CODE: Coherence Based Decision Boundaries for Feature Correspondence

Wen-Yan Lin, Fan Wang, Ming-Ming Cheng, Sai-Kit Yeung,
Philip H. S. Torr, Minh N. Do, Fellow, IEEE, and Jiangbo Lu, Senior Member, IEEE

Abstract—A key challenge in feature correspondence is the difficulty in differentiating true and false matches at a local descriptor level. This forces adoption of strict similarity thresholds that discard many true matches. However, if analyzed at a global level, false matches are usually randomly scattered while true matches tend to be coherent (clustered around a few dominant motions), thus creating a coherence based separability constraint. This paper proposes a non-linear regression technique that can discover such a coherence based separability constraint from highly noisy matches and embed it into a correspondence likelihood model. Once computed, the model can filter the entire set of nearest neighbor matches (which typically contains over 90% false matches) for true matches. We integrate our technique into a full feature correspondence system which reliably generates large numbers of good quality correspondences over wide baselines where previous techniques provide few or no matches.

Index Terms—Feature matching, wide-baseline matching, visual correspondence, RANSAC

1 INTRODUCTION

Correspondence between image pairs involve finding the projections of the same scene points in both images. By linking multiple images together, correspondence is a critical first input for many vision systems [1], [2], [3]. As applications vary wildly, correspondence algorithms must accommodate a wide range of baselines and scenarios, e.g. Internet images, noisy infrared images, high resolution images, low-resolution video frames, etc. In addition, the desired stability of downstream systems requires correspondence algorithms to find as many matches as possible while keeping false matches to a minimum. The simultaneous need for large numbers of matches, robustness and flexibility places a heavy strain on correspondence algorithms and has motivated intensive research along this direction.

To date, feature matching [4], [5], [6] is the correspondence solution of choice for many computer vision systems. While lacking the correspondence density offered by optical flow alternatives [7], [8], [9], [10], feature matchers provide an attractive blend of wide baselines, fast speed and fine localization. The goal of feature matching is to correspond sparsely scattered, distinctive key-points. Distinctiveness is enhanced by consolidating local patch information into transformation invariant descriptors and matching decisions are based on descriptor comparison. While generally effective, feature correspondence typically discards many true matches to suppress the number of false matches [11], [12], [13]. This can cause a paucity of matches which negatively impacts downstream algorithms. The problem is especially severe at the extremes of feature correspondence’s working range shown in Fig. 1A). Lowering match acceptance thresholds provides many more (sometimes by a few orders of magnitude [11]) true matches. However, false matches increase more rapidly in number, leading to the mess shown in Fig. 1B).

Despite appearances, matches in Fig. 1B) may actually be separable into true and false matches. This is because true matches tend to be coherent, with neighboring pixels sharing similar motions, while wrong matches are usually randomly scattered. This leads us to propose a novel approach termed COherence based DEcision boundaries or CODE, which computes a coherence based partition of the potential correspondence space into true and false regions. At first glance, this approach appears self-contradictory, since coherence estimation and thus the partition is intrinsically coupled with the unknown (or highly imperfectly estimated) correspondence. However, we observe that a coarse coherence estimate is sufficient as we can eventually rely on the feature’s spatial localization for fine matching. This permits quasi-decoupling of the coherence and correspondence steps. In CODE, we develop a robust, non-linear regression formulation by treating feature matches as noisy data points. This is used to model a coherence cost (or the likelihood) for every possible motion. As the regression model only needs to be coarsely estimated, it can be effectively approximated from very noisy point correspondences obtained from low acceptance thresholds. Once estimated, the model forms a coherence based decision boundary for verifying correspondence hypotheses. The result is efficient, coherence enforcement for feature correspondences.

The proposed approach has some distinctive theoretical and practical advantages. 1) Global motion modeling: By applying our optimization on the bilateral domain which contains both spatial and motion coordinates $[x, y, u, v]$,
A) Traditional A-SIFT feature matches

B) A-SIFT feature matches with lowered acceptance thresholds

C) A-SIFT feature correspondence with CODE regression

Fig. 1: Feature matching for an example image pair. A) A-SIFT [6], a highly regarded wide-baseline feature matcher. B) Relaxed acceptance thresholds result in many more true matches hidden in a pile of false matches. C) Our fusion of A-SIFT with the proposed technique termed CODE. CODE regression can model the motion adequately even under such noisy matches in B). Note that only the right side of the main arch structure is visible in the second image.

a smooth curve can model spatially discontinuous motion (conceptually similar to edge preserving bilateral filters [14]). This dispenses with the need to handle discontinuities by weakening the smoothing function or demanding explicit discontinuity detection. Instead, a global as-smooth-as-possible function accommodates a wide variety of motions, making modeling both flexible and robust. 2) Computational tractability: We show that an as-smooth-as-possible regression function can be formulated as a convex cost with a guaranteed global minimum. This avoids initialization problems and facilitates mathematically elegant solutions. 3) Soft modeling: Given our input feature matches’ noisy nature, modeling motion as a hard one-to-one mapping assignment between images is unlikely to be accurate. The proposed likelihood model expresses this ambiguity, and it adds a layer of robustness since (unlike in one-to-one modeling). As a result, a mistakenly high score for an incorrect motion need not affect the score of the true motion, and this allows graceful degradation on difficult scenes. 4) Efficiency: Naive correspondence likelihood modeling is potentially very expensive. If the likelihood of every possible motion for every pixel is stored as a matrix, even a small $640 \times 480$ image, with potential horizontal and vertical motion ranges of $[-640, 640]$ and $[-480, 480]$, would require a matrix of size $640 \times 480 \times (2 \times 640) \times (2 \times 480)$. If expressed as a likelihood based regression model, we can estimate every matrix entry with a few hundred variables. This makes correspondence likelihood modeling computationally tractable.

In practice, CODE consists of a number of smooth likelihood decision boundaries for validating feature correspondences. These can be rapidly computed from noisy correspondences hypotheses (50% wrong matches are acceptable at this stage). Once computed, the decision boundaries can filter the entire nearest-neighbor matching set (which often has over 90% outliers) for inliers, with the one-to-one correspondence forming a natural complement to the likelihood based CODE boundaries. This process retains many more correspondences than traditional descriptor based thresholding methods while eliminating nearly all wrong matches. Our regression can be computed in a few seconds and since verification is cheap, the proposed technique scales efficiently to large numbers of features. In practice, CODE filtering time is $O(N)$ where $N$ is the pre-defined number of feature matches used in the regression.

To make a complete feature correspondence package, we fuse CODE with our re-implementation of GPU A-SIFT [15] and Muja and Lowe’s [16] fast approximate nearest neighbor matching. As CODE’s overhead is low, our full system is faster than the original A-SIFT implementation [6], while procuring many more matches. An example is shown in Fig. 1C).

This current work extends our previous conference papers [11], [17]. In this paper, we provide a thorough exposition of the proposed technique, and improve its scalability to handle large numbers of matches through a theoretical approximation. In addition, we perform a number of new and expanded experiments to validate the proposed technique against a range of state-of-the-art methods. In summary, the main contributions are listed as follows:

- We propose CODE, a principled approach expressing the motion smoothness constraints as a matching likelihood estimate of every possible motion of every image pixel. CODE can be efficiently computed from highly noisy feature matches and forms an effective mechanics of discerning the difference between true and false matches.
- We integrate CODE with A-SIFT to create a feature matching system which yields high quality matches at wide baselines where previous techniques provide few or no matches. Despite the aggressive matching, it avoids matching images of different scenes/objects. We share our optimized implementation, in the form of executable files, to the research community at: http://www.kind-of-works.com/home/code
- We demonstrate our high numbers of reliable correspondences bring compelling improvements when integrated into the existing structure-from-motion systems.

1.1 Related Work

This paper owes a significant debt to an array of pioneering research on affine invariant, wide-baseline features [6], [18], [19]. However, the original works could not fully exploit their features’ invariance due to difficulty of identifying false matches. This fact forced these methods to adopt very strict similarity criteria which discarded many true match hypotheses. By modeling the true underlying motion from very noisy matches, our CODE technique can reliably differentiate true and false matches. This results in a flood of good matches shown in Fig. [...]
Similar to optical flow [7], [8], [9], [10], [20], we use motion smoothness to facilitate correspondence matching. However, we retain feature correspondence’s design philosophy, trading matching density and fineness in favor for wide-baselines, higher speeds and robustness. As such our technique may ignore subtle motion details which might be retained by optical flow algorithms. If desired, density and fine motions may be recovered through a subsequent re-computation, but this is beyond the current paper’s scope.

Our formulation builds on the motion coherence framework [21], [22]. Unlike the smoothness prior enforced in most optical flow algorithms which directly penalizes motion differences, motion coherence seeks the smoothest continuous motion field consistent with observed data. The global data integration has repeatedly demonstrated impressive levels of stability [23], [24], [25]. However, the original formulation does not accommodate motion discontinuities and is vulnerable to local minima. By adapting the motion coherence into a motion discontinuity-preserving, global regression technique, we avoid these problems while maintaining its robustness.

Edge-preserving bilateral filters [14] were a major inspiration for this work. However, unlike bilateral filtering which is a local operation, our regression on the bilateral domain computes a global model. This allows us to connect information from across the image for robustness to outliers. In addition, while bilateral filters return a single value at each pixel, CODE will return a continuous function encoding a distribution of likelihoods for all possible motion values a pixel can take.

We also draw inspiration from previous attempts to fuse smoothness constraints with a sparse feature correspondence. These include graph matching [13], [25], [26], [27], [28], bounded distortion [29] and mesh-based correspondence reasoning [30], with a number of approaches [29], [30] explicitly focused on removing false matches. However, the graph and distortion based techniques [29], [30] are vulnerable to local minima. Further, their computational cost increases with the number of features, making the techniques less scalable to high resolution imagery. While scalability and non-convexity are problems with most correspondence formulations, assuming pre-computed, albeit noisy correspondence creates a sub-problem which we show is amenable to convex regression modeling, thus avoiding both these issues.

In practical terms, a family of RANSAC techniques [31], [32], [33], including the recent branch-and-bound formulation [34], share our goal of removing false matches. However, the basic RANSAC formulation is designed for small (e.g. 10 - 20 variables) linearizable models, making them application specific. Our general smoothness based constraints cannot be enforced with RANSAC, thus motivating our investigation of regression techniques. In practice, it is best to use our method to boost the number of true matches first, then followed by RANSAC to estimate application specific parameters when appropriate.

Also worth mentioning are region growing correspondence algorithms [35], [36] which provide quasi-dense correspondence while handling large baselines and occlusions. These techniques provide many more correspondences than ours. However our solution retains the feature based techniques’ innate advantage in speed, correspondence accuracy and handling large scale changes.

Finally we emphasize that different from the objectives of SIFT flow [37] or object matching [38], we do not desire correspondence between different objects. Given two images of physically different objects, our desired result is no correspondence.

2 Our Approach

This section outlines our general formulation for coherence based data regression. We begin by designing a multifunction regression in Sect. 2.1 based on motion coherence [21], [22]. Sect. 2.2 explains the bottlenecks of this approach and proposes an accelerated approximation. We provide a succinct summary of the regression process in Sect. 2.3. Based on the theoretical foundation laid in this section, Sect. 3 applies the regression techniques to the correspondence problem.

2.1 Generalized Coherence for Data Regression

The problem is formulated as fitting a smooth function to observed data points. The fitting function is $f : \mathbb{R}^D \mapsto \mathbb{R}$, where $p \in \mathbb{R}^D$ and $q$ are its domain and co-domain. We assume $f(p)$ is a linear combination of $K$ smooth functions $\{f_k(p)\}_{k=1,2,...,K}$, such that

$$q = f(p) = \sum_{k=1}^{K} a_k(p)f_k(p) ,$$

where $a_k(p)$ are known weighting functions over the $D$-dimensional domain of $p$, which provide formulation flexibility we exploit later.

The observed data consists of $N$ noisy scalar values $\{\tilde{q}_j\}$ at corresponding locations $\{p_j\}$, which are assumed to be noisy observations of $f(p)$, such that

$$\tilde{q}_j = f(p_j) + n_j = \sum_{k=1}^{K} a_k(p_j)f_k(p_j) + n_j ,$$

with $n_j$ representing noise. With respect to Fig. 2, p is a spatial point on the two-dimensional X-Y domain. The regression function $f(p)$ is represented by a continuous valued surface, while $\tilde{q}_j$ is the observed value for a spatial point $p_j$.

Each individual $f_k(\cdot)$ function is composed of two terms:

$$f_k(p) = H_k + \phi_k(p) .$$

1. The formulation is unchanged if $q$ is a vector. However for computational speed, $q$ is a scalar throughout this paper.
where $\bar{\phi}_k(\omega)$ denotes the Fourier transform of a function $\phi_k(\cdot)$, while $\tilde{g}(\omega)$ is the Fourier transform of a Gaussian function with a spatial standard deviation $\gamma$. Hence, Eqn. (4) achieves smoothness by penalizing high frequency terms.

Our goal is to find the smoothest possible set of $f_k(\cdot)$ functions consistent with the given data points $\{\mathbf{p}_j, \hat{q}_j\}$. This is expressed as the following energy minimization:

$$E = \sum_{j=1}^{N} C(\hat{q}_j - f(\mathbf{p}_j)) + \lambda \sum_{k=1}^{K} \Psi_k$$

(5)

$$= \sum_{j=1}^{N} C(\hat{q}_j - \sum_{k=1}^{K} a_k(\mathbf{p}_j)f_k(\mathbf{p}_j)) + \lambda \sum_{k=1}^{K} \Psi_k.$$  

Here, $C(\cdot)$ represents the Huber function as used in [17]:

$$C(z) = \text{Huber}(z) = \begin{cases} z^2 & \text{if } |z| \leq \epsilon \\ 2\epsilon |z| - \epsilon^2 & \text{if } |z| > \epsilon \end{cases}$$

and $\lambda$ is the weight given to the smoothness constraint $\Psi_k$.

Directly minimizing $E$ with respect to functions $f_k(\cdot)$ appears intractable as $\{f_k(\cdot)\}$ are continuous functions. However, we can reduce the problem to an optimization over a finite number of variables as shown below.

Note the Fourier transform relation, $\phi_k(\mathbf{p}) = \int_{\mathbb{R}^D} \bar{\phi}_k(\omega)e^{2\pi i \langle \mathbf{p}, \omega \rangle} d\omega$. We know that at the minimum of the energy $E$ of Eqn. (5), its derivative is zero. Hence,  

$$\frac{\delta E}{\delta \bar{\phi}_k(\mathbf{z})} = 0, \forall \mathbf{z} \in \mathbb{R}^{D}, k \in \{1, 2, \ldots, K\}.$$  

(7)

$$\int_{\mathbb{R}^D} \frac{\delta^2}{\delta \bar{\phi}_k(\mathbf{z})} \bar{\phi}_k(\mathbf{z})e^{2\pi i \langle \mathbf{p}, \omega \rangle} d\omega = 0$$

$$\sum_{j=1}^{N} \mathbf{w}_k(j)e^{2\pi i \langle \mathbf{p}_j, \mathbf{z} \rangle} + 2\lambda \frac{\bar{\phi}_k(-\mathbf{z})}{\bar{\mathbf{g}}(\mathbf{z})} = 0$$

where $\mathbf{w}_k$ is a $N \times 1$ vector that serves as a place-holder for more complicated terms.

Rearranging the terms gives

$$\bar{\phi}_k(\mathbf{z}) = \bar{\mathbf{g}}(\mathbf{z}) \sum_{j=1}^{N} \mathbf{w}_k(j)e^{-2\pi i \langle \mathbf{p}_j, \mathbf{z} \rangle},$$

(8)

Taking the inverse Fourier transform of Eqn. (8), we can write our continuous functions $\{\phi_k(\mathbf{p})\}$ in terms of a finite set of $N$-dimensional vectors $\{\mathbf{w}_k\}$,

$$\phi_k(\mathbf{p}) = \sum_{j=1}^{N} \mathbf{w}_k(j)g(\mathbf{p}, \mathbf{p}_j) = \sum_{j=1}^{N} \mathbf{w}_k(j)e^{-\frac{i\mathbf{p} \cdot \mathbf{p}_j}{\sigma^2}},$$

(9)

\[ \forall k \in \{1, 2, \ldots, K\}, \]

where $g(\mathbf{p}, \mathbf{p}_j)$ is a Gaussian radial basis function and $\{\mathbf{w}_k(\cdot)\}$ are unknown variables. Thus Eqn. (9) states that the smoothest possible functions $\{\phi_k(\cdot)\}$ which minimize the energy in Eqn. (5) must lie in the space spanned by $N$ radial basis functions $\{g(\mathbf{p}, \mathbf{p}_j)\}$.

For the smoothness constraint, substituting Eqn. (8) into Eqn. (5) allows the continuous regularization function $\Psi_k$ to be expressed in terms of $\mathbf{w}_k$

$$\Psi_k = \mathbf{w}_k^T \mathbf{G} \mathbf{w}_k,$$

(10)

where $G$ is a symmetric matrix with its elements given as

$$G(i, j) = g(\mathbf{p}_i, \mathbf{p}_j) = e^{-\|\mathbf{p}_i - \mathbf{p}_j\|^2/2\gamma^2}.$$  

(11)

Substituting Eqns. (9) and (10) into Eqn. (5) yields

$$\arg \min_{\{f_k(\mathbf{p})\}} \sum_{j=1}^{N} C(\hat{q}_j - f(\mathbf{p}_j)) + \lambda \sum_{k=1}^{K} \Psi_k$$

$$= \arg \min_{\{f_k(\mathbf{p})\}} \sum_{j=1}^{N} C\left(\hat{q}_j - \sum_{k=1}^{K} a_k(\mathbf{p}_j)f_k(\mathbf{p}_j)\right) + \lambda \sum_{k=1}^{K} \Psi_k$$

$$= \arg \min_{\{\mathbf{w}_k, H_k\}} \sum_{j=1}^{N} C\left(\hat{q}_j - \sum_{k=1}^{K} a_k(\mathbf{p}_j)\left(H_k + \sum_{i=1}^{N} \mathbf{w}_k(i)g(\mathbf{p}_j, \mathbf{p}_i)\right)\right) + \lambda \sum_{k=1}^{K} \mathbf{w}_k^T \mathbf{G} \mathbf{w}_k.$$  

(12)

where the energy is dependent only on a finite number of variables, i.e., $\{\mathbf{w}_k\}$ and $H_k$. Since $G$ is a Gram matrix [22], this makes the coherence term $\Psi_k$ in Eqn. (10) convex. As the Huber loss function $C(\cdot)$ is also convex and the sum of convex functions is convex, the overall energy minimization problem in Eqn. (12) is convex. Therefore, a gradient descent minimization of Eqn. (12) will lead to a guaranteed global minimum.

Based on the estimated variables $\{\mathbf{w}_k\}$ and $H_k$, the global regression function $f(\mathbf{p})$ is constructed from $\{f_k(\mathbf{p})\}$:

$$f(\mathbf{p}) = \sum_{k=1}^{K} a_k(\mathbf{p})f_k(\mathbf{p}) = \sum_{k=1}^{K} a_k(\mathbf{p})(H_k + \phi_k(\mathbf{p}))$$

$$= \sum_{k=1}^{K} a_k(\mathbf{p})\left(H_k + \sum_{i=1}^{N} \mathbf{w}_k(i)g(\mathbf{p}_j, \mathbf{p}_i)\right).$$

(13)

This formulation mirrors that developed in our previous conference paper [17]. For applications of this global regression approach to data fitting and image warping, we encourage interested readers to refer to [17].

### 2.2 Accelerated Coherence Based Global Regression

Examining the formulation presented in Sect. 2.1 one can find that the length of $\mathbf{w}_k$ vectors and hence the number of variables in Eqns. (8) and (12) increase linearly with $N$, the number of observed data points under consideration. This creates a computational burden when $N$ is large. Motivated to address this challenge, we propose an approximation that decouples this linear computational dependency, with its full derivation given in Appendix A.

Consider $f_k(\mathbf{p})$ in Eqn. (13), the $k$-th smooth function to be estimated. If many data points in the input set $\{\mathbf{p}_i\}$

2. This may partially explain the success of radial basis functions in learning networks [39].
are adjacent, a good approximation of $f_k(p)$ can be given as follows,

$$f_k(p) \approx \tilde{f}_k(p) = H_k + \sum_{j=1}^{M} \tilde{w}_k(j)g(p, \tilde{p}_j),$$  \hspace{1cm} (14)

where $\tilde{w}_k$ is an $M$-dimensional vector ($M \ll N$), and $\{\tilde{p}_j\}$ are $M$ representative cluster centroids distributed across the space occupied by the original points $\{p_j\}$. These centroids are usually obtained by K-means clustering [40] on $\{p_j\}$. If most of the $p_j$ points have near duplicates in the set $\{\tilde{p}_j\}$, the function $\tilde{f}_k(p)$ will be able to closely approximate all possible values of the original $f_k(p)$ function in Eqn. (13).

Interestingly, the approximated regression functions $\{\tilde{f}_k(p)\}$ can be easily integrated into the global energy function [12], which becomes

$$\arg\min_{\{\tilde{w}_k, H_k\}} \sum_{j=1}^{N} C \left( \tilde{q}_j - \sum_{k=1}^{K} a_k(p_j)\tilde{f}_k(p_j) \right) + \lambda \sum_{k=1}^{K} \tilde{w}_k^T \tilde{G} \tilde{w}_k,$$  \hspace{1cm} (15)

where $\tilde{G}$ is an $M \times M$ matrix, with $\tilde{G}(i,j) = g(\tilde{p}_i, \tilde{p}_j)$ and the overall cost remains convex. Note that in this approximation, the values of $\lambda$ and $M$ are not coupled, i.e., the same $\lambda$ can be properly used for different settings of $M$.

This clustering-based approximation is actually useful in accelerating many computer vision tasks, where closely located pixels create many redundant variables. Apart from this paper, we believe our approximation can also aid the original motion coherence [23] and its derivations [23], [24].

### 2.3 Summary of the Coherence Based Regression

To provide a succinct summary of the regression process, we give a recap of it here. Given $N$ noisy scalar values $\{\tilde{q}_j\}$ at corresponding locations $\{p_j\}$, where $p_j$ is a $D$-dimensional vector, we can explain the observed data $\{p_j, \tilde{q}_j\}$ with a continuous function $\tilde{f}(p)$. The function $\tilde{f}(p)$ is composed of a linear combination of $K$ smooth functions, which returns a value for each query point $p$,

$$q = \tilde{f}(p) = \sum_{k=1}^{K} a_k(p) \tilde{f}_k(p),$$  \hspace{1cm} (16)

where $\{a_k(p)\}$ are known weighting functions over the $D$-dimensional domain $p$. The function $\tilde{f}_k(p)$ is in turn parametrized by $\{\tilde{w}_k, H_k\}$ variables:

$$\tilde{f}_k(p) = H_k + \sum_{j=1}^{M} \tilde{w}_k(j)g(p, \tilde{p}_j) = H_k + \sum_{j=1}^{M} \tilde{w}_k(j)e^{-\|p - \tilde{p}_j\|^2/\gamma^2},$$  \hspace{1cm} (17)

which can be estimated by a gradient descent minimization of the convex cost

$$\arg\min_{\{\tilde{w}_k, H_k\}} \sum_{j=1}^{N} C \left( \tilde{q}_j - \tilde{f}(p_j) \right) + \lambda \sum_{k=1}^{K} \tilde{w}_k^T \tilde{G} \tilde{w}_k =$$

$$\arg\min_{\{\tilde{w}_k, H_k\}} \sum_{j=1}^{N} C \left( \tilde{q}_j - \sum_{k=1}^{K} a_k(p_j)\tilde{f}_k(p_j) \right) + \lambda \sum_{k=1}^{K} \tilde{w}_k^T \tilde{G} \tilde{w}_k,$$  \hspace{1cm} (18)

A Huber loss function $C(.)$ is adopted to evaluate the data fitting quality. A summary of the estimated variables and user-defined parameters are given in Table I.

### 3 Coherence Based Decision Boundaries

With the general formulation for coherence based data regression presented in Sect. 2, we now apply the regression techniques to the correspondence problem. The key idea is to estimate a coarse coherence based model from a sparse set of very noisy feature matches. This model quantifies the coherent matching evidence for every possible motion, providing a coherence based decision boundaries for verifying correspondence hypotheses. More specifically, we achieve this through two regression functions which act as a chain of cascaded filters for eliminating wrong feature matches. We describe individual modules below and elaborate on design decisions and properties in Sect. [51].

**Matching likelihood boundaries**: What constitutes coherent motion? We consider a local motion is coherent if it satisfies either of these criteria: a) a concentrated cluster of local features making the motion; and b) many features over a large spatial extent making the motion. A visual illustration is shown in Fig. [3].

To enforce this coherence, we formulate a regression based likelihood function in which each given feature match is a noisy observed data of the form:

$$\{p_j = [x_j, m_j, x_j + m_j, o_j], \tilde{q}_j = 1\},$$  \hspace{1cm} (19)

where each match indexed by $j$ hypothesizes a "1" value at location $p_j$. Here $x_j = [x_j, y_j]$ and $m_j = [u_j, v_j]$ are two-dimensional vectors representing image coordinates and motion vectors respectively, while $o_j$ is a $4 \times 1$ vector representing the relative affine orientation of the matched feature descriptor. From Eqn. (17), we choose a single function $K = 1$ without the $H_k$ bias variable, and set $a_k(p) = 1, \forall p$. This leads to a regression function

$$\tilde{f}(p) = \sum_{j=1}^{M} \tilde{w}(j)g(p, \tilde{p}_j).$$  \hspace{1cm} (20)

Substituting the regression function (20) into the cost (18), we attain

$$\arg\min_{\tilde{w}} \sum_{j=1}^{N} C(1 - \tilde{f}(p_j)) + \lambda \tilde{w}^T \tilde{G} \tilde{w},$$  \hspace{1cm} (21)

whose minimization estimates the parameters of $\tilde{f}(p)$.

Observe that the smoothness cost $\tilde{w}^T \tilde{G} \tilde{w}$ in Eqns. (4) and (10) not only penalizes un-smooth motions but penalizes all motions by encouraging $\tilde{w}$ to be zero. Thus, the lack

<table>
<thead>
<tr>
<th>Estimated</th>
<th>${H_k}$</th>
<th>optional bias variable in $f_k(p)$</th>
<th>${\tilde{w}_k}$</th>
<th>function variables in $\tilde{f}_k(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>User-defined</td>
<td>$\lambda$</td>
<td>weight of smoothness terms</td>
<td>$\gamma$</td>
<td>std. deviation of smoothing Gaussian</td>
</tr>
<tr>
<td></td>
<td>$\epsilon$</td>
<td>threshold in the Huber function $C(.)$</td>
<td>$K$</td>
<td>number of smooth functions for $\tilde{f}(p)$</td>
</tr>
<tr>
<td></td>
<td>${a_k(p)}$</td>
<td>user defined weighting function</td>
<td>$M$</td>
<td>length of $\tilde{w}_k$</td>
</tr>
<tr>
<td></td>
<td>${\tilde{p}_j}$</td>
<td>representative subsampled points</td>
<td>$N$</td>
<td>number of data points</td>
</tr>
<tr>
<td>Others</td>
<td>$\tilde{f}(p)$</td>
<td>cost regression function in Eqn. (16)</td>
<td>given in Eqn. (18)</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 3: Coherence based separation of true and false matches. Motions are considered coherent if (a) many local points make similar motions or (b) there is broad spatial support for the motion. This is enforced via the likelihood function in Eqn. (21). Correspondingly, feature matches in (c) and (d) do not give coherent motions, as the matches are not consistent in (c), while there are insufficient smoothly moving points to justify a long-range motion coherence model in (d).

of a bias term in \( \tilde{f}(p) \) given in Eqn. (20) means the likelihood function stays at zero, unless forced upward by the data. The likelihood function rises towards one if there exist a cluster of adjacent matches whose collective pull justifies an “un-smooth” peak. Alternatively, if a motion has significant supports across a wide area but with no strong local pull, the likelihood function can also rise to one, since a smooth surface over a large extent can be fitted for these matches by incurring a low smoothness penalty. In contrast, randomly scattered, incorrect matches do not exert sufficient collective pull and are ignored by the regression function. A simplified matching likelihood function is illustrated in Fig. 4.

From noisy feature correspondences, we use Eqn. (21) to fit a robust likelihood function given by \( \tilde{f}(p) \). The acceptance condition for a feature matching hypothesis \( p_i \) is

\[
\text{accept}(p_i) = \text{true} \quad \text{if} \quad \tilde{f}(p_i) > \epsilon_{\text{likelihood}}
\]  

This mechanics can take very noisy data and remove most gross matching errors. However, it lacks fine spatial awareness and some erroneous matches will remain.

Bilaterally varying affine motion boundaries: The matching likelihood function can be considered as a blind cluster discovery mechanics. However, we know that correct motions not only cluster but tend to approximate a piecewise smoothly varying affine model \([23]\). We enforce this knowledge through a bilaterally varying affine regression function, which computes a motion likelihood by checking a hypothetical motion’s consistency against its estimated model.

We focus on the X motion direction first. Observed data take the form:

\[
\text{observed data} = \{p_j, q_j = x_j + u_j\},
\]  

where \( x_j, u_j \) are obtained from the given feature correspondence. The definition of \( p_j \) remains unchanged from the likelihood function formulation in Eqn. (19).

By setting \( K = 3 \) in Eqn. (16) and setting \( a_1(p) = x, a_2(p) = y, a_3(p) = 1 \), we have

\[
\tilde{f}_x(p) = \tilde{f}_1(p)x + \tilde{f}_2(p)y + \tilde{f}_3(p),
\]  

where each \( \tilde{f}_x(p) \) is \( H_k + \sum_{j=1}^{M} \tilde{w}_k(j)g(p, \tilde{p}_j) \) represents an affine motion parameter for the location \( p \). To estimate \( \tilde{f}_k(p) \), from Eqn. (17), we optimize the following cost function

\[
\text{arg } \min_{\{\tilde{w}_k, H_k\}} \sum_{i=1}^{N} C(x_i + u_i - \tilde{f}_x(p_i)) + \lambda \sum_{k=1}^{3} \tilde{w}_k G \tilde{w}_k.
\]  

Similarly for the Y direction, we have a bilaterally varying affine model as follows,

\[
\tilde{f}_y(p) = \tilde{f}_4(p)x + \tilde{f}_5(p)y + \tilde{f}_6(p),
\]  

which can be computed from the cost

\[
\text{arg } \min_{\{\tilde{w}_k, H_k\}} \sum_{i=1}^{N} C(y_i + v_i - \tilde{f}_y(p_i)) + \lambda \sum_{k=4}^{6} \tilde{w}_k G \tilde{w}_k.
\]

While not a direct likelihood function, the bilaterally varying affine model can be applied to distinguish correct and wrong feature matches through a thresholding step:

\[
\text{accept}(p_i) = \text{true} \quad \text{if} \quad \sqrt{(\tilde{f}_x(p_i) - x_i - u_i)^2 + (\tilde{f}_y(p_i) - y_i - v_i)^2} < \epsilon_{\text{spatial}}.
\]  

In practice, we find this stage gives more refined estimates but lacks the robustness of the preceding likelihood module. Thus, we cascade the affine motion boundary estimation after the likelihood boundary, which has removed most of the grossly incorrect matches.

3.1 Important Design Considerations and Properties

Having outlined our basic formulation, we will now delve into the properties and design considerations of our coherence-based regression functions.

Global smoothness: It is enforced in all the functions. Unlike a local smoothness penalization which directly penalizes (motion) differences between neighbors, global smoothness enforced in our formulation seeks the globally smoothest function. By integrating information from across
Fig. 5: CODE applied to A-SIFT features [6]. A and B demonstrate CODE’s ability to handle large motion discontinuities and occlusions, while C demonstrates its ability to handle large rotations. In alternative display, we color code feature matches according to their spatial locations on the left image. The matching transfers these color codes to the right image. To better illustrate the large rotation, C uses a vector display.

the image, global functions provide a high level of robustness, allowing the regression functions to be computed from very noisy input as shown in Fig. 4.

Domain choice: The choice of the domain for \( p \) in the as-smooth-as-possible regression formulation can be arbitrary. We are motivated to choose the domain shown in Eqn. (19) out of three concerns:

Discontinuity preservation: Typically, global smoothness interferes with discontinuity preservation. Consider a function \( f_k(p) \) where \( p \) represents pixel coordinates, while the range of \( f_k(p) \) represents motion. The smoothness (or infinite differentiability) implies a continuity constraint:

\[
\lim_{\Delta p \to 0} f_k(p + \Delta p) - f_k(p) = 0, \quad \forall p = [x, y]^T \in \mathbb{R}^2,
\]

which forces the function value in the neighborhood of \( p \) to be similar. This means smooth functions must incur large errors at discontinuous motion boundaries.

We tackle this problem by using the bilateral domain spanning both the spatial and motion dimensions, i.e., \( p = [x, y, u, v]^T \), for conceptual understanding, we use a simpler form than that in (19). This redefines neighbors such that spatially neighboring points with different velocities are no longer adjacent. Therefore, we can assign very different function values to points with adjacent spatial coordinates, while retaining the constraint that \( f_k(p) \) must be smooth. If the motion difference \( (\Delta u, \Delta v) \) between two points in the bilateral domain \( p \) tends to infinity, the point separation also tends to infinity, significantly reducing their influence on each other. This separation holds irrespective of the spatial coordinates, and allows motion discontinuity to be accommodated by a smooth function shown in Fig. 4. Examples of this discontinuity handling are shown in Fig. 5A, B).

Affine smoothness: It has been noted in [23] that because of the local generalization property of affines, motions are smoother when over-parameterized in the affine domain. Thus, we concatenate the affine parameters \( o \) to \( p \) in (19). This allows us to better handle circular motions such as in Fig. 5C).

Symmetry: Finally, for the sake of mathematical symmetry between left and right images, we concatenate the arguably redundant \( x_j + m_j \) term to \( p \) in (19).

Note that we find many different domain choices give good results and the domain’s chosen in this paper serve as guidelines rather than canonical selections.

Multiple motion hypotheses: Computing regressions on the bilateral domain means estimating a function value for every \([x, y, u, v]^T\). This forces the function to consider every possible value that each pixel can take. Thus, unlike bilateral filters which estimate a single value for each pixel, CODE regression functions encode the likelihood for every possible motion each pixel can take. This soft-modeling adds a significant layer of robustness, as mistakenly giving a high score to an incorrect motion need not affect the score of the true motion. The computed regression functions and the resulting decision boundaries of CODE are subsequently used to verify feature correspondence hypotheses, with the hard, sub-pixel accurate feature matches forming a natural complement to the soft likelihood models.
Convex cost: The cost functions for estimating likelihood and bilaterally varying affine boundaries are convex. This avoids the need for initialization and allows our algorithm to handle challenging motions such as the large rotation shown in Fig. 5C).

3.2 Implementation

Our algorithm begins with feature matching. View-invariant features are computed through a GPU A-SIFT with nearest-neighbor matching computed by FLANN. Typically, feature matchers choose to accept matches based on a ratio test with a threshold setting \( \tau = 0.6 \). i.e., a match is accepted if its nearest-neighbor difference is at least 0.6 times smaller than its second nearest neighbor. However, this can lead to few matches as shown in Fig. 1A). In this paper, we take selected matches obtained with a weaker threshold setting \( \tau = 0.86 \) as the input. While matching results appear like a mess in Fig. 1B), the weaker threshold provides many true matches, from which information can be mined.

The selected matches are used to compute likelihood boundaries through the regression step in Eqn. (21). The current set is filtered by the likelihood boundaries and acceptable matches are used to compute bilateral affine motion boundaries in Eqns. (25) and (27). Once the regression functions and the resulting decision boundaries of CODE are computed, all matches obtained by setting \( \tau = 1.0 \) are passed through the cascaded filters. The finally accepted feature matches are output matches. The minimizer used is the Ceres solver and the process is illustrated in Fig. 6.

In theory, better results can be obtained through iterations, where the output matches form a new set of selected matches. In practice, we find the gain is small and hence restrict the study to a single iteration in this paper.

Parameter settings: To accommodate images of different sizes, the Hartley normalization is applied to the selected matches such that they have zero mean and average distance from the center of \( \sqrt{2} \). The user-defined parameters in Table 1 for the likelihood boundaries are \( \lambda = 1, \gamma = 1, \epsilon = 0.1, K = 1, M = 100, N = 30,000 \), with the set \( \{\bar{p}_i\} \) being centroids obtained from K-means clustering of the selected data. Matches are accepted with an \( \epsilon_{\text{likelihood}} = 0.6 \) in Eqn. (22). The user-defined parameters for the bilateral affine boundaries are \( \lambda = 1, \gamma = 1, \epsilon = 0.1, K = 3, M = 100, N = 1000 \). Matches are accepted with an \( \epsilon_{\text{spatial}} = 0.01 \) in Eqn. (28). The design details of the regression function \( f(p) \) and weighting functions \( \{a_k(p)\} \) are provided in Eqns. (20), (24) and (26).
Fig. 7: Comparison between the original A-SIFT matcher [6] and A-SIFT w CODE, applying the proposed CODE regression to A-SIFT. For this test, we constructed a challenging dataset of static scene images. It includes representative images from a variety of prior papers [11], [13], [24], [29]. Set A has limited out-of-plane rotation, while Set B features very large viewpoint changes. CODE recovers many true feature matches that A-SIFT discards while maintaining high precision. This gain is especially noticeable for wide-baseline image pairs contained in Set B.

algorithm can interpolate and extrapolate to obtain more correspondences. To simulate this possibility, each match is dilated with a radius of 12 pixels (on its 640 × 480 rescaled image) to define a matching area. The total matching area covered by the A-SIFT is then measured against that of A-SIFT w CODE, and the resulting area ratio is reported as Area %. Even on this measure, CODE improves A-SIFT quite significantly and allows to cover an appreciably larger area. Note that this measure under-weights more difficult scenes which have small areas of overlap. Therefore, we also report Total Failures which gives the fraction of the tested scene pairs with no matches returned from an algorithm. In the absence of ground-truth, an approximate Recall is given as (1 - Total Failures). Finally, F-Number is calculated as a summary statistic:

\[ F_{-\text{number}} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \]  

Av. Time and Max. Time of 1.96 and 2.76 seconds reflect the computational overhead of integrating CODE. In fact, such a computational overhead is marginal when compared with the runtime of the original A-SIFT matcher (about 20 seconds) and our re-implementation (about 10 seconds). Runtimes are measured on a laptop with Intel i7CPU (2.4GHz), 8GB RAM and NVIDIA GPU GeForce GTX 660M.

4.2 Structure from Motion

In the last subsection, we performed a direct comparison between the A-SIFT matcher and our integration of CODE in feature matching. The experiments justified CODE significantly improves A-SIFT. In this subsection, we choose to evaluate the two different feature matchers by integrating them into a complete end-to-end Structure from Motion (SfM) system, where feature matching is a fundamental early processing module.

Structure from Motion (SfM) seeks to infer 3D models from 2D images. It typically comprises two main stages: a) camera position computation where camera poses are inferred from feature matches such as A-SIFT, and b) a multi-view stereo reconstruction step, where images with known relative positions are fused into a full 3D model. An interesting fact is a test dataset, designed more for one of the stages, can be difficult for the other. For example, images from the multi-view stereo database [43] permit high quality dense multi-view reconstruction, if camera positions are assumed known; but the same set of images are less amenable to camera position recovery, due to difficulty in computing reliable feature correspondences. Hence, most SfM systems fail to reconstruct significant 3D scene sections. We attempted to reconstruct these scenes by integrating A-SIFT w CODE in the well-regarded Visual SfM system [3], [44], [45], [46], [47]. While reconstruction results are not always perfect, we can recover a complete 3D structure for all multi-view scenes. An example is shown in Fig. 8.

This experiment serves two purposes. First, it demonstrates that CODE’s high quality feature matching results can translate into meaningful improvements on other computer vision tasks. Second, feature correspondence is indeed critical to the final 3D model’s quality. Our experiment
provides an indirect visualization of feature correspondence quality on challenging multi-view stereo datasets like [43]. Here each of the six image sequences has twenty-four images, leading to a total of 1656 pairwise matches. [Lu: Since the 3D reconstruction task here is computationally much heavier than all the other tests, we ran this comparative experiment on a desktop PC with an NVIDIA GeForce GTX980-Ti GPU.] The complete 3D reconstruction results for A-SIFT with and without CODE are shown in Appendix C in the supplementary material.

### 4.3 Comparison to Other Correspondence Algorithms

Thus far evaluations have focused on CODE’s ability to facilitate A-SIFT feature correspondence. We now turn our attention to comparing our full A-SIFT w CODE to other matching techniques. In this section, we drop the cumbersome name A-SIFT w CODE and call our technique CODE. The most related algorithms are those which seek to remove or refine incorrect feature matches. Two recent examples are bounded distortion (BD) by Yaron et al. [29], and Mode Seeking by Chao et al. [13], [26]. At the time of writing, they represent a fair representation of the state-of-the-art. For experimental completeness, we also evaluate quasi-dense correspondence NRDC by HaCohen et al. [36] and agglomerative clustering ACC by Cho et al. [38]. As NRDC seeks dense correspondence, computation time on large images is prohibitively long. Because of this, unlike other algorithms, NRDC is run on down-sampled images of 640 × 480 resolution. As stated in [36], NRDC requires different settings for evaluation on the co-recognition dataset [49]. However, this impacts its general performance. Thus, we evaluate NRDC using its default parameters, and also NRDC(t) tuned for the co-recognition dataset. Next, we conduct evaluations for these competing algorithms on three datasets, and we present the evaluation protocol and experimental results for each of these datasets one by one.

1) **Co-recognition dataset** [49] consists of six image pairs where significant sections of the scene are re-shuffled. It tests an algorithm’s ability to handle large independent motions while providing large numbers of correspondence. Ground-truth segmentation boundaries are provided. A match is considered correct if it lies within a 7-pixel radius of the ground-truth segment boundary. The fraction of true matches is tabulated as Av. Precision in Fig. 9. We compute a correspondence area by dilating each match with a 12-pixel boundary [36]. The percentage of the total area overlapping with ground-truth segmentation is the Precision (area), a stricter measure than Av. Precision. Recall (area) is the percentage of ground-truth segmentation detected after dilation. F-number from Eqn. (30) provides a performance summary. When comparing computational time, we exclude the runtime of feature detection and nearest-neighbor matching (our implementation is faster), and focus only on that of a core, correspondence evaluation/filtering algorithm. As NRDC does not use features, we provide its full computational time, and acknowledge this difference by indicating its runtime with a superscript ‘∗’. A more detailed discussion on timing is provided in Sect. 4.4. We report all the results based on the above metrics in Fig. 9.

The area based statistics used for co-segmentation evaluation favors NRDC, a quasi-dense matching technique. However, our sparse feature matching method CODE remains surprisingly competitive, with its Recall (area) performance comparable to NRDC(t). In addition, it retains the computational speed and precision traditionally associated with feature matchers.

2) **Our dataset** consists of 20 images provided with the code of [13], [26], [29] with some additions. The images chosen have wide variations in scale, illumination, viewpoint change and image resolution. This is also the dataset used in Sect. 4.1 except that images are no longer downsampled. Image pairs are divided in two sets. Set A has no large out-of-plane rotational motion and is in the working range of all matching algorithms. Images in Set B have large out-of-plane rotational motions and cannot be meaningfully handled by [29] which utilizes SIFT features. Results are...
Fig. 9: Evaluation on the co-recognition dataset [49]. We use a Recall (area) metric which favors dense correspondence over feature matchers. Despite this, CODE is surprisingly competitive at the co-segmentation task. This is a testament to CODE’s number of correspondences and ability to handle multiple motion models. Example results are visualized in the top panel.

Fig. 11: The ETHZ-toys dataset [50], [51]. We measure the object-retrieval correctness. Mode Seeking [13] tends to find correspondence on image pairs even when there exists no common objects. This produces high recall but low precision. NRDC [36] represents a better balance. In contrast, our algorithm is aggressive in match retrieval but also strict in enforcing correctness, leading to a high F-number.
4.4 Computational Time

It is true that computational efficiency is not the most critical issue when the primary concern is on estimating quality feature matches for an image pair. However, applications like Structure from Motion and image stitching typically require all-pair feature correspondence, where the complexity scales quadratically with the dataset size. As a result, an innocuous 30-second runtime scales to 50 minutes for all-pair matching on a small 10-image set. In this respect, feature matchers have a significant advantage over dense matchers as the heavy cost absorbed in the feature computation is only incurred once per image. This leaves a relatively low recurring time shown in Table 2. Observe that our theoretical time on a 10-image dataset is 12.4 minutes, while NRDC’s is 45 minutes. In addition, our timings reported in Table 2 include many high resolution images, while NRDC used down-sampled 640 × 480 images. Such a clear runtime advantage enables a practical algorithm to leverage the enhanced accuracy offered by high resolution imagery to benefit challenging tasks such as Structure from Motion.

Next, we compare our computational time against other match verification or refinement alternatives like BD [29] and Mode Seeking [13]. From Fig. 9 and Fig. 10 it can be seen that CODE’s average time is typically 5 to 6 times lower than alternatives. However, CODE’s maximum time can be one to two orders of magnitude lower, suggesting a scalability advantage. The picture is clearer in Fig. 12 which shows CODE’s computational cost scales well to handling large numbers of feature matches. This is due to CODE’s generalization ability. As proposed in Sect. 2.2, the computationally expensive regression step can be computed on a subset of M matches, where M is pre-defined. The subsequent match verification step is fast and scales efficiently with the increasing number of input matches N (N ≫ M). This makes the overall cost O(M) ⋅

<table>
<thead>
<tr>
<th>Algo</th>
<th>Av. Precision</th>
<th>Av. Match Num</th>
<th>Area %</th>
<th>Av. Time/s</th>
<th>Max time/s</th>
<th>Total Failures</th>
<th>F-number</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD [29]</td>
<td>0.84</td>
<td>616.76</td>
<td>13.5%</td>
<td>27.51</td>
<td>64.88</td>
<td>2/7</td>
<td>0.657</td>
</tr>
<tr>
<td>Mode Seeking [13]</td>
<td>0.84</td>
<td>416.4</td>
<td>22.07%</td>
<td>8.14</td>
<td>9.98</td>
<td>3/7</td>
<td>0.680</td>
</tr>
<tr>
<td>NRDC(t) [36]</td>
<td>0.95</td>
<td>1843e+03</td>
<td>14.29%</td>
<td>25.69*</td>
<td>31.17*</td>
<td>5/13</td>
<td>0.747</td>
</tr>
<tr>
<td>NRDC [36]</td>
<td>0.95</td>
<td>4855e+04</td>
<td>112.5%</td>
<td>24.7*</td>
<td>31.4*</td>
<td>4/13</td>
<td>0.803</td>
</tr>
<tr>
<td>ACC [38]</td>
<td>0.96</td>
<td>65</td>
<td>18.8%</td>
<td>3.68</td>
<td>6.07</td>
<td>5/13</td>
<td>0.751</td>
</tr>
<tr>
<td>CODE, Ours</td>
<td>0.96</td>
<td>4.41e+03</td>
<td>100%</td>
<td>2.97</td>
<td>6.27</td>
<td>1/13</td>
<td>0.940</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algo</th>
<th>Av. Precision</th>
<th>Av. Match Num</th>
<th>Area %</th>
<th>Av. Time/s</th>
<th>Max time/s</th>
<th>Total Failures</th>
<th>F-number</th>
</tr>
</thead>
<tbody>
<tr>
<td>BD [29]</td>
<td>0.84</td>
<td>616.76</td>
<td>14.29%</td>
<td>25.69*</td>
<td>31.17*</td>
<td>5/7</td>
<td>0.439</td>
</tr>
<tr>
<td>Mode Seeking [13]</td>
<td>0.84</td>
<td>606.07</td>
<td>40.64%</td>
<td>40.39</td>
<td>270.23</td>
<td>3/13</td>
<td>0.815</td>
</tr>
<tr>
<td>NRDC(t) [36]</td>
<td>0.95</td>
<td>4.949e+04</td>
<td>105.8%</td>
<td>24.9*</td>
<td>29.6*</td>
<td>5/13</td>
<td>0.747</td>
</tr>
<tr>
<td>NRDC [36]</td>
<td>0.95</td>
<td>4.855e+04</td>
<td>112.5%</td>
<td>24.7*</td>
<td>31.4*</td>
<td>4/13</td>
<td>0.803</td>
</tr>
<tr>
<td>ACC [38]</td>
<td>0.96</td>
<td>65</td>
<td>18.8%</td>
<td>3.68</td>
<td>6.07</td>
<td>5/13</td>
<td>0.751</td>
</tr>
<tr>
<td>CODE, Ours</td>
<td>0.96</td>
<td>4.41e+03</td>
<td>100%</td>
<td>2.97</td>
<td>6.27</td>
<td>1/13</td>
<td>0.803</td>
</tr>
</tbody>
</table>

Fig. 10: Evaluation of different feature matching techniques with the proposed CODE. A challenging image dataset and a set of metrics as described in Sect. 4.1 are used for this evaluation. The dataset contains full-resolution, representative images primarily from a variety of papers [13], [24], [26], [29]. Set A has limited out-of-plane rotation while Set B has very large viewpoint changes. CODE with constant parameters excels in a wide variety of scenarios as can be seen from the low number of Total Failures and high Precision.
current experiments are limited to sparse feature matching with an emphasis on speed and reliability. As such our solution ignores fine motion details and small independent motions. More visual illustrations are shown in Appendix. However, by reliably modeling general motion, we believe our technique lays a solid foundation for subsequent works to enhance the density and fineness of motion estimates.

ACKNOWLEDGMENTS

We acknowledge Siying Liu and Nguyen Tan-Dat’s help.

REFERENCES

References


Wen-Yan Lin received his PhD degree from the National University of Singapore in 2012, supervised by Prof. Loong-Fah Cheong and Dr. Dong Guo. He subsequently worked for the Institute of Infocomm Research Singapore and Prof. Philip Torr. He is currently a post-doc at the Advanced Digital Sciences Center Singapore.

Fan Wang received the B.Eng. degree in computer science and technology from Lanzhou University, Lanzhou, China, in 2012 and M.S. degree in electrical engineering from the National University of Singapore, Singapore, in 2013. He is with the Advanced Digital Sciences Center, Singapore, as a Software Engineer. His research interests include 3D computer vision and 3D reconstruction.

Ming-Ming Cheng received the PhD degree from Tsinghua University in 2012, supervised by Prof. Shi-Min Hu. Then he did two years research fellow, with Prof. Philip Torr in Oxford. He is currently an associate professor at Nankai University. His research interests include computer graphics, computer vision, and image processing. He has received the Google PhD fellowship award, the IBM PhD fellowship award, etc.

Sai-Kit Yeung is currently an Assistant Professor at the Singapore University of Technology and Design (SUTD), where he leads the Vision, Graphics and Computational Design (VGD) Group. He was also a Visiting Assistant Professor at Stanford University and MIT. Before joining SUTD, he had been a Postdoctoral Scholar in the Department of Mathematics, University of California, Los Angeles (UCLA). He was also a visiting student at the Image Processing Research Group at UCLA in 2008 and at the Image Sciences Institute, University Medical Center Utrecht, the Netherlands in 2007. He received his PhD in Electronic and Computer Engineering from the Hong Kong University of Science and Technology (HKUST) in 2009. He also received his BE degree (First Class Honors) in Computer Engineering and MPhil degree in Bioengineering from HKUST in 2003 and 2005 respectively.

Philip H. S. Torr received his PhD degree from Oxford University. After working for another three years at Oxford, he worked for six years as a research scientist for Microsoft Research, first in Redmond, then in Cambridge, founding the vision side of the Machine Learning and Perception Group. He is now a professor at Oxford University. He has won awards from several top vision conferences, including ICCV, CVPR, ECCV, NIPS and BMVC. He is a senior member of the IEEE, Royal Society Wolfson Research Merit Award holder, and program co-chair of ICCV 2013.

Minh N. Do was born in Vietnam in 1974. He received the B.Eng. degree in computer engineering from the University of Canberra, Australia, in 1997, and the Dr.Sc. degree in communication systems from the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland, in 2001. Since 2002, he has been on the faculty at the University of Illinois at Urbana-Champaign (UIUC), where he is currently a Professor in the Department of Electrical and Computer Engineering, and hold joint appointments with the Coordinated Science Laboratory, the Beckman Institute for Advanced Science and Technology, and the Department of Bioengineering. His research interests include image and multi-dimensional signal processing, wavelets and multiscale geometric analysis, computational imaging, augmented reality, and visual information representation.

He received a Silver Medal from the 32nd International Mathematical Olympiad in 1991, a University Medal from the University of Canberra in 1997, a Doctorate Award from the EPFL in 2001, a CAREER Award from the National Science Foundation in 2003, and a Young Author Best Paper Award from IEEE in 2008. He was named a Beckman Fellow at the Center for Advanced Study, UIUC, in 2006, and received a Xerox Award for Faculty Research from the College of Engineering, UIUC, in 2007. He was a member of the IEEE Signal Processing Theory and Methods Technical Committee, Image, Video, and Multidimensional Signal Processing Technical Committee, and an Associate Editor of the IEEE Transactions on Image Processing.
Jiangbo Lu received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering, Katholieke Universiteit Leuven, Leuven, Belgium, in 2009.

From April 2003 to August 2004, he was with VIA-S3 Graphics, Shanghai, China, as a Graphics Processing Unit (GPU) Architecture Design Engineer. In 2002 and 2005, he conducted visiting research at Microsoft Research Asia, Beijing, China. Since October 2004, he has been with the Multimedia Group, Interuniversity Microelectronics Center, Leuven, Belgium, as a Ph.D. Researcher. Since September 2009, he has been working with the Advanced Digital Sciences Center, Singapore, which is a joint research center between the University of Illinois at Urbana-Champaign, Urbana, and the Agency for Science, Technology and Research (A*STAR), Singapore, where he is leading a few research projects. His research interests include computer vision, visual computing, image and video processing, interactive multimedia applications and systems, and efficient algorithms for various architectures. Dr. Lu is an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). He received the 2012 TCSVT Best Associate Editor Award.
**APPENDIX A**

**APPROXIMATE REGRESSION**

This section provides an approximation that allows for rapid non-linear regression. We begin with a brief overview of the problem. In main body, the regression problem is reduced to an energy minimization given by Eqn. (12). Here, \{p_j, q_j\} are \(N\) given data points. The minimization’s goal is to fit a curve \(f(p)\), which best explains the given data. As \(f(p)\) is parametrized by \{\(w_k, H_k\)\} terms, the minimization in Eqn. (12) is

\[
\arg\min_{\{w_k, H_k\}} \sum_{j=1}^{N} C(q_j - f(p_j)) + \lambda \sum_{k=1}^{K} w_k^T G w_k ,
\]

where

\[
G(i, j) = g(p_i, p_j) = \exp\left(-\frac{(p_i - p_j)^2}{\tau^2}\right),
\]

\[
f(p) = \sum_{k=1}^{K} a_k(p) f_k(p),
\]

\[
f_k(p) = H_k + \sum_{i=1}^{N} w_k(i) g(p_i, p_i).
\]

Here \(H_k\) is a scalar and \(w_k\) is an \(N\)-dimensional vector.

We seek to approximate the cost of Eqn. (31) with a significantly shorter \(\tilde{w}_k\) vector by leveraging the fact that there are many similar \(p_i, p_j\) data points. We assume without loss of generality that the last \(p_{N-t}, \ldots, p_N\) data points are similar, and remove \(t-1\) of them to create a shrunk dataset \{\(\tilde{p}_j\)\}, with \(M = N - t + 1\) elements. We then approximate the curve \(f_k(p)\) in Eqn. (32) with the shorter \(M\)-dimensional vector \(\tilde{w}_k\):

\[
f_k(p) \approx \bar{f}_k(p) = H_k + \sum_{i=1}^{M} \tilde{w}_k(i) g(p_i, p_i)
\]

where

\[
\tilde{w}_k(i) = \begin{cases} w_k(i), & \text{if } i < M, \\ \sum_{j=M}^{N} w_k(j), & \text{if } i = M. \end{cases}
\]

\[\tilde{G}(i, j) = g(\tilde{p}_i, \tilde{p}_j)\]

where \(\tilde{G}\) is a \(M \times M\) matrix.

The shorter \(\tilde{w}_k\) vector will approximate the original \(f_k(p)\) function better if the values of \(p_{N-t}, \ldots, p_N\) are more similar, and will do so perfectly if \(p_{N-t}, \ldots, p_N\) are identical.

Replacing \(w_k\) with \(\tilde{w}_k\) not only affects the regression surface \(f(p)\) but also the smoothness cost \(w_k^T G w_k\) in Eqn. (31). This can be theoretically accommodated by changing the smoothness weight \(\lambda\) appropriately. Thus we seek to discover the interaction between \(\lambda\) and \(w_k\)’s vector length.

We observe that because of the values taken by \(\tilde{w}_k\) in Eqn. (34),

\[\tilde{G}(i, :)\tilde{w}_k \approx \begin{cases} G(i, :)w_k, & \text{if } i < M, \\ \sum_{j=M}^{N} w_k(j)G(i, :) & \text{if } i = M, \end{cases}\]

this leads to the following convenient equivalence

\[w_k^T G w_k \approx \tilde{w}_k^T \tilde{G}\tilde{w}_k .\]

As in Eqns. (33), this approximation becomes perfect, if \(p_{N-t}, \ldots, p_N\) are identical.

The result of the approximations in Eqn. (33) and Eqn. (36) is that the curve \(f(p)\) in Eqn. (31) can be approximated as

\[
f(p) \approx \bar{f}(p) = \sum_{k=1}^{K} a_k(p) \bar{f}_k(p),
\]

and the cost function in Eqn. (31) can be approximated as

\[
\arg\min_{\{w_k, H_k\}} \sum_{j=1}^{N} C(\tilde{q}_j - \bar{f}(p_j)) + \lambda \sum_{k=1}^{K} \tilde{w}_k^T \tilde{G}\tilde{w}_k .
\]

While the above proof is for the consolidation of a single cluster, the same proof allows \(\tilde{w}_k\) to be formed from recursive consolidation of multiple clusters. This makes \(\tilde{w}_k\) substantially shorter than \(w_k\), and hence Eqn. (38) is much cheaper to minimize compared to Eqn. (31).

Note that recursive consolidation presented here is mainly for theoretical understanding. In practice, \{\(\tilde{p}_j\)\} is directly computed through K-means clustering \[40\] of \{\(p_j\)\} data points. Due to the finding in Eqn. (36), the size of the match clusters need not to be explicitly kept and only their centroid positions are used.
**APPENDIX B**

**NON-RIGID MOTION**

One of the strengths of our CODE formulation is that it naturally accommodates handling large non-rigid motion scenes. However, CODE’s formulation means that small independent motions are invariably classified as incorrect and hence are ignored. Quantitative evaluation of this property is shown in Fig. 9 and more visual illustration is shown in Fig. 13.

**APPENDIX C**

**STRUCTURE FROM MOTION RESULTS**

This sections tabulates all results for the 3-D reconstruction dataset [43], which are given in Fig. 14, Fig. 15, Fig. 16, Fig. 17, Fig. 18 and Fig. 19. For each dataset, reconstruction was achieved with VisualSfM [3], with A-SIFT and A-SIFT w CODE as a feature matcher. Apart from A-SIFT, we also applied other feature matchers to the dataset, but we find that nearly all previous techniques leave large gaps in the reconstruction. Our results are not perfect, still with notable errors in the Dinosaur and Matrix multi-view stereo image sequences in Fig. 16 and Fig. 17 but they represent a significant improvement over the prior state-of-the-art. For readers versed in Structure from Motion, we highlight that no prior camera calibration was assumed. Better reconstruction results can be obtained by using the focal length parameters provided with the dataset [43].

**APPENDIX D**

**OUR DATASET**

Thumbnails of the twenty image pairs from our dataset are shown in Fig. 20.
Fig. 13: Correspondence on non-rigid scenes. CODE’s formulation accommodates large non-rigid motions such as object permutation or folding shown in (a-e). However, CODE ignores small independent motions, because CODE’s formulation invariably considers them as wrong matches with low motion coherence. This is illustrated in (f). Depending upon the application, this can be an asset that facilitates motion segmentation or a limitation as small objects cannot be well tracked.
Fig. 14: *Predator* image sequence.

Fig. 15: *Warrior* image sequence.
Input view  | A-SIFT  | A-SIFT w CODE
---|---|---
Fig. 16: *Dinosaur* image sequence.

Input view  | A-SIFT  | A-SIFT w CODE
---|---|---
Fig. 17: *Matrix* image sequence.
Fig. 18: Soldier image sequence.

Fig. 19: Mummy image sequence.
Fig. 20: Image pairs used in our dataset.