Monocular Scale

alex.kreimer

May 2016

1 The dataset

We train and test on the KITTI dataset. There are 11 sequences with the vehicle path ground truth and 10 more sequences without the ground truth (total of 43552 stereo images). Since the cameras and the GPS sensor are only roughly synchronized we do not use the ground truth provided, but rather run a SLAM algorithm (ORB-SLAM2) to produce an alternative camera motion data.

2 Corner Extraction and Matching

Note: by corners we mean the image corners and by features we denote the machine learning features.

We use Harris corners and square 11×11 patches as corner descriptors. SSD is used as a distance measure with the winning pair declared a match. To prune the outliers we fit the fundamental into the matched corner sets and remove the corners that do not agree with this model.

3 Feature Extraction

We bin each image into $M \times N$ grid. For each bin in the image we compute the histogram of corner disparities.By disparity we denote the displacement of the corner in the image plane, e.g., if $f_1 = (x_1, y_1)$ and (the matching) $f_2 = (x_2, y_2)$ the disparity d is:

$$d = \|f_1 - f_2\| = \|(x_1, y_1) - (x_2, y_2)\|$$

We cross validated the different grid sizes ($N = 4, 5, 6$ and $M = 2, 3, 4$) and the bin sizes $nb = 100, 200, 300, 400, 500$). Grid size of 6×4 with 300 histogram bins produced the best results (e.g. the total feature vector length is $6 * 4 * 300 = 7200$).

Some statistics of the features is presented in the Figure 2. We expect the peaks, that correspond to a closer image regions have a distributions shifted to the right (i.e., larger displacements) and the peaks that correspond to a regions farther away should be closer to zero. This behavior can be observed especially

well for the feature vectors that correspond to larger camera displacements (e.g. Figure 2c).

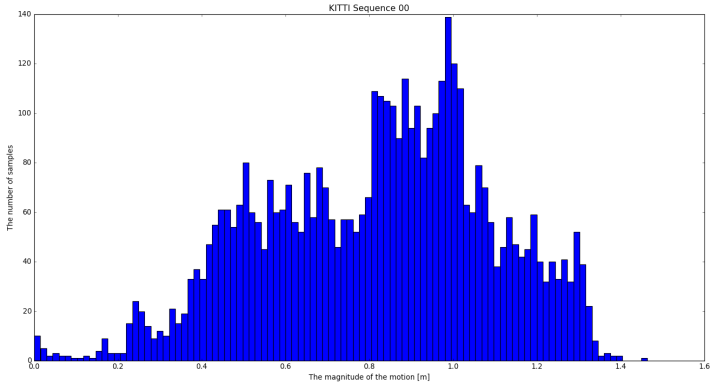


Figure 1: The distribution of the camera translation magnitudes for the sequence '00'

4 Regression Models

We train two different random forest models: the extremely random tree forest (ERT) and the linear regression in the leafs ERT (LRERT). The difference is that the former averages the values in the leaf at test time, while the latter fits a linear regression at each leaf, which is used as a final predictor. The implementation of the basic ERT is taken from the sklearn library, while we implemented the the linear regression in leafs features.

5 Model Evaluation

Figure 3 depicts the distribution of the translation magnitudes in the training set. This training set was produced by taking subsequent pairs of images. Figure 4 shows the model evaluation results.

Note: ERT outperforms the linear regression in leafs ERT.

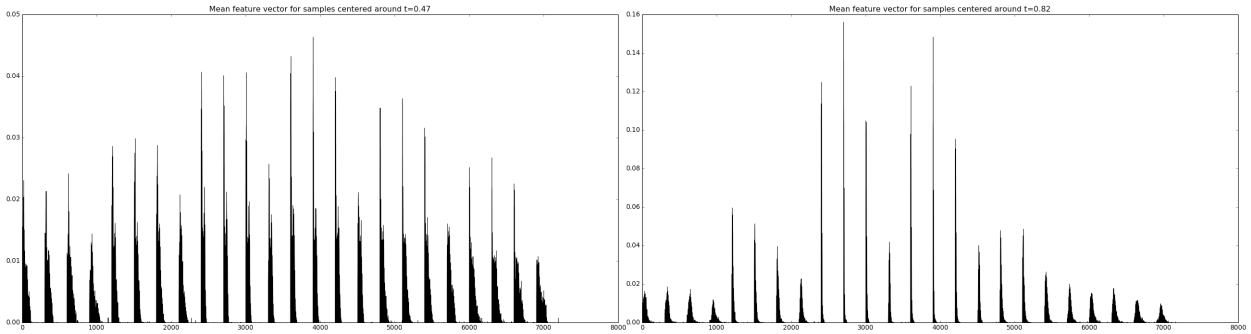
6 Path Scale prediction

Here we evaluate of the scenario where the monocular SLAM algorithm produces a sequence of camera motion estimations which are correct up to a single global scale (e.g. the relative scale of the subsequent measurements is available).

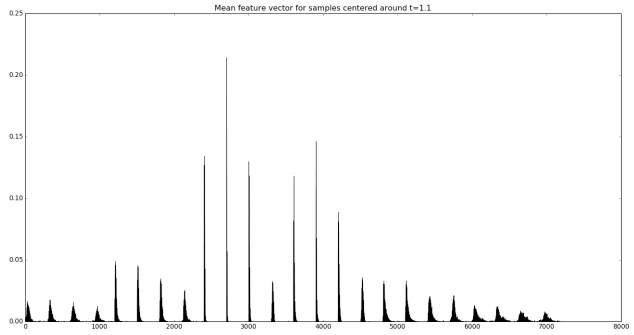
Given a set of the relative translation magnitude measurements (produced by the SLAM) t_i and the set of the predicted (by the trained models) scales $\hat{t}_i$ we search for the global scale s_P s.t.

$$s_p = \underset{s}{\operatorname{argmin}} \quad \sum_i (st_i - \hat{t}_i)^2$$

The results of the path scale prediction are summarized in the Figure 7



(a) (b)



(c)

Figure 2: Average feature vectors for samples centered around the specific camera translation magnitude. Each peak corresponds to a grid cell (e.g. here the grid is 4 rows by 6 columns by 300 bins, so the feature vector is of the dimension 6*4*300=7200). The grid is sampled in a column-major mode. So the first four peaks correspond to the leftmost column of the image grid.

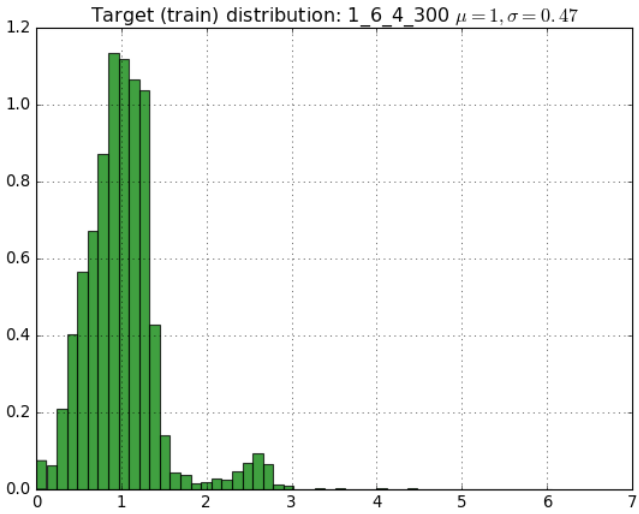
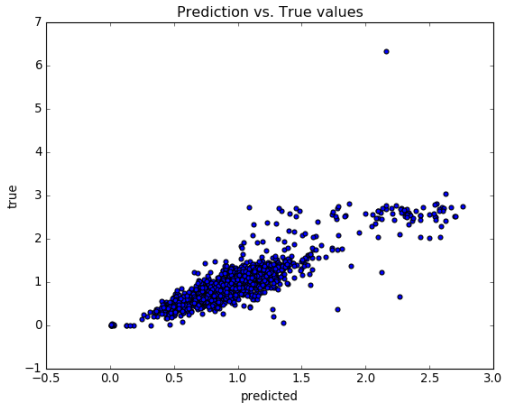
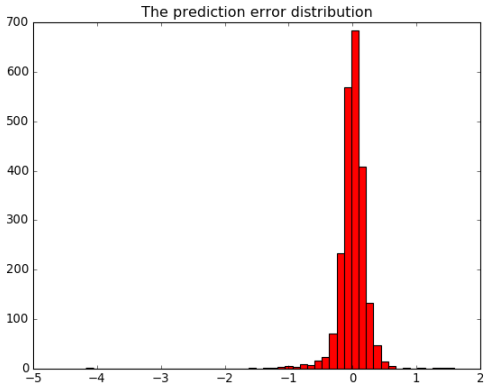


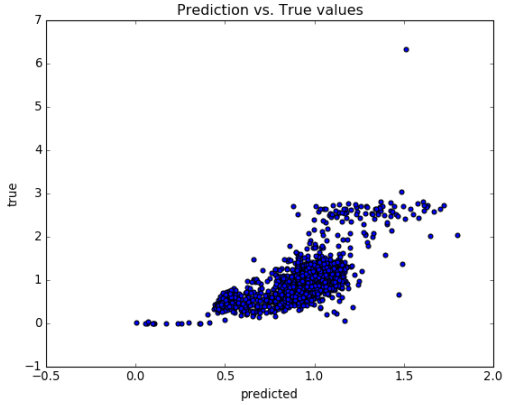
Figure 3: The distribution of the camera motion magnitudes for subsequent image pairs



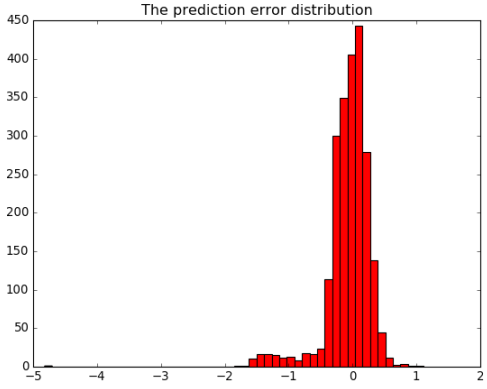
(a) ERT: RMSE=0.23



(b) ERT



(c) LRERT: RMSE=0.36



(d) LRERT

Figure 4: Models evaluation over the validation set

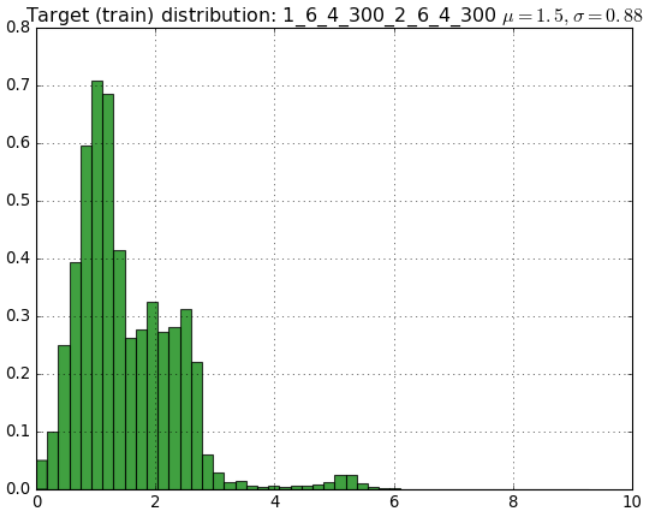
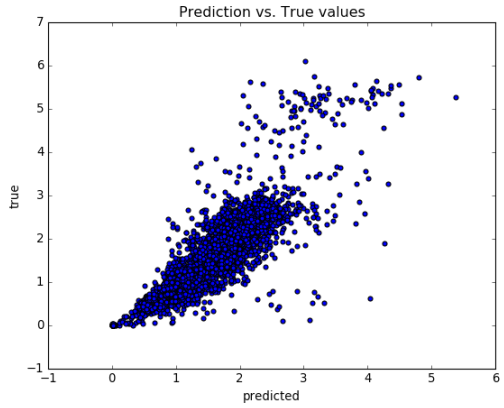
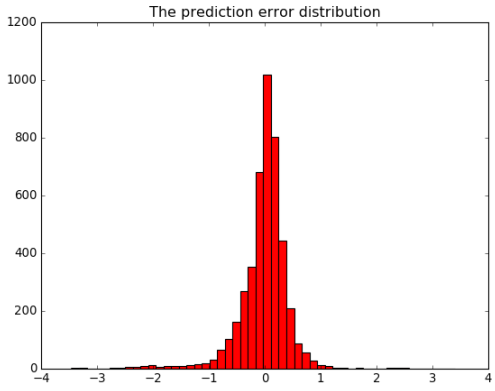


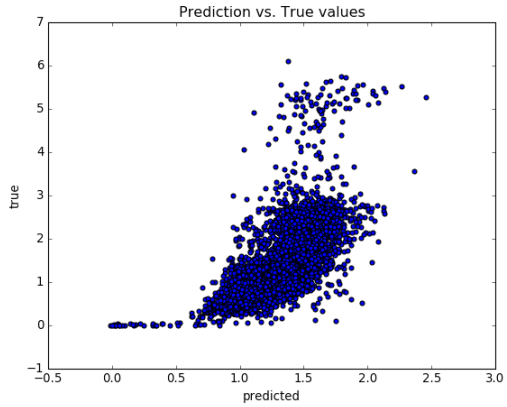
Figure 5: Dataset 2: contains the translation magnitudes for triplets of images (e.g. 2 subsequent camera motions)



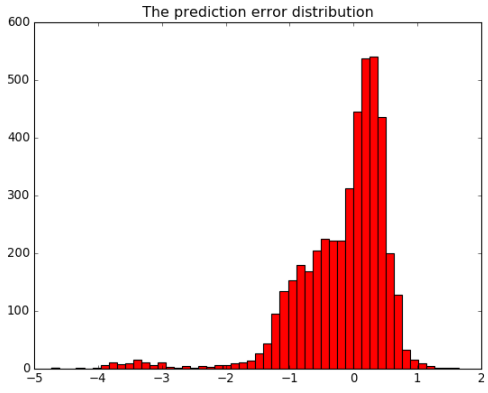
(a) ERT: RMSE=0.46



(b) ERT



(c) LRERT: RMSE=0.75



(d) LRERT

Figure 6: Results for the dataset that contains image triplets

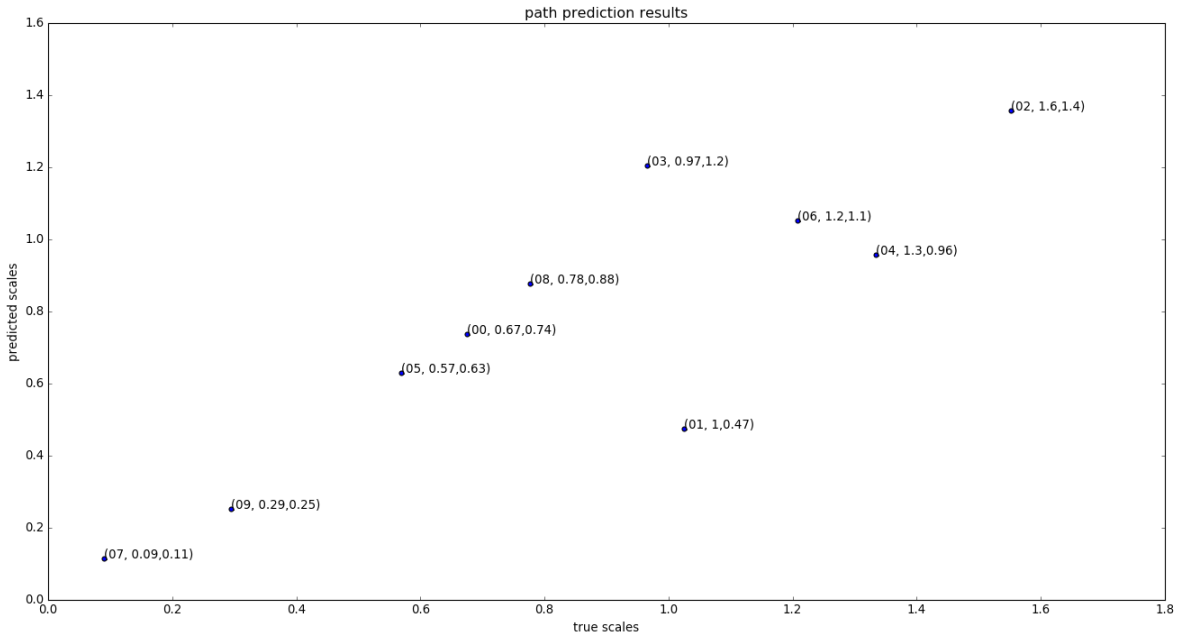


Figure 7: A path scale estimation results. Each point is annotated with a triplet (KITTI sequence number, true scale, predicted scale)