Pattern Recognition 63 (2017) 279–290

Contents lists available at ScienceDirect

Pattern Recognition

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

A dense flow-based framework for real-time object registration under compound motion

Songfan Yang[a,b], Le An[c], Yinjie Lei[a,*], Mingyang Li[d], Ninad Thakoor[e], Bir Bhanu[e], Yiguang Liu[f]

[a] College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, China
[b] FaceThink Inc., Beijing, China
[c] National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan, Hubei, China
[d] Google Inc., Mountain View, USA
[e] Center for Research in Intelligent Systems, University of California, Riverside, USA
[f] School of Computer Science, Sichuan University, Chengdu, Sichuan, China

ARTICLE INFO

Keywords:
Object registration
Spontaneous facial expression
SIFT flow
Optical flow
Super-resolution

ABSTRACT

A moving object often has elastic and deformable surfaces (e.g., a human head). Tracking and measuring surface deformation while the object itself is also moving is a challenging, yet important problem in many video analysis tasks. For example, video-based facial expression recognition requires tracking non-rigid motions of facial features without being affected by any rigid motions of the head. In this paper, we present a generic video alignment framework to extract and characterize surface deformations accompanied by rigid-body motions with respect to a fixed reference (a canonical form). We propose a generic model for object alignment in a Bayesian framework, and rigorously show that a special case of the model results in a SIFT flow and optical flow based least-square problem. We demonstrate that dynamic programming can be used to speed up the computation of our algorithm. The proposed algorithm is evaluated on three applications, including the analysis of subtle facial muscle dynamics in spontaneous expressions, face image super-resolution, and generic object registration. Experimental results, in terms of both qualitative and quantitative measures, demonstrate the efficacy of the proposed algorithm, which can be executed in real time.

1. Introduction

Video registration is an important topic in video processing, computer vision and pattern recognition. It has various applications such as face recognition [1], facial expression recognition [2], image stitching [3], color demosaicking [4], etc. Depending upon different applications, there can be specific requirements for the registration techniques [5,6]. Broadly speaking, in the process of registration, most algorithms overlay objects spatially via motion estimation and compensation.

Video registration becomes a more challenging problem if there are object surface deformations which are further compounded by rigid-body motions or/and camera motion; in particular, if subtle surface non-rigid motions have to be detected and precisely characterized in applications such as medical imaging and facial expression. To appreciate the difficulties in precisely characterizing surface deformation amidst complex compound motion, let us examine a concrete example: the human facial expression analysis, in which the non-rigid muscle motion is of the central focus. Accurate facial expression analysis is hampered by the following complications:

1. Facial expression comprises non-rigid muscle motion and rigid head motion.
2. The head pose comprises both in-plane rotation and out-of-plane rotation.
3. The muscle motion is subtle in spontaneous expressions.
4. The data are streaming instead of being in a batch form.
5. The consecutive frames should comply with temporal smoothness constraint for micro-expression analysis.
6. The imaging condition varies, such as the illumination or resolution of the face region.

In this paper, we propose a new video registration approach, termed SIFT and Optical Flow Image Transform (SOFIT), that tackles the aforementioned challenges in aligning object features through video frames in the presence of compounded surface deformation

* Corresponding author.
E-mail address: yinjie@scu.edu.cn (Y. Lei).

http://dx.doi.org/10.1016/j.patcog.2016.10.015
Received 14 March 2016; Received in revised form 11 October 2016; Accepted 12 October 2016
Available online 14 October 2016

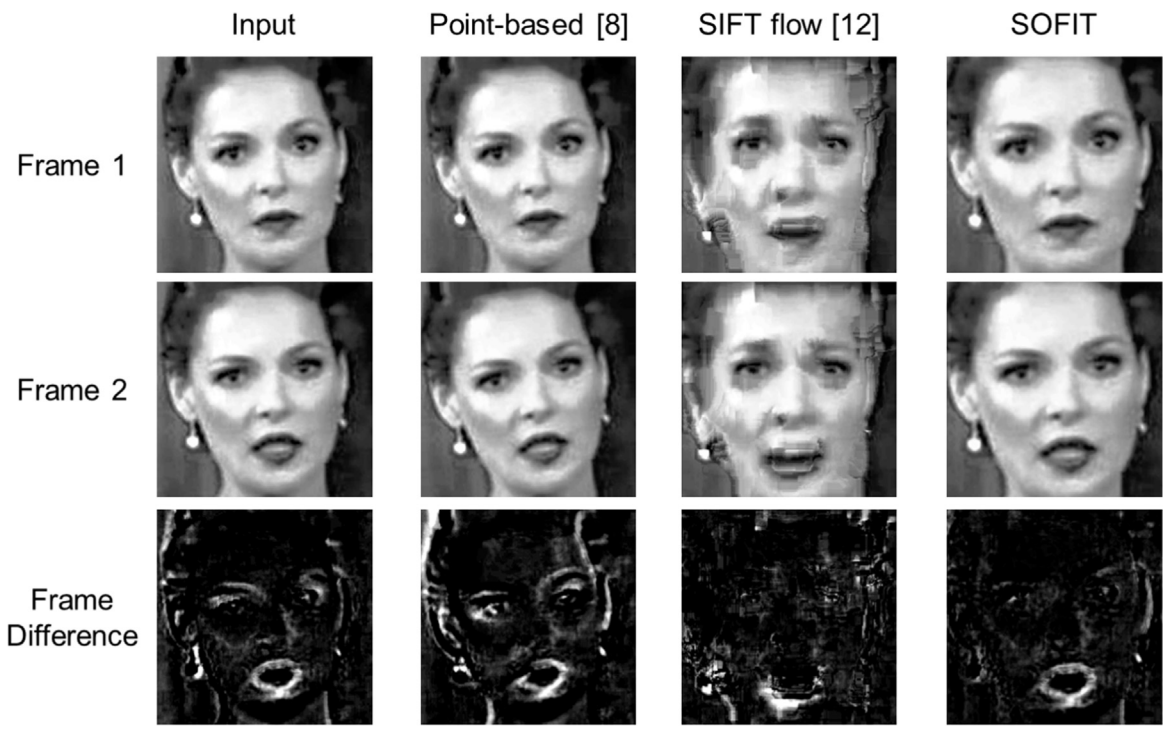

Fig. 1. Comparison of registration results. Row 3 is the absolute difference between frame 1 and frame 2. Column 2 is the point-based affine registration method used in [7–11], where affine (or piece-wise affine) transformation is computed from 83 facial feature points generated by the state-of-the-art detector [8]. Column 3 uses SIFT flow [12] to align with the Avatar Reference face model from [13]. Ideally, we would like the frame difference to show only at locations where the non-rigid motion is present (mouth area in this case). The proposed method, SOFIT, achieves the most plausible result.

and rigid motion.

In various tasks such as recognition, super-resolution, video compression, the deformed object should be aligned with respect to a canonical form or reference model. For instance, such a reference model is instrumental in facial expression analysis [13].

Facial muscle motion is similar for the same expression irrespective of the person [14], but the facial feature location (such as eyes, nose, mouth) of different people varies. Thus, finding a canonical reference feature location for all the faces is favorable for analyzing the dynamics of facial features across population. In other words, face registration is critical to facial expression recognition. In the proposed SOFIT approach, we need to transform every frame of the streaming video data into a canonical pose by neutralizing the effects of rigid body motion on the deformable object.

To further clarify the aforementioned design objective, let us examine, via Fig. 1, how different video registration methods behave when being applied to registering frame 2 with respect to frame 1. All methods in Fig. 1 are able to account for the in-plane head rotation. However, as illustrated by the frame difference images (row 3) for the point-based affine (or piece-wise affine) transformation (column 2) and the SIFT flow transformation (column 3), there is motion on most parts of the face. This is similar to the unaligned face image (column 1) where the image is the output of Viola-Jones face detector [15]. This suggests us to impose the temporal smoothness constraint so that the frame difference is small for areas with no motion; while for areas with motion (mouth area in this case), the frame difference should capture this change, as demonstrated by the results of the proposed method (column 4).

In this paper, we model the alignment-of-compound-motion problem in three steps. *First*, each frame is aligned with respect to a reference frame in a general distance measure, which is then instantiated to the SIFT flow criterion, thereafter. *Second*, our model enforces a smoothness constraint on adjacent frames. It is realistic for the consecutive frames to comply with the smoothness constraint. We realize this by depending this current transformation estimation on a number of previous frames in an optical flow criterion. *Third*, large transformation is penalized to prevent over-fitting. We also extend this approach to register many other types of objects and demonstrate applications in areas such as image super-resolution. More results can be found on our project website.[1]

The rest of the paper is organized as follows. After reviewing the related work and highlighting our contribution in Section 2, Section 3 presents our general model as well as the solution to the registration problem using the dense flow approximation to estimate the affine transformation parameters. The experimental results and discussions are provided in Section 4. Finally the conclusion is drawn in Section 5.

2. Related work and contributions

2.1. Related work

Video registration has been a fundamental topic in computer vision and image processing. Recent successful object retrieval and recognition methods, such as [16,17], have made progressive achievements, while accurate registration can further prompt the performance on these applications. As an object may undergo a complex motion (rigid and/or non-rigid), conventional video registration methods [5,6] attempt to correct both types of motion. On the contrary, we attempt to remove the rigid motion while retaining and characterizing the non-rigid motion. Such problem occurs when a moving object has deformable surface, which may contain crucial information (e.g. facial expression).

To analyze facial expressions, behavioral scientists have developed Facial Action Coding System (FACS) [14] as an objective standard to describe the muscle motion. According to FACS, human (coders) can decompose every possible facial behavior into Action Units (AU), which roughly correspond to the muscles that produce them. Automatic AU

[1] http://www.ee.ucr.edu/~syang/sofit/index.html.