Supplementary Materials for "Multi View Registration for Novelty/Background Separation"

Omid Aghazadeh, Josephine Sullivan and Stefan Carlsson

Contents

1 Data Set **1**

2 More Qualitative Results **2**

3 Technical Details **2**
3.1 Feature Vectors Used in $\tilde{P}$ 3
3.2 Normalized Bagged Similarity 4
3.3 Estimating $\tilde{P}$ 6

1 Data Set

Figure 1 shows 7 examples of one of the places in our data set. Despite the fact that we had more than these number of images per place, in this paper, we did not use all of them: we used 4 images for training/cross validation and 3 for testing purposes. Note that using 4 images for training means that we have $4 \times \binom{6}{5} = 24$ different choices for training using 5 reference images, and similarly, $3 \times \binom{6}{5} = 18$ different choices for the testing purposes per place, which is more than sufficient for training / testing purposes. The main reason for this is the combinatorial costs in increasing the maximum number of images e.g. considering 8 images per place in total and dividing it into 4 training and 4 testing images, we would have had $4 \times \binom{7}{5} = 84$ choices for training and the same for testing - per place. This number increases to $6 \times \binom{12}{5} = 4752$ in case of using all the 12 images and 5 reference images which would have been much more expensive to deal with. For other numbers of reference images we randomly picked the same number of training and testing cases (24, 18) e.g. in case of 3 reference images - from the possible $4 \times \binom{6}{3} = 80$ training cases - we randomly picked 24.

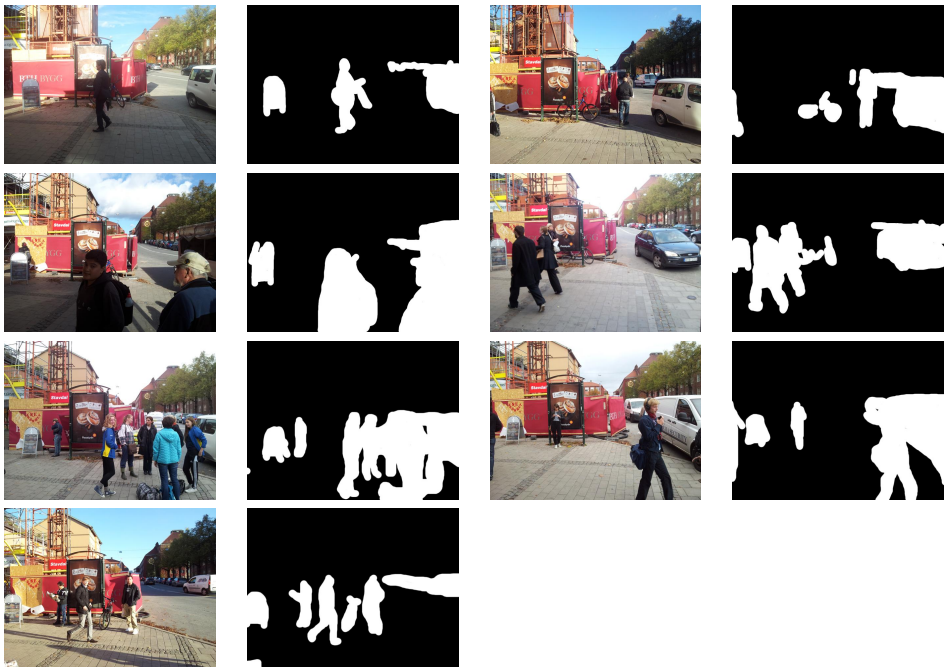

Figure 1: 7 images of one of the places from our data set and the manually defined ground truth for each image

2 More Qualitative Results

Figure 2 depicts more qualitative results. The same behavior as in the results in the paper can be observed: In addition to the general appealing behavior of the algorithm, we have some occasional missing novelties and false detections. We expect the results to improve if temporal information is additionally considered or a true multi view registration - which satisfies the projective geometry in all the views simultaneously - is formulated and solved for.

The last column in Figure 2 compares the segmentations in the 3rd column to our manually labelled ground truth. Note the over extension of the manual labellings with respect to the exact boundaries of novelties.

3 Technical Details

Here, we clarify the meaning of the "log"column in Table 3 in the paper. It is 1 if the negative log of the posterior was used in the data term of the energy function (Similar to Equations 4, 6 and 8 in the paper) and 0 if the posterior itself was used. From the table it is evident that good solutions can be found without the use of the sensitive log operator if proper priors are considered to convert likelihoods to posteriors and if proper bandwidths are used in the

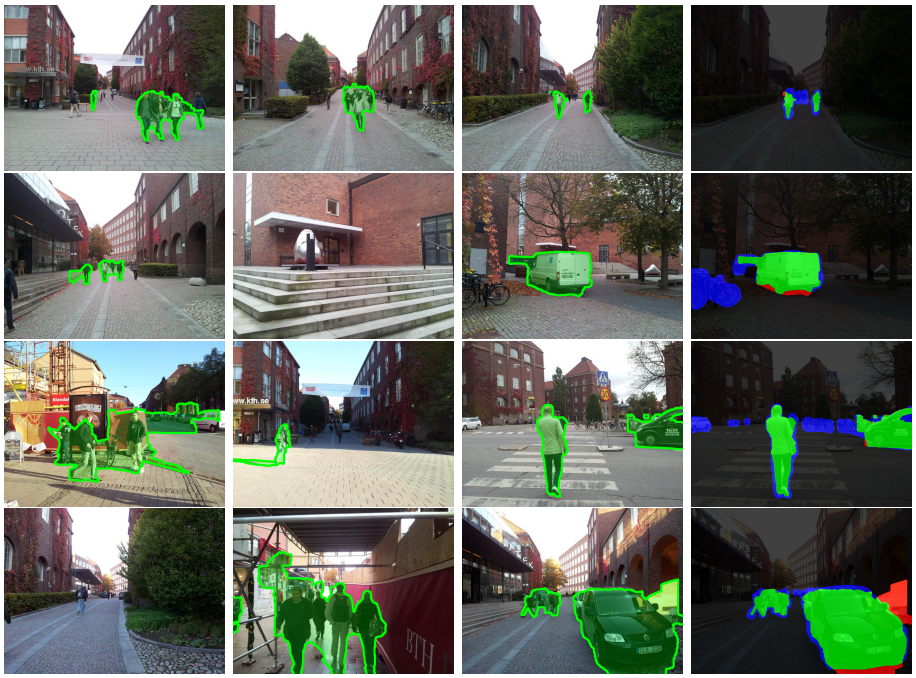

Figure 2: More qualitative results on the first 3 columns. The last column depicts the segmentation of the 3rd column imposed on the ground truth: black and green represent correctly detected background and novelty and red and blue represent background detected as novelty and novelty detected as background respectively. Note the over extended definition of novelties in our ground truth.

likelihood estimations. However, as expected, large bandwidths leading to very smooth likelihoods e.g. $h > 0.66$, require the sensitive log operator to be able to discriminate between the novelties and the background i.e. to guide the segmentation to converge to the desired solutions. As smaller bandwidths can prevent over-smooth likelihood maps, they can be discriminative without relying on the log operator.

3.1 Feature Vectors Used in $\tilde{P}$

Algorithm *FeatureExtract* shows the feature extraction process for $\tilde{P}$.

Algorithm *FeatureExtract*
Input: I_q, $R = \{I_{r_1}, ..., I_{r_n}\}$, $R_{\to q} = \{I_{r_1 \to q}, ..., I_{r_n \to q}\}$,$\Sigma_a = \{\sigma_{a_1}, ..., \sigma_{a_{na}}\}$, $\Sigma_s = \{\sigma_{s_1}, ..., \sigma_{s_{ns}}\}$, σ_{SF}
Output: F_q
1. **for** $I_r \in R$
2. $\bar{F}_{q,r} \leftarrow \emptyset$
3. **for** $\sigma_a \in \Sigma_a$

4. $\bar{F}_{q,r} = \bar{F}_{q,r} \times G_{\sigma_a} * \|S_q^{(\sigma_{SF})} - (S_r^{(\sigma_{SF})})_{\to q}\|$
5. $\bar{F}_{q,r} = \bar{F}_{q,r} \times G_{\sigma_a} * \|I_q - I_{r\to q}\|$
6. **for** $\sigma_s \in \Sigma_s$
7. $\bar{F}_{q,r} = \bar{F}_{q,r} \times G_{\sigma_a} * \|S_q^{(\sigma_s)} - S_{r\to q}^{(\sigma_s)}\|$
8. **for** $\sigma_s \in \Sigma_s$
9. $\bar{F}_{q,r} = \bar{F}_{q,r} \times \sum_c QC_{0.5}^A \left(H_{SI}^{(\sigma_s)}(I_q, ., c) - H_{SI}^{(\sigma_s)}(I_{r\to q}, ., c)\right)$
10. $\bar{F}_{q,r} = \bar{F}_{q,r} \times \sum_c QC_{0.5}^A \left(H_{SV}^{(\sigma_s)}(I_q, ., c) - H_{SV}^{(\sigma_s)}(I_{r\to q}, ., c)\right)$
11. $F_q^{AM} = \frac{1}{|R|}\sum_r \bar{F}_{q,r}$
12. $F_q^{HM} = \frac{|R|}{\sum_r \frac{1}{\bar{F}_{q,r}}}$
13. $F_q^{M} = \min_r \bar{F}_{q,r}$
14. $F_q = F_q^{AM} \times F_q^{HM} \times F_q^{M}$

where

- we used $F_1 \times F_2$ to refer to the concatenation of feature vectors F_1 and F_2 ,
- $G_{\sigma_a} * X$ refers to the convolution of X with a Gaussian kernel with standard deviation σ_a,
- σ_{SF} is the scale the SIFT feature vectors in Sift Flow were computed on,
- A is the similarity matrix for Quadratic Chi kernel. We used the following band limited similarity matrix $A_{i,j} = \frac{1}{1+|i-j|}[|i-j| < 4]$ where the [] is the Iverson bracket,
- $H_{SI}^{(s)}$ and $H_{SV}^{(s)}$ denote the shift invariant and shift variant histograms - of intensities inside a square region of size $(2s+1) \times (2s+1)$ - respectively,
- Σ_a and Σ_s define window sizes (standard deviations for Gaussian windows and window length for histogram computations) for spatial aggregation and scale computations respectively,
- S_q^s defines the sift vector on scale s computed on I_q.

We used $\Sigma_a = \{2, 4, 8, 16\}$ and $\Sigma_s = \{2, 4, 8\}$ in the paper which results in $3|\Sigma_a| = 12$ dimensions for I^{err}, $3|\Sigma_a|(1 + |\Sigma_s|) = 48$ dimensions for S^{err} and $3|\Sigma_s|2 = 18$ for H^{err} feature (a total of 78 dimensions). Note the superior performance of the NBS feature (H^{err}) compared to the rest of the Sift based features (S^{err}) despites its lower dimensionality (Figure 6 (left) in the paper).

3.2 Normalized Bagged Similarity

The computation of NBS can be made very efficient using Integral Histograms. Normalizing channels between $[0, 1]$ and quantizing each channel into $N = 32$

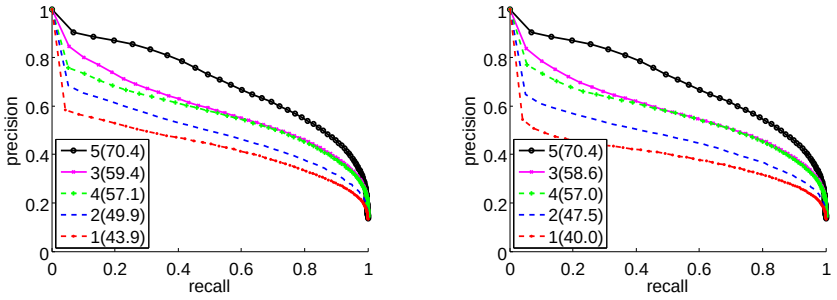

Figure 3: Evaluation of $\tilde{P}$ when trained on the data using different reference image numbers and tested on the corresponding testing set (left) and (right) when trained using the train data with 5 reference images and tested on testing data with different number of reference images.

bins and using linear interpolation, we compute the IH of the image and compute the histogram of a given width centered around a given point by 2 histogram additions and 2 subtractions.

In order to build invariance into NBS, we compute the statistics of the regions on which the histograms are obtained (from the histograms themselves)

$$\begin{array}{lllllll} \mu(H) & \approx & E[i] & = & \sum i\,P(i) & \approx & \sum_{n=0}^{N-1} \frac{n}{N-1} H(n) \\ \sigma^2(H) & \approx & E[(i-E[i])^2] & = & \sum (i-\mu)^2\,P(i) & \approx & \sum_{n=0}^{N-1} \left(\frac{n}{N-1} - \mu(H)\right)^2 H(n) \end{array} \quad (1)$$

where $E[i]$ denotes the first moment of the intensities of the pixels inside a region described by the histogram H. To make NBS Shift Invariant, we shift (each bin of) the histogram by the approximated first moment ($\mu(H)$) and interpolate the target - in the re-sampled bin locations from $[-1, 1]$ - by linear interpolation. The same approach is used for the affine invariant version but, the target is normalized by the second moment as well and the bins are then re-sampled from $[-3, 3]$[1]. However, as the discriminativeness of the measure becomes less as the invariance level increases, we did not include affine invariant version of NBS in the computation of $\tilde{P}$ and instead, we used both Shift Invariant and Shift Variant versions of the NBS in the feature pool. We also experimentally found out that the shift variant version is more discriminative and suits our problem more.

It is also possible to exhaustively search for a shift in one of the histograms that minimizes a distance measure between the two. However, we found such an approach to be computationally more demanding - specially if the distance measure is expensive to evaluate e.g. non diagonal similarity matrices in Quadratic

[1]with the assumption of Gaussian distribution of intensities, 3 standard deviation covers 0.997 of the space.

Chi kernels - without any specific advantages.

3.3 Estimating $\tilde{P}$

Figure 3 (left) shows the results of training the logistic regression function using different number of reference images (the same as Figure 6 (middle) in the paper) and Figure 3 (right) shows the result of the logistic function learnt using the training set with 5 reference images but evaluated on the test sets using different number of reference images. It can be seen that the decrements in the performance gets smaller and smaller when the number of reference images are increased (3.9, 2.4, 0.8, 0.1) which suggests that

- The logistic function being learnt gets more and more independent of the training data as the reference image set size increases.
- The regression process is perhaps converging to an optimal function irrespective of the number of reference images (in training and testing times) as enough data is provided to the method. The figure provides strong support for this idea.