Robust Frame-to-Frame Hybrid Matching

Lei Chen\(^1\), Z. L. Wang\(^2\) and Yunde Jia\(^1\)

\(^1\)Beijing Laboratory of Intelligent Information Technology, School of Computer Science
Beijing Institute of Technology, Beijing 100081, P.R. China
\(^2\)School of Electronics and Information Engineering
Beijing Jiaotong University, Beijing 100044, P.R. China
\{chenlei,jiayunde\}@bit.edu.cn,zlwang@bjtu.edu.cn

Abstract—In this paper, we propose a hybrid approach for addressing feature-based matching problem. We aim to obtain robust and accurate correspondence between features from image frames under unknown and unstructured environments. The approach incorporates image texture analysis, 2-D analytic signal theory and color modeling. It takes advantage of geometric invariant property in texture and monogenic signal information as well as photometric invariant property in HSV color information. The detected features are well localized with high accuracy and the selected matches are robust to changes in scale, blur, viewpoint, and illumination. Experiments conducted on a standard benchmark dataset demonstrate the effectiveness and reliability of our approach.

Keywords—hybrid matching; monogenic signal; color entropy;

I. INTRODUCTION

Establishing frame-to-frame correspondence is a fundamental task in computer vision. However, it is difficult to achieve robust and accurate result via a single approach alone, even with some iterative process. Therefore, instead of seeking a single perfect matching method, it is natural to develop a hybrid approach which produces superior results under various imaging conditions and overcomes the weakness of each individual method.

Basically, the search for image correspondence can be treated as a two-stage process including feature detection and matching. Local (invariant) feature is widely used in this case for its notable repeatability and distinctiveness. A variety of local feature detectors are reported in the literature. These methods can be mainly classified into three categories: corner detector [1], blob detector [2], [3] and region detector [4]. The latest published overview of the state-of-the-art local feature detectors is given by Tuytelaars and Mikolajczyk [5], and they show that corner detectors are typically more accurate in their localization than other features while their scale is not well estimated. In general, blob and region structures are suitable for dealing with category-level of object recognition. However, they suffer from two main problems: the number of detected features is huge and these features typically represent homogeneous regions lacking distinctiveness. For feature-based matching, in most cases, the region centered on every detected feature point is converted into a feature vector (descriptor), and features are matched based on the distance measurement between descriptors, e.g. the Mahalanobis or Euclidean distance. Distribution-based Scale Invariant Feature Transform (SIFT) descriptor [3] has been shown to outperform the others. However, it increases time cost on matching [6].

In this paper, we propose a feature-based approach for frame-to-frame matching under unknown and unstructured environments. Our approach enables to locate features in sub-pixel precision and finds stable matches over time in spite of changes in imaging conditions. An improved Harris detector is adopted to automatically extract feature points with high localization accuracy. Texture, monogenic signal and HSV color entropy are efficiently integrated in a hybrid matching procedure for obtaining geometric and photometric invariant matches. The obtained low number of matches with high correct match rate and high localization accuracy are especially suitable for time- and accuracy-sensitive practical applications.

II. FEATURE-BASED HYBRID MATCHING APPROACH

Matching procedure searches for correspondences of a set of feature points between two frames which are related by geometric and photometric transformations such as scale, blur, illumination, etc. Our approach consists of the following two stages: (1) Extract features on each frame through an improved Harris corner detector. (2) Find putative matches based on a complementary combination of the texture, monogenic signal and color entropy.

A. Feature detection

We use simple Harris corner detector [1] for interest point extraction in each frame. Harris corner detector is based on the auto-correlation matrix \(A\) which describes the gradient distribution in a local neighborhood of a point \(x\):

\[
A = g(\sigma) \ast \left[ \begin{array}{cc}
I^2_1(x, \sigma_D) & I_1(x, \sigma_D)I_2(x, \sigma_D) \\
I_1(x, \sigma_D)I_2^*(x, \sigma_D) & I^2_2(x, \sigma_D)
\end{array} \right] (1)
\]

with
\[
I_i(x, \sigma_D) = \frac{\partial}{\partial x} g(\sigma_D) * I(x)
\]
(2)

\[
g(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}}
\]
(3)

where \(I(x)\) is the intensity value of point \(x\).

The first order horizontal and vertical derivatives \(I_x, I_y\) of the image are computed by convolution with the mask of \([-2 \ 1 \ 0 \ 1 \ 2]\) and Gaussian kernel of scale \(\sigma_D = 1\) (the differentiation scale). The derivatives are then smoothed in the neighborhood of the point with a Gaussian window of scale \(\sigma_I = 2\) (the integration scale) to get \(I_{sx}, I_{sy}\) and \(I_{sy}\). Since the major drawback of [1] is the existence of manual tuning on parameter \(\alpha\), we choose a Harris function [7] to compute the strength \(s\) of each image pixel:

\[
s = \frac{I_{sx}I_{sy} - I_{sx}I_{sy}}{I_{xx} + I_{yy} + \varepsilon}
\]
(4)

where a small constant \(\varepsilon\) is used to avoid a zero denominator.

After the strength of all image points is acquired, non-maximum suppression is then applied to select the interest point at each pixel where its strength is stronger than at all other pixels in a \(3 \times 3\) windowed area. Finally, the interest point whose strength is higher than a threshold is chosen for matching, and the threshold is set to 1% of the maximum interest point strength in this paper. Additionally, sub-pixel coordinates of features are obtained through quadratic approximation of the parabola function in the neighborhood of a local maximum.

**B. Hybrid matching**

The feature matching procedure is the key part in our approach. In contrast to SIFT [3] or Speeded-Up Robust Features (SURF) [2] which build descriptors to distinguish detected points and recognize correspondences, we prefer direct matching for reducing computation cost. To improve the robustness and reliability of matching result, a multi-step processing strategy is employed including three steps (Let \(N_1\) and \(N_2\) denote the quantities of detected points in the first and second image respectively):

1. **Texture based processing.** *Normalized cross-correlation* (NCC) over two \(9 \times 9\) patches centered on every two detected feature points from two images is calculated, then a correlation matrix \(M_1\) represents the similarity between each pair of points is acquired. To select putative matching pairs, we use a mutual consistency check [8] on the correlation matrix \(M_1\), in which feature matching is performed twice, from the first to the second and from the second to the first. The point \(P_1\) in the first image and point \(P_2\) in the second image are treated as a potential matching pair if and only if the check result is consistent in both ways, which means that the correlation value \(M_1(P_1, P_2)\) is the maximum in both of the row \(P_1\) and column \(P_2\) in \(M_1\). For each selected points pair, the corresponding correlation value and the coordinates of two points are stored.

2. **Analytic signal based processing.** The monogenic signal [9] is a 2-D analytic signal based on the Riesz transform and preserves the split of identity, i.e., invariance-equivariance property. The image is interpreted in terms of the local phase, the local orientation and the local magnitude under the monogenic signal framework. A correlation matrix \(M_2\) is generated by monogenic phase data within two \(11 \times 11\) windows surrounding every two detected feature points from two images [10]. The candidate matching pairs are collected following the same scheme described in texture based processing.

3. **Color entropy based processing.** We choose HSV color space instead of common RGB color space for feature matching. \(H(hue), S(saturation), V(value)\) are not inter-correlated and are changing in a comparable steady range, which make HSV color space resistant to the effect of illumination change. The color entropy is given by

\[
E(x, y) = -\sum_{i \in R} w_i \log_2 w_i
\]
(5)

where \(E(x, y)\) is the entropy value of feature point \((x, y)\), \(i\) represents the \(H/S/V\) value in a \(5 \times 5\) local region \(R\) centered on \((x, y)\), and \(w_i\) denotes the weight of value \(i\). \(w_i\) is defined as the ratio of the number of pixels with value \(i\) to the number of all pixels (equals to 25 in this case) in region \(R\). The potential matching pair is valid if the difference of all three color channel entropy value between two points are relatively small.

The feature pairs which are selected in all the three steps are accepted as the final matching pairs. In practical implementation, the matching process suffices the following three constraints: (1) The candidate pairs with a correlation value below a certain threshold \(T1\) are discarded in texture based processing; (2) When monogenic signal phase data are applied for matching, we reject the candidate pairs whose correlation value are smaller than a predefined threshold \(T2\); (3) In HSV entropy based processing, for the selected pair, the maximal values of entropy difference for \(H, S\) and \(V\) color channel are predefined as \(TH, TS\) and \(TV\) respectively. The constraints are tunable via thresholds and the trade-off is between quality and quantity of the matching pairs, so we set the parameters in matching procedure to \(T1 = 0.85, T2 = 0.5, TH = 0.3, TS = 0.4, TV = 0.3\).

**III. IMPROVEMENT**

For every feature point in the first image, a square window of data is collected and correlated with a window corresponding to every feature point in the second image to find putative matches in the aforementioned matching approach. For speed consideration, the generation of matrix \(M_1\) and \(M_2\) is the most computationally demanding part.
in the whole matching process. The key to achieving fast matching is therefore to minimize the computation cost on pairwise correlation. In practice, all the elements in the correlation matrix are initialized as minus infinity, and correlation values are recalculate only for points that are within a distance range from each other. The search range can be predicted according to the specific application, and this technique provides a significant speedup for matching procedure. Furthermore, texture and monogenic signal based processing are totally independent from each other and can be performed in parallel.

There exist unavoidable mismatches among the final matched pairs, the Least Median of Squares (LMedS) method [11] can be adopted for result refinement. When our method is applied in real applications, LMedS is used to find a consensus group of point pairs which are consistent with an instantiation of the model (the homography matrix).

### IV. Experiments

We have evaluated the robustness and accuracy of the proposed approach on the standard benchmark dataset collected by Mikolajczyk and Tuytelaars [12], [4]. This reference dataset contains eight series of six images under different imaging conditions and ground truth of homographies between image pairs are available. The description of four image series (‘boat’, ‘bikes’, ‘wall’ and ‘leuven’) used for evaluation is listed in Table.I.

Let $P_1$ in the first image and $P_2$ in the second image denote a pair of matching points, the residual $e$ of the matching pair $P_1P_2$ is formulated by the absolute Euclidean distance

$$e_i = |X_i - \hat{X}_i| \quad i \in [1, N]$$

(6)

where $N$ is the number of the final matching pairs, $X_i$ is the actual measurement coordinates of $P_2$, and $\hat{X}_i$ is the ground truth coordinates of $P_2$ which is obtained by the measurement coordinates of $P_1$ and homography matrix between two images. $P_1P_2$ is defined as a mismatch if its corresponding residual is larger than 1.

In Table.II, the second column is the number of final selected matches. The third column shows the quantity of mismatches remaining in the final matching result. The average residual errors of the final result is given in the last column, and it shows that the overall matching accuracy of our approach is high. A comparison of correct match rate produced by our approach, SIFT and SURF is shown in Table.III. Comparisons are performed using available well-known implementation \(^1\)\(^2\) without changing the default parameters.

It is obvious that our approach achieves a low number of matches with high correct match rate and high localization accuracy. The comparison shows that our approach outperforms widely used SIFT and SURF method with a very high and stable correct match rate over all detected matches. These valuable characteristics are especially suitable for time- and accuracy-sensitive practical applications and will effectively decrease the ambiguity and computational complexity of tracking and recognition tasks.

Figure.1 displays points correspondence and their coordinates in the first/left and second/right images. It shows that our method is reliable under a variety of common situations, e.g., viewpoint change, scale change, image blur and illumination change in realistic scenarios.

### V. Conclusion

In this paper, we present a feature-based approach to obtain robust and accurate matching results under various image conditions. The feature points are extracted with an improved Harris corner detector. The proposed hybrid approach is able to obtain a set of matches which are robust against photometric and geometric transformations through an efficient integration of texture, monogenic signal and HSV color information. Moreover, the mismatch ratio is significantly decreased, while high precision in feature localization is maintained. Our work will be further optimized and applied for mobile robot online simultaneous localization and mapping (SLAM) applications.

---

ACKNOWLEDGEMENT

This work was supported by the Natural Science Foundation of China under Grant No90920009 and No60905006, and the Chinese High-Tech Program under Grant 2009AA01Z323, and was partially supported by the State Key Laboratory of Robotics and System, Harbin Institute of Technology(SKLRS-2009-ZD-01).

REFERENCES


