Object Recognition Robust to Imperfect Depth Data

David F. Fouhey, Alvaro Collet, Martial Hebert, Siddhartha Srinivasa

Robotics Institute, Carnegie Mellon University, USA
{dfouhey,acollet,hebert,siddh}@cs.cmu.edu

Abstract. In this paper, we present an adaptive data fusion model that robustly integrates depth and image only perception. Combining dense depth measurements with images can greatly enhance the performance of many computer vision algorithms, yet degraded depth measurements (e.g., missing data) can also cause dramatic performance losses to levels below image-only algorithms. We propose a generic fusion model based on maximum likelihood estimates of fused image-depth functions for both available and missing depth data. We demonstrate its application to each step of a state-of-the-art image-only object instance recognition pipeline. The resulting approach shows increased recognition performance over alternative data fusion approaches.

Despite its tremendous potential, dense depth estimation has fundamental limitations that must be addressed for robust performance. In many realistic scenes, depth sensors fail to compute depth measurements on portions of the associated color data (as shown in Fig. 1). We refer to this phenomenon of missing depth data as *depth fading*. Objects close to the camera, reflective or specular surfaces, poor lighting conditions, and surfaces seen at oblique angles often suffer from depth fading. These issues arise from fundamental physical limitations in depth perception, and affect all depth estimation approaches to varying degrees.

Although some problems are naturally robust to depth fading (e.g., by easily factoring into subproblems with and without depth [1]), many authors [2–4] resort to interpolating depth data and then propagating the interpolated values. Common interpolation methods include the recursive median filter [2], inpainting [3], or optimization techniques that minimize curvature [4]. These methods are effective when used for interpolation (the small holes in Fig. 1(top)), but produce severely inaccurate results when used for extrapolation (the moderate fading in Fig. 1(bottom). This inaccuracy has consequences for end-to-end performance: our results demonstrate that not distinguishing between interpolated and measured depth can lead to worse performance than not using depth at all.

In contrast to propagating interpolated data, this paper proposes to address the limitations of depth sensors at an algorithmic level by making the perception system aware of interpolated data. We propose a general model to adaptively combine depth and image measurements. We derive joint image-depth measurements as the maximum likelihood estimate given the independent image and

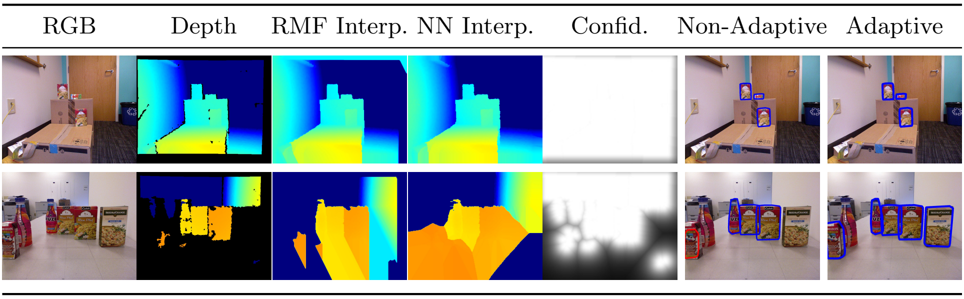


Fig. 1. Recursive median filter and nearest neighbor succeed at removing small depth fading (top row), but fail in scenes with stronger depth fading (bottom row). Our adaptive model yields robust performance under both conditions.

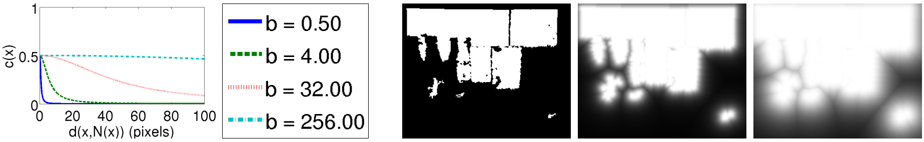
depth measurements. We combine a depth-filling technique with per-pixel confidences to extend depth measurements to areas with depth fading. In this way, we adaptively combine image and depth measurements for every pixel. To demonstrate the flexibility and effectiveness of our model, we integrate it into each of the components of a state-of-the-art image-only object instance recognition system, MOPED [5]. We use it to derive an adaptive distance metric for feature-based pose estimation; a prior generation technique based on adaptive 2D/3D similarity; and a depth-adaptive feature matching scheme.

We evaluate all contributions on a set of realistic scenes with up to 10 objects per scene, with depth fading ranging from 15% to 85% of the image. Each proposed algorithm is validated independently, and the adaptive model is evaluated in comparison to non-adaptive approaches via an integrated system that uses all algorithms. Our adaptive model demonstrates substantial gains over the image-only system with little depth fading present and a seamless transition to image-only performance with severe depth fading. In contrast, we demonstrate that under less favorable conditions, non-adaptive approaches (such as simply propagating interpolated data) can perform substantially worse than not using depth altogether (a decrease of 15% recall) while not performing any better than an adaptive approach under favorable conditions.

1 Adaptive 2D/3D Measurement Model

We propose a general model to adaptively combine image and depth measurements into a joint image-depth function. Given partial observations $x_{2D} \in \mathcal{X}_{2D}$ (image only) and $x_{3D} \in \mathcal{X}_{3D}$ (depth only), let $\phi_{2D} : \mathcal{X}_{2D} \to \mathbb{R}$ be an image-only function, and $\phi_{3D} : \mathcal{X}_{3D} \to \mathbb{R}$ a depth-only function. Defining the full observation $x = \{x_{2D}, x_{3D}\} \in \mathcal{X} = \mathcal{X}_{2D} \times \mathcal{X}_{3D}$, the joint image-depth function $\phi : \mathcal{X} \to \mathbb{R}$ is a combination of the partial functions ϕ_{2D}, ϕ_{3D}.

The functions ϕ, ϕ_{2D} and ϕ_{3D} are general functions which can model a wide array of processes, as we show in Sections 4, 5, and 6. For the remainder of this paper, we assume that $\phi_{2D}(x_{2D})$ and $\phi_{3D}(x_{3D})$ are measured in the same units.



(a) (b)

Fig. 2. The role of confidence values. (a) $c(x)$ vs $d(x, N(x))$ for $\psi = 1$ and varying b; (b) $c(x)$ plotted for Fig. 1(bottom) with $\psi = 0$ and $b = 1, 10, 32$ (l. to r.).

We also assume that $\phi_{2D}(x_{2D})$ and $\phi_{3D}(x_{3D})$ are noisy observations, conditioned on $\phi(x)$, of the true values $\bar{\phi}_{2D}(x_{2D})$, $\bar{\phi}_{3D}(x_{3D})$, corrupted with i.i.d. noise with distributions $\mathcal{N}(0, \sigma_{2D}(x_{2D}))$ and $\mathcal{N}(0, \sigma_{3D}(x_{3D}))$ respectively.

Our goal is to find $\phi(x)$ that maximizes $P(\phi(x)|\phi_{2D}(x_{2D}), \phi_{3D}(x_{3D}))$. Using Bayes' Rule and assuming no prior on the function distributions, the optimal fused function $\phi^*(x)$ corresponds to the Maximum Likelihood Estimate (MLE)

$$\phi^*(x) = \underset{\phi(x)}{\arg\max} \;\; P(\phi_{2D}(x_{2D}), \phi_{3D}(x_{3D})|\phi(x)). \tag{1}$$

Following [6], the MLE $\phi^*(x)$ is computed as

$$\phi^*(x) = \frac{\sigma^2_{2D}(x_{2D})}{\sigma^2_{2D}(x_{2D}) + \sigma^2_{3D}(x_{3D})}\phi_{3D}(x_{3D}) + \frac{\sigma^2_{3D}(x_{3D})}{\sigma^2_{2D}(x_{2D}) + \sigma^2_{3D}(x_{3D})}\phi_{2D}(x_{2D}). \tag{2}$$

Parameterizing Eq. 2 in terms of a depth confidence function $c(x)$, and defining the ratio of variances $\gamma^2(x) = \frac{\sigma^2_{3D}(x_{3D})}{\sigma^2_{2D}(x_{2D})}$, then

$$c(x) = \frac{1}{1 + \gamma^2(x)}, \qquad \phi^*(x) = c(x)\phi_{3D}(x_{3D}) + [1 - c(x)]\phi_{2D}(x_{2D}). \tag{3}$$

In this formulation, $c(x)$ reflects the confidence of the depth function relative to the image function. The value of $c(x)$, and thus $\phi^*(x)$, depends only on the ratio of variances γ^2 of the noise distributions of $\phi_{2D}(x_{2D})$, $\phi_{3D}(x_{3D})$.

We now extend the model in Eq. 3 to depth-filling methods (e.g., nearest neighbor, recursive median filter) by estimating the ratio of variances $\gamma^2(x)$ in areas with depth fading. Let INT be an interpolation method which computes depth z_x for a point x with missing depth from a set of support datapoints $N(x)$ with depth measurements, such that $z_x = \text{INT}(N(x))$. We assume that the variance $\sigma^2_{3D}(x)$ for an interpolated datapoint x is higher than for support datapoints $N(x)$, i.e., $\sigma^2_{3D}(x) > \sigma^2_{3D}(N(x))$. Then, given that the variance $\sigma^2_{2D}(x)$ remains constant, the ratio of variances $\gamma^2(x) > \gamma^2(N(x))$, and thus $c(x) < c(N(x))$. The resulting dense depth map contains depth values and confidences for every datapoint, with decreasing confidences $c(x)$ for areas with depth fading.

We estimate the ratio of variances γ^2 of interpolated datapoints from ones with measured depth. For a known datapoint: $\gamma^2(x) = \psi^2(x)$ where $\psi^2(x)$ is from a sensor model or is a fixed value (effectively a prior). For an interpolated datapoint x,

$$\gamma^2(x) \approx \psi^2(N(x)) + d(x, N(x))^2/b^2, \tag{4}$$

where $d(x, N(x))$ is a distance function between the interpolated datapoint x and its support datapoints $N(x)$ in the image plane. The scalar parameter b encodes the trust in interpolated values with respect to measured values.

The confidence model $c(x)$ for dense depth depends exclusively on the interpolator INT, the scalar b, and $\psi^2(x)$ for known datapoints. For the remainder of this paper, we use simple Nearest Neighbor as our interpolator, but other alternatives (e.g., recursive median filter) are also possible.

The ratio of variances for known datapoints ψ^2 depends on the particular task and sensor. ψ^2 can be estimated empirically in some cases, but it is hard in the general case. The alternative is to set ψ^2 to a fixed value: $c(x)$ is constant for known datapoints, but $c(x)$ adapts for datapoints with unknown depth. If both noise distributions are believed to have equal variances, $\psi^2 = 1$; then $c(x) = 0.5$ for all datapoints with *known* depth, weighing both functions equally. An alternative, useful when the depth function is more informative than the image function, is $\psi^2 = 0$; then $c(x) = 1$ for all datapoints with known depth, thus exclusively using the depth function when depth is known, and adaptively reverting to the image function when depth is not available.

We illustrate common values of parameters ψ^2 and b in Fig. 2. In Fig. 2(a), we show $c(x)$ for a variety of values of b as the interpolation distance increases. The scalar b parameterizes the confidence decrease rate as a function of interpolation distance; analytically, for a fixed ψ^2, a point b pixels away from the known value has confidence $(\psi^2 + 1)/(\psi^2 + 2)$. Fig. 2(b) shows maps of $c(x)$ plotted for Fig. 1 (bottom) for b ranging from 1 to 32 pixels.

Setting ψ and b for a new algorithm is straightforward. Sensible values of ψ are $\psi = 1$ (for balanced image and depth measurements), $\psi = 0$ (for no information in the image), or values from a depth sensor model. To find a suitable b, we perform a logarithmic grid search over the validation set for a fixed ψ. We establish the stability of b in extensive experiments detailed in Section 7.

2 Problem Formulation

We demonstrate the application of our model to each step of a feature-based object instance recognition system [5]. In this section, we briefly introduce the structure of its pipeline. The input to the system is a calibrated RGBD image $I = \{\text{Rgb}, \boldsymbol{z}\}$ such that each color pixel value has a corresponding depth z_i. The output of the system is a set of object hypotheses $\mathbf{H}$ represented by an object identity and the pose of the object in the camera's reference frame. Each object model to be recognized is represented as a set of features $\mathbf{F}_o$; each feature is represented by a 3D point location $P = [X,\ Y,\ Z]^T$ in the object's reference frame and a feature descriptor D, (e.g., SIFT [7]).

The general approach is to find object poses by solving the Perspective-N-Point (PnP) problem on a set of image-model matches. The system first detects features and matches them to the stored database of model. Nearby matches are grouped together to generate priors for object poses in order to efficiently handle multiple object instances and reject outliers. Given that objects are mostly continuous in image space, clustering feature locations in image space produces good object priors. Using these priors as an initialization, the resulting PnP problems given the image-model matches are solved using RANSAC and non-linear minimization of the reprojection error.

In the image-only system [5], matching is done with the standard 2-nearest neighbors ratio test [7], clustering with mean-shift in the image plane, and pose estimation optimizes the reprojection error [8] with Levenberg-Marquardt minimization. In Sections 4-6, we add depth data to each approach with our adaptive model: we adapt the match acceptance threshold, integrate depth features into prior generation, and derive an adaptive pose estimation error function.

3 Data Sets

Scenes in common RGBD datasets [3, 4] are captured by a human operator in the sensor's recommended operating range, but still miss up to 85% [3] and 83% [4] of their depth data in some cases. In fact, $2.5-3.8\%$ of the scenes in [3, 4] are missing over half of their depth data. In applications where sensor position and scene composition cannot be controlled (e.g., in mobile robotics), depth fading becomes a critical factor to address for robust perception [9].

We replicate this spectrum with two datasets with varying depth quality. We captured all scenes using a Microsoft Kinect RGBD sensor with registered RGB and depth data. Additionally, we gathered a smaller training set that contains the same objects in 79 scenes of various environments.

Offices Dataset: In this first set, we aim to represent optimal operating conditions for RGBD sensors; most scenes show little or no depth fading. These scenes only depict small gaps due to partial occlusion or shadowing, with an average depth coverage of 66% of the image; most fading is due to registering the depth and color images and does not fall on the objects. We captured 350 scenes with an average of 4.4 objects per scene.

Tables Dataset: In the second set, we captured 200 scenes with large sections of depth fading. These scenes show, among other anomalies, missing surfaces due to steep viewing angles and objects at short distances away from the sensor which cause heavy depth fading. These 200 scenes contain an average of 5.2 objects and have 35% average depth coverage.

4 Depth-Adaptive Pose Estimation

In the first demonstration of the model, we derive an adaptive algorithm for feature-alignment-based pose estimation. In general, the estimation of an object pose given a set of 2D/3D correspondences between image features and a known model is the well-known PnP problem. The most accurate solutions are usually found by non-linear least squares minimization of pose parameters using the reprojection or backprojection errors [5]. To adaptively use depth information, we introduce a 2D/3D distance metric, or ϕ.

Given a set of features $\boldsymbol{F}$ with 2D positions p_i and 3D positions P_i, we parameterize P_i as a line L_i through p_i and the camera center, as well as a depth z_i. Given a pose hypothesis with transformation T, we define $P_{T;i}$ as the position of the corresponding feature of the hypothesis that matches P_i. Let $\hat{P}_{T;i}$ be the projection of $P_{T;i}$ onto L. Using $\hat{P}_{T;i}$ we derive two common image-only errors: the backprojection error is the 3D distance $||\hat{P}_{T;i} - P_{T;i}||^2$ and the reprojection error is the distance when projected on the image plane.

To formulate our adaptive model ϕ, we select the backprojection error as ϕ_{2D}, and introduce an orthogonal penalty in depth $||P_i - \hat{P}_{T;i}||^2$ to serve as ϕ_{3D}. Our objective function is the sum of $\phi^*(i) = c(i)\phi_{3D}(i) + [1 - c(i)]\phi_{2D}(i)$, $\forall\ i$. We set $\psi(i) = 1$ and use our validation set to determine $b = 0.1$, which is held constant throughout the experiments.

Validation: On the Offices data set, optimizing the adaptive error results in lower average relative translation error (1.5% vs. 3.7%) and rotation error (5.6° vs. 7.6°) with respect to the ground-truth when compared to optimizing the standard reprojection error.

5 Depth-Adaptive Priors for Object Recognition

Model-based object recognition and pose estimation from local features requires solving two sub-problems: data association and pose optimization. In simple scenes, RANSAC and a PnP solver are sufficient. However, realistic scenarios may contain large numbers of outliers and multiple instances of the same object. In this case, it is vital to compute object priors to limit the otherwise overwhelming search space of potential hypotheses. A number of approaches have been used to generate priors, both from depth and image data. Collet *et al.* use estimate spatial feature density with Mean Shift to find plausible regions for objects in [10]. Other approaches [11, 12] use an horizontal plane detector to generate priors.

In this work, we use clustering for prior generation akin to [5]. We partition the set of matches for a object into clusters and search only within the poses supported by these clusters. An ideal cluster contains only matches supporting one object instance and no outliers.

To provide a prior generation algorithm that is robust to depth fading, we propose an agglomerative clustering scheme based on 2D/3D feature similarity. Here, we extend our model to pairs of measurements and fuse a 2D-similarity function $\phi_{2D} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with a 3D-similarity function $\phi_{3D} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. To model the confidence of a pair of points, we assume their independence and set $c(\{i, j\}) = c(i)c(j)$. Our similarity function is then:

$$\phi^*(i, j) = c(i)c(j)\phi_{3\mathrm{D}}(i, j) + [1 - c(i)c(j)]\phi_{2D}(i, j) \tag{5}$$

We build each term of our similarity function using simple features. We denote p_i and P_i as the 2D and 3D positions of image feature F_i, respectively; $\hat{P}_i$ is the 3D position of the corresponding feature in the model; and $d(\cdot, \cdot)$ is the Euclidean distance between two vectors.

Spatial Proximity. Objects are generally continuous; we thus use the feature $S_{2\mathrm{E}}(i,j) = \exp(-d(p_i, p_j)^2/\sigma_{2\mathrm{D}})$, and equivalently S_{3E} in 3D using P_i, P_j.

Depth Discontinuity. Two matches are unlikely to belong to the same object if there is a significant depth discontinuity between them. We formalize this intuition by sampling N points along the line through p_i and p_j in the image plane, and measuring the change in depth as an angle over each segment compared to the global change in depth θ of line $\overline{p_i p_j}$. This yields a penalty on strong changes in depth: $S_{\mathrm{D}}(i,j) = \exp(-\max_{1 \leq k \leq N}\{(\theta_k - \theta)\}/\sigma_{\mathrm{disc}})$.

Distance Consistency. The availability of depth measurements enables us to check the consistency of the distance between points in the world and their locations in the model coordinate frame: we can prevent the clustering of two matches if they are at inconsistent distances and thus cannot support the same pose. The consistency similarity function is defined as:

$$S_{\mathrm{C}}(i,j) = \exp(-\left(|d(P_i, P_j) - d(\hat{P}_i, \hat{P}_j)|/d(\hat{P}_i, \hat{P}_j)\right)/\sigma_{\mathrm{cons}}). \tag{6}$$

To combine these features into similarity functions, we choose $\phi_{2D}(i,j) = S_{2E}(i,j)$ and $\phi_{3D}(i,j) = \frac{1}{2} S_C(i,j)(S_{\mathrm{D}}(i,j) + S_{3\mathrm{E}}(i,j))$. We enforce the S_C feature more strongly as it is a hard requirement for clustering, rather than a preference (such as spatial proximity). We set $\psi(i) = 1$ and use our validation set to determine $b = 25$, which is held constant throughout for all experiments. We use Agglomerative Clustering with group-average linkage to cluster.

Validation: We define a match as an inlier if the match has reprojection error 2 pixels or less with regards to a correctly detected object. We define *cluster precision* as the fraction of matches in any clusters that are inliers and *cluster recall* is the fraction of inlier matches appearing in any cluster. On the Offices data set, while obtaining similar cluster recall, the proposed adaptive clustering approach obtains 82% cluster precision, in comparison to 28% with Mean Shift in the image plane; further, since it can reject mismatches on objects, it even achieves 2% higher precision than using the ground-truth outlines of objects.

6 Adaptive Feature Matching

We demonstrate the application of our method to situations in which the function depends only on 3D. For instance, depth can constrain the scale at which one searches for objects (e.g., [13]). However, such techniques are problematic when depth data is largely absent. Our model transitions between aggressive scale constraining during optimal conditions and cautious behavior when depth data is unavailable. To achieve this, we let ϕ_{2D} be a constant function and set $\psi(i) = 0$.

One way to constrain the search scale for objects is to adjust the number of feature matches that are accepted by adjusting the threshold used in the ratio test. The ratio test [7] is a common criterion to evaluate whether a pair of local features are sufficiently similar to be considered a match. A match between a

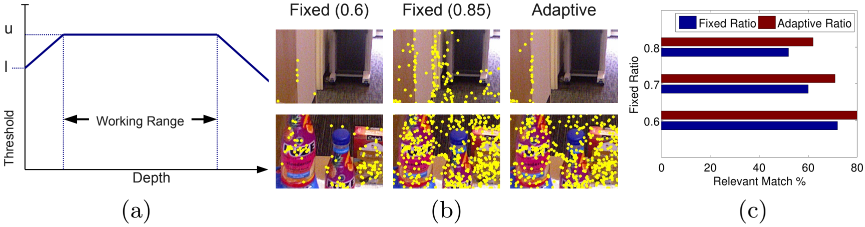


Fig. 3. (a) The threshold $\theta(z)$ as a function of depth. (b) A qualitative illustration: our adaptive approach provides more matches in relevant regions while not increasing the number of matches in the background. (c) Quantitative results on the Offices data set; when one replaces each fixed threshold with an adaptive one that retains the same number of relevant matches, one achieves a higher fraction of relevant matches.

local feature F_i and its nearest neighbor $N(F_i)$ in a database is only made if the ratio between that distance and the distance to the second nearest neighbor $d(F_i, N_2(F_i))$ is less than a threshold τ. The only parameter in the ratio test is the threshold τ. In the absence of *a priori* information, an educated guess is used. With RGBD sensors, however, we have depth measurements for each local feature, and we know the scale of the objects in the database. Thus, we can use the working ranges at which we expect to detect our objects based on the object's size and density of features.

To maximize recognition and speed performance, our goal is to find just enough matches for an object to be detected throughout the object's entire working range (determined by physical size), and no matches outside this range. We approximate this behavior by replacing the fixed τ with a function ϕ^* that depends on the depth measurement z_i and the confidence score $c(i)$. When depth information is present, a function $\theta(z)$ maps depth z to a threshold. The form of $\theta(z)$ is illustrated in Fig. 3(a). To make this approach robust to imperfect depth data, we set $\phi_{3D}(i)$ to $\theta(z_i)$ and $\phi_{2D}(i)$ to a default ratio $\theta(z_0)$, where z_0 is a default depth. In our experiments, z_0 is fixed at $1m$. Since ϕ_{3D} completely determines ϕ with known depth, we set $\psi(i) = 0$; we use our validation set to determine $b = 75$, which is held constant throughout the experiments. Parameters l, u are fixed using grid search to maximize recall on a validation set; they are not sensitive to particular values, and are in the vicinity of usually used values.

Validation: We evaluate the matching performance by counting the fraction of relevant matches; we define a feature as *relevant* if it falls on any ground-truth object. Higher thresholds yield more relevant features at the cost of more irrelevant ones. Since our approach reverts to a per-object fixed threshold in the absence of depth data, we only consider matches with sufficiently high confidence (above 0.5) to focus the evaluation on the depth-dependent scheme. Fig. 3(b,c) demonstrates that one can replace a fixed ratio with an adaptive ratio that yields the same number of relevant matches, but fewer irrelevant matches.

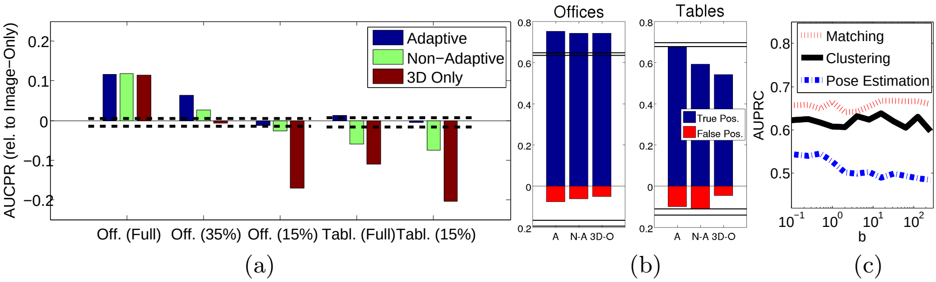


Fig. 4. (a) Average area under the precision-recall curve (AUPRC) for adaptive vs. non-adaptive approaches, relative to image-only. The range of the image-only system on 10 runs is plotted with black lines. (b) True and false positive rate at maximum F1 score for Adaptive (A), Non-Adaptive (NA), and 3D-Only (3DO). (c) Average AUPRC for individual algorithms for fixed ψ^2 and varying b (log scale) on the validation set.

7 Integrated Testing

We evaluate our adaptive model applied to an integrated system using all proposed algorithms, which we refer to as **MOPED-RGBD**. To evaluate the adaptive model, we compare to using two non-adaptive models on the integrated system: **3D Only**, which uses only data points for which there is depth data, and **Non-Adaptive**, which just propagates interpolated values, thus trusting all depth values (interpolated and measured) equally. All results are reported relative to the image-only system, **MOPED** [5].

We process each image 5 times to account for the non-determinism of RANSAC. A detection is correct if translation and rotation to the ground truth is less than 10 cm and 20°. *Precision* is the fraction of correct objects in the hypotheses, and *recall* is the fraction of ground-truth objects with a correct hypothesis.

Our results demonstrate the importance of an adaptive approach to depth fading. These results are summarized in Fig. 4(a,b) via area under the precision-recall curve (AUPRC) relative to the image-only system MOPED, as well as true and false positive rates; detailed graphs and qualitative examples apppear in the supplementary materials. Under optimal conditions (Fig. 4(a) Offices Full), all approaches to depth fading perform similarly, outperforming the image-only system. With moderate to severe depth fading (Tables Full), **the non-adaptive approaches (including propagating interpolated values) perform substantially worse than the image-only baseline.** In contrast, our adaptive model performs comparably.

To distinguish performance changes due to depth fading from those due to scene composition, we perform experiments, shown in Fig. 4(a), with synthetically degraded data. We remove data within circles with varying radii (5-30 pixels) to produce scenes with 35% depth coverage (average coverage of Tables) and 15% depth coverage (the lowest coverage in our data and in [3]) for each RGBD image in each dataset, yielding Offices 15%, etc. Again, only the adaptive approach consistently performs comparably to the image-only baseline.

Finally, we show the stability of b for each algorithm by showing the AUPRC of a system using only each algorithm for varying b and fixed $\psi(i)$ in Fig. 4(c). Performance is not sensitive to specific values, and only the order of magnitude is relevant. The best-performing b are small ($b < 1$) for tasks needing precise values (pose estimation), and larger for those needing rough values (e.g., clustering).

8 Conclusions

We have introduced an adaptive model to robustly integrate depth data into image-only perception and have applied it to object recognition. The adaptive model outperforms the image-only baseline under optimal conditions and transitions to image-only performance under severe depth fading; in contrast, non-adaptive approaches lead to worse performance than not using depth altogether.

Acknowledgments: This work was supported in part by the NSF under ERC Grant No. EEEC-0540865, by a NSF Graduate Research Fellowship to DF, and the Caja Madrid Fellowship to AC.

References

1. Cadena, C., McDonald, J., Leonard, J.J., Neira, J.: Place recognition using near and far visual information. In: IFAC World Congress. (2011)
2. Lai, K., Bo, L., Ren, X., Fox, D.: A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In: ICRA. (2011)
3. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV. (2012)
4. Janoch, A., Karayev, S., Jia, Y., Barron, J., Fritz, M., Saenko, K., Darrell, T.: A category-level 3-d object dataset: Putting the kinect to work. In: Workshop on Consumer Depth Cameras in Computer Vision (in conjunction with ICCV). (2011)
5. Collet, A., Martinez, M., Srinivasa, S.S.: The MOPED framework: Object Recognition and Pose Estimation for Manipulation. International Journal of Robotics Research **30** (2011) 1284 – 1306
6. Hackett, J., Shah, M.: Multi-sensor fusion: a perspective. In: ICRA. (1990)
7. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60** (2004) 91–110
8. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer (2011)
9. Chiu, W.C., Blanke, U., Fritz, M.: Improving the kinect by cross-modal stereo. In: BMVC. (2011)
10. Collet, A., Berenson, D., Srinivasa, S.S., Ferguson, D.: Object recognition and full pose registration from a single image for robotic manipulation. In: ICRA. (2009)
11. Lai, K., Fox, D.: A Scalable Tree-based Approach for Joint Object and Pose Recognition. In: Conference on Artificial Intelligence. (2011)
12. Kootstra, G., Kragic, D.: Fast and Bottom-Up Object Detection, Segmentation, and Evaluation using Gestalt Principles. In: ICRA. (2011)
13. Helmer, S., Lowe, D.: Using stereo for object recognition. In: ICRA. (2010)