concatenating the histograms for each spatial bin, i.e., $f_A = [\mathbf{a}_1; \mathbf{a}_2; \cdots; \mathbf{a}_N]$. Similar notation is used for patch $B$, which is extracted around a keypoint detected with errors. Errors in location ($\Delta x$ and $\Delta y$), orientation ($\Delta\theta$) and scale ($\Delta s$) are defined as:

$$\Delta x = x_B - x_A \tag{2a}$$
$$\Delta y = y_B - y_A \tag{2b}$$
$$\Delta\theta = \theta_B - \theta_A \tag{2c}$$
$$\Delta s = \frac{s_B}{s_A} - 1 \tag{2d}$$

where $x_A, y_A, \theta_A, s_A$ correspond to the 2D locations, orientation and scale for patch $A$, and similarly $x_B, y_B, \theta_B, s_B$ for patch $B$.

Consider the problem of comparing the descriptors $f_A$ and $f_B$. We use a distortion measure based on a $L_p$-norm, such as $\|f_A - f_B\|_p^p = \sum_{n=1}^{N}\sum_{d=1}^{D} |a_n[d] - b_n[d]|^p$. Our objective is to characterize how the errors $\Delta x$, $\Delta y$, $\Delta\theta$, and $\Delta s$ give rise to a distortion $\|f_A - f_B\|_p^p$. Our goal is to capture the most important effects for descriptors that are based on histogram of gradients – so we do not take into account some optimizations that are used in practice, such as Gaussian weighting, gradient magnitude weighting, $L_2$ normalization, etc.

## 3. MODELING DESCRIPTOR DISTANCE

Let us denote the overlap region of the spatial bins $A_n$ and $B_n$ as $O_n$, and the non-overlap regions as $A'_n$ and $B'_n$. A normalized histogram $\mathbf{a}_n$ can be decomposed into a contribution from $O_n$ and a contribution from $A'_n$, denoted as $\mathbf{o}_n$ and $\mathbf{a}'_n$, respectively. For each component of $\mathbf{a}_n$ and $\mathbf{b}_n$:

$$a_n[d] = \alpha_n o_n^A[d] + (1-\alpha_n) a'_n[d] \tag{3a}$$
$$b_n[d] = \beta_n o_n^B[d] + (1-\beta_n) b'_n[d] \tag{3b}$$

where $\mathbf{o}_n^A$ and $\mathbf{o}_n^B$ are histograms calculated from $O_n$ with respect to the orientations from patches $A$ and $B$, respectively, and $\alpha_n$ and $\beta_n$ are the proportions of overlap areas for $A_n$ and $B_n$:

$$\alpha_n = \frac{|O_n|}{|A_n|} = \frac{|O_n|}{|O_n| + |A'_n|} \tag{4a}$$
$$\beta_n = \frac{|O_n|}{|B_n|} = \frac{|O_n|}{|O_n| + |B'_n|} \tag{4b}$$

As in Fig. 1, $|A_n| = wh$, and $|B_n| = wh(1+\Delta s)^2$. Therefore, $\alpha_n = (1+\Delta s)^2 \beta_n$.

We would like to model the component-wise difference $(a_n[d] - b_n[d])$ in terms of the detection errors. Hence, we compute the difference of (3a) and (3b) and substitute $\alpha_n$ in terms of $\beta_n$ to obtain:

$$\begin{aligned} a_n[d] - b_n[d] &= \alpha_n o_n^A[d] - \beta_n o_n^B[d] \\ &\quad + (1-\alpha_n) a'_n[d] - (1-\beta_n) b'_n[d] \qquad (5a) \\ &= (1-\beta_n)(a'_n[d] - b'_n[d]) \\ &\quad + \beta_n (o_n^A[d] - o_n^B[d]) \\ &\quad + \beta_n (2\Delta s + \Delta s^2)(o_n^A[d] - a'_n[d]) \qquad (5b) \end{aligned}$$

With this expression, we obtain:

$$\begin{aligned} \|f_A - f_B\|_p^p = \sum_{n=1}^{N}\sum_{d=1}^{D} & |(1-\beta_n)(a'_n[d] - b'_n[d]) \\ & + \beta_n (o_n^A[d] - o_n^B[d]) \\ & + \beta_n (2\Delta s + \Delta s^2)(o_n^A[d] - a'_n[d])|^p \end{aligned} \tag{6}$$

Thus, the $L_p$ distance of descriptors $f_A$ and $f_B$ can be expressed as a function of three terms: (a) Difference of histograms of non-overlap regions $(a'_n[d] - b'_n[d])$, magnified by the proportion of non-overlap area $(1-\beta_n)$, (b) Difference of histograms in $O_n$, $(o_n^A[d] - o_n^B[d])$, magnified by the proportion of overlap area $\beta_n$, and (c) Difference of histograms within $A_n$, $(o_n^A[d] - a'_n[d])$, magnified by the proportion of overlap area and the error in scale, $\beta_n(2\Delta s + \Delta s^2)$.

**Translation error only.** In the following, we consider in more detail the case where translation errors dominate. With $\Delta s \approx 0$, $\alpha \approx \beta$, $\Delta\theta \approx 0$, $o_n^A \approx o_n^B$, the component-wise difference (5b) can be simplified as:

$$z_{d,n} = a_n[d] - b_n[d] = (1-\beta_n)(a'_n[d] - b'_n[d]) \tag{7}$$

Using this, the descriptor distortion can be expressed as:

$$\|f_A - f_B\|_p^p = \sum_{n=1}^{N}\sum_{d=1}^{D} |z_{d,n}|^p = \sum_{n=1}^{N}\sum_{d=1}^{D} |(1-\beta_n)(a'_n[d] - b'_n[d])|^p \tag{8}$$

Thus, in the case of pure translation error, descriptor distances depend mainly on the distance between histograms of non-overlapping regions. For a translation error $\mathbf{\Delta v} = (\Delta x, \Delta y)$, we can write $\beta_n$ as:

$$\beta_n = \begin{cases} \frac{wh - (|\Delta x|h + |\Delta y|w - |\Delta x||\Delta y|)}{wh}, & \text{for } -w \le \Delta x \le w, -h \le \Delta y \le h \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

This can be simplified as $\beta_n = wh \cdot \mathrm{Pyr}(\mathbf{\Delta v})$, where the pyramid function $\mathrm{Pyr}(.)$ is defined as:

$$\mathrm{Pyr}(x,y) = \begin{cases} \frac{(w-|x|)(h-|y|)}{w^2h^2}, & \text{for } -w \le x \le w, -h \le y \le h \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Thus, the component-wise difference (7) can be expressed as:

$$z_{d,n} = (1 - wh \cdot \mathrm{Pyr}(\mathbf{\Delta v}))(a'_n[d] - b'_n[d]) \tag{11}$$

In the following, we consider in more detail $L_2$ distances, for two cases: (i) fixed translation errors, and (ii) uniformly-distributed translation errors. We start by focusing on $z_{d,n}^2$, the component-wise squared distance, and characterize its distribution. We obtain closed-form expressions for the expected value of $\|f_A - f_B\|_2^2$ in both cases.

### 3.1. Fixed translation error

The expected value of squared component-wise differences for a given translation error $\Delta\mathbf{v}$ can be written as:

$$E[z_{d,n}^2 | \mathbf{\Delta v}] = (1 - wh \cdot \mathrm{Pyr}(\mathbf{\Delta v}))^2 E[(a'_n[d] - b'_n[d])^2] \tag{12}$$

We develop $E[(a'_n[d] - b'_n[d])^2]$ further:

$$E[(a'_n[d] - b'_n[d])^2] = E[a'_n[d]^2] - 2E[a'_n[d] b'_n[d]] + E[b'_n[d]^2] \tag{13}$$

Now, we make two assumptions: (i) $a'_n[d]$ and $b'_n[d]$ are identically distributed, (ii) $a'_n[d]$ and $b'_n[d]$ are uncorrelated. This assumption is reasonable since the non-overlapping regions are separated and lie on opposite sides of spatial bin $n$. Thus, (13) can be simplified as:

$$\begin{aligned} & E[(a'_n[d] - b'_n[d])^2] \\ & = E[a'_n[d]^2] - 2E[a'_n[d]]E[a'_n[d]] + E[a'_n[d]^2] \\ & = 2 \cdot \mathrm{Var}[a'_n[d]]. \end{aligned} \tag{14}$$

To simplify the rest of the analysis, we consider a discretized grid of pixels within a patch. Similar to (1), we can expand $a'_n[d]$ as:

$$a'_n[d] = \frac{|\mathcal{A}'^d_n|}{|\mathcal{A}'_n|} = \frac{1}{|\mathcal{A}'_n|} \sum_{\substack{x,y \\ [x,y] \in \mathcal{A}'_n}} g_d[x,y] \tag{15}$$

where $\mathcal{A}'_n$ is the set of pixels in the non-overlapping region of spatial bin $A_n$ and $|\mathcal{A}'_n|$ is its cardinality. We introduce the binary mask $g_d[x,y]$, which is 1 if the gradient at $x,y$ has orientation that falls into bin $d$, and 0 otherwise. We can expand $|\mathcal{A}'_n|$ as:

$$\begin{aligned} |\mathcal{A}'_n| &= wh(1-\beta_n) \\ &= wh(1 - wh \cdot \mathrm{Pyr}(\boldsymbol{\Delta v})) \end{aligned} \tag{16}$$

Denote $\mathrm{Prob}(g_d[x,y]=1) = p_d$ and assume that $g_d[x,y]$ is stationary. In this case, the variance of $a'_n[d]$ can be computed using (15):

$$\begin{aligned} \mathrm{Var}[a'_n[d]] &= \frac{1}{|\mathcal{A}'_n|^2} \sum_{\substack{x,y \\ [x,y] \in \mathcal{A}'_n}} \mathrm{Var}[g_d[x,y]] \\ &+ \frac{1}{|\mathcal{A}'_n|^2} \sum_{\substack{x,y \\ [x,y] \in \mathcal{A}'_n}} \sum_{\substack{u,v \\ [u,v] \in \mathcal{A}'_n \\ [u,v] \neq [x,y]}} \mathrm{Cov}[g_d[x,y], g_d[u,v]] \\ &= \frac{1}{|\mathcal{A}'_n|} p_d(1-p_d) \\ &+ \frac{1}{|\mathcal{A}'_n|^2} \sum_{\substack{x,y \\ [x,y] \in \mathcal{A}'_n}} \sum_{\substack{u,v \\ [u,v] \in \mathcal{A}'_n \\ [u,v] \neq [x,y]}} \mathrm{Cov}[g_d[x,y], g_d[u,v]] \\ &= \frac{1}{|\mathcal{A}'_n|} p_d(1-p_d) + \frac{1}{|\mathcal{A}'_n|^2} \sum_{\substack{i,j \\ [i,j] \neq \mathbf{0}}} \mathrm{N}_{i,j}(\boldsymbol{\Delta v}) \sigma^{(d)}_{i,j} \end{aligned} \tag{17}$$

where $\sigma^{(d)}_{i,j} = \mathrm{Cov}[g_d[t_1,t_2], g_d[t_1-i, t_2-j]]$ and $\mathrm{N}_{i,j}(\boldsymbol{\Delta v})$ is the number of pairs of pixels in the non-overlapping region whose locations are separated by $[i,j]$. Note that $\sigma^{(d)}_{i,j}$ is independent of $t_1$ and $t_2$ since we assume stationarity. With this, the expected value of the component-wise difference conditioned on the translation error $\boldsymbol{\Delta v}$ can be computed by combining (16), (17), (14) and (12):

$$\begin{aligned} E[z^2_{d,n}|\boldsymbol{\Delta v}] &= \frac{2}{wh} p_d(1-p_d)(1 - wh \cdot \mathrm{Pyr}(\boldsymbol{\Delta v})) \\ &+ \frac{2}{w^2h^2} \sum_{\substack{i,j \\ [i,j] \neq \mathbf{0}}} \mathrm{N}_{i,j}(\boldsymbol{\Delta v}) \sigma^{(d)}_{i,j}. \end{aligned} \tag{18}$$

The values of $\sigma^{(d)}_{i,j}$ can be estimated from training data. We can use an approximation for $\mathrm{N}_{i,j}(\boldsymbol{\Delta v})$, which is discussed in detail in the supplemental material[1] (Appendix A). From (8), we can derive:

$$E[\|f_A - f_B\|^2_2 | \boldsymbol{\Delta v}] = \sum_{n=1}^{N} \sum_{d=1}^{D} E[z^2_{d,n}|\boldsymbol{\Delta v}] \tag{19}$$

Using (18), we can then substitute for $E[z^2_{d,n}|\boldsymbol{\Delta v}]$ in (19) to obtain a closed-form expression for the conditional expectation of the squared $L_2$ distance between local descriptors under a given translation error.

We can further characterize the distribution of $z^2_{d,n}$ given $\boldsymbol{\Delta v}$. Note that $a_n[d]$ is a normalized sum of many identically distributed variables, similar to (15). From the Central Limit Theorem for correlated random variables, $a_n[d]$ tends to a Gaussian distribution. Since we assume that $a_n[d]$ and $b_n[d]$ are identically distributed, $z_{d,n}$ is approximately zero-mean Gaussian, from (7). From the properties of Gaussian distributions, if follows that $z^2_{d,n}$ is approximately distributed as a Gamma distribution with shape parameter $\frac{1}{2}$ and scale parameter $2 \times E[z^2_{d,n}|\boldsymbol{\Delta v}]$. Thus, (18) also enables us to obtain an approximate distribution for $z^2_{d,n}$ in closed-form.

[1] Supplemental material is available on the authors' websites.

### 3.2. Uniformly-distributed translation error

Given a distribution of translation errors and using (18), we obtain:

$$\begin{aligned} E[z^2_{d,n}] &= E_{\boldsymbol{\Delta v}}[E[z^2_{d,n}|\boldsymbol{\Delta v}]] \\ &= \frac{2}{wh} p_d(1-p_d)(1 - wh \cdot E_{\boldsymbol{\Delta v}}[\mathrm{Pyr}(\boldsymbol{\Delta v})]) \\ &+ \frac{2}{w^2h^2} \sum_{\substack{i,j \\ [i,j] \neq \mathbf{0}}} E_{\boldsymbol{\Delta v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta v})] \sigma^{(d)}_{i,j} \end{aligned} \tag{20}$$

When $\boldsymbol{\Delta v}$ follows a uniform distribution, the term $E_{\boldsymbol{\Delta v}}[\mathrm{Pyr}(\boldsymbol{\Delta v})]$ can be computed in closed-form, and the term $E_{\boldsymbol{\Delta v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta v})]$ can be derived in closed-form using an approximation. These derivations are presented in detail in the supplemental material (Appendix A), due to space limitations. Once again, we can derive from (8):

$$E[\|f_A - f_B\|^2_2] = \sum_{n=1}^{N} \sum_{d=1}^{D} E[z^2_{d,n}] \tag{21}$$

Using (20), we can substitute for $E[z^2_{d,n}]$ in (21) to obtain a closed-form expression for the expectation of the squared $L_2$ distance between local descriptors under uniformly-distributed translation errors.

### 3.3. Model using IID assumption

We refer to the model developed in the previous subsections as M-S, where "S" refers to the stationarity assumption. We also consider a simpler model, which makes a stronger assumption, denoted as M-IID: $g_d[x,y]$ are assumed to be independent and identically distributed. M-IID can be seen as a variation of M-S: the expressions for $E[\|f_A - f_B\|^2_2|\boldsymbol{\Delta v}]$ and $E[\|f_A - f_B\|^2_2]$ in this case can be obtained by setting $\sigma^{(d)}_{i,j}$ to zero in (18) and (20). Our objective is to evaluate different variations of the model, with different model complexities, to find which one is most suitable.

### 3.4. Applicability of the model

The developed models might be applicable to any descriptor which uses histograms to summarize pixel-level statistics – in other words, the model is not restricted to histograms based on gradient orientations. For example, the expressions developed in this section would be applicable to histograms based on derivatives of any order, if these derivatives are roughly stationary. In this case, $p_d$ and $\sigma^{(d)}_{i,j}$ would be replaced by the equivalent probabilities and covariances, respectively. Note that some expressions (e.g., (6)) do not assume stationarity and are even more general.

## 4. EXPERIMENTS

We conduct experiments to validate the developed models, using two datasets: (i) Stanford Mobile Visual Search (SMVS) dataset [24]: 65,593 keypoints are extracted from database images using a Difference-of-Gaussians detector [15]; (ii) CNN2h dataset [25]: 78,452 keypoints are extracted from database video frames using a Temporally-Coherent Detector [26]. These datasets present very different visual contents: the SMVS dataset contains clean images of objects, while the CNN2h dataset contains video frames extracted from a CNN newscast. By using these datasets, with two different keypoint detectors, our goal is to validate our models in different cases. We divide each dataset randomly into two even splits, and use one exclusively for learning model parameters, and the other exclusively for comparing model predictions to data.

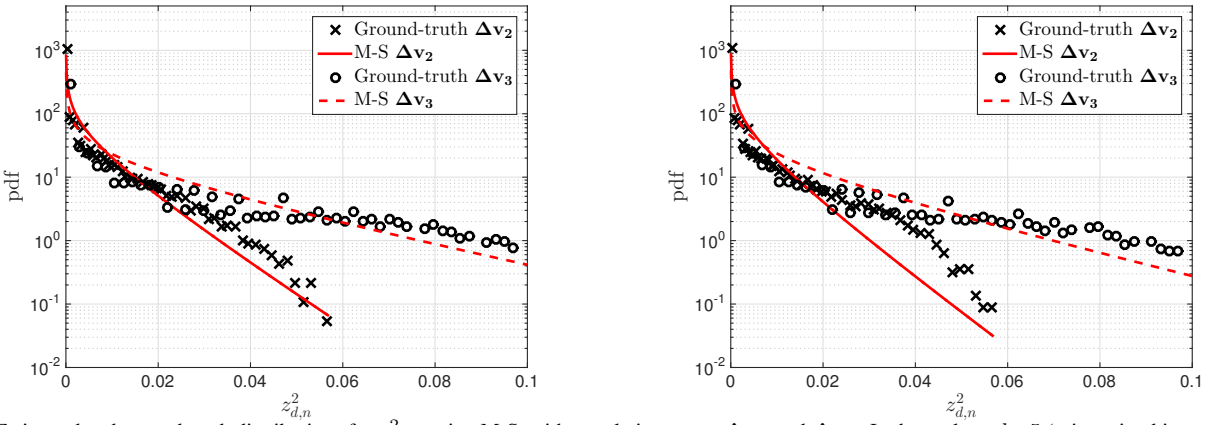


**Fig. 2**: Estimated and ground-truth distributions for $z^2_{d,n}$, using M-S, with translation errors $\Delta\mathbf{v}_2$ and $\Delta\mathbf{v}_3$. In these plots, $d = 5$ (orientation bin centered at 180°) and $n = 4$ (spatial bin on top-right corner). Left: SMVS dataset. Right: CNN2h dataset.

| | (a) Fixed errors | | | (b) Uniform distribution | | |
|---|---|---|---|---|---|---|
| | $\Delta\mathbf{v}_1$ | $\Delta\mathbf{v}_2$ | $\Delta\mathbf{v}_3$ | $U_1$ | $U_2$ | $U_3$ |
| **SMVS** | | | | | | |
| M-IID | 13.85 | 6.58 | 3.65 | 14.85 | 10.02 | 5.62 |
| M-S | **86.97** | **75.36** | **83.76** | **80.49** | **79.58** | **78.45** |
| **CNN2h** | | | | | | |
| M-IID | 14.21 | 6.68 | 3.73 | 14.48 | 9.78 | 5.48 |
| M-S | **80.62** | **68.30** | **77.33** | **80.09** | **78.94** | **77.59** |

**Table 1**: Model accuracy in percent, calculated using (22). The expected value of the squared $L_2$ descriptor distance $E[\|f_A - f_B\|_2^2]$ is computed from data and compared to the models. The results are summarized for two cases: (a) fixed keypoint errors, i.e., the expectation conditioned on $\Delta\mathbf{v}$, and (b) uniform distributions of keypoint errors. The M-S model provides a much better approximation of the empirically measured data.

For descriptor extraction and learning of model parameters, we extract a canonical patch around the detected keypoint, normalized in terms of scale and orientation, of size $64 \times 64$ pixels. We compute the descriptor using $4 \times 4$ spatial bins (i.e., $w = h = 16$) and 8 gradient orientation bins, as in SIFT. The covariances $\sigma^{(d)}_{i,j}$ are estimated for $0 \leq |i|, |j| \leq 10$ (and considered zero otherwise). To evaluate quantitatively the estimates given by our models, we first compute the relative error (RE), defined as the ratio of the absolute error to the true value. The results are then presented as accuracy, i.e.:

$$\mathrm{Acc}(y, \hat{y}) = 1 - \mathrm{RE}(y, \hat{y}) = 1 - \frac{|y - \hat{y}|}{y} \qquad (22)$$

where $y$ denotes the ground-truth value and $\hat{y}$ the estimate. The range of $\mathrm{Acc}(y, \hat{y})$ varies from $-\infty$ (poor estimate) to 1 (perfect estimate).

**Models for $z^2_{d,n}$ given $\Delta\mathbf{v}$.** We validate our models for fixed translation errors using three different error directions and magnitudes: $\Delta\mathbf{v}_1 = [1, 1]$, $\Delta\mathbf{v}_2 = [-1, 3]$ and $\Delta\mathbf{v}_3 = [4, -4]$. To estimate the ground-truth distribution for a given $\Delta\mathbf{v}$, we shift each patch by the specified $\Delta\mathbf{v}$ and compute the distribution of $z^2_{d,n}$ over the entire test set. Fig. 2 presents some ground-truth and estimated distributions of $z^2_{d,n}$, with translation errors $\Delta\mathbf{v}_2$ and $\Delta\mathbf{v}_3$, using M-S. It can be seen that the Gamma distribution fits the data well for M-S. Further results for the estimated distributions of $z^2_{d,n}$ are provided in supplemental material (Appendix B), due to space limitations. Tab. 1 part (a) summarizes the accuracy of our estimates. M-S estimates $E[\|f_A - f_B\|_2^2|\Delta\mathbf{v}]$ with mean accuracy of 78.72%. M-IID, on the other hand, generates poor estimates.

**Models for $z^2_{d,n}$ with uniform $\Delta\mathbf{v}$.** We validate our models for uniform translation errors, $-\frac{U}{2} \leq \Delta x, \Delta y \leq \frac{U}{2}$, using three cases: $U_1 = 2$, $U_2 = 4$, and $U_3 = 8$. To estimate the ground-truth expected values in this case, we use a Monte Carlo simulation, where each patch is shifted by a translation vector drawn from the uniform distribution, and the results are aggregated over the entire test set. Tab. 1 part (b) shows accuracy results. Again, M-IID estimates are poor, suggesting that the IID assumption is too severe in this case. M-S estimates $E[\|f_A - f_B\|_2^2]$ with much higher mean accuracy: 79.19%.

**Discussion.** While in both cases M-S explains most of the variation of ground-truth values, the required accuracy level depends on the application under consideration. In many applications, descriptor matching uses a ratio test, where a putative match is considered correct if the ratio of distances between the first and second nearest neighbors in the database is small enough. In this case, this ratio is more important than the actual distance values. This was exploited in [15], using approximate nearest neighbor (ANN) methods for substantial speedup. We perform an experimental evaluation to test if M-S is useful in such applications: we compute the accuracy (22) when estimating ratios of expected values of squared $L_2$ distances. First, for fixed translation errors, we compute the ratio between the estimated expected value given $\Delta\mathbf{v}_1$ and the estimated expected value given $\Delta\mathbf{v}_3$: compared to the ground-truth ratio of expected values, the accuracy is 97.24% for the SMVS dataset, and 97.32% for the CNN2h dataset. Second, for uniform translation errors, we compute the ratio between the estimated expected value using $U_1$ and the estimated expected value using $U_3$: compared to the ground-truth ratio of expected values, the accuracy is 98.32% for the SMVS dataset, and 97.96% for the CNN2h dataset. For a comparison, M-IID obtains negative accuracies for the estimates of these ratios (i.e., the estimates are poor, leading to large relative errors). We conclude that the M-S models are useful for such applications, as the ratio of estimated values is very similar to the ratio of ground-truth values.

## 5. CONCLUSION

Our mathematical analysis considers local image descriptor similarity as a function of keypoint detection errors. For descriptors that are based on gradient orientation histograms, we show that the $L_p$ distance between two descriptors can be expressed in closed-form for general translation, scale and orientation detection errors. In the case where translation errors dominate, we show that the expected value of the squared differences can be estimated in closed-form. Experimental results validate that the derived analytical models approximate the ground-truth well, for a model that assumes stationarity of the gradient orientation. On the other hand, with an IID assumption for the gradient orientation, the model accuracy is not satisfactory.

## 6. REFERENCES


[1] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod, "Rotation-Invariant Fast Features for Large-Scale Recognition and Real-Time Tracking," *Signal Processing: Image Communication*, vol. 28, no. 4, 2013.

[2] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato, "SIFT Features Tracking for Video Stabilization," in *Proc. International Conference on Image Analysis and Processing (ICIAP)*, 2007.

[3] I. Skrypnyk and D. Lowe, "Scene Modelling, Recognition and Tracking with Invariant Image Features," in *Proc. ISMAR*, 2004.

[4] L.-Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, "Overview of the MPEG-CDVS Standard," *IEEE Transactions on Image Processing*, vol. 25, no. 1, 2016.

[5] R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Locality in Generic Instance Search from One Example," in *Proc. CVPR*, 2014.

[6] G. Schindler, M. Brown, and R. Szeliski, "City-Scale Location Recognition," in *Proc. CVPR*, 2007.

[7] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, "Mobile Product Search with Bag of Hash Bits and Boundary Reranking," in *Proc. CVPR*, 2012.

[8] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition," *International Journal of Computer Vision*, vol. 103, no. 1, 2013.

[9] M. Douze, H. Jégou, and C. Schmid, "An Image-based Approach to Video Copy Detection with Spatio-Temporal Post-filtering," *IEEE Transactions on Multimedia*, vol. 12, no. 4, 2010.

[10] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. CVPR*, 2005.

[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010.

[12] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. CVPR*, 2006.

[13] S. Yan, X. Xu, D. Xu, S. Lin, and X. Li, "Beyond Spatial Pyramids: A New Feature Extraction Framework with Dense Spatial Sampling for Image Classification," in *Proc. ECCV*, 2012.

[14] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification," in *Proc. ECCV*, 2006.

[15] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[16] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, 2008.

[17] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, "Compressed Histogram of Gradients: A Low-Bitrate Descriptor," *International Journal of Computer Vision*, vol. 96, no. 3, 2012.

[18] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in *Proc. CVPR*, 2004.

[19] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, 2005.

[20] M. Ambai and Y. Yoshida, "CARD: Compact and Real-Time Descriptors," in *Proc. ICCV*, 2011.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A Comparison of Affine Region Detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, 2005.

[22] V. Chandrasekhar, *Low-Bitrate Image Retrieval with Compressed Histogram of Gradients Descriptors*, Ph.D. thesis, Stanford University, 2013.

[23] D. Chen and B. Girod, "A Hybrid Mobile Visual Search System with Compact Global Signatures," *IEEE Transactions on Multimedia*, vol. 17, no. 7, 2015.

[24] V. Chandrasekhar, D. Chen, S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod, "The Stanford Mobile Visual Search Data Set," in *Proc. ACM MMSys*, 2011.

[25] A. Araujo, M. Makar, V. Chandrasekhar, D. Chen, S. Tsai, H. Chen, R. Angst, and B. Girod, "Efficient Video Search Using Image Queries," in *Proc. ICIP*, 2014.

[26] M. Makar, V. Chandrasekhar, S. Tsai, D. Chen, and B. Girod, "Interframe Coding of Feature Descriptors for Mobile Augmented Reality," *IEEE Transactions on Image Processing*, vol. 23, no. 8, 2014.

# MODELING THE IMPACT OF KEYPOINT DETECTION ERRORS ON LOCAL DESCRIPTOR SIMILARITY

## *SUPPLEMENTAL MATERIAL*

*André Araujo, Haricharan Lakshman, Roland Angst, Bernd Girod*

Department of Electrical Engineering, Stanford University, CA

**ABSTRACT**

We provide further derivations and experimental results. First, we present detailed derivations to obtain approximate closed-form expressions for the expected squared $L_2$ distance when translation errors are known, or when they are uniformly distributed. Second, we provide more experimental results to validate the Gamma distribution model derived for component-wise squared $L_2$ distances.

## Appendix A: Supplemental derivations

In this section, we first present a useful approximation of $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$, which works well in practice. With this expression, we can write $E[\|f_A - f_B\|_2^2|\boldsymbol{\Delta}\mathbf{v}]$ in closed-form. Then, we present the derivation of the expected values of $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ and $\mathrm{Pyr}(\boldsymbol{\Delta}\mathbf{v})$, using a uniform distribution of translation errors. This allows us to obtain a closed-form expression for $E[\|f_A - f_B\|_2^2]$.

### Approximation of $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$

$\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ denotes the number of pixel pairs within the non-overlap region of a given spatial bin which are separated by an $[i, j]$ displacement, when the spatial bin is shifted by $\boldsymbol{\Delta}\mathbf{v}$ with respect to the ground-truth spatial bin. Note that we are not interested in $\mathrm{N}_{0,0}(\boldsymbol{\Delta}\mathbf{v})$, as the summations that use $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ explicitly discard the $[0, 0]$ displacement. This happens because the $[0, 0]$ displacement gives rise to the variance of the random variable under consideration, which can be calculated in closed-form in our model.

Fig. 1 gives three examples of pixel grids with some displacements, with $w = h = 4$. The blue grid corresponds to the grid in the correct position, while the red one corresponds to the grid with the incorrect keypoint detection. To give some examples of the values we want to compute, the case (a) has:

- $\mathrm{N}_{1,0}([1,-1]) = \mathrm{N}_{-1,0}([1,-1]) = 3$
- $\mathrm{N}_{0,1}([1,-1]) = \mathrm{N}_{0,-1}([1,-1]) = 3$
- $\mathrm{N}_{1,1}([1,-1]) = \mathrm{N}_{-1,-1}([1,-1]) = 1$
- $\mathrm{N}_{2,2}([1,-1]) = \mathrm{N}_{-2,-2}([1,-1]) = 1$
- $\mathrm{N}_{3,3}([1,-1]) = \mathrm{N}_{-3,-3}([1,-1]) = 1$

Clearly, we see that $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v}) = \mathrm{N}_{-i,-j}(\boldsymbol{\Delta}\mathbf{v})$, so we only need to compute $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ for half of all possible $[i, j]$.

It is difficult to express $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ exactly as a function of $i, j$ and $\boldsymbol{\Delta}\mathbf{v}$, so we use an approximation. We divide the non-overlap region into two regions, as in Fig. 2: (i) a "vertical" region (green), which is generated by horizontal shifts and (ii) a "horizontal" region (orange), which is generated by vertical shifts. We calculate $\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v})$ for each of these regions, and subtract the contribution from their intersection:

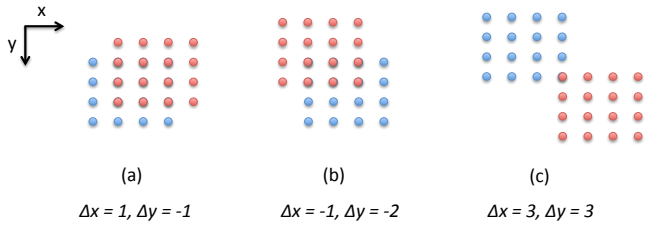


**Fig. 1**: Three different examples of pixel grids with $w = h = 4$. The samples in blue represent the grid in the correct position, while the red ones represent the grid used for the incorrect keypoint detection.

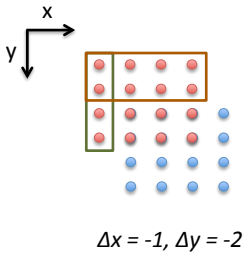


**Fig. 2**: Example of pixel grids with $w = h = 4$, showing the two regions into which the non-overlap area is divided: vertical (green) and horizontal (orange). The samples in blue represent the grid in the correct position, while the red ones represent the grid used for the incorrect keypoint detection.

$$\begin{aligned}\mathrm{N}_{i,j}(\boldsymbol{\Delta}\mathbf{v}) \approx\ & \max(0, (h - |j|))\max(0, |\Delta x| - |i|) \\ & + \max(0, (w - |i|))\max(0, |\Delta y| - |j|) \\ & - \max(0, |\Delta x| - |i|)\max(0, |\Delta y| - |j|) \end{aligned} \quad (1)$$

Considering Fig. 2, we can see that this expression is an approximation since we are not taking into account the pairs formed by, say, a pixel on the right of the orange region and a pixel on the bottom of the green region. More concretely, $\mathrm{N}_{-3,3}([-1,-2]) = 1$, but this approximation gives $\mathrm{N}_{-3,3}([-1,-2]) = 0$. However, this expression works well because the pixel pairs which are not taken into consideration are usually the ones which are the most distant, and the covariance between distant pixels is weaker. In this example, our approximation gives $\mathrm{N}_{0,1}([-1,-2]) = 6$, $\mathrm{N}_{1,1}([-1,-2]) = 3$, $\mathrm{N}_{2,0}([-1,-2]) = 4$ $\mathrm{N}_{2,1}([-1,-2]) = 2$, which are all correct.

We can then use (1) to compute the final closed-form expression for $E[\|f_A - f_B\|_2^2|\boldsymbol{\Delta}\mathbf{v}]$.

### Expected value of $\mathrm{N}_{i,j}(\boldsymbol{\Delta v})$

We consider a discrete uniform distribution of translation errors. In this case, the distribution is separable with probability mass function of $\frac{1}{U^2}$ at each point within $[-\frac{U}{2}, \frac{U}{2}-1] \times [-\frac{U}{2}, \frac{U}{2}-1]$, with $U > 0$ and multiple of 2.

Using the approximation (1), we obtain:

$$\begin{aligned} & E_{\boldsymbol{\Delta v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta v})] \\ & \approx \max(0,(h-|j|))E_{\Delta x}[\max(0,|\Delta x|-|i|)] \\ & + \max(0,(w-|i|))E_{\Delta y}[\max(0,|\Delta y|-|j|)] \\ & - E_{\Delta x}[\max(0,|\Delta x|-|i|)]E_{\Delta y}[\max(0,|\Delta y|-|j|)] \end{aligned} \tag{2}$$

The calculation of $E_{\boldsymbol{\Delta v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta v})]$ thus requires the computation of $E_{\Delta x}[\max(0,|\Delta x|-|i|)]$. Using the discrete uniform distribution:

$$\begin{aligned} & E_{\Delta x}[\max(0,|\Delta x|-|i|)] = \\ & = \frac{1}{U}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1} \max(0,|\Delta x|-|i|) \\ & = \frac{1}{U}\left[\max\left(0,\frac{U}{2}-|i|\right) + 2\times\sum_{\Delta x=0}^{\frac{U}{2}-1}\max(0,|\Delta x|-|i|)\right] \end{aligned} \tag{3}$$

In the case where $|i| \leq \frac{U}{2}-1$, we can derive:

$$\begin{aligned} \sum_{\Delta x=0}^{\frac{U}{2}-1}\max(0,|\Delta x|-|i|) &= \sum_{\Delta x=|i|}^{\frac{U}{2}-1}(|\Delta x|-|i|) \\ &= \frac{\left(\frac{U}{2}-|i|\right)\left(\frac{U}{2}-1-|i|\right)}{2} \end{aligned} \tag{4}$$

In the case where $|i| > \frac{U}{2}-1$, clearly $\sum_{\Delta x=0}^{\frac{U}{2}-1}\max(0,|\Delta x|-|i|) = 0$. Finally, we obtain:

$$\begin{aligned} & E_{\Delta x}[\max(0,|\Delta x|-|i|)] = \\ & = \frac{1}{U}\left[\max\left(0,\frac{U}{2}-|i|\right) + \left(\frac{U}{2}-|i|\right)\max\left(0,\frac{U}{2}-1-|i|\right)\right] \end{aligned} \tag{5}$$

The final expression for $E_{\Delta y}[\max(0,|\Delta y|-|j|)]$ is very similar, naturally. Thus, we can easily compute $E_{\boldsymbol{\Delta v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta v})]$ by using (2).

### Expected value of $\mathrm{Pyr}(\boldsymbol{\Delta v})$

Consider the same discrete uniform distribution of translation errors as before: a separable distribution with probability mass function of $\frac{1}{U^2}$ at each point within $[-\frac{U}{2}, \frac{U}{2}-1] \times [-\frac{U}{2}, \frac{U}{2}-1]$, with $U > 0$ and multiple of 2. In this case:

$$\begin{aligned} & E_{\boldsymbol{\Delta v}}[\mathrm{Pyr}(\boldsymbol{\Delta v})] \\ & = \frac{1}{U^2}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1}\sum_{\Delta y=-\frac{U}{2}}^{\frac{U}{2}-1}\mathrm{Pyr}(\Delta x,\Delta y) \\ & = \frac{1}{U^2}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1}\sum_{\Delta y=-\frac{U}{2}}^{\frac{U}{2}-1}\frac{(w-|\Delta x|)(h-|\Delta y|)}{w^2h^2} \\ & = \frac{1}{U^2w^2h^2}\sum_{\Delta x=-\frac{U}{2}}^{\frac{U}{2}-1}(w-|\Delta x|)\sum_{\Delta y=-\frac{U}{2}}^{\frac{U}{2}-1}(h-|\Delta y|) \\ & = \frac{1}{wh}\left[1-\frac{U(w+h)}{4wh}+\frac{U^2}{16wh}\right] \end{aligned} \tag{6}$$

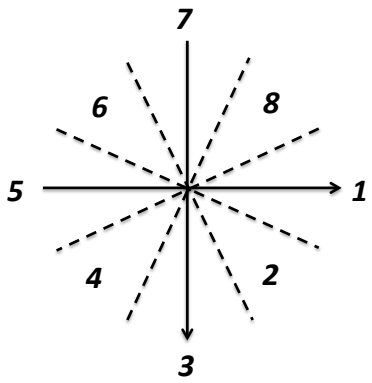


**Fig. 3**: Gradient orientation bin ($d$) numbering convention used in our work.

| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 |

**Fig. 4**: Spatial bin ($n$) numbering convention used in our work.

where the last step is obtained by using the fact that $\sum_{k=0}^{K} = \frac{K(K+1)}{2}$, and standard mathematical derivations. In the case of $w = h$, we can further simplify (6) to:

$$E_{\boldsymbol{\Delta v}}[\mathrm{Pyr}(\boldsymbol{\Delta v})] = \frac{1}{w^2}\left(1-\frac{U}{4w}\right)^2 \tag{7}$$

Together with the expression for $E_{\boldsymbol{\Delta v}}[\mathrm{N}_{i,j}(\boldsymbol{\Delta v})]$ derived in the previous subsection, the expression for $E_{\boldsymbol{\Delta v}}[\mathrm{Pyr}(\boldsymbol{\Delta v})]$ can then be used to derive the final closed-form expression for $E[\|f_A - f_B\|_2^2]$.

## Appendix B: Supplemental experimental results

In this section, we present further experimental results for the distribution of $z_{d,n}^2$ given $\boldsymbol{\Delta v}$. Our objective is to provide further evidence that the Gamma distribution approximation for $z_{d,n}^2$, using M-S, works well. To clarify which spatial and orientation bins are used, Fig. 3 and Fig. 4 present the conventions we use for numbering them. Fig. 5 and Fig. 6 present estimated distributions for different orientation and spatial bins, using both the SMVS and the CNN2h datasets, plotted against ground-truth distributions measured from data. These figures show the estimated distributions using M-IID and M-S, plotted for $\boldsymbol{\Delta v}_1$, $\boldsymbol{\Delta v}_2$ and $\boldsymbol{\Delta v}_3$ (as described in the experimental section of the main part of this paper). We can infer that the M-IID models (in green) estimate the distributions poorly. The M-S model (in red) estimates the ground-truth distributions much better, certainly capturing the main trends.

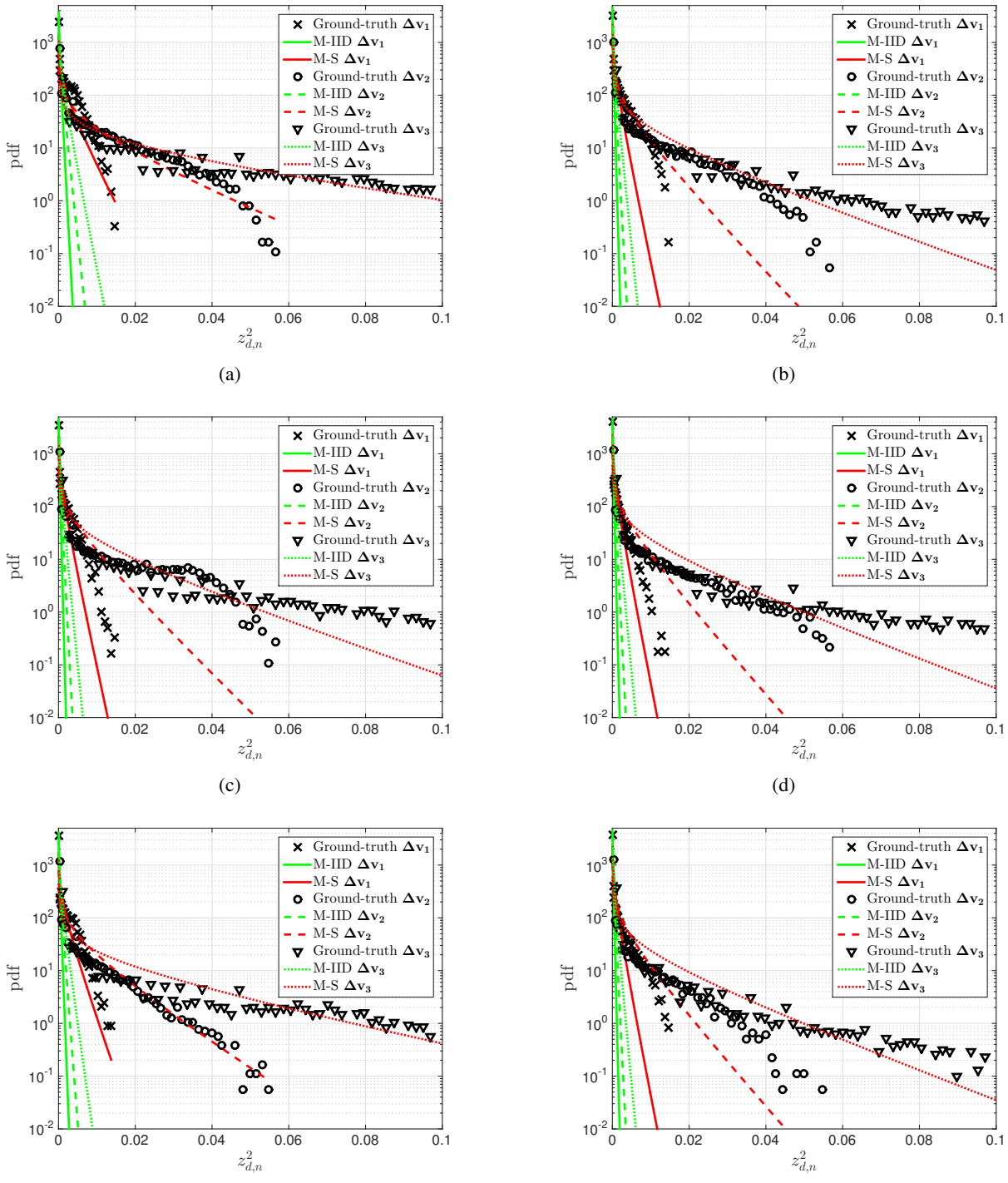


**Fig. 5**: Estimated and ground-truth distributions for $z^2_{d,n}$, using M-IID and M-S, with translation errors $\boldsymbol{\Delta}\mathbf{v}_1$, $\boldsymbol{\Delta}\mathbf{v}_2$ and $\boldsymbol{\Delta}\mathbf{v}_3$, using the SMVS dataset. In these plots, (a) $d = 1$, (b) $d = 2$, (c) $d = 3$, (d) $d = 4$, (e) $d = 5$, and (f) $d = 6$, all with $n = 3$.

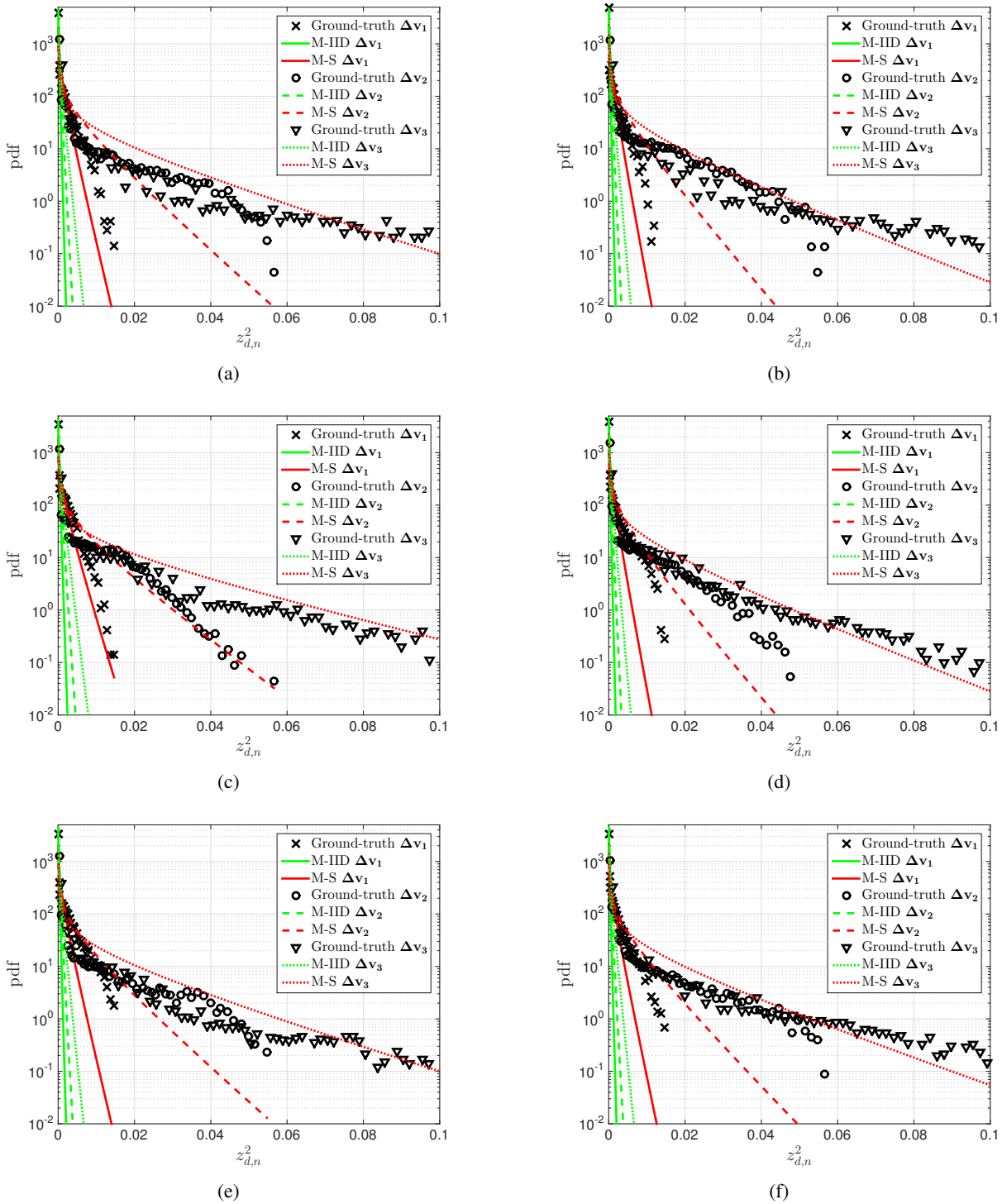


**Fig. 6**: Estimated and ground-truth distributions for $z^2_{d,n}$, using M-IID and M-S, with translation errors $\mathbf{\Delta v}_1$, $\mathbf{\Delta v}_2$ and $\mathbf{\Delta v}_3$, using the CNN2h dataset. In these plots, (a) $d = 3$, (b) $d = 4$, (c) $d = 5$, (d) $d = 6$, (e) $d = 7$, and (f) $d = 8$, all with $n = 10$.