Pruning Local Feature Correspondences Using Shape Context

Gustavo Carneiro and Allan D. Jepson
Department of Computer Science,
University of Toronto ,Toronto, ON, Canada.
{carneiro,jepson}@cs.utoronto.ca

Abstract

We propose a novel approach to improve the distinctiveness of local image features without significantly affecting their robustness with respect to image deformations. Local image features have proven to be successful in computer vision tasks involving partial occlusion, background noise, and various types of image deformations. However, the relatively high number of outliers that have to be rejected from the correspondences set, formed during the search for similar features, still plagues this approach. The task of rejecting outliers is usually based on estimating the global spatial transform suffered by the features in the correspondences set. This presents two problems: i) it cannot properly deal with non-rigid objects, and ii) it is sensitive to a high number of outliers. Here, we address these problems by combining typical local features [2, 7] with shape context [1]. A performance evaluation shows that this new semi-local feature generally provides higher distinctiveness and robustness to image deformations, thus potentially increasing the inlier/outlier ratio in the correspondences set. Also, we show that in wide baseline stereo matching, and non-rigid motion applications, the use of the novel semi-local feature not only provides robustness to non-rigid deformations, but also produces a higher inlier/outlier ratio than the standard Hough clustering of the global spatial transform of parameters.

1. Introduction

Highly distinctive and robustly detectable local features [2, 6, 7, 12, 11, 13] have been shown to be useful in several computer vision applications. Doubtless, the main applications involving these types of local features are those handling partial occlusion, background noise, and several types of image deformations. Examples of such applications include: wide baseline stereo [10, 14], long range motion [2, 7, 12], and object recognition with a limited set of model images [8]. However, there is still a common problem affecting all the systems above, which is the relatively small number of inliers present in a typical correspondences set built during the feature similarities search. The task of rejecting outliers, while keeping the inliers, then becomes of supreme importance in such systems.

Various approaches envisioned for outlier rejection have focused mainly on systems that strongly depend on inferring the global spatial transform of local features [3]. Unfortunately, two issues affect these methods: a) they cannot deal with non-rigid objects, and b) they are sensitive to high number of outliers in the correspondences set. Here we propose a novel approach to solve these problems, which is based on adding semi-local geometric information to the feature vector. A somewhat similar approach to filter out outliers from the correspondences set is described in [12], where a fixed number of local features around a given feature is used to determine its semi-local structure. On the other hand, our approach considers all the features in a tunable neighborhood to build the semi-local structure of a given local feature.

While the distinctiveness is clearly improved, care must be taken so that the high robustness of local features is not degraded. The semi-local feature proposed here is implemented using typical local feature approaches (here, we consider the methods [2, 7]) and a variation of the shape context method [1]. This variation is proposed to improve the robustness of the shape context feature in terms of partial occlusion, rotation, and scale changes, and it is as follows: a) nearby neighboring features have higher weight than features that are farther away during the construction of the shape context histogram; b) boundary effects are reduced by spreading a single vote over a small region of the histogram; c) invariance to rotation is achieved by rotating the histogram axis according to the main orientation of the feature; and d) scale invariance is reached by re-scaling the distance measures.

We study how the inclusion of the shape context variation affects the performance of the local features proposed in [2, 7] using the performance evaluation method introduced in [2]. We observe that the performance is considerably improved in terms of distinctiveness while the robustness is not significantly affected by the changes. We show that the use of this new semi-local feature in wide baseline matching and non-rigid motion problems generally produces a set of correspondences that is robust to non-rigid deformations and that has a higher inlier/outlier ratio than Hough clustering, which is a common approach that uses global pose to reject outliers.

2. Semi-local Image Features

The local features proposed in the literature (e.g., [2, 7, 11, 12, 13]) are in fact formed not only from values at some particular image location, but also from values extracted from neighboring pixels. What makes them local is the small support region, which generally comprises four to 16 sub-sampled pixels around the feature location. Usually, increasing the support region size improves the feature distinctiveness, but degrades the feature robustness to changes. Here, we propose a method to increase the support region size of a local feature, thus improving its distinctiveness, but without strongly affecting its robustness.

2.1. Local Image Features

Local image features suitable for local image representation must have three properties: a) distinctiveness, b) detectability, and c) robustness to image deformations. In [2, 9] it is empirically shown that both the multi-scale phase based [2] and SIFT [7] features are suitable for this task since they generally have those properties. Those features are extracted using the following two steps: a) the 'where' step selects interest points that are robustly localizable under common image deformations forming the set of locations $\mathcal{I}_i = \{\mathbf{x}_l\}$ at the following set of wavelengths (in pixels): $\{\Lambda_o = 4(2^{i/4}), i = 0, 1, .., 12\}$; b) the 'what' step extracts a feature vector describing the image structure in the neighborhood of an interest point, say $\mathbf{f}_l = \mathbf{f}(\mathbf{x}_l) = [m_l, \theta_l, \sigma_l, \mathbf{v}_l]$. Here m_l is the model identification, θ_l is the main orientation of the location $\mathbf{x}_l$ (see [4]), $\sigma_l = \frac{\lambda_l}{4.26}$ represents the feature scale, and $\mathbf{v}_l$ is the feature vector values.

The features extracted from an image I_i are then represented by $\mathcal{O}_i = \{\mathbf{f}(\mathbf{x}_l) | \mathbf{x}_l \in \mathcal{I}_i\}$. The similarity between local features $\mathbf{f}_l$ and $\mathbf{f}_o$ is computed according to the methods described in [2, 7], and we denote such similarity function by $s_f(\mathbf{f}_l, \mathbf{f}_o) \in [0, 1]$.

2.2. Variation of Shape Context

The shape context feature proposed in [1] is based on a log-polar space histogram as shown in Fig. 1. Although the log-polar space makes the descriptor more sensitive to positions of nearby features than to those farther away, we added the following additional properties in order to improve its robustness to occlusion, to reduce boundary effects, and also to make it robust to rotation and scale changes. A vote in a specific histogram bin is weighted by the following function that decreases with distance: $w(\mathbf{f}_l, \mathbf{f}_o) = e^{\frac{-0.5\mathcal{D}^2(\mathbf{f}_l, \mathbf{f}_o)}{L^2}}$, where $\mathcal{D}(\mathbf{f}_l, \mathbf{f}_o) = \frac{\|\mathbf{x}_l - \mathbf{x}_o\|}{\sqrt{\sigma_l^2 + \sigma_o^2}}$ is the scale invariant distance measure, and

$$L = \frac{\text{maximum model diameter in pixels}}{L_{\text{div}}}, \qquad (1)$$

where L_{div} is a variable. This results in an approach that prioritizes the votes of nearby features. Moreover, in order to reduce boundary effects, each neighboring feature votes for the two closest bins in each dimension. Finally, we make the shape context robust to rotation changes by rotating the histogram axis according to the main orientation of the feature. For all the cases below, we set $L_{\text{div}} = 100$.

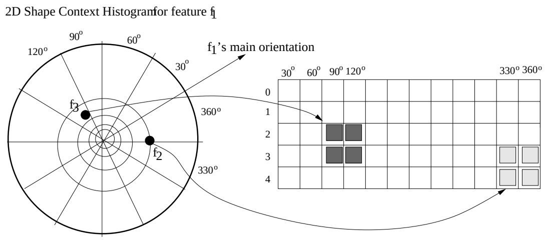

Figure 1. Shape context of local feature $\mathbf{f}_1$. As in [1], we also use five bins for log(distance) and 12 bins for relative orientation. Note that we modify the original shape context method as follows: a) the histogram is rotated according to the main orientation of $\mathbf{f}_1$, b) the votes of neighboring features $\mathbf{f}_2$ and $\mathbf{f}_3$ are weighted in terms of their distance to $\mathbf{f}_1$ (darker cell means higher weight), c) each vote spans four histogram bins to reduce boundary effects, and d) the distance is scaled to make it robust to scale changes.

The shape context similarity between feature histograms $h(\mathbf{f}_l)$ and $h(\mathbf{f}_o)$ is computed using the $\mathcal{X}^2(h(\mathbf{f}_l), h(\mathbf{f}_o))$ test statistic defined in [1]. The similarity between two histograms is then defined by $s_h(h(\mathbf{f}_l), h(\mathbf{f}_o)) = e^{\frac{-0.5\mathcal{X}^2(h(\mathbf{f}_l), h(\mathbf{f}_o))}{(0.5)^2}}$.

3. Performance Evaluation

The performance evaluation utilizes a database of images $\{I_i\}_{i \in \{1,...,112\}}$, where $\{I_i\}_{i \in \{1,...,100\}}$ are used to compute the true positive rates (TP), and the remaining 12 images form the database of random images used for the false positive rate (FP) calculation. The TP is computed by taking the proportion of features that $s_f(\mathbf{f}_l, \tilde{\mathbf{f}}_l) > \tau_f$ and $s_h(h(\mathbf{f}_l), h(\tilde{\mathbf{f}}_l)) > \tau_s$ such that $\|(M(d)\mathbf{x}_l + t(d)) - \tilde{\mathbf{x}}_l\| < \epsilon$, where $(M(d)\mathbf{x}_l + t(d))$ is the transformed position of feature $\mathbf{f}_l$ in the deformed test image, according to spatial deformation d. Here, $\mathbf{f}_l \in \mathcal{O}_i$ and $\tilde{\mathbf{f}}_l \in \tilde{\mathcal{O}}_i$, where $\tilde{\mathcal{O}}_i$ is the feature set from $\tilde{I}_i$, which is I_i after a known deformation d is applied. On the other hand, the FP is computed by taking the proportion of random image features in the set $\{\mathcal{O}_j\}_{j \in \{101,...,112\}}$ that $s_f(\tilde{\mathbf{f}}_l, \mathbf{f}_o) > \tau_f$ and $s_h(h(\tilde{\mathbf{f}}_l), h(\mathbf{f}_o)) > \tau_s$. Note that the database of random images has approximately 10^4 features. We generate the ROC curves by varying the feature similarity threshold $\tau_f \in [0, 1]$ and then evaluating TP and FP using the following values

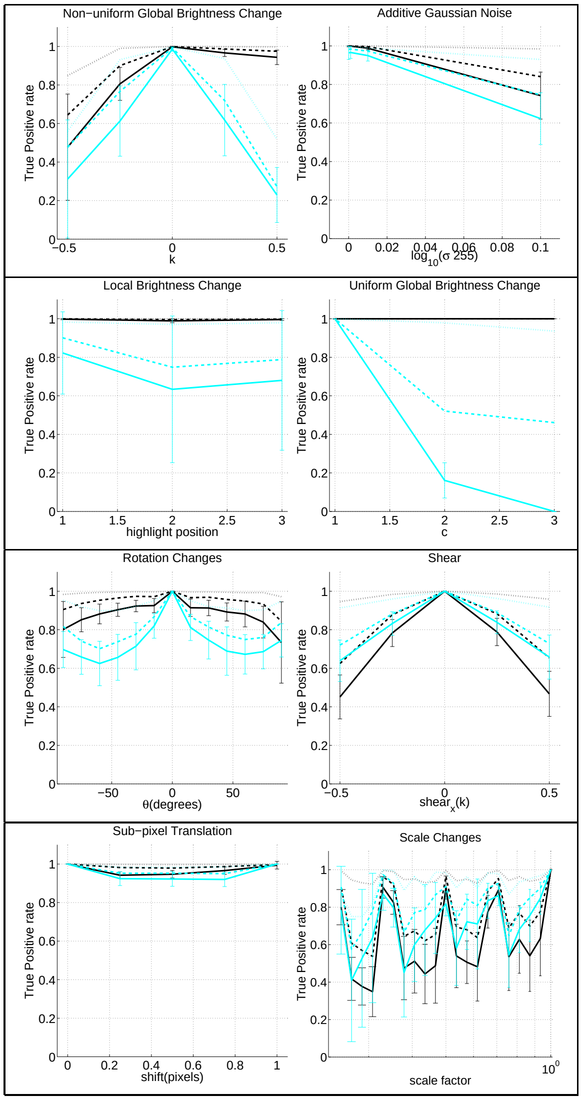

Figure 2. The TP rate curves in terms of eight different image deformations are obtained by fixing the FP rate at 0.1% in the ROC curves generated by the evaluation experiment. Black curves are the phase local feature [2] without shape context (solid), with shape context such that $\tau_s = 0.6$ (dashed), and $\tau_s = 0.8$ (dotted). Cyan curve shows the performance of SIFT [7] without shape context (solid), with shape context such that $\tau_s = 0.6$ (dashed), and $\tau_s = 0.8$ (dotted). Note that the error bars are omitted for the dashed and dotted curves for clarity, but are roughly the same size as the ones we show.

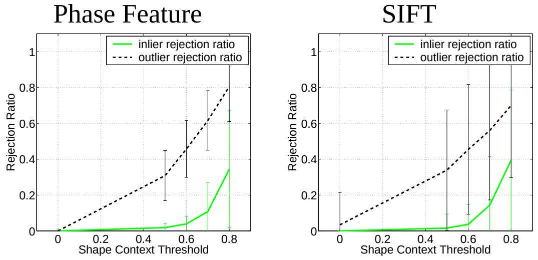

Figure 3. Inlier and outlier rejection ratios.

for $\tau_s \in \{0, .5, .6, .7, .8\}$. Notice that when $\tau_s = 0$, we are not using the shape context.

Fig. 2 shows the TP rates for a FP rate of 0.1% for eight different image deformation types described in [2]. With shape context, the smaller FP rate is quite easy to achieve for the lower shape context threshold and is nearly guaranteed for the higher threshold. As a consequence a very loose matching threshold on phase correlation can be used to achieve this false positive rate, allowing the reported TP rate to be close to one. In order to resolve the performance of features with shape context more precisely, we would need more features in our database.

The inliers and outliers that are rejected from the correspondences set as τ_s increases (with FP=0.1%) are shown in Fig. 3. The inlier rejection is computed as $\frac{N_{in}(0)-N_{in}(\tau_s)}{N_{in}(0)}$, where $N_{in}(\tau_s)$ is the number of inliers (see computation of TP rate above) for a given τ_s, while the outlier rejection is calculated as $\frac{(N_{tot}(0)-N_{in}(0))-(N_{tot}(\tau_s)-N_{in}(\tau_s))}{(N_{tot}(0)-N_{in}(0))}$, where $N_{tot}(\tau_s)$ is the total number of features in the correspondences set for a given τ_s. From these curves it is clear that the use of shape context (for $\tau_s \leq 0.7$) rejects many outliers while keeping most of the inliers in the correspondence set.

4. Applications

In order to assess the distinctiveness and robustness of the semi-local feature proposed in this work, we consider the following two applications: wide baseline stereo matching, and non-rigid motion. We only combine the shape context described in Section 2.2 with the multi-scale phase-based local feature [2] since it produces the overall best results in the performance evaluation experiment. We built a system that is divided into feature extraction, searching, and verification steps. The feature extraction is as described in Section 2.1. The searching forms the correspondences set $\{(\mathbf{f}_l, \tilde{\mathbf{f}}_l) | \mathbf{f}_l \in \mathcal{O}_i, \tilde{\mathbf{f}}_l \in \tilde{\mathcal{O}}_i, s_f(\mathbf{f}_l, \tilde{\mathbf{f}}_l) > 0.75\}$. Note that $\tilde{\mathcal{O}}_i$ is the feature set from image $\tilde{I}_i$, which contains an unknown deformed version of image I_i. In the experiments below, we compare the following two possibilities to reject outliers: a) our method using shape context; and b) Hough clustering using the same configuration as in [8], where we select the group with the highest number of features.

4.1. Wide Baseline Stereo Matching

Here, we take the set provided by the outlier rejection methods and compute the $\mathbf{F}$ matrix [5] using RANSAC [15]. We are interested in computing the proportion of in-

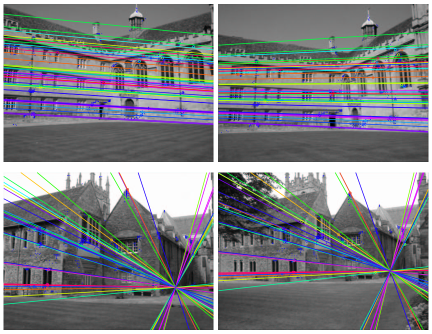

Figure 4. **Wide-baseline stereo matching. Top row shows frames 1 and 5 of the Wadham set of images, and bottom row presents frames 1 and 3 from the Merton set. The lines represent corresponding epipolar lines.**

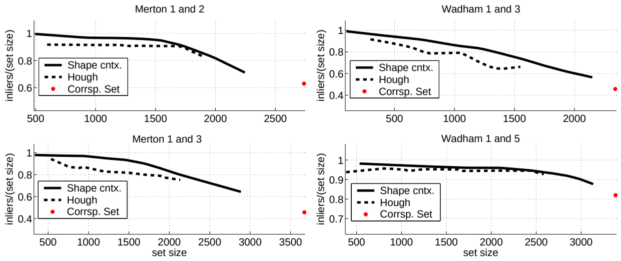

Figure 5. **Proportion of inliers from the sets provided by the outlier rejection methods.**

liers within this set. An inlier is considered to be a feature that lies within four pixels (approximately the spatial resolution of the local features used) of the epipolar line computed from the $\mathbf{F}$ matrix. Fig. 4 shows two examples of the epipolar lines computed from the image pairs using the semi-local features such that $\tau_s = .7$ (images available from Oxford's Visual Geometry Group's webpage). Fig. 5 presents the proportion of inliers in terms of the set size provided by the outlier rejection methods, where the curves were obtained by varying τ_s in our method and varying the bin sizes of the Hough transform. Notice that for sets of equal size, the use of shape context for rejecting outliers provides a higher inlier ratio than Hough clustering.

4.2. Non-rigid Motion

We also consider the problem of non-rigid motion in Fig. 6, where we show the correspondences provided by the outlier rejection methods. For this problem, one wants to find a good compromise between distinctiveness and robustness when deciding on the values of τ_s and the histogram bin sizes for the semi-local feature and Hough, respectively. We chose those values based on the curves in Fig. 5, and they are $\tau_s = 0.7$ for our method, and for Hough we have 15^o for rotation bin size, a factor of 2 for scale, and 0.15 times the maximum model diameter for translation. Note how the use of shape context not only allows for a higher inlier ratio than Hough clustering, but also finds inliers over the whole 'snake of cans' model that suffered a non-rigid deformation.

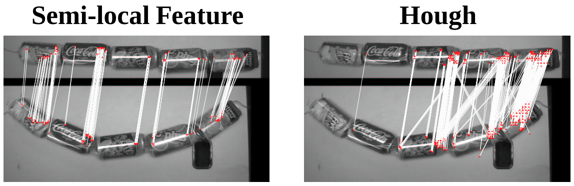

Figure 6. **Non-rigid motion using the 'snake of cans' model. The left image shows the correspondences (white lines) from the semi-local features, and the right depicts the correspondences from Hough clustering.**

Acknowledgement We wish to thank Sven Dickinson for helpful discussions.

References

[1] S. Belongie et al. Shape matching and object recognition using shape contexts. *IEEE PAMI*, 24(24):509–522, 2002.

[2] G. Carneiro and A. Jepson. Multi-scale phase-based local features. In *IEEE CVPR*, Madison, Wisconsin, USA, 2003.

[3] S. Dickinson et al. From volumes to views: An approach to 3-d object recognition. *VGIP: Image Understanding*, 55(2):130–154, 1992.

[4] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE PAMI*, 13(9):891–906, 1991.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[6] M. Lades et al. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.

[7] D. Lowe. Object recognition from local scale-invariant features. In *IEEE ICCV*, pages 1150–1157, Corfu, Greece, 1999.

[8] D. Lowe. Local feature view clustering for 3d object recognition. In *IEEE CVPR*, 2001.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *IEEE CVPR*, Madison, Wisconsin, USA, 2003.

[10] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, pages 754–760, 1998.

[11] B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31–50, 2000.

[12] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE PAMI*, 19(5):530–535, 1997.

[13] A. Shokoufandeh et al. View-based object recognition using saliency maps. *Image and Vision Computing*, 17:445–460, 1999.

[14] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *ECCV*, pages 814–828, 2000.

[15] P. Torr and D. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24(3):271–300, 1997.