Optics and Lasers in Engineering 95 (2017) 69–77

Contents lists available at ScienceDirect

Optics and Lasers in Engineering

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

A fast bottom-up approach toward three-dimensional human pose estimation using an array of cameras

Maryam Ghaneizad[a], Zahra Kavehvash[a,*], Khashayar Mehrany[a], S.M. Ali Tayaranian Hosseini[b]

[a] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
[b] Department of Electrical Engineering, Amirkabir University of Technology, Iran

ARTICLE INFO

Keywords:
3D human pose estimation
3D reconstruction
Integral imaging

ABSTRACT

In this paper, employing recorded images of multiple cameras, we propose a novel fast approach for three-dimensional (3D) human pose reconstruction. Opening a new framework to the pose estimation application, the proposed method is inspired by optical 3D reconstruction algorithms conventionally used for integral imaging. Thanks to the fact that the pose estimation can be carried out by using only key features of the raw recorded images, the computation time and the complexity of our method are considerably reduced. Furthermore, utilizing the here proposed algorithm, rapid variations in actions can be easily tracked. The validity of the proposed method is demonstrated by several experimental results in different postures.

1. Introduction

3D human pose estimation is an important problem with many potential applications from gaming and human-computer interaction to smart care centers and surveillance systems [1–3]. The ability to recognize different human postures in real-time can also be very useful for semantic labeling of various environments; that is to perceive the environment in which cameras are built in by extracting the humans' body pose and their activities [4].

Many different methods have been proposed for human pose estimation. Among them, a variety of methods have focused on the 3D reconstruction of the body pose from monocular image sequences [5,6]. Although such methods can successfully recover human pose in different scenarios, there are some situations where the intrinsic inability of monocular cameras to fully explain 3D objects impedes successful reconstruction of human pose. Human body self-occlusion and non-observability of motions along axial direction counts among the most noticeable reasons for the failure of human pose estimation by using monocular image sequences [7]. In recent years, some approaches have been developed based on the depth image analysis. Depth images could provide depth information by encoding an image into different colors [8–13]. For example, in [8–11], depth images provided by Kinect have been used for human detection or pose estimation in indoor environments. Recent approaches on human motion analysis using depth data are reviewed in [12,13]. Nevertheless, all methods using depth images suffer from strong noise added to the depth data and also the self-occlusion problem.

There are several other methods [14–16] for human pose estimation, which employ multiple video-streams recorded from different viewing angles and therefore are not burdened with the aforementioned problems. These methods can be generally divided into *top-down* and *bottom-up* categories [18].

In top-down methods [14,19–21], a 3D pose model is first assumed, which is usually referred to as the previous time model. A similarity evaluation function is then defined between the 3D model and 2D images. Next, the sought-after 3D model is iteratively obtained as the similarity evaluation function reaches the extremum value [14]. However, searching for the optimal 3D model is a complicated and time-consuming process, which necessitates a nonlinear hierarchical optimization algorithm. Furthermore, the validity of such methods relies heavily upon the validity of time consistency in postures. In other words, in case of relatively rapid body movements, the optimization algorithm will be likely to converge to a local optimum and hence the algorithm fails to lead to the accurate 3D model at the current time [14]. The resultant error can further propagate in future frames' model estimation and end up with a large amount of error.

In bottom-up methods [23–25]; however, features of interest, e.g. silhouettes, edges or body parts, are first detected in 2D images. These features are then assembled into the 3D model. Thus, the accuracy of such methods do not rely on the validity of time consistency in postures. The assembly of the detected features into a certain 3D configuration is a heavy computational task. This computational load may, however, be decreased by using optical imaging principles. To this end, in this paper, for the first time, as far as we know, an optically inspired fast

* Corresponding author: Tel.: +98-9127107687
E-mail address: kavehvash@sharif.edu (Z. Kavehvash).

http://dx.doi.org/10.1016/j.optlaseng.2017.03.012
Received 5 January 2017; Received in revised form 18 March 2017; Accepted 29 March 2017
Available online 17 April 2017

bottom-up approach for multi-camera image fusion is used in the pose estimation application. Thanks to the fact that the proposed optically inspired 3D image fusion technique doesn't require solving any optimization problem, its time, accuracy and simplicity are much improved compared to conventional top-down frameworks. To the best of our knowledge, the idea of using integral imaging principles has only been utilized in human hand gesture recognition where optical principles are used for the mere purpose of blurring the background [17]. In contrast, the here proposed optically inspired method is employed to provide the 3D human posture which is to be used in whole body gesture analysis.

Given that the proposed approach does not depend on the preceding frames information, any sudden movement of the body can be accurately tracked. Moreover, the speed of pose estimation can be further increased not only because the iterative steps are no longer necessary but also because whole body images conventionally used for pose estimation can be substituted by the image of seed points, i.e. the 3D location of hands, head and shoulders. The remainder of the paper is structured as follows. Section 2 describes the proposed optical fusion algorithm for human pose estimation. Section 3 discusses the mathematical formulation for our proposed framework. Section 4 confirms the proposed structure through experimental results and finally, Section 5 concludes the paper.

2. The proposed multi-camera image fusion algorithm

The proposed framework is based on an intuitive algorithm for fusion of multi-camera images based on closed-form equations used in geometrical optics. The goal is to locate the 3D position of human seed points; hands, head and shoulders, from which the 3D skeleton may be reconstructed. The proposed bottom-up algorithm is very fast and accurate and hence can be a suitable alternative for the complicated vision-based algorithms used for 3D pose estimation. Having no dependencies on preceding frames information, the model provides two advantages: First, the pose reconstruction is very fast and the algorithm can be implemented in real-time. Second, the algorithm is highly immune to estimation errors and can preclude any probable error from further propagating in successive frames. The proposed algorithm contains the following steps:

(i) Recording multiple 2D images from different viewpoints by ordinary 2D cameras.
(ii) Extracting the seed points on each image based on the image processing algorithm discussed in [18].
(iii) Fusing seed points (images) through optically inspired relations.
(iv) Post-processing the reconstructed 3D images to derive the 3D location of seed points and building the final 3D skeleton.

Let us explain each of the aforementioned steps more rigorously.

At the first step of the algorithm, we need to record multiple 2D intensity images from the intended person. To this end, we have proposed a curved arrangement of cameras in addition to the standard linear arrangement. In both arrangements, we are able to use a fairly small number of cameras. Furthermore, both arrangements are one-dimensional and no vertical shift is introduced between camera positions. For the curved arrangement, we may further reduce the number of cameras compared to the linear arrangement for a similar performance. This is mainly due to the fact that in the curved arrangement, the diversity between different viewpoint images is increased. Both arrangements will be further discussed in Section 3. The second step is a supplementary step which is included to further improve the efficiency of the proposed algorithm. In our application, i.e. the human pose reconstruction, a lot of information of the human body and the background scene is excessive and needs not to be included in the reconstruction process. Therefore, we have included an image data reduction step in our algorithm to further improve depth resolution, the system speed, the energy consumption and the system available memory. After recording 2D images by the cameras, each image may be separately processed to yield the only required information for the algorithm which is the positions of so-called seed points. Detection of seed points can be done very easily by the image processing algorithm described in [18]. Doing so enables us to quantize the depth information of the object which lowers down the required depth resolution. This low resolution can be achieved with only a few cameras making the system more cost effective. In other words, using a limited number of cameras, as is the case herein, the depth resolution is very low and without the aforementioned image data reduction, the algorithm will fail to reconstruct depth information. The novelty to remove the excessive information and keep only seed points in 2D images, helps us to get the most benefit from the available axial resolution. Doing so, we quantize the human body information in 2D images. Since the seed points are apart from each other and the variations in depth are not abrupt, there will remain only a few intensity points with sufficient distances along the axial direction out of a large number of neighboring points with nearly the same depth values. Thus, the available depth resolution, though it is low, will be enough to differentiate these points in the axial direction. Here after, we may call the resultant binary images, containing one at the location of a seed point and zero elsewhere, seed images. Seed images can be represented as the following:

$$\hat{I}_i = \sum_{j=1}^{5} \delta[x - x_j^i, y - y_j^i] \tag{1}$$

where (x_j^i, y_j^i) are the coordinates (of the image) of jth seed point in the ith seed image, δ is the 2D Kronecker delta function and five seed points are considered to be the 3D location of head, two shoulders and two hands. At the next step of the algorithm, the proposed optically inspired reconstruction algorithm is used for the fusion of seed images based on the formulation which will be discussed in Section 3. As will be seen there, considering a specific seed point, e.g. head, in all seed images and fusing the images based on optical principles to reconstruct the 3D image at different depth planes, lead to a plain output at depth z_0 where the head has been originally located and blurred outputs at other depths. The amount of blurring in the reconstructed output at an arbitrary depth plane is proportional to its axial distance from z_0. Therefore, reconstructing at z_0, the plain image of the head is accompanied by the adverse blurred images of all other seed points. Although we have drastically reduced this undesirable effect by substituting the whole recorded image at the first step by only a few seed points at the second step, a post-processing is also needed to completely eliminate it.

The final step of the algorithm is the aforementioned post-processing to clear the discussed undesirable effect in reconstructed images and derive a clean 3D image of the 3D locations of seed points, i.e. $(x_j, y_j, z_j); j = 1, \ldots, 5$. Joining these virtual seed points together in a volumetric space, we are able to reconstruct the 3D pose of the human.

Thus, the proposed algorithm is a combination of the optically inspired relations and the image processing algorithms to further improve the efficiency and the performance of the pose estimation algorithm. The detailed formulation of the proposed algorithm for both mentioned camera configurations will be discussed in the next section.

3. Mathematical formulation

In this section, the proposed optically inspired multi-camera image fusion technique is mathematically formulated. Let us consider the viewpoint images recorded by each camera as $I_i(x', y');\ i = 1, \ldots, N$, with x' and y' in the range of the image size and N the total number of cameras and let us represent the seed images as $\hat{I}_i(x', y')$ each being a sparse binary matrix. The seed images are the required input for the reconstruction process. In this process, 2D positions of each seed point