Journal of Visual Communication and Image Representation 55 (2018) 253–262

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Visual adaptive tracking for monocular omnidirectional camera☆

Yazhe Tang[a,*], Zhi Gao[a,*], Feng Lin[a], Y.F. Li[b], Fei Wen[c]

[a] Temasek Laboratories, National University of Singapore, Singapore 117411, Singapore
[b] Department of Mechanical and Biometical Engigneering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong Special Administrative Region
[c] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Keywords:
Visual tracking
Omnidirectional camera
Multi-feature integration

ABSTRACT

This paper presents a sophisticated patch-based visual tracking algorithm using an omnidirectional camera with distortion adaptation. The omnidirectional camera is modeled using the equivalent projection theory, so that a nonlinear deformed neighbourhood can be accurately estimated in the image plane, which significantly facilitates feature coding. In order to improve the omnidirectional tracking performance, a patch-based multi-feature matching method is proposed under a probability framework. In particular, the distributions of patches covering key parts of the target are weighted adaptively according to their joint-feature response, which is able to track target robustly and filter out the outliers effectively. Extensive experiments have been conducted to verify the performance of the proposed omnidirectional tracking algorithm, which obtains promising results on challenging datasets and outperforms many state-of-the-art methods.

1. Introduction

Omnidirectional cameras have drawn significant interest because of its wide field of view from a single image, which enables it to obtain much more information. Catadioptric omnidirectional camera can provide an omnidirectional visual sensing with a compact structure. With the merit of big field of view, catadioptric omnidirectional sensors have been popularly applied in a variety of applications, including surveillance [7], intelligent vehicles [28], and robotics [4], etc.

Central catadioptric sensors may work as a combination of mirrors and lenses, such as paraboloid, hyperboloid, and ellipsoid systems. These sensors meet the definition of general single viewpoint system, thus can be referred as central vision system [2]. This paper mainly focuses on central catadioptric camera since the imaging geometry is much easier to be modeled than that of non-central one. Due to the serious distortion introduced by the catadioptric devices, the geometry of central catadioptric vision systems cannot be described by the conventional pinhole model. Consequently, conventional algorithms cannot be directly applied for Catadioptric Omnidirectional Vision (COV) systems. The widely known alternative approaches are based on image unwarping, which can remove distortion from the omnidirectional imaging device [13]. However, these solutions could increase the computational cost significantly for the rectification and interpolation operations. Moreover, such operations usually introduce much noise in image [8].

Alternatively, distortion involved neighborhood definitions on raw image are explored for efficiency concern. Svoboda et al. [25] explained why conventional rectangular neighborhood was not appropriate for omnidirectional vision, and introduced a method to define the neighborhood on a mirror, followed by projecting a small patch on the image plane for distorted neighborhood estimation. Ieng et al. [17] discussed the difficulty of matching local visual features in omnidirectional images for different resolutions. They proposed to extract patches of different angular apertures for the same feature to address the matching problems. Geyer and Daniilidis [14] applied equivalent sphere theorem to formulate the imaging process of vision system, which can be applied to the camera with single viewpoint constraint, such as perspective and catadioptric sensors. Benefiting from equivalent projection, the conventional visual methods can be extended to omnidirectional domain [9,3] for image processing and tracking applications. Similarly, Rameau et al. [23] integrated color feature with equivalent projection to track the moving object in distorted omnidirectional vision.

In feature coding level, tracking methods can be roughly divided into holistic-based methods [35] and part-based ones [34]. The first category performs feature coding holistically and models object densely over image locations. However, this model may lead to poor adaptability for occlusion [32]. To overcome this problems, local features coding has been extensively studied [12,6], and superior performance of such methods has been reported. When target is partially occluded, the remaining visible parts can still provide reliable cues for tracking.

☆ This paper has been recommended for acceptance by Xinchao Wang.

* Corresponding authors.
E-mail addresses: yztang2008@yahoo.com (Y. Tang), gaozhinus@gmail.com (Z. Gao), linfeng@nus.edu.sg (F. Lin), meyfli@cityu.edu.hk (Y.F. Li), wenfei@whu.edu.cn (F. Wen).

https://doi.org/10.1016/j.jvcir.2018.06.015
Received 30 November 2017; Received in revised form 23 March 2018; Accepted 14 June 2018
Available online 18 June 2018

Most of these tracking methods can be viewed as part-based object matching over time. The part-based tracker of [1] was one of the pioneering methods in this trend. In their tracker, parts corresponding to arbitrary patches voted for target positions and scales in a competitive manner. Erdem et al. [11] addressed the part-based tracking by differentiating the object patches based on their reliability. Each patch contributed to the target state prediction according to its reliability, achieving a better accuracy. More recently, Liu et al. [18] presented a part-based tracker using kernel correlation filters to efficiently model the local patches, in which the authors exploited motion constraint between frames. In addition, Makris et al. [20] developed a probabilistic local part-based object recognition framework using intensity and depth information. Therein, depth information is leveraged to facilitate the disambiguation of occlusion.

In this paper, we propose a promising solution for robust visual tracking in omnidirectional vision. Conventional methods are hard to achieve a satisfactory performance in catadioptric sensors due to its inherent distortion of image. To realize distortion adaptive representation, our work employs equivalent projection to model the geometric imaging of omnidirectional vision. Benefiting from equivalent projection, distortion involved neighborhood of local patch can be defined in omnidirectional image. Then, the distortion-adaptive feature coding can be developed to integrate multiple feature information, which can significantly improve the tracking performance [19]. For robust feature matching, a features joint model is proposed to integrate multiple feature matching into a single probability framework. Instead of using holistic representation, the target is divided into patches to improve flexibility. Given the system geometric model, a distortion adaptive patch-wised multi-feature joint representation is proposed for robust tracking. This framework provides a flexible solution to address the challenge of occlusion. Relying on the adaptive patch-based scheme, tracker can focus on the reliable patches and ignore the corrupted patches. This paper sets human as target for tracking experiment since it is nonrigid and representative. The performance of proposed method is validated in the challenging nonrigid target motion experiments. The main contributions of this paper are as follows:

(1) A flexible distortion involved patch-based tracking framework is proposed for omnidirectional image. The distortion adaptive neighborhood is calculated according to the equivalent projection transformation. A probabilistic model is proposed to integrate the patch units for a flexible tracking. Different from the previous works [27] which formulated distortion-invariant feature under the hypothesis of geometric constraint, the proposed equivalent projection can be extended to more general applications.
(2) A distortion adaptive multi-feature fusion method is proposed to enhance the robustness of our algorithm. Multi-feature integration is a key problem in feature matching, which might significantly improve the performance of algorithm. Given the proposed neighborhood definition, distortion adaptive multi-feature coding is designed to formulate the spatial constraint multi-feature joint response for enhanced target matching. To adapt complex tracking, online multi-feature scheme is presented in feature formulation.
(3) A weight estimation mechanism is presented to adaptively adjust the contribution. Patch-based algorithm provides a flexible way to handle difficult scenarios. The reliable patches will dominate the tracking guidance and unreliable ones will decrease its function.

The rest of this paper is organized as follows. Section 2 introduces the geometric model of catadioptric camera using equivalent projection and defines the distortion adaptive neighborhood for feature coding. Section 3 elaborates on the proposed patch-wised multi-features integration approach which formulates joint-feature response within a single probabilistic model. Section 4 presents experiments that verify the performance of the proposed algorithm and conclusions are made in Section 5.

2. Catadioptric omnidirectional image formation

The adaption of neighborhood is essential to guarantee the accuracy of visual tracking. Usually, the neighborhood of a given point is simply defined as the square region centered at this point. However, central catadioptric omnidirectional camera exhibits serious nonlinear distortion due to the involvement of a quadratic reflection mirror. Therefore, conventional definition of neighborhood is not appropriated for catadioptric images because such special distortion of image is not considered.

2.1. Central catadioptric image formation

Catadioptric omnidirectional sensors can be classified into two categories depending on the number of viewpoints [2]. Central catadioptric sensors with single viewpoint can be geometrically corrected and reconstructed using the original catadioptric image. These kinds of sensor consists of a parabolic mirror associated with an orthographic camera and hyperbolic, elliptic and plane mirrors with a perspective camera. Non-central catadioptric sensors with several viewpoints have much less significant geometric properties and may be made of the other possibilities of associations between mirrors and cameras. This paper focuses only on the former category. As discussed in Geyer and Daniilidis [14], catadioptric omnidirectional vision can be modeled by a unified projection model, which is equivalent to a two-step projection via a unitary sphere centered on the focus of the mirror. Firstly, a 3D point X_w is projected to the sphere centered at O_c, as shown in Fig. 1(a). Subsequently, the point on the sphere X_s is projected to the image plane to obtain a pixel point X_i. Such equivalence is very interesting since it allows performing image processing in a new space in which deformation has been considered.

2.2. Geometric projection from image to sphere

Fig. 1 shows the equivalent projection model which depicts a unit sphere centered at O_c. The dotted line represents the reflection mirror for catadioptric camera, which is also centered at O_c. The point O_p is located at the vertical axis of camera with distance ξ from center point O_c. The point O_p is between the center point O_c and surface of unit sphere, depending on the shape of mirror. The equivalent image plane I is placed in a distance φ from center point O_c as shown in Fig. 1. Then, we have:

$$\frac{x_w}{x_s} = \frac{y_w}{y_s} = \frac{z_w}{z_s} = \gamma \tag{1}$$

With the given condition of unitary sphere coordinate (x_s, y_s, z_s), we have $x_s^2 + y_s^2 + z_s^2 = 1$. Based on Eq. (2), we obtain:

$$\gamma = \sqrt{x_w^2 + y_w^2 + z_w^2} \tag{2}$$

According to the perspective projection principle, the coordinates of point X_i on the image plane are then obtained from the intersection point between ray $\overrightarrow{O_p X_i}$ and image plane I passing through point X_s. Then, the geometric relationship between world point X_w and image point X_i can be formulated as

$$\begin{cases} \frac{x_s}{\xi + z_s} = \frac{x}{\xi + \varphi} \\ \frac{y_s}{\xi + z_s} = \frac{y}{\xi + \varphi} \\ z = \varphi \end{cases} \tag{3}$$

where is the coordinate of point X_i in the normalized camera plane. According to Eqs. (2) and (2), we have: