J. Vis. Commun. Image R. 23 (2012) 397–408

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

3D information extraction using Region-based Deformable Net for monocular robot navigation

Khaled M. Shaaban [a,*], Nagwa M. Omar [b]

[a] Electrical Engineering Department, Faculty of Engineering, Assiut University, Assiut, Egypt
[b] Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut, Egypt

ARTICLE INFO

Article history:
Received 14 May 2011
Accepted 6 December 2011
Available online 14 December 2011

Keywords:
Robot navigation
Monocular vision
Stereo vision
Correspondence problem
Video segmentation
Deformable contours
3D information extraction
Depth information extraction

ABSTRACT

This paper extends the Region-based Deformable Net (RbDN) technique described in [1] to extract the 3D information of all the objects in the scene from a single moving camera. The technique is used for segmenting real-time video sequences captured from a single moving camera. The deformation process tracks the changes in the location and the shape of the segments across the frames. These changes along with the camera displacement are used to estimate the 3D information. The algorithm is completely autonomous and does not require pre-knowledge, training, or assumption about the contents of the sequence. It can handle the difficult case where the motion of the camera is parallel to its optical axis. It can also estimate the distances to objects that are more than 100 m away as long as the camera displacement is over 10% of the expected distance to the objects.

1. Introduction

Robot navigation based upon self-measurements like odometer for moved distance and compass for angle of rotations leads to accumulative error in the final position. This error grows with time until the robot loses orientation. Observing landmarks then estimating the position relative to them does not suffer from this error accumulation. That is why it seems natural to seek navigation using Machine Vision [2–8]. Navigation requires estimating the 3D information of the objects in the scene relative to the position of the camera. Three bases for extracting this information are available: Time-of-Flight cameras, Stereo Vision and Monocular Vision.

In the Time-of-Flight depth estimation technique, the time required for the light to complete a round-trip from the camera to the object is used to estimate depth information. Usually a modulated light (typically infrared) is transmitted toward the object and its reflection is received by the camera. The phase shift between the transmitted and the received light depends upon the distance travelled. The ability to separate the incoming signal depends upon the intensity of the reflected light, the intensity of the background light and the dynamic range of the sensor. The Sun light has more than 20 W/m^2 and the light source is typically Light Emitting Diodes (LEDs) with power in the milliwatt range, which makes detecting the incoming light a hard task. Therefore the range of accurate distance measurements is usually 1–4 m [9]. Also the cyclic characteristic of the modulated light results in ambiguity beyond a certain distance. For example, this distance is 7.5 m if the modulation frequency is 20 MHz [10]. Lowering the modulation frequency increases the distance but decreases the resolution of the depth estimation. These factors in addition to the multi reflection error limit the applications of the Time-of-Flight camera.

The second technique is Monocular Vision [5,8,11–16], in which the 3D information is extracted from a sequence of images acquired under a relative motion of the camera. The last technique is Stereo Vision [17–20], in which the 3D information is obtained from two separate views of the same scene. Stereo Vision accuracy decreases rapidly with the increase of the distance to the object compared to the baseline distance separating the two views. In this case Monocular Vision strategy provides a better solution.

In Monocular Vision navigation, the recovery of the 3D information requires reliable and continuous Video Segmentation in real-time. Video Segmentation is the process of partitioning the first frame into its segments then tracking these segments across the subsequent frames [21–25]. This tracking information along with the camera displacement provides the means to estimate the distance to the objects represented by these segments. Tracking is an active area of research with a broad range of concepts [26–29]. These concepts could be classified into four main categories [27]: model-based, region-based, feature-based, and contour based methods.

* Corresponding author. Fax: +20 88 2327254.
E-mail addresses: kshaaban@hotmail.com (K.M. Shaaban), nagwa_omar@hotmail.com (N.M. Omar).

1047-3203/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2011.12.001

Model-based tracking methods use detailed prior knowledge of the object shapes for the matching process. It needs accurate geometrical models for the objects to be found in the scene. This approach suffers from two drawbacks [27]. First, it cannot detect objects not included in the database. Second, the construction of the model is complex and for some objects is impossible.

Region-based methods track an object using the 2-D shape of the connected region that represents its projection in the image. This tracking approach depends upon information provided by the entire region pixels such as motion, color, or texture. The participation of all pixels usually makes this class of algorithms more robust but expensive [27]. Also moving objects with fast variations in their projection in the image is harder to track.

Feature-based methods track the extracted selected features of each object. Several feature-based matching techniques have been proposed, but they are not specifically designed for video object tracking. An adaptation to object tracking is presented in [27]. In [27] the video objects are tracked by tracking their corners. Tracking of the object features independently results in easy and stable tracking algorithms. However, the problem of grouping the features to determine the group that belong to the same object is expensive especially for scene with large number of objects [27].

Contour-based methods track the contour of the object instead of tracking the whole set of object pixels. These methods use the motion information to estimate an anticipated contour in the next frame. Then the shape and location of the anticipated contour is enhanced to improve the fit to the object. Furthermore the actual change in the contour location is used to update the motion information [27]. The Deformable Contour Method [26,28–35] is a good example for algorithm belonging to this class but it faces some drawbacks as will be illustrated in Section 2.

In [1] we proposed a new segmentation technique for still images that belongs to the Deformable Contour Method (DCM). We called this technique Region-based Deformable Net (RbDN). This technique, as will be summarized in the Section 2, is capable of performing entire image segmentation based on the homogeneity of the color distribution within the regions. This technique has a short segmentation time and a small number of segments. These qualities lend itself to Video Segmentation particularly if we consider the small time separating successive frames. The small image variation between frames leads to fast convergence of the deformation process. Therefore tracking the changes in the image is relatively fast when compared with the classical feature matching usually needed in Stereo Vision systems. This speed allows for the real-time performance necessary for robotic application. In this paper we will extend this segmentation technique to be used for Video Segmentation and test its applicability for Monocular Vision navigation.

The rest of this paper is organized as follows: Section 2 reviews the RbDN technique. Section 3 describes the use of the RbDN technique to segment a video sequence. Section 4 explains using the RbDN technique to extract the objects 3D information. Section 5 shows some of the experimental results. Section 6 concludes this work.

2. RbDN technique review

In general, Deformable Contour Methods (DCMs) are energy minimizing techniques that deform a single or multiple independent contours under the influence of internal and external forces [26,28–36]. The internal forces impose the contour smoothness and the external forces attract the contour to the object boundary. DCMs try to minimize the integration of these forces around the contour. Parametric deformable models [31,36], geometric deformable models [30,35] and Prototype-based deformable models [32–34] are popular examples. Another type of deformable contours depends upon indirect representation of the regions' boundaries in the form of Level Sets [35]. Despite the advantages of the Level Set methods, their computational cost is high, which limits their use for real-time application [30].

The current use of DCMs is limited to a single object tracking or few independent ones [29]. When DCMs are used to track more than one object, multiple independent contours are used. Such multiple contour strategies suffer from problems [29] such as:

- Multiple contours drift into each other and occasionally track the same image features. These drifts lead to meaningless results.
- The number of contours initialized over the image fixes the number of segments to be tracked. Therefore, the algorithm cannot handle removed or newly added objects to the scene.
- The tracking success is highly sensitive to initialization. Thus it is difficult to start tracking process automatically.
- The deformation results have no topological information to describe how the features are related to each other spatially. Such information is of great value to higher scene understanding necessary for navigation.

As an attempt to solve the above problems the Region-based Deformable Net technique was introduced in [1]. RbDN technique deforms a single planar net that we call Deformable Net (DN). This mathematical structure is capable of comprising all information regarding the segmented regions. The RbDN uses the deformable contour strategy to enhance the fit of this mathematical structure to the contours of the segments in the image. In the next sections a brief description of the DN and its deformation will be given.

2.1. Deformable Net (DN) structure

The DN combines concepts from the Contour Representation and Region Adjacency Graph (RAG) [37]. It is simply a plane graph that consists of a group of vertices, V, connected by edges, E. Each vertex, v, is represented by a point in the Euclidian plane and each edge, e, is a line segment that connects two vertices. The plane graph has a unique characteristic: it can be sketched on a piece of paper in such a way that no edges meet in a point other than the common ends (the vertices). The set of edges is divided into subsets each represents a polygon in the graph. Every edge contributes in exactly two polygons except the edges at the outer boundary. In general, there is a large number of ways in which the edges can be ordered into polygons. A unique order is to use polygons with the smallest possible area. That is to minimize the overlapping of polygons. Therefore, the DN is a way to partition the graph into a set of polygons. As a result, the mathematical notation for the Deformable Net could be written as $DN = (V, E, P)$ where V is the set of all vertices, and E is the set of all edges and P is the set of all polygons satisfying the given condition.

When this structure is used to represent a segmented image the vertices are aligned with the segments corners and the polygons approximate to the contour of the segments. Fig. 1 represents an example of such structure.

This mathematical representation is necessary to introduce the concept of deformation to the process of image segmentation. One can easily imagine the process of deformation as the process of adjusting the location of the vertices to coincide the segments' corners in the real image. Further more, the process of Video Segmentation could be considered as the process of maintaining the correspondence between the DN vertices and the objects' corners as they change location across the frames.

The DN as a mathematical description of the segmented image is useless without a set of operations to manipulate this structure. The following is a list of the essential operations that can be performed on this description: