MISMATCH REMOVAL VIA COHERENT SPATIAL MAPPING

Jiayi Ma¹, Ji Zhao¹, Yu Zhou², Jinwen Tian¹

¹Institute for Pattern Recognition and Artificial Intelligence
²Department of Electronics and Information Engineering
Huazhong University of Science and Technology, Wuhan, 430074, China.
{jyma2010, zhaoji84, zhouyu.hust}@gmail.com, jwtian@mail.hust.edu.cn

ABSTRACT

We propose a method for removing mismatches from given putative point correspondences in image pairs. Our algorithm aims to recover the underlying coherent spatial mapping which related to inliers. The thin-plate spline (TPS) is chosen to parameterize the coherent spatial mapping, and we formulate the solution of it as a maximum likelihood problem. The mismatches could be successfully removed after the EM algorithm, which we used for solving the problem, converges. The quantitative results on various experimental data demonstrate that our method outperforms many state-of-the-art methods. Moreover, the proposed method is also able to handle the case that image pairs contain non-rigid motions.

Index Terms— Mismatch removal, thin-plate splines, nonlinear mapping, outlier, point correspondence

1. INTRODUCTION

This paper focuses on establishing accurate point correspondences between two images of the same scene. Many of the computer vision tasks such as building 3D models, camera self-calibration, registration, object recognition, and structure and motion recovery [1] start by assuming that the point correspondences and two-view image relations have been successfully recovered.

Point correspondences are in general established by comparing the distances of keypoints’ local features. This may result in a number of mismatches (outliers) due to viewpoint changes, occlusions, repeated structures, etc. The existence of mismatches is usually enough to ruin the traditional estimation methods. In this case, robust estimators are developed to provide reliable point correspondences [2]. These methods search for sets of matches which consistent with some global geometric constraints.

During the last decades, various robust estimators have been proposed in the statistics and computer vision literatures. Here we briefly review some of which that are widely used. Among the statistics community, two representatives are Maximum-likelihood estimators (M-estimators) [3] and Least Median of Squares (LMedS) estimator [4]. The former minimizes the sum of symmetric, positive-definite functions of residuals with a unique minimum at zero, while the latter minimizes the median of squared residuals. In the computer vision community, RANSAC [2] and MLESAC [5] are two widely used robust estimators. They both try to get a minimum outlier-free subset to estimate a given parametric model by resampling, and the difference is that MLESAC chooses the solution which maximizes the likelihood rather than the inlier count as in RANSAC. Recently, there appear some new non-parametric model based methods, such as Identifying point correspondences by Correspondence Function (ICF) [6] and Vector Field Consensus (VFC) [7]. The former rejects mismatches though learning a correspondence function pair which map points in one image to their corresponding points in the other image, while the latter converts the mismatch removal problem into a robust vector field learning problem, and learns a smooth field to fit the potential inliers as well as estimates a consensus inlier set.

In this paper, we present a new method for mismatch removal. Notice the fact that an image pair contain the same 3D scene, and then it will in general exist a smooth nonlinear spatial mapping which could fit the correct point correspondences well. And the mismatches will be easily removed if we recover this mapping. Motivated by this, we parameterize the mapping by a general purpose spline tool — the thin-plate spline (TPS) and focus on recovering it. Experimental results on various image data show the effectiveness of this method.

2. METHOD

Given a set of putative image point correspondences \( S = \{(x_n, y_n)\}_{n=1}^{N} \) in two views which may be perturbed by noises and outliers, the goal is to remove outliers contained in the point set to establish reliable point correspondences.

Without loss of generality, a nonlinear mapping \( f : y_n = f(x_n) \) is adopted to characterize the underlying coherent spatial relations of the inliers. And we call this mapping coherent spatial mapping. Obviously, if we successfully recover the mapping \( f \), then the mismatches can be easily removed. However, estimating the coherent spatial mapping also requires reliable point correspondences. To solve this dilemma,
we formulate the problem as a maximum likelihood problem, and then solve it under an EM framework.

2.1. A maximum likelihood formulation

We give a maximum likelihood formulation for computing the coherent spatial mapping \( f \). In the following we make the assumption that the noise on inliers is Gaussian with zero mean and uniform standard deviation \( \sigma \), and the outlier distribution is uniform [5, 7]. Thus, the likelihood is a mixture model as:

\[
p(Y|X, \theta) = \prod_{n=1}^{N} \left( \frac{\gamma e^{-\frac{|y_n - f(x_n)|^2}{2\sigma^2}} + \frac{1-\gamma}{a} \right),
\]

where \( \theta = \{f, \sigma^2, \gamma\} \) is the set of unknown parameters, \( \gamma \) is the percentage of inliers and \( a \) is just a constant, i.e. the area of the image. \( X = (x_1; \cdots; x_N) \) and \( Y = (y_1; \cdots; y_N) \) are matrices of size \( N \times 3 \), due to the use of homogeneous coordinates for the image point set, i.e. \( x = (x_x, x_y, 1) \).

2.2. An EM solution

Generally speaking, the true parameter set \( \theta \) maximizes likelihood (1). Now we give a maximum likelihood estimation of \( \theta \), i.e. \( \theta^* = \arg \max_\theta p(Y|X, \theta) \). The well known EM algorithm provides a natural framework for solving this problem. The E-step basically estimates the responsibility indicating to what degree a sample belonging to inlier under the given coherent spatial mapping \( f \), while the M-step updates \( f \) based on the current estimate of the responsibility. Following standard approach, we simply summarize the EM iteration as follows.

**E-step:** denote \( v_n = f(x_n) \), we update the responsibility \( p_n \) as

\[
p_n = \frac{\frac{\gamma}{2\pi\sigma^2}e^{-\frac{|y_n - v_n|^2}{2\sigma^2}} + \frac{1-\gamma}{a}}{\frac{\gamma}{2\pi\sigma^2}e^{-\frac{|y_n - f(x_n)|^2}{2\sigma^2}} + \frac{1-\gamma}{a}}.
\]

**M-step:** parameters \( \sigma^2 \) and \( \gamma \) are updated as

\[
s^2 = \frac{(Y-V)^T \mathbf{P}(Y-V)}{2 \cdot \text{tr}(\mathbf{P})},
\]

\[
\gamma = \frac{\text{tr}(\mathbf{P})}{N},
\]

where \( \mathbf{P} = \text{diag}(p_1, \ldots, p_N) \), \( V = (v_1; \cdots; v_N) \), and \( \text{tr}() \) is the trace. Notice that \( v_n \) is a homogeneous coordinate, we should normalize it with scale 1 before computing \( p_n \) and \( \sigma^2 \). To complete the EM algorithm, the mapping \( f \) should be estimated in the M-step. This is the key step in our method, and we will discuss in the next section.

2.3. Estimation of the coherent spatial mapping

According to the complete negative log-likelihood of equation (1), the mapping \( f \) is estimated by minimizing a weighted empirical error \( \frac{1}{2\sigma^2} \sum_{n=1}^{N} p_n |y_n - f(x_n)|^2 \). This is in general ill-posed since the mapping \( f \) is not unique. Notice that the responsibility \( p_n \) is a posterior probability indicating to what degree the sample \( n \) belongs to inlier. When \( p_n = 0 \), the point correspondence \( (x_n, y_n) \) is considered as an outlier and will not involve in estimating the coherent spatial mapping. Thus we consider \( p_n \) as a soft decision as its continuous value over the interval [0, 1].

To generate a smooth mapping fitting for the image point correspondences, we choose the TPS for parametrization. The TPS is a general purpose spline tool which produces a smooth functional mapping for supervised learning [8]. It has no free parameters that need manual tuning, and also has a close-form solution which can be decomposed into a global linear affine motion and a local non-affine warping component controlled by coefficients \( \mathbf{A} \) and \( \mathbf{W} \) respectively:

\[
f(x) = x \cdot \mathbf{A} + \tilde{K}(x) \cdot \mathbf{W},
\]

where \( \tilde{K}(x) \) is a \( 1 \times N \) vector defined by the TPS kernel, i.e. \( K(r) = r^2 \log r \), and each entry \( \tilde{K}_n(x) = K(|x - x_n|) \). Define the kernel matrix \( \mathbf{K}_{N \times N} = \{K_{ij}\} \) where \( K_{ij} = K(|x_i - x_j|) \). With a regularization parameter \( \lambda \), the nonlinear spatial mapping \( f \) can be then estimate by minimizing a TPS energy function as:

\[
\mathcal{E}(\mathbf{A}, \mathbf{W}) = \frac{1}{2\sigma^2} \| \mathbf{P}^{1/2}(Y - XA - KW) \|^2 + \frac{\lambda}{2} \text{tr}(\mathbf{W}^T \mathbf{K} \mathbf{W}).
\]

The second smoothness term is the standard TPS regularization term, and it is the bending energy which has a physical explanation and is independent on the linear component of the coherent spatial mapping. To solve the TPS parameter pair \( \mathbf{A} \) and \( \mathbf{W} \), we use a QR decomposition [8, 9], i.e.

\[
P^{1/2}X = [Q_1 Q_2] \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}
\]

Minimizing the energy function (6), we obtain

\[
\mathbf{W} = Q_2(\mathbf{S}^T \mathbf{S} + \lambda \sigma^2 \mathbf{T} + \mathbf{c} \mathbf{I})^{-1} \mathbf{S}^T Q_2^T \tilde{Y},
\]

\[
\mathbf{A} = \mathbf{R}^{-1} Q_1^T (\tilde{Y} - \mathbf{P}^{1/2} \mathbf{K} \mathbf{W}),
\]

where \( \mathbf{S} = Q_1^T \mathbf{P}^{1/2} \mathbf{K} Q_2, \mathbf{T} = Q_2^T \mathbf{K} Q_2, \) and \( \mathbf{c} \mathbf{I} \) is used for numerical stability.

Once the EM algorithm converges, we get the coherent spatial mapping \( f \). The mismatches are then able to be removed by checking whether they are consistent with the estimated mapping. This is equivalent to obtaining the inlier set \( I \) from the responsibility \( p_n \) with a predefined threshold \( \tau: I = \{np_n > \tau, n = 1, \cdots, N\} \). The mismatch removal algorithm is summarized in algorithm 1.

**Relation to VFC:** From the perspective of mismatch removal, our method is related to the VFC algorithm [7]. On the one hand, both the two algorithms try to seek a mapping \( f \) fitting the inliers well under a Bayesian framework and use the EM approach to solve it. On the other hand, our algorithm
is different from VFC. The VFC algorithm convert the correspondence problem into vector field learning problem, and learning a smooth field in a reproducing kernel Hilbert space (RKHS). While in our method, the spatial mapping related to inliers is parameterized by TPS, and then the mapping can be clearly decomposed into linear and nonlinear components. Moreover, the bending energy minimized by TPS has a specific physical explanation. This may be beneficial in the case of image pairs with non-rigid motions.

3. EXPERIMENTAL RESULTS

To test the mismatch removal performance of our algorithm, we performed experiments on a wide range of real images. Four additional mismatch removal methods are used for comparison: RANSAC, MLESAC, ICF and VFC.

There are mainly two parameters in our algorithm: the TPS regularization parameter $\lambda$ and inlier threshold $\tau$. In practice, we find that our method is not very sensitive to parameter tuning. We set $\lambda = 800$ and $\tau = 0.75$ throughout this paper. The open source VLF toolbox\(^1\) [10] is used to determine the initial matches of SIFT [11], and the match correctness is determined using the same criterion as in [7].

**Results on several image pairs.** We first tested the mismatch removal performance of our method on several image pairs, including wide baseline image pairs (*Mex* and *Tree*) and image pairs of non-rigid object (*Peacock* and *T-shirt*). The results are presented in Fig. 1. The performance is characterized by precision and recall. For the *Mex* pair, as shown on the top row of Fig. 1, there are 158 initial correspondences with 76 mismatches; the correct match percentage is 51.90%; after using our method to remove mismatches, 84 matches are preserved, including all the 82 correct matches. The precision-recall pair is about (97.62%, 100.00%). The rest three rows of Fig. 1 present similar results on image pairs of *Tree*, *Peacock* and *T-shirt*.

\(^1\)available at: http://www.vlfeat.org/

Fig. 1: Mismatch removal results of our method on image pairs of *Mex*, *Tree*, *Peacock* and *T-shirt*. The initial correct match percentages are 51.90%, 56.29%, 71.61% and 60.67% respectively. After using our method to remove mismatches, we obtain precision-recall pairs (97.62%, 100.00%), (98.85%, 91.49%), (100.00%, 99.41%) and (100.00%, 99.45%). For each group of results, the left pair denotes the identified suspect correct matches, and the right pair denotes the removed suspect mismatches.

Now we give a performance comparison on these image pairs with the other four mismatch removal methods, as shown in Table 1. We see that MLESAC has slightly better precisions than the RANSAC with the cost of producing slightly lower recalls. The recall of ICF is quite low, although it has a satisfactory precision. Compared to these three methods, VFC and our method have satisfactory performance due to the simultaneous high precision and high recall. However, in the case of image pair containing non-rigid object, our method has even better performance. The average run time of our method on these four image pairs is about 0.5 seconds on an Intel Pentium 2.0 GHz PC with Matlab code.

Notice that we did not compare to RANSAC and MLESAC on the image pairs of non-rigid object, since in this case the two view relation, i.e. fundamental matrix, modeled in RANSAC and MLESAC is no longer exist. In general, Our method is effective for mismatch removal not only on image pairs related by rigid motions but also on image pairs with non-rigid motions.

**Results on a dataset.** We also tested our method on the dataset of Mikolajczyk et al [12], which contains image pairs of large view angle, image rotation and affine transformation,

**Algorithm 1: The Mismatch Removal Algorithm**

**Input:** Putative correspondence set $S = \{x_n, y_n\}_{n=1}^N$, parameter $\lambda$, $\tau$, and kernel $K$.

**Output:** Inlier set $I$

1. **Initialization:**
2. Construct kernel matrix $K$ using the definition of $K$;
3. **repeat**
   1. **E-step:**
   2. Update $P = \text{diag}(p_1, \ldots, p_N)$ by equation (2);
   3. **M-step:**
   4. Update mapping $f$ by using equations (7) and (8);
   5. Update $\sigma^2$ and $\gamma$ by equations (3) and (4);
4. **until** some stopping criterion is satisfied;
5. The inlier set is determined by $I = \{i | p_i > \tau\}$.\n
---

\(^1\)available at: http://www.vlfeat.org/
etc. We use all the 40 pairs, and for each pair, we set the SIFT distance ratio threshold \( t \) to 1.5, 1.3 and 1.0 respectively (the greater value of \( t \) indicates the smaller amount of matches with higher correct match percentage). The initial average precision of all image pairs is 69.58%, and nearly 30 percent of the training sets have correct match percentage below 50%.

Fig. 2 gives the results of four methods on this dataset, and each scattered dot represents a precision-recall pair on an image pair. The average precision-recall pairs are (93.84%, 98.50%), (94.99%, 92.40%), (93.95%, 62.69%), (98.57%, 97.75%) and (98.02%, 98.14%) for RANSAC, MLESAC, ICF, VFC and our method respectively. Note that the performances of VFC and our method are quite close, thus we omit the result of VFC in the figure for clarity. As shown, RANSAC and MLESAC have good performance on most of the image pairs; still, due to the low initial correct match percentage, the effect is not so satisfactory on several image pairs. ICF usually has high precision or recall, but not simultaneously. Our method (and VFC) has the best precision-recall trade-off, and the scattered dots almost concentrate on the upper right corner. These results demonstrate that the mismatch removal capability of our method is not affected by low initial correct match percentage, large view angle, image rotation and affine transformation since these cases are all contained in the dataset.

4. CONCLUSION

Within this paper a novel mismatch removal method based on estimating coherent spatial mapping on inliers has been shown. It alternately fits a smooth spatial mapping for inliers and detects outliers under an EM framework. The experimental results on benchmark datasets show that our method outperforms state-of-the-art methods such as RANSAC. Moreover, the effective results achieved in the non-rigid motion case show its potential value in the area of image retrieval or image-based non-rigid registration.

5. REFERENCES


