Computer Vision and Image Understanding 118 (2014) 71–83

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Spatio-temporal weighting in local patches for direct estimation of camera motion in video stabilization ☆

Soo Wan Kim *, Shimin Yin, Kimin Yun, Jin Young Choi

School of Electrical and Computer Engineering, ASRI, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 17 August 2011
Accepted 19 September 2013
Available online 30 September 2013

Keywords:
Video stabilization
Camera motion estimation
Spatio-temporal information
Object function minimization

ABSTRACT

This paper presents a robust video stabilization method by solving a novel formulation for the camera motion estimation. We introduce spatio-temporal weighting on local patches in optimization formulation, which enables one-step direct estimation without outlier elimination adopted in most existing methods. The spatio-temporal weighting represents the reliability of a local region in estimation of camera motion. The weighting emphasizes regions which have the similar motion to the camera motion, such as backgrounds, and reduces the influence of unimportant regions, such as moving objects. In this paper, we develop a formula to determine the spatio-temporal weights considering the age, edges, saliency, and distribution information of local patches. The proposed scheme reduces the computational load by eliminating the integration part of local motions and decreases accumulation of fitting errors in the existing two-step estimation methods. Through numerical experiments on several unstable videos, we verify that the proposed method gives better performance in camera motion estimation and stabilization of jittering video sequences.

1. Introduction

The number of cameras on mobile platforms is significantly increasing, and access to them is much easier these days. However, the videos captured by those cameras are easily exposed to the instability problem because of hand-trembling or the absence of tools to maintain the camera path, such as camera dollies and steadicam. Under these conditions, the quality of video declines and it is hard to watch or understand scenes. The problem of unstable video not only occurs in the mobile platform case. Even mounted cameras on a pole or building wall can be jittered by wind or resonance of mounting structures. Without the aid of costly hardware devices, such as gyros and corresponding control units, the video stabilization process is adopted as a tool to transform acquired shaking videos to smooth sequences and enhance their quality. The enhanced sequences will be able to improve significantly the performance in higher image analysis, e.g. object detection, recognition and visual surveillance [1].

Generally, the video stabilization algorithm is composed of three steps: (1) motion estimation (2) motion compensation, and (3) image composition [2]. The camera motion between two consecutive frames is estimated in the motion estimation step, and the result is passed to the other two steps. Motion estimation is the most important step in video stabilization because the remaining two steps significantly depend on the performance of this step. In the motion compensation part, the estimated motion is divided into the intentional motion and the shaky motion to find the actual unwanted motion. This compensation step is required when the camera itself is moving and shaking at the same time, but can be omitted in the case of the static jittering camera. In the image composition step, the input frames are synthesized to build stabilized sequences, using the estimated and compensated motion. In this paper, we focus on motion estimation which is the most crucial step for the final performance of the stabilization algorithm.

Various approaches have been developed for camera motion estimation in video stabilization. Among them, the feature tracking based approach is well known for its good performance. One of the features being widely used is the SIFT feature [3]. SIFT feature based methods [2,4–8] extract feature descriptors in images, match them to find corresponding pairs in the consecutive frames, and estimate camera motion from the motions of features. SIFT feature trajectories are combined with several techniques such as considering local neighborhood [2], decreasing accumulation errors [4], and adopting particle filter scheme [5,6]. These approaches suffer from non-repeatability of SIFT features, large computation to extract and match features every frame, and the existence of occlusion/foreground objects. In [9], Chang et al. proposed an optical flow based approach and Censi et al., in [10], tracked corner features by the Kalman filter. These methods are faster than SIFT

☆ This paper has been recommended for acceptance by Tinne Tuytelaars.

* Corresponding author. Fax: +82 2 888 4182.
E-mail addresses: soowankim@snu.ac.kr (S.W. Kim), simcom79@gmail.com (S. Yin), kmyun@neuro.snu.ac.kr (K. Yun), jychoi@snu.ac.kr (J.Y. Choi).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.09.005

based methods, but they also cannot handle scenes with foreground objects and feature tracking failure. Recently, Liu et al. [11] introduced a method to smooth the camera path by adding subspace constraints on feature trajectories. However, their method works only on video, not in online sequences, and requires long feature tracks over multiple frames. Grundmann et al.'s method in [12] does not show good performance for the scene with a small number of features.

Another category for the motion estimation is the block matching based approach [13–16]. Rather than using a single representative point, this approach selects several blocks in images and finds their local motions by comparing similarities of corresponding blocks between consecutive frames. The similarity can be measured by edge information [13], gray coded bit plane [14], intensity mean absolute difference [15], and memory filter [16]. These approaches are fast and simple, but do not guarantee robust performance if the extracted blocks are not discriminative enough or those blocks are on foreground objects. This is because a homogeneous region has no information, and foreground objects have their own motions which are not related to the camera motion. The domain transformation methods, such as polar coordinates [17], Hilbert Huang Transform [18], and phase domain [19,20], got more robust results in certain situations. However, the effect of the domain transformation is limited to specific situations, and approaches in this category still have problems arising from foreground objects and dynamic scenes.

Other types of the existing methods, such as integral projection methods [21,22] and the complex Markov random field (MRF) model [1], also have limitations. The projection methods cannot provide accurate results when the actual camera motion model is complex, and an MRF based approach requires too much computation, even though it shows robust performance. Liu et al., in their work [23], proposed an energy minimization concept which shows robust stabilization but it is computationally expensive for 3D camera motion estimation using the structure from motion method. Most existing approaches to estimate global camera motion from local motions require additional outlier elimination algorithms, such as RANSAC [24], which are generally heavy to compute.

In this paper, we aim to achieve light computational load as well as high accuracy in motion estimation even when containing foreground objects and dynamic scenes. For this purpose, we design a weighted optimization formulation with spatio-temporal information on local patches. Depending on the spatio-temporal information contained in the local patches, we assign weights to emphasize the effect of reliable patches and to reduce that of unimportant patches in the optimization formulation. To assign the weights, we develop a formula calculating the spatio-temporal weights considering the age, edges, saliency, and distribution information of the local patches. Because patches on homogeneous regions and foreground objects are less weighted by the formula, camera motion can be estimated more robustly even in scenes with foreground objects. This spatio-temporal weighting concept enables a direct estimation of global motion parameters without additional outlier elimination algorithms adopted by most existing schemes. To verify the performance of the proposed scheme, numerical experiments on several unstable videos are conducted in the sense of computational load and accuracy in camera motion estimation and video stabilization.

2. Problem statements

The camera motion between two consecutive frames can be estimated from background motions in those frames because background regions do not have their own motion and their views are changed only by camera motion. For this reason, the background motions can provide information about the camera motion, even though the amount of information may vary with different characteristics. Generally, camera motion induces 2D image motion, which can be described by a 3×3 motion parameter matrix. The matrix represents the relationship between two image frames, and the matrix has different forms according to its various degrees of freedom (DOF). While a lower DOF model cannot describe complex camera motion, a higher DOF model requires heavy computation time, and might cause overfitting errors, e.g. [14,17] use a 3 DOF model and [4–6,13,16] adopt a 4 DOF model. In the proposed method, a 6 DOF model is adopted to handle complex camera motions.

With the defined motion parameter matrix, we adopt a minimization scheme for camera motion estimation. From Lucas Kanade Tracking (LKT) in [25], the following object function is minimized.

$$\min_H O(H) = \min_H \sum_{i=1}^{m} \{I_t(X_i) - I_{t+1}(f(X_i, H))\}^2, \quad (1)$$

where m is the number of extracted points, X_i is the ith discriminative point ($X_i = (x_i, y_i)$) on the tth frame, and $f(X_i, H)$ is the warping function that transforms X_i with H which is the motion parameter matrix. $I_t(X_i)$ is the intensity value of X_i on the tth frame, and $I_{t+1}(f(X_i, H))$ is the intensity value at the position of $f(X_i, H)$ in the $t+1$th frame (the position corresponds to the previous extracted point in the tth frame). Any minimization algorithm can be adopted, but, in our method, an iterative gradient descent method is used to solve the minimization problem. This formulation, however, does not consider different amount of information in each local point and weigh all points X_i's equally. To handle this problem, adopting an outlier rejection algorithm, such as RANSAC [24], is quite a standard solution to remove unreliable points but it is computationally expensive. To avoid high computational outlier rejection algorithms, we aim to establish a novel optimization formulation with spatio-temporal weighting on each local point in this paper.

3. Motion estimation with spatio-temporal weighting

3.1. Overview

The overview of the proposed method is illustrated in Fig. 1. Initially, on the first frame, local patches are generated randomly so that the center point of each patch lies on an edge (to be described in Section 3.2). After initialization, a new patch is generated only when an existing patch is discarded due to its low weight. The total number of patches is maintained by discard and re-generation. The weight of each patch is calculated by the proposed method from their spatio-temporal information (to be described in Section 3.4). Then, camera motion is estimated by the minimization of the

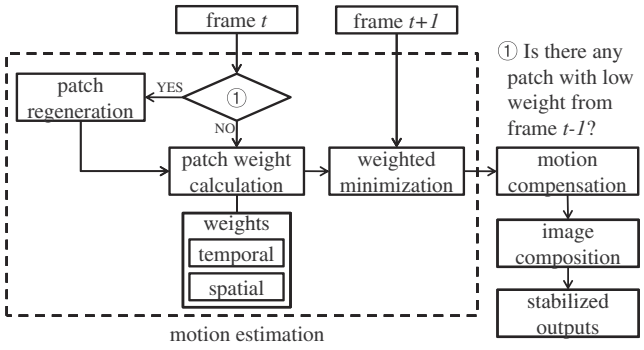

Fig. 1. The basic flow of the proposed algorithm.