applied sciences

MDPI

Article

A 3D Human Skeletonization Algorithm for a Single Monocular Camera Based on Spatial–Temporal Discrete Shadow Integration

Jie Hou *, Baolong Guo, Wangpeng He and Jinfu Wu

School of Aerospace Science and Technology, Xidian University, No. 2 Taibai Rd, Xi'an 710071, China;
blguo@xidian.edu.cn (B.G.); hewp@xidian.edu.cn (W.H.); wjf505@gmail.com (J.W.)
* Correspondence: jie.hou.xdu@gmail.com; Tel.: +86-137-5993-1558

Academic Editor: Takayoshi Kobayashi
Received: 8 May 2017; Accepted: 29 June 2017; Published: 3 July 2017

Abstract: Three-dimensional (3D) human skeleton extraction is a powerful tool for activity acquirement and analyses, spawning a variety of applications on somatosensory control, virtual reality and many prospering fields. However, the 3D human skeletonization relies heavily on RGB-Depth (RGB-D) cameras, expensive wearable sensors and specific lightening conditions, resulting in great limitation of its outdoor applications. This paper presents a novel 3D human skeleton extraction method designed for the monocular camera large scale outdoor scenarios. The proposed algorithm aggregates spatial–temporal discrete joint positions extracted from human shadow on the ground. Firstly, the projected silhouette information is recovered from human shadow on the ground for each frame, followed by the extraction of two-dimensional (2D) joint projected positions. Then extracted 2D joint positions are categorized into different sets according to activity silhouette categories. Finally, spatial–temporal integration of same-category 2D joint positions is carried out to generate 3D human skeletons. The proposed method proves accurate and efficient in outdoor human skeletonization application based on several comparisons with the traditional RGB-D method. Finally, the application of the proposed method to RGB-D skeletonization enhancement is discussed.

Keywords: 3D model; shadow information; skeletonization; spatial–temporal integration

1. Introduction

The development of three-dimensional (3D) human skeleton extraction contributes enormously to prospering fields like virtual reality and somatosensory human–computer interaction. However, current 3D human skeletonization algorithms require specified acquisition equipments including RGB-Depth (RGB-D) cameras and wearable sensors, or a specific experimental setup like ring illuminator array. RGB-D cameras like Microsoft Kinect are designed to perform human skeletonization in a short range [1,2]. Wearable sensors only perform effective skeletonization on human subjects wearing experimental tags. Ring illuminator array requires precise subject position and illuminator array setup during the 3D modelling and skeletonization procedures. These setup restrictions of traditional human skeletonization methods bring great limitation on the outdoor applications.

Instead of deploying algorithms on traditional specified platforms, this work pays attention to the commonest projection of the human body on the ground. Shadow is the projection of a opaque object on a certain surface, containing single-view silhouette information of the object. Multiple methods have been developed to extract information from shadow. Current methods mainly focus on the recovery of mesh model [3–5] or point clouds [6,7] of static objects based on partial shadow information [8]. In this paper, a silhouetted shadow-based skeleton extraction (SSSE) method is proposed. The proposed SSSE method deploys shadow information extraction algorithm to the field of human skeletonization [9–11].

Based on the proposed SSSE method, six 3D joint positions in the human skeleton can be precisely extracted in outdoor scenarios with a normal monocular camera. Compared with current indoor 3D human skeleton extraction methods based on RGB-D cameras like Kinect, the proposed SSSE method reduces constraints on input device choice and application environment setup.

This work is motivated by the procedure of taking a silhouette photo. During this procedure, the human body blocks a part of light from reaching film or sensor, leaving a body sketch on the silhouette photo. Human shadow on the ground, from the aspect of silhouette imaging, can be regarded as a silhouette photo of the human body on a special giant film. The ground surface plays the role of film. For captured frames containing human shadows, each shadow on the ground can provide extra human contour information from a unique observation angle view other than the camera view.

This paper mainly focuses on the extraction and aggregation of the extra silhouette information from spatial–temporal discrete human shadows on the ground, aiming to perform 3D human skeletonization with a monocular camera in outdoor scenarios. Based on the aggregation of multiple shadows from discrete spatial–temporal coordinates, SSSE is capable of launching 3D human skeletonization even in outdoor scenes where the scale is too large for traditional methods [12–14] to handle [6,15,16]. The main contributions of this paper are related to three aspects:

(1) The 3D human skeletonization is realized with a normal monocular camera based on the proposed SSSE method.
(2) The proposed SSSE method achieves 3D human skeletonization in a large-scale outdoor scene.
(3) The proposed SSSE method deploys the aggregation of temporal–spatial discrete two-dimensional (2D) shadow information in a 3D human skeletonization procedure

The remaining sections of this paper are organized as follows: In Section 2, the basic theory for shadow-based single frame human skeletonization is introduced first, followed by the advanced SSSE method aggregating temporal–spatial discrete shadow information to recover complete skeleton sequences. In Section 3, a five-step method is introduced to deploy the proposed SSSE method in large-scale outdoor scenarios with a monocular camera. In Section 4, the effective range and precision of the SSSE skeletonization results are evaluated in comparison with the skeletonization result of traditional RGB-D method. Additionally, a fusion application of the SSSE and RGB-D skeletonization method is achieved in Section 5, providing much wider effective range in outdoor scenarios. Eventually, the advantages and potential applications of the proposed SSSE method are illustrated in Section 6.

2. Basic Theory

This section presents the basic theory of the silhouetted shadow-based 3D human skeletonization method. To illustrate our method clearly, the basic theory under multiple light source scenarios is introduced first. Then the advanced theory designed to aggregate temporal–spatial discrete shadow information is introduced to achieve skeleton recovery under single light source scenarios.

2.1. Skeleton Simulation in Multi-Light-Source Scenarios

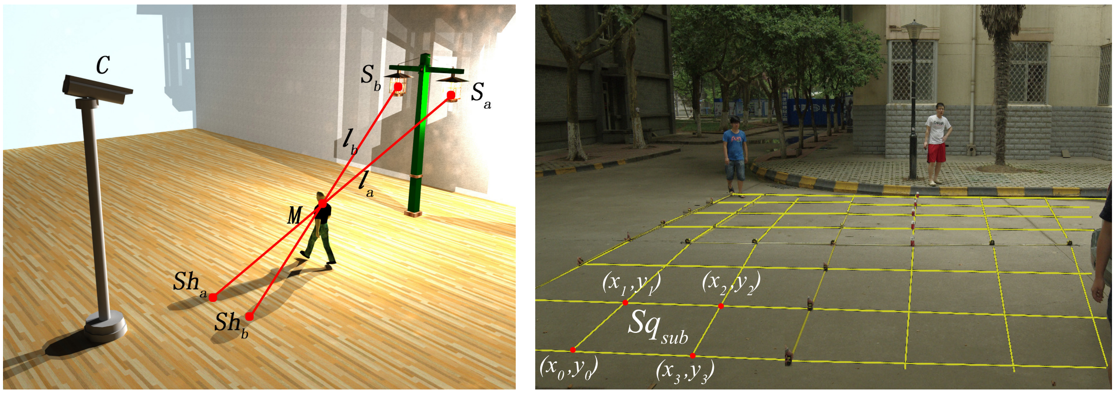
In a multiple light source scenario, contour of each human shadow on the ground is decided by two factors:

(1) human contour shape.
(2) positional relationship between the light source and the human.

Since each single shadow on the ground is restricted in a 2D plate, it is impossible to reproduce 3D information from any single shadow image. However, multiple shadows generated by different light sources can carry contour information from multiple 3D view angles, allowing the reproduction of 3D information.

A 3D voxel model of an object can be simulated from shadows generated by a annular set of light sources [6]. However, out of the laboratory environment, accurate manual arrangement of light source positions is elusive. Thus SSSE is designed to be adaptive to the posterior combination of random light source positions. With two or more shadows generated by different light sources, our method is capable of simulating 3D human skeleton information.

In a multiple light source scenario shown in Figure 1a, Sh_a and Sh_b are shadows of human body M, generated by light sources S_a and S_b respectively. The scenario is captured by camera C, and the human body M and shadows Sh_a, Sh_b are captured as P, Shc_a, Shc_b in the frame, respectively. With the captured frame, 3D position of the certain joint part $M_p \in M$ can be extracted based on the following three steps.



(**a**) (**b**)

Figure 1. Demo of silhouetted shadow-based skeleton extraction (SSSE) in a multi light source scenario: (**a**) A simulated dual light source scenario; (**b**) Scenario reconstruction.

2.1.1. 3D Scenario Reproduction

The silhouettes of human shadow are projected on the ground. To locate 2D shadow areas corresponding to different human joints, 3D scenario reproduction is launched to extract original 2D silhouette information Sh from the corresponding images Sh_c captured by camera C. The extraction is launched through the a two step perspective transformation between the ground surface plane $S(u, v)$ and the camera coordinate plane $C(x', y')$. Due to the progressive road engineering and partial patching, the height levels between different road parts are normally discontinuous. Instead of deploying global perspective transformation between the ground surface plane $S(u, v)$ and image coordinate plane $Im(x, y)$, this work proposes a block matrix-based projection transformation optimized for uneven road surfaces.

Before the 3D scenario reproduction procedure, the projection transformation parameter matrices are calculated once. Then frame-by-frame block matrix-based projection transformations are launched to extract original silhouette information.

In order to illustrate the block matrix-based projection transformation clearly, traditional plane-to-plane projection transformation is presented first. Then the block matrix-based projection transformation is introduced along with the optimized extraction solution for parameter matrices. Based on the extracted parameter matrices, the simplified Equation (15) for frame-by-frame projection transformation is presented.

Plane-to-Plane Projection Transformation

During the imaging process of a monocular camera, the projection transformation from ground surface plane to image coordinate plane is carried out in two steps. Firstly, each point (u, v) on ground surface plane is projected to corresponding coordinates (x', y') on the camera coordinate plane.

Secondly, a linear transformation happens inside the camera, transforming coordinates (x', y') to pixel coordinates (x, y) on the image coordinate plane.

The projection transformation between the point coordinates (u, v) on the ground surface plane and the corresponding coordinates (x', y') on the camera coordinate plane is presented as below:

$$[x', y', w'] = [u, v, 1]\, A \tag{1}$$

w' is a fixed camera internal parameter that affects the linear transformation from the camera coordinate plane to the image coordinate plane. Noticeably, A is the projection transformation calibration matrix that defines the relationship between ground surface plane and camera coordinate plane. Multiple transformations are taken into consideration in architecting the projection transformation calibration matrix A.

- Rotation transformation. Most surveillance cameras are not precisely set up at the horizontal angle which is parallel with the ground surface. The non-horizontal installation attitude brings a rotated field of view. The rotation transformation is introduced to calibrate the rotated field of view, ensuring the calibrated field of view parallel with the ground surface.
- Scale transformation. The coordinate system of the ground surface plane is measured in centimeters. However, pixel is the basic unit of measurement in the image coordinate plane. Thus the scale transformation is introduced to bridge two different units of measurement, extracting ground surface plane coordinates from the pixel coordinates.

Both rotation transformation and scale transformation are linear transformations. The coordinates of both transformations are combined into the linear parameter matrix L.

- Translation transformation. For the image coordinate plane, the origin of the coordinate system is fixed at the bottom left corner. For each captured frame, the origin of the coordinate system on the ground surface plane does not necessarily coincide with the origin of image coordinate plane. The translation transformation is introduced to calibrate the translation between two coordinate systems. The detailed parameters for translation transformation are given in parameter matrix T.
- Perspective transformation. Instead of the flat view, a perspective view is captured by each monocular surveillance camera in each frame. Thus, the perspective transformation is introduced to recover the flat ground surface plane from the captured perspective view. The detailed perspective transformation parameters are given in parameter matrix P.

For linear transformation parameter matrix L, the scale transformation parameters c_x and c_y and rotation angle θ are included.

$$L = \begin{bmatrix} c_x \cos\theta & c_x \sin\theta \\ -c_y sin\theta & c_y \cos\theta \end{bmatrix} \tag{2}$$

The translation transformation parameter matrix T is made up of translate values t_x and t_y in different axis directions.

$$T = \begin{bmatrix} t_x & t_y \end{bmatrix}^T \tag{3}$$

The perspective transformation parameter matrix P is made up of perspective values p_x and p_y in different axis directions.

$$P = \begin{bmatrix} p_x & p_y \end{bmatrix}^T \tag{4}$$

Based on the detailed transformation parameter matrices T, L and P, the projection transformation matrix A can be presented as:

$$A = \begin{bmatrix} L & P \\ T^T & 1 \end{bmatrix} = \begin{bmatrix} c_x \cos\theta & c_x \sin\theta & p_x \\ -c_y sin\theta & c_y \cos\theta & p_y \\ t_x & t_y & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix} \tag{5}$$

Based on the Equations (1) and (5), coordinates x',y' and w' on the camera coordinate plane can be presented by coordinates u, v on the ground surface plane and sub-parameters of matrix A.

$$x' = a_{11}u + a_{21}v + a_{31} \tag{6a}$$

$$y' = a_{12}u + a_{22}v + a_{32} \tag{6b}$$

$$w' = a_{13}u + a_{23}v + 1 \tag{6c}$$

Then, a linear transformation carried out to calculate pixel coordinates x and y in image coordinate plane. The transformation is controlled by the camera internal parameter w'.

$$(x,y)^T = \left(\frac{x'}{w'}, \frac{y'}{w'}\right)^T \tag{7}$$

Eventually, the pixel coordinates x and y can be presented by ground surface plane coordinates u, v and sub-parameters of projection transformation calibration matrix A.

$$x = \frac{a_{11}u + a_{21}v + a_{31}}{a_{13}u + a_{23}v + 1} \tag{8a}$$

$$y = \frac{a_{12}u + a_{22}v + a_{32}}{a_{13}u + a_{23}v + 1} \tag{8b}$$

Additionally, if the human shadow pixel coordinates x and y and projection calibration matrix A is acknowledged, the real-world coordinates u and v of the human shadow can be extracted based on solving the Equations (8a) and (8b). The procedure of solving real-world coordinates u and v is simplified in Equation (9).

$$(u,v)^T = f((x,y)^T, A) \tag{9}$$

Block Matrix-Based Projection Transformation Parameter Calculation

The traditional plane-to-plane projection transformation is designed for ideal scenarios with continuous flat ground surface. Nevertheless, the realistic scenarios contain uneven ground surfaces with discontinuous pavement levels. Thus, the single projection transformation calibration matrix A is not capable of ensuring precise projection transformation for all sub-blocks of the uneven ground surface.

In order to deploy the projection transformation on realistic scenarios with high precision, a block matrix-based projection transformation is proposed in this part. Instead of deploying imprecise plane-to-plane global transformation, the proposed method launches a set of precise sub-transformations. Each single sub-transformation covers only one partially flat sub-block on the ground surface, ensuring the precise projection transformation between a surface sub-block and the corresponding image subset. For each sub-block, the unique projection transformation calibration matrix A_{sub} is non identical with the parameter matrices belonging to other sub-blocks.

The parameter matrices A_{sub} of different sub-blocks are calculated separately based on Equations (8a) and (8b). To solve the unique calibration matrix of each sub-block, four pairs of marked point coordinates on ground surface plane and their corresponding pixel coordinates on image coordinate plane are required. However, manipulating massive markers to calculate parameter matrices of all sub-blocks will bring a heavy workload.

In order to simplify the setup, the optimized block matrix based parameter calculation procedure is designed to be marker coordinates multiplexable and parallel computing friendly. From the top-view angle, the ground surface is divided into a matrix consisting of multiple intensive square sub-blocks as shown in Figure 1b.

Each sub-block is a unit square area Sq_{sub} defined by the four corner markers, occupying one meter square area on the ground surface as shown in Figure 1b. The coordinate set of four markers on ground surface is defined as $M^S_{sub} = \{(u_i, v_i), i = 1, 2, 3, 4\}$, their corresponding image coordinate set is $M^{Im}_{sub} = \{(x_i, y_i), i = 1, 2, 3, 4\}$.

For each sub-block Sq_{sub}, a set of auxiliaries is introduced to simplify the calculation of parameter matrices A_{sub} based on Equation (10). The scale auxiliary parameters set includes Δx_1, Δx_2, Δy_1, and Δy_2.

$$\begin{cases} \Delta x_1 = x_2 - x_3 \\ \Delta x_2 = x_4 - x_3 \\ \Delta y_1 = y_2 - y_3 \\ \Delta y_2 = y_4 - y_3 \\ \Delta x_3 = x_1 - x_2 + x_3 - x_4 \\ \Delta y_3 = y_1 - y_2 + y_3 - y_4 \end{cases} \tag{10}$$

Additionally, the parallel auxiliary parameters Δx_3 and Δy_3 are introduced as Equation (10) as well. If both auxiliary parameters Δx_3 and Δy_3 approach zero, the field of camera view is regarded as parallel with the sub-block.

The translation parameter T_{sub}, perspective parameter P_{sub} and linear parameter L_{sub} in each calibration matrix A_{sub} can be solved as:

$$T_{sub} = \begin{bmatrix} a_{31} & a_{32} \end{bmatrix}^T = \begin{bmatrix} x_1 & y_1 \end{bmatrix}^T \tag{11a}$$

$$P_{sub} = \begin{bmatrix} a_{13} & a_{23} \end{bmatrix}^T = [\tfrac{\Delta x_3 \Delta y_2 - \Delta x_2 \Delta y_3}{\Delta x_1 \Delta y_2 - \Delta x_2 \Delta y_1}, \tfrac{\Delta x_1 \Delta y_3 - \Delta x_3 \Delta y_1}{\Delta x_1 \Delta y_2 - \Delta x_2 \Delta y_1}]^T \tag{11b}$$

$$L_{sub} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} (x_2 - x_1 + a_{12} x_2) & (y_2 - y_1 + a_{13} y_2) \\ (x_4 - x_1 + a_{12} x_3) & (y_4 - y_1 + a_{23} y_4) \end{bmatrix} \tag{11c}$$

The extraction procedure of the block matrix based projection transformation calibration matrix can be simplified as:

$$A_{sub} = P(M^S_{sub}, M^{Im}_{sub}) = \begin{bmatrix} L_{sub} & P^T_{sub} \\ T_{sub} & 1 \end{bmatrix} \tag{12}$$

Block-Matrix Based Projection Transformation Deployment

Based on Equation (9) and the calculated parameters in matrix A_{sub}, real-world coordinates (u, v) of each point in one sub-block area can be calculated from the corresponding pixel coordinates (x, y). The presentation of extraction procedure can be simplified as:

$$(u, v)^T = f((x, y)^T, A_{sub}) \tag{13}$$

Noticeably, different from the original global calibration matrix A, each sub-block calibration matrix A_{sub} is only deployed on the restricted regional transformation between the sub-block area on the ground and the corresponding pixel range in the image.

Once all parameter matrices A_{sub} for different sub-blocks are extracted through the block matrix-based parameter calculation procedure, coordinates (x, y) of pixels belonging to different sub-blocks can be transformed to corresponding real-world coordinates (u, v) inside the sub-block Sq_{sub}. Block matrix-based projection transformation is deployed based on the parallel computation of sub-transformations illustrated in Equation (13). The deployment algorithm of a sub-transformation is illustrated in Algorithm 1.

Algorithm 1: Block matrix based projection transformation deployment Algorithm

```
Input: $M^S_{sub} = \{(u_i, v_i), i = 1,2,3,4\}$: coordinates set of marker positions for sub-block $Sq_{sub}$;
       $M^{Im}_{sub} = \{(x_i, y_i), i = 1,2,3,4\}$ :corresponding pixel coordinates set of $M^S_{sub}$ on image
              coordinate plane $Im(x, y)$;
       $(x, y)$: image coordinates of captured pixel in human shadow silhouette
Output: $(u, v)$: corresponding real-world coordinates of $(x, y)$
foreach $Sq_{sub}$ do
    $[\Delta x_1, \Delta x_2, \Delta y_1, \Delta y_2] = [x_2 - x_3, x_4 - x_3, y_2 - y_3, y_4 - y_3]$
    $[\Delta x_3, \Delta y_3] = [x_1 - x_2 + x_3 - x_4, y_1 - y_2 + y_3 - y_4]$
    $T_{sub} = \begin{bmatrix} a_{31} & a_{32} \end{bmatrix}^T = \begin{bmatrix} x_1 & y_1 \end{bmatrix}^T$
    $P_{sub} = \begin{bmatrix} a_{13} & a_{23} \end{bmatrix}^T = [\frac{\Delta x_3 \Delta y_2 - \Delta x_2 \Delta y_3}{\Delta x_1 \Delta y_2 - \Delta x_2 \Delta y_1}, \frac{\Delta x_1 \Delta y_3 - \Delta x_3 \Delta y_1}{\Delta x_1 \Delta y_2 - \Delta x_2 \Delta y_1}]^T$
    $L_{sub} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^T = \begin{bmatrix} (x_2 - x_1 + a_{12} x_2) & (y_2 - y_1 + a_{13} y_2) \\ (x_4 - x_1 + a_{12} x_3) & (y_4 - y_1 + a_{23} y_4) \end{bmatrix}^T$
    $A_{sub} = \begin{bmatrix} L_{sub} & P_{sub} \\ T^T_{sub} & 1 \end{bmatrix}$;
end
$(u, v)^T = f((x, y)^T, A_{sub})$
```

For each sub-block area Sq_{sub}, a distinctive sub-transformation thread is launched based on the specific calibration matrix A_{sub}. The parallel computation of block matrix-based projection transformation contains multiple sub-transformation threads. For the simplicity of the parallel computation presentation, A_{mat} is introduced as the collection of all calibration sub-matrices $\{A_{sub}\}$ for different sub-blocks. The overall transformation is simplified as Equation (14).

$$(u, v)^T = F((x, y)^T, A_{mat}) \tag{14}$$

Based on Equation (14), the real-world coordinates (u, v) of human shadow silhouette Sh can be extracted from corresponding pixel coordinates $(x', y') \in Sh_c$ captured by a monocular camera. The block matrix-based projection transformation between the captured human shadow silhouette Sh_c and the corresponding real-world shadow silhouette Sh is illustrated in Equation (15).

$$Sh = F(Sh_c, A_{mat}) \tag{15}$$

The benefits of the block matrix based projection transformation are obvious:

- The positions of markers can be reused to simplify the scenario set up. For a scenario containing a $m \times n$ square meter area, the number of markers is reduced from $(4 \times m \times n)$ to $(m + 1) \times (n + 1)$.
- Parallel sub-transformations on different sub-blocks can be processed synchronously to accelerate the overall projection transformation procedure.
- Only when the position of camera is moved or the ground surfaced is repaved, will partial recalibration work be necessary for the affected sub-block Sq_{sub}.

Overall, all parameter matrices A_{sub} for different sub-blocks only need to be calculated once. Then all pixel coordinates in video frames can be transformed into the real-world coordinates on the ground surface plane. The block matrix-based structure also simplifies the parameter maintenance procedure when changes occur in the scenario.

2.1.2. Silhouette Information Extraction

For the extracted human shadow contour Sh on the ground surface, joint positions are extracted through an optimized method based on the silhouette contour extreme point seeking method. Comparing with traditional human segmentation methods, only silhouette information is available for shadow contour segmentation in our work. In order to perform an efficient joint position extraction based on precise silhouette contour segmentation [17], a two-step algorithm is presented in this section.

Human Shadow Silhouette Contour Preprocess

Firstly, a survey for global peak points on the shadow contour is launched to locate most obvious joint positions on the human shadow contour. In this step, the gravity center coordinate $(\overline{u}, \overline{v})$ of human shadow contour Sh is calculated first. For human shadow contour Sh containing N contour points (u_m, v_m), the gravity center $(\overline{u}, \overline{v})$ can be extracted based on the Equation (16).

$$(\overline{u}, \overline{v}) = \left(\frac{1}{N} \sum_{m=1}^{N} u_m, \quad \frac{1}{N} \sum_{m=1}^{N} v_m \right) \tag{16}$$

Then, the the distance curve D between contour points $(u_m, v_m) \in Sh$ and the gravity center $(\overline{u}, \overline{v})$ is calculated for the localization of global peak points. The value of each point on the distance curve D is calculated based on Equation (17). The Cartesian distance is applied in the Equation (17) as a linearized approximation for the value of each point on the distance curve.

$$D\,(m) = \sqrt{(u_m - \overline{u})^2 + (v_m - \overline{v})^2} \tag{17}$$

In order to reduce the interference of grainy ground surface in the joint position extraction procedure, the distance curve D is denoised based on Equation (18). The smooth length unit η is set as 10 in our experiment. In the next step, the localization procedure of major joint positions is based on the denoised distance curve $\overline{D}$.

$$\overline{D}\,(m) = \frac{1}{\eta+1} \sum_{l=-\frac{\eta}{2}}^{\frac{\eta}{2}} D\,(u_{m+l}, v_{m+l}) \tag{18}$$

The global peak points including head and two feet appear at the maximum point on the distance curve. Based on the denoised distance curve $\overline{D}$, the major joint positions can be located through seeking peak points. The normalized distance curve extraction procedure is illustrated from Equation (16) to Equation (18) and simplified in the stage Equation (19). In order to simplify the subsequent presentations, function Pre is introduced to cover the extraction procedure for the normalized distance curve $\overline{D}$ based on the human shadow contour Sh.

$$\overline{D} = Pre(Sh) \tag{19}$$

Localization of Major Joint Positions on Human Shadow Silhouette Contour

In the second step, a quick localization of global maximum peaks is launched first to locate the positions of head and both feet, then elaborate local search for major joints including hands, shoulders and knees is carried out.

(1) Localization of Global Convex Areas

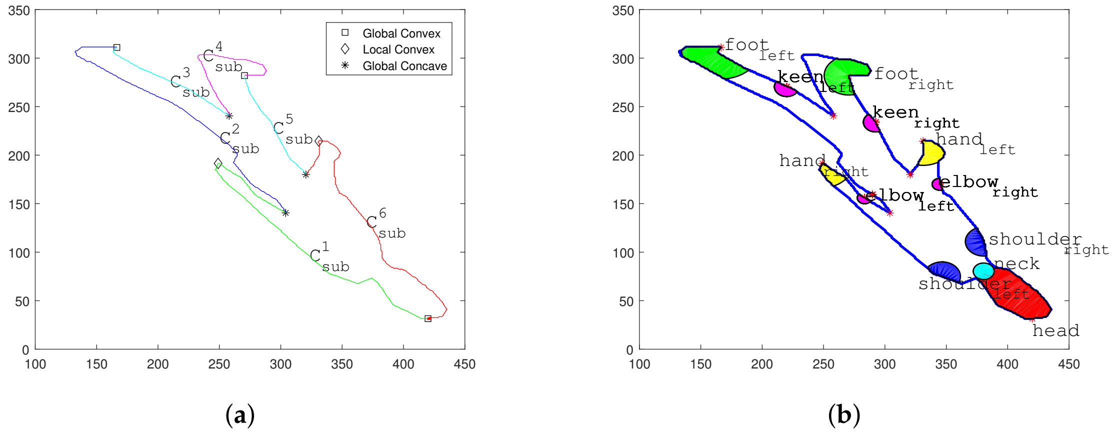
Three global maximum peaks of denoised curve $\overline{D}$ is marked in corresponding positions on Figure 2a with square symbols. The marked positions indicate precise global convex area on the human shadow silhouette contour, including head Sp^{head}, left foot $Sp^{foot_{left}}$ and right foot $Sp^{foot_{right}}$. As shown in Figure 2b, the area containing the head are marked in red, and areas containing the feet are marked in green.

(2) Localization of Auxiliary Anchor Points

Based on the acknowledged major joint positions including head and feet, the positions of rest joints are calculated through locating the local peak and nadir points.

Based on the three major joint positions, the shadow contour is divided into three sub-curves. Each sub-curve contains one auxiliary anchor point at the corresponding local nadir position on curve $\overline{D}$. The auxiliary anchor points are markered with star symbols in Figure 2a.

- The sub-curve between two feet joints contains the position of hip center Sp^{hip} at the local nadir position.
- The sub-curves between the head position and two feet positions contain positions of two oxters at local nadir positions, respectively.



(**a**) (**b**)

Figure 2. Silhouette information analyses and joint position extraction. (**a**) Sub-curve segmentation; (**b**) two-dimensional (2D) joint position extraction on the shadow area.

(3) Localization of Remaining Major Joint Positions

Based on the three global peaks and three auxiliary anchor points, the shadow contour is subdivided into six new sub-curves. To illustrate the extraction of remaining joint positions clearly, six sub-curves are marked as C^i_{sub}. The indicated i ranges from 1 to 6 as shown in Figure 2a. C^1_{sub} to C^6_{sub} cover the shadow contour in a clockwise direction , initiating from the head position.

- Sub-curves C^1_{sub} and C^6_{sub} cover the contour ranges of left arm and right arm. Thus the local peak positions of these two sub-curves are hand positions. Their local nadir positions are located between the cervical vertebra position Sp_{neck} and two shoulders.
- The local peak positions of C^2_{sub} and C^4_{sub} indicate the positions of two keens $Sp^{keen_{left}}$ and $Sp^{keen_{right}}$ in the shadow area.
- Similarly, the nadir positions of C^3_{sub} and C^5_{sub} can assist the positioning of both keens $Sp^{keen_{left}}$ and $Sp^{keen_{right}}$.

The major joint position localization procedure is illustrated in the three steps above and simplified in stage Equation (20). In order to simplify the subsequent presentations, function *Loc* is introduced to cover the localization procedure for major joint position set $\{Sp^{joint}\}$ based on the human shadow contour *Sh* and the corresponding distance curve $\overline{D}$.

$$\{Sp^{joint}\} = Loc\left(Sh, \overline{D}\right) \tag{20}$$

Noticeably, the joint position localization procedure can also be adopted in the joint position extraction from a normal human pose contour. The human pose classification illustrated in Section 2.2.2 is based on the joint position extraction procedure illustrated in Equation (20).

2.1.3. 3D Joint Position Estimation and Skeleton Synthesis

In a multiple light source scenario, more than one human shadow is projected on the ground surface at the same time. In order to identify shadow areas generated by different light sources, 2D human shadow contour Sh and joint position region Sp^{joint} are footnoted with corresponding light source identifier i as shown in Equation (21). Additionally, the point coordinates (u, v) inside the each region Sp_i^{joint} are footnoted as $(u_i^{joint}, v_i^{joint})$.

$$Sp_i^{joint} \subset Sh_i \tag{21}$$

In order to estimate the 3D joint position Mp^{joint} of each major joint, the light beams L_i^{joint} from different light sources S_i blocked by Mp^{joint} are reconstructed first. Then, the 3D position of Mp^{joint} is calculated based on allocating the shared voxel area between multiple reconstructed light beams L_i^{joint}. Finally, the human skeleton is synthesized based on the calculated 3D joint position set $\{Mp^{joint}\}$.

For the first step, each light beam L_i^{joint} is generated as a 3D cone with its vertex on the light source position $S_i = (u_i, v_i, h_i)$. The underside of each cone is the joint area Sp_i^{joint} in the shadow.

While any light beam L_i^{joint} from light source S_i is blocked by the certain joint part Mp^{joint} of human body M, the joint shadow area $Sp_i^{joint} = \{(u_i^{joint}, v_i^{joint})\}$ is produced on the ground. Thus, the direction of blocked light beam L_i^{joint} leads to shadow area $Sp_i^{joint} \subset Sh_i$, going through human body part Mp^{joint}. If $w \in [0, h_i]$ is introduced as the height component in the cone expression of L_i^{joint}, the 3D space caused by L_i^{joint} can be presented as Equation (22).

$$L_i^{joint} = (\frac{(h_i - w)u_i^{joint}}{h_i}, \frac{(h_i - w)v_i^{joint}}{h_i}, w) \tag{22}$$

The 3D light beam shape extraction procedure presented by Section 2.1.3 is simplified in the stage Equation (23). For the simplicity of the subsequent presentations, function Occ is introduced to cover the 3D light beam shape extraction procedure for the occupied 3D cone shape L_i^{joint} based on the corresponding joint position Sp_i^{joint} and the light source position S_i.

$$L_i^{joint} = Occ(Sp_i^{joint}, S_i) \tag{23}$$

Then, for multiple light beams L_i^{joint} generated by different light sources S_i, Mp^{joint} is the shared subset for all reconstructed L_i^{joint}. Thus, the 3D position Mp^{joint} can be generated based on the intersection of all reconstructed L_i^{joint} as shown in Figure 3a. sum_l is the sum of light sources in the scenario.

$$Mp^{joint} = \bigcap_{i=1}^{sum_S} L_i^{joint} \tag{24}$$

Figure 3a demonstrates the recovery of neck joint area Mp^{neck} based on two related shadow areas Sp_a^{neck} and Sp_b^{neck} generated by light sources S_a and S_b.

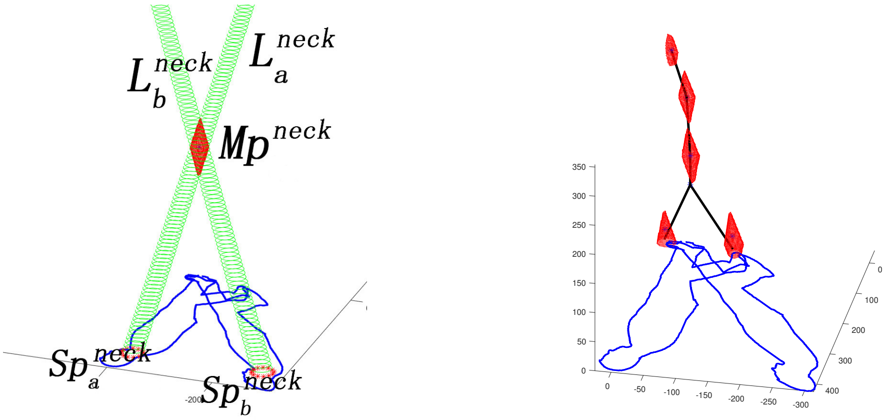
Finally, 3D human skeleton Sk with multiple joint positions is synthesized by calculating Mp^{joint} joint by joint.

Figure 3b presents the skeleton synthesis procedure of a human being based on human shadow information under a multi light source scenario. The illustrated human skeleton synthesis procedure is presented in Algorithm 2 and simplified in Equation (25).

$$Sk = Syn(\{Sh_i\}, \{S_i\}) \tag{25}$$

However, there are two restrictions for the deployment of the basic theory:

- **Condition (1)** Two or more light sources are required in the scene.
- **Condition (2)** Relative angular positions between human body and different light sources should be different.



(**a**) (**b**)

Figure 3. Demo of SSSE in a multiple light source scenario: (**a**) A simulated dual light source scenario. Sp_a^{neck} and Sp_a^{neck} are the joint areas of the neck position in the shadows projected by light source *a* and *b*, respectively. Similarly, light beams L_a^{neck} and L_b^{neck} are generated by light sources *a* and *b*, respectively. The enclosure Mp^{neck} is the intersection area of L_a^{neck} and L_b^{neck}. (**b**) Scenario reconstruction.

Algorithm 2: Human skeleton synthesis procedure under multiple light source scenario

```
Input: $\{Sh_i\}$: 2D human shadow contour on the ground surface.
       $\{S_i\}$ :3D positions of multiple light sources $\{(u_i, v_i, h_i)\}$.
Output: $Sk$: 3D human skeleton synthesis based on seven major joint positions.
foreach $S_i = (u_i, v_i, h_i)$ do
    $\overline{D_i} = Pre(Sh_i)$
    $\{Sp^{joint}\} = Loc\left(Sh_i, \overline{D}_i\right)$
    $L_i^{joint} = (\frac{(h_i - w)u_i^{joint}}{h_i}, \frac{(h_i - w)v_i^{joint}}{h_i}, w) \quad , w \in [0, h_i]$
end
foreach $joint$ do
    $Mp^{joint} = \bigcap_{i=1}^{sum_l} L_i^{joint}$
end
$Sk = \{Mp^{joint}\}$
```

2.2. Skeleton Simulation in Single-Light-Source Scenario

The basic theory introduced in Section 2.1 is only effective in scenes containing two or more shadows generated by multiple light sources. For single light source scenarios, only one shadow is generated in each captured frame. In order to extend the proposed basic theory in single light source scenarios, a video sequence instead of a single frame is taken into consideration. Human shadow contours are footnoted with time coordinate *t* in this part. The extension solution is introduced below.

2.2.1. Theoretic Proof of the Extension Solution in a Single Light Source Scenario

For every video sequence, the extension solution is based on two facts:

Temporal Distinguished Relative Position between Light Source and Human Body

In a sequence, the relative position between a moving human and a fixed light source keeps changing. In other words, temporal discrete shadows Sh_t are generated by the light sources from different relative positions θ_t towards the human. The temporal neighboring human shadows Sh_t and Sh_{t+1} are distinguished from each other because of different relative positions between the light source S and the human body. For neighboring frames at time coordinates t and $t+1$, it is clear that $\theta_t \neq \theta_{t+1}$ and $Sh_t \neq Sh_{t+1}$.

Temporal Discrete Shadows for Same Human Pose

In order to categorize different frames based on the human poses, the 2D contour of human body captured by a monocular camera is regarded as the human pose P_t at the time coordinate t. As shown in Figure 4b, same human pose P_0 appears repeatedly during an activity sequence. Since each relative position between the light source and human body is different frame by frame, multiple frames sharing the same human pose P_0 can be found. Each frame owns different shadows Sh_t and projection angles $s\theta_t$. In an activity sequence, all the human shadows Sh_t sharing the same human pose P_0 are categorized into the set $\{Sh_t\}$. For different human shadows $Sh_t \in \{Sh_t\}$, their corresponding relative position angles θ_t are distinguished from each other. If multiple human shadows in $\{Sh_t\}$ with different projection angles θ_t are integrated in one single frame as shown in Figure 4b, condition (2) of launching the basic theory proposed in Section 2.1 is satisfied. Through applying translation transformations on each integrated frames to make the all human poses P_t spatially coincide with the central pose P_{j_c}, an artificial multiple light source scenario satisfying conditions (1) and (2) is established as shown in Figure 4c.

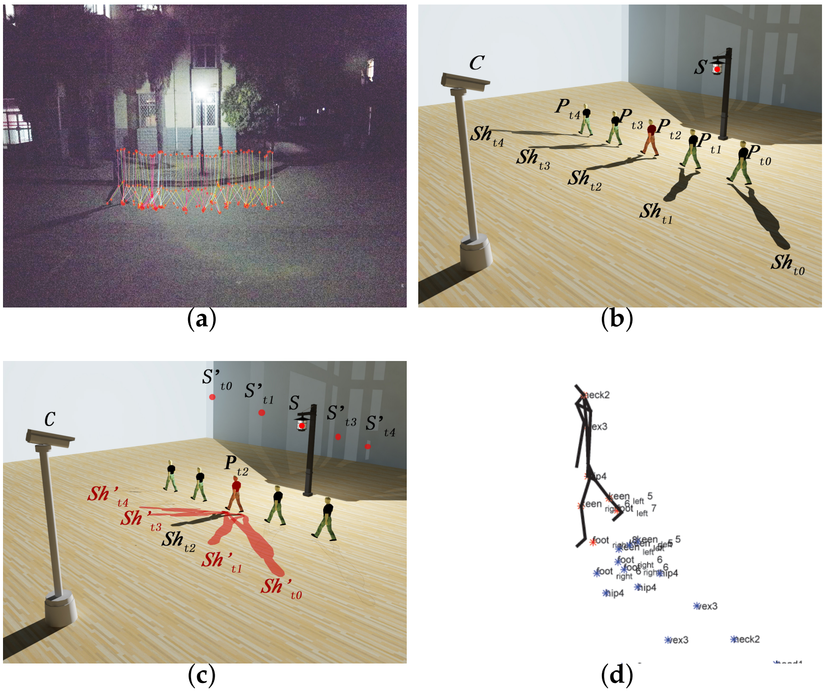


Figure 4. Demo of SSSE procedure in a single light source scenario. (**a**) Pose classification based on major joint positions; (**b**) Spatial–temporal discrete human poses belonging to same class; (**c**) Temporal–spatial aggregation; (**d**) Three-dimensional (3D) human skeletonization.

The simulated scenario makes it feasible to recover the skeleton of the shared pose P_o in a single light source scenario based on the basic theory proposed in Section 2.1.

2.2.2. Temporal–Spatial Aggregation Method

Before the deployment of human skeletonization, it is necessary to find shadows that share the same human pose P_0, yet have distinctive projection angles θ_t. Human pose classification and temporal–spatial shadow aggregation are deployed to fit spatial coordinates of shadows in $\{Sh_t\}$ with the chosen central pose position P_{t_c}.

Human Pose Classification

In order to analyses the human pose P_{t_i} at each time coordinate t_i, the denoised distance curve $\overline{D_{t_i}}$ between the human pose contour and human pose gravity center is extracted based on the same method illustrated in Equation (19). Similarly, the stage Equation (26) covers the normalized distance curve extraction procedure illustrated from Equation (16) to Equation (18). The function Pre presents the extraction procedure for the normalized distance curve $\overline{D_{t_i}}$ based on the contour curve P_{t_i}.

$$\overline{D_{t_i}} = Pre(P_{t_i}) \tag{26}$$

Based on the distance curve $\overline{D_{t_i}}$, major peak point set $\{Jp^k_{t_i}|k = 1,2,3\}$ including head and two feet are extracted from the captured human contour based on the same procedure presented in Equation (20). Similar to the stage Equation (20), stage Equation (27) covers the major joint position localization procedure illustrated in the Section 2.1.2. The function Loc presents the human joint position extraction procedure for major joint position set $\{Jp^k_{t_i}\}$ based on the human contour Sh_{t_i} and the corresponding distance curve $\overline{D_{t_i}}$.

$$\{Jp^k_{t_i}\} = Loc(Sh_{t_i}, \overline{D_{t_i}}) \tag{27}$$

The positions of three peak points of the human pose contour are combined into a star feature to describe the human pose in each frame [10,18]. Then unsupervised classification is adopted to assort each frame with corresponding pose category label based on the star feature [19].

$$C_{t_i} = Lab(\{Jp^k_{t_i}\}) = j|\arg\min_j(\sum_{k=1}^{3}\sum_{Jp^k_j \in \mathrm{P}_j}\left\|Jp^k_{t_i} - Jp^k_j\right\|^2) \tag{28}$$

Temporal–Spatial Shadow Aggregation

During the temporal-spatial shadow aggregation procedure shown in Figure 4c, temporal discrete light sources are aggregated in a single frame. Normally, for multiple human poses, the human pose P_{t_i} with median time coordinate t_i is chosen as the central pose P_{jc}.

For each human pose $P_{t_i} \in \{P_j\}$, the translation transformation parameter $T_{t_i \to jc}$ is defined by the vector between corresponding joint points in P_{t_i} and P_{jc}, satisfying the spatial transformation from P_{t_i} to P_{jc}.

Noticeably, the joint positions $Jp^k_{t_i}$ and Jp^k_{jc} are captured in the image coordinate plane. Before calculating the translation transformation $T_{t_i \to j}$ in real-world coordinates, it is necessary to transform the joint coordinates into the real- world coordinates $Sp^k_{t_i}$ and Sp^k_{jc} based on the stage projection transformation $F((x,y), A_{mat})$ presented in Equation (14).

$$Sp^k_{t_i} = F(Jp^k_{t_i}, A_{mat}) \tag{29a}$$

$$Sp^k_{jc} = F(Jp^k_{jc}, A_{mat}) \tag{29b}$$

$$\overrightarrow{T_{(t_i,j_c)}} = \frac{1}{2}\sum_{k=2}^{3}\overrightarrow{Sp_{t_i}^k Sp_{j_c}^k} = \Delta u_{(t_i,j_c)}\alpha + \Delta v_{(t_i,j_c)}\beta \tag{29c}$$

$$A_{t_i \to j_c} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \Delta u_{(t_i,j_c)} & \Delta v_{(t_i,j_c)} & 1 \end{bmatrix} \tag{29d}$$

As shown in Equation (29a), $Sp_{t_i}^k$ is the extracted real-world human joint coordinates at the original captured position. In Equation (29b), $Sp_{j_c}^k$ is the human joint coordinates at the destination position. Then the translation vector $\overrightarrow{T_{t_i \to j_c}}$ is calculated based on the averaged horizontal translation vectors from the original position to the destination position. As shown in Equation (29c), α and β are two vertical unit vectors of the real-world coordinate system on the ground surface. The translation vector $\overrightarrow{T_{t_i \to j_c}}$ is presented as a combination of translation components $\Delta u_{(t_i,j_c)}$ and $\Delta v_{(t_i,j_c)}$ in two vertical directions. Based on the translation components $\Delta u_{(t_i,j_c)}$ and $\Delta v_{(t_i,j_c)}$, a translation transformation matrix $A_{t_i \to j_c}$ can be established for the translation calculation as shown in Equation (29d).

For the convenience of further illustration, the extraction procedure of matrix $A_{t_i \to j_c}$ is simplified in Equation (30). The function *Par* is introduced to present translation matrix extraction procedure illustrated from Equation (29a) to Equation (29d).

$$A_{t_i \to j_c} = Par(\{Jp_{t_i}^k\}, \{Jp_{j_c}^k\}) \tag{30}$$

In order to maintain a consistent expression system, the translation transformation is presented in the same format with Equations (14) and (15). When the translation transformation in Equations (31) and (32) is deployed synchronously on the light source S_{t_i} and human shadow contour Sh_{t_i} for each frame, all the transformed human shadows Sh'_{t_i} fit the spatial coordinates of the central human pose P_{t_c} in each simulated scenario.

$$Sh'_{t_i} = F(Sh_{t_i}, A_{t_i \to j_c}) \tag{31}$$

The position of light source S_{t_i} applies the same transformation $T_{t_i \to j_c}$ along with the related shadow Sh_{t_i}, simulating multiple light sources $S'{}_{t_i}$ in the single frame.

$$S'_{t_i} = f(S, A_{t_i \to j_c}) \tag{32}$$

The temporal–spatial aggregation procedure illustrated above is presented in Algorithm 3. With more than two positional distinctive light sources simulated in the same frame, the skeleton synthesis procedure presented in Equation (24) can be applied on the simulated human shadow set $\{Sh'_{t_i}\}$ and the corresponding light source set $\{S'_{t_i}\}$. Based on Equation (33), the skeleton Sk_{j_c} of pose P_{j_c} can be synthesized under a single light source scenario as shown in Figure 4d. The detailed human skeleton synthesis procedure under a single light source scenario is illustrated in Section 3.

$$Sk_{j_c} = Syn(\{Sh'_{t_i}\}, \{S'_{t_i}\}) \tag{33}$$

```
Algorithm 3: Temporal–spatial aggregation procedure
    Input: $t_i$: time coordinate for each frame;
           $Sh_{t_i}$:human shadow on the ground surface in frame $t_i$;
           $P_{t_i}$: human pose in frame $t_i$;
           $S$: light source position;
           $\{Jp^k_{jc}\}$: joint position set of the central pose on the aggregation destination ;
    Output: $Sh'_{t_i}$: integrated human shadow $Sh_{t_i}$ in the simulated scenario.
            $S'_{t_i}$: integrated light source position in correspondence with $Sk'_{t_i}$.
    foreach time coordinate $t_i$ do
        $\overline{D_{t_i}} = Pre(P_{t_i})$
        $\{Jp^k_{t_i}\} = Loc(Sh_{t_i}, \overline{D_{t_i}})$
        $C_{t_i} = Lab(\{Jp^k_{t_i}\})$
        $A_{t_i \to jc} = Par(\{Jp^k_{t_i}\}, \{Jp^k_{jc}\})$
        $Sh'_{t_i} = F(Sh_{t_i}, A_{t_i \to jc})$
        $S'_{t_i} = f(S, A_{t_i \to jc})$
    end
```

3. Proposed Method

Based on the basic theory and its extension introduced in Section 2, a normal single light source scenario can support 3D human skeletonization. In this section, a five-step algorithm is proposed according to the illustrated theory as shown in Figure 5. The procedure of the proposed method is shown in Algorithm 4.

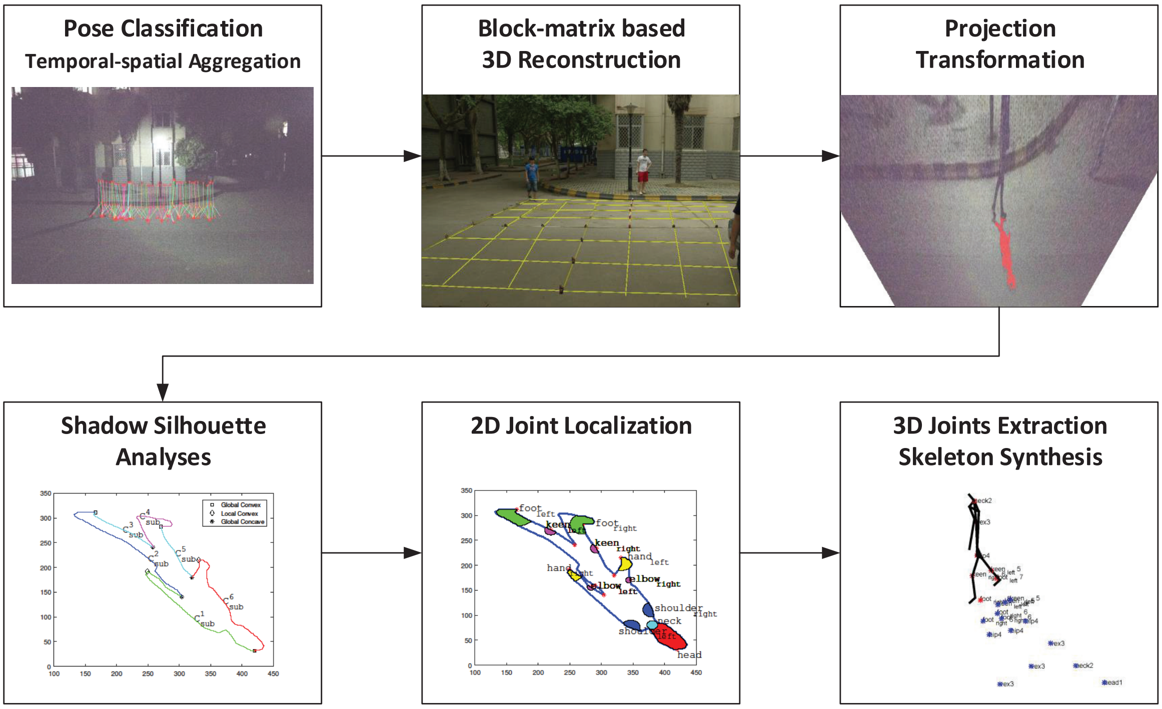


Figure 5. The flow chart of skeleton synthesis procedure based on SSSE.

Algorithm 4: Skeleton synthesis procedure

```
Input: $t_i$: time coordinate for each frame;
$Shc_{t_i}$:captured human shadow in frame $t_i$;
$P_{t_i}$: human pose in frame $t_i$;
$S$: light source position;
$\{Sq_{sub}\}$: the set of sub-blocks on the ground surface plane $S$
$\{M^S_{sub}\}$: the marker position coordinate sets for sub-block $Sq_{sub}$ on $S$;
$\{M^{Im}_{sub}\}$: the pixel coordinates set of $\{M^S_{sub}\}$ on the image coordinate plane.
Output: $Sk_{t_i}$:3D human skeleton corresponding to $Sh_{t_i}$ at time coordinate $t_i$
foreach $Sq_{sub}$ do
    $A_{sub} = Par(M^S_{sub}, M^{Im}_{sub})$
end
$A_{mat} = \{A_{sub}\}$.
foreach $P_{t_i}$ at $t_i$ do
    $\overline{D_{t_i}} = Pre(P_{t_i})$
    $\{Jp^k_{t_i}\} = Loc(Sh_{t_i}, \overline{D_{t_i}})$
    $C_{t_i} = Lab(\{Jp^k_{t_i}\})$
    $A_{t_i \to jc} = Par(\{Jp^k_{t_i}\}, \{Jp^k_{jc}\})$
end
foreach $Shc_{t_i}$ at $t_i$ do
    $Sh_{t_i} = F(Shc_{t_i}, A_{mat})$
    $Sh'_{t_i} = F(Sh_{t_i}, A_{t_i \to jc})$
    $S'_{t_i} = f(S, A_{t_i \to jc})$
end
foreach Pose Category C do
    $Sk_{jc} = Syn(\{Sh'_{t_i}\}, \{S'_{t_i}\})$
end
foreach $t_i$ do
    $Sk_{t_i} = T_{jc \to t_i}(Sk_{jc})$
end
```

Pose Classification

In a human activity sequence captured under a single light source scenario, frames at different time coordinates are classified based on human poses [8] on the captured frames. For each captured human pose P_{t_i} at time coordinate t_i, the denoised distance curve between contour points and the gravity center of P_{t_i} can be extracted based on the method presented in Equation (19). The deployment of the extraction method on the human pose P_{t_i} is presented in Equation (34).

$$\overline{D_{t_i}} = Pre(P_{t_i}) \tag{34}$$

Based on the method presented in Equation (20), a major peak joint position set $\{Jp^k_{t_i}\}$ is extracted from the human contour P_{t_i}, including the head position $Jp^1_{t_i}$, the left foot position $Jp^2_{t_i}$ and right foot position $Jp^3_{t_i}$.

$$\{Jp^k_{t_i}\} = Loc(Sh_{t_i}, \overline{D_{t_i}}) \tag{35}$$

Based on the normalized peak joint positions, raw frames containing same class human poses P_{t_i} is aggregated to the human pose category P_j based on the automatic unsupervised clustering illustrated in Equation (28). C_{t_i} is the category label of human pose P_{t_i} as shown in Equation (36).

$$C_{t_i} = Lab(\{Jp^k_{t_i}\}) \tag{36}$$

Preprocess

The preprocess procedure transforms the captured shadow contour pixel coordinates Shc_{t_i} into the real-world coordinates Sh_{t_i}.

Before the preprocess of the first shadow contour Shc_{t_i}, all the $A_{sub} \in A_{mat}$ are calculated and saved for further preprocess procedures. For each square unit area Sq_{sub}, the related projection parameter matrices A_{sub} are calculated based on four real-world coordinates $\{M_{sub}^{S}\}$ and their corresponding imaging coordinates $\{M_{sub}^{Im}\}$ based on Equation (12).

Based on the calibration matrix set $A_{mat} = \{A_{sub}\}$, the global projection transformation $F(Shc_{t_i}, A_{mat})$ can be figured out. Based on the projection transformation presented in Equation (37), captured human shadow contour pixel coordinates Shc_{t_i} can be extracted from each of the raw frames and transformed into the real-world coordinates Sh_{t_i}.

$$Sh_{t_i} = F(Shc_{t_i}, A) \tag{37}$$

Temporal–Spatial Aggregation

Preprocessed shadow contours Sh_{t_i} are aggregated according to category P_j of corresponding human pose P_{t_i}. Nevertheless, the real-world coordinates of $P_{t_i} \in P_j$ are spatially dispersed due to the human movement as shown in Figure 4a. Thus it is necessary to aggregate shadow contours Sh_{t_i} of the same central human pose P_{jc} to deploy precise joint position estimation.

For each pose category, one central human pose P_{jc} is set up as the aggregating destination for other human shadow Sh_{t_i} related with $P_{t_i} \in P_j$.

The translation of each human shadow Sh_{t_i} is based on the translation transformation calibration matrix $A_{t_i \to j_c}$. The translation transformation matrix is calculated based on the Equation (30). Since the major peak joint position sets $\{Jp_{t_i}^{k}\}$ and $\{Jp_{j_c}^{k}\}$ are obtained in the pose classification step, the translation transformation calibration matrix $A_{t_i \to j_c}$ can be extracted as shown in Equation (38).

$$A_{t_i \to j_c} = Par(\{Jp_{t_i}^{k}\}, \{Jp_{j_c}^{k}\}) \tag{38}$$

Along with the 2D translation of each Sh_{t_i}, the corresponding 3D light source position S_i is moved with the identical translation as shown in Equation (39b). The aggregated human shadow Sh'_{t_i} and light source S'_{t_i} offer the ideal multiple light source situation for 3D joint position estimation.

As shown in Figure 4b, when P_{t_2} is setup as the aggregating destination, other Sh_{t_i} are aggregated to the aggregating destination through the 2D translation as shown in Equation (39a).

$$Sh'_{t_i} = F(Sh_{t_i}, A_{i_t \to j_c}) \tag{39a}$$

$$S't_i = f(S, A_{i_t \to j_c}) \tag{39b}$$

Joint Position Estimation and Skeleton Synthesis

For each aggregated human shadow contour Sh'_{t_i}, joint area estimation is launched based on the algorithm introduced in the basic theory section. First of all, the gravity center G'_{t_i} of curve Sh'_{t_i} is calculated. Then, the denoised distance curve $\overline{D'_{t_i}}$ between each point $(x'_{t_i}, y'_{t_i}) \in Sh'_{t_i}$ and G'_{t_i} is available based on the preprocess procedure illustrated in Section 2.1.2.

$$\overline{D'_{t_i}} = Pre(Sh'_{t_i}) \tag{40}$$

The 2D positions of major joint areas including head $Sp_i'^1$, neck $Sp_i'^2$, hip center $Sp_i'^3$, left keen $Sp_i'^4$, right keen $Sp_i'^5$, left foot $Sp_i'^6$ and right foot $Sp_i'^7$ can be obtained through locating the peak and nadir points in $\overline{D'_{t_i}}$.

$$\{Sp'^{k}_{t_i}\} = Loc(Sh'_{t_i}, \overline{D'_{t_i}}) \tag{41}$$

In each simulated scenario, silhouette information extraction is applied to each joint area Sh'_{i_k}. In order to estimate the 3D joint position based on Sp'_{t_i}, the ray set $L'^{k}_{t_i}$ connecting light source S'_{t_i} and joint shadow area Sp'_{t_i} is simulated.

$$L'^{k}_{t_i} = Occ(Sp'^{k}_{t_i}, S'_{t_i}) \tag{42}$$

Since more than two simulated light sources S'_{t_i} exist in the scenario, silhouette information of single joint area is extracted separately for each light sources. Based on all ray sets $L'_{t_i\_k}$ targeting at the same joint, 3D joint position $M^{k}_{p_j}$ can be calculated based on Equation (43).

$$Mp^{k}_{j} = \bigcap_{i=1}^{sum_l} L'^{k}_{t_i} \tag{43}$$

Repeating steps above for each major joints, 3D joint position set $\{Mp^{k}_{j}\}$ containing all joint positions can be figured out. Then joint positions can be synthesized based on the combination Equation (44).

$$Sk_{j_c} = \{Mp^{k}_{j}\} \tag{44}$$

In order to simplify the presentation in Algorithm 4, the illustrated joint position estimation and skeleton synthesis procedure is simplified into Equation (45).

$$Sk_{j_c} = Syn(\{Sh'_{t_i}\}, \{S'_{t_i}\}) \tag{45}$$

Frame Integration

Repeating the above steps, synthesized 3D human skeletons Sk_{j_c} can be generated for all human poses category by category. The kinematic model of skeleton Sk_{p_j} contains seven major joints, including head, neck, hip, both keens and both feet. The bones connecting particular joints are regarded as rigid objects. Based on the pose classification result in the step **(1)**, the time coordinate t_i of each $P_{t_i} \in P_j$ can be tracked. Then, reassign synthesized human skeleton Sk_{j_c} to frame t_i as Sk_{t_i} based on the reverse translation transformation.

$$Sk_{t_i} = F^{-1}(Sk_{j_c}, A_{t_i \to j_c}) \tag{46}$$

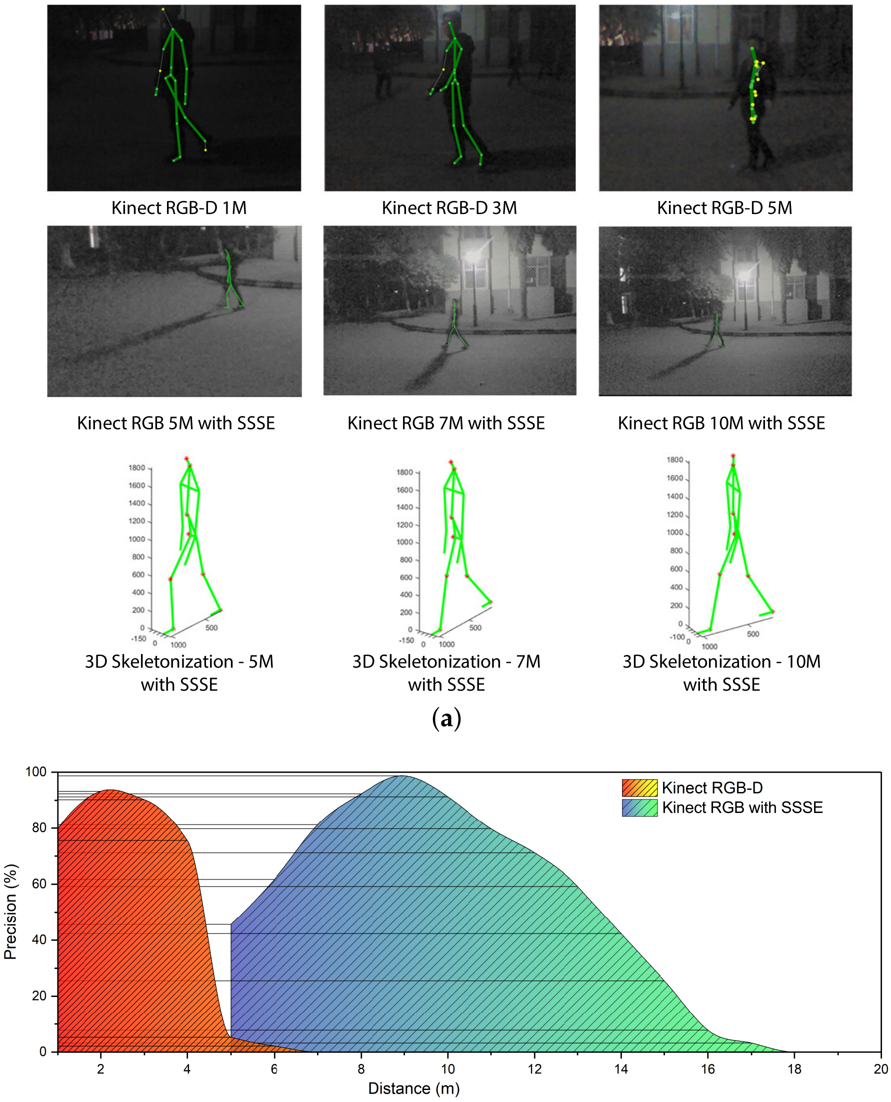
4. Experimental Validation

In this section, the experimental data source and settings are illustrated first. Then the effective range and precision of the proposed method are validated in comparison with the RGB-D based method.

4.1. Data Source Description and Experimental Settings

The experiments are launched based on data captured by a Kinect RGB-D camera, containing daily human activities. Captured sequences include both RGB frames and normal depth frames captured by Kinect. Kinect extracts human skeleton automatically based on the combined information of RGB frames and depth frames [15]. However, SSSE is deployed only on RGB frames captured by the monocular RGB camera on Kinect.

In each sequence captured for effective range validation, Kinect is set up at a static distance from the human subject. The photographic distance increases from 1 m to 20 m with a fixed step of 1 m. Sampled skeletonization results based on both methods at different distances are presented in Figure 6a.



(a)

(b)

Figure 6. Experimental results. (**a**) A comparison of tracking results; (**b**) Effective ranges of RGB-D-based results and SSSE-based results.

4.2. Effective Range and Precision Analyses

In order to validate the effectiveness of the proposed SSSE method and traditional RGB-D method, two aspects including effective range and precision are evaluated. In the following, the effective distance range is marked first. Then, the precision of six major 3D joint positions extracted by SSSE is evaluated.

Effective Range

Effective range is defined as the distance between the sensor and human, which allows effective human skeleton extraction. Effective human skeleton extraction in the effective range generates valid human joint positions. For the RGB-D based method, each extracted joint position comes with a confidence index. Valid joints are joints with confidence above 0.7. For the SSSE method, valid joints are extracted joints not affected by sheltering. In the following experiments, frames with all valid simulated human joints are defined as effective frames. In order to obtain the effectiveness–distance relationship of both methods, the shares of effective frames at different distance levels are measured. In addition, 1000 to 1200 frames containing 3D human skeletons sampled at each photographic distance from 1 to 20 m are evaluated for each method. For the RGB-D-based skeletonization procedure, effective frames are automatically labeled based on the corresponding joint confidence. For an SSSE-based procedure, effective frames are chosen based on the number of valid joints in each skeleton. The effectiveness of both methods at the same distance can be represented by the shares of effective frames among all frames. For each method, effective range covers photographic distances whose effectiveness exceed a specified threshold.

The official parameter of Kinect [1,15] indicates the effective range of state-of-art Kinect result is from 0.8 m to 3.5 m. Thus, the range of distance where effectiveness is above 0.8 is regarded as the effective range. As shown in Figure 6b, the effective range of SSSE is 7–10 m. Note that the effectiveness of SSSE decreases when photographic distance exceeds 10 m because of the limitation of camera resolution. The experimental result in Figure 6a shows that SSSE can provide reliable 3D human skeletonization at an effective range of 7–10 m, while Kinect is unable to extract human skeleton information when the photographic distance exceeds 5 m.

Precision Evaluation

As with the effectiveness evaluation result mentioned above, the RGB-D-based method and SSSE provide effective skeleton extraction results at different distance ranges. Precisions of all extracted joints by SSSE are determined by the deviation values relative to corresponding ground truth joint positions. In the precision evaluation procedure, two Kinects are setup for different purposes. Kinect No.1 is set up 9 m way from human object, capturing RGB frames for human skeletonization based on SSSE. Kinect No.2 is setup 3 m away from human object, capturing RGB-D frames along with 3D human skeletons simultaneously. Since 3 m is inside the effective range of the RGB-D-based 3D skleletonization, the 3D joint positions captured by Kinect No.2 are valid joints, providing ground truth for the deviation calculation. Based on the experimental scenario setup, 1546 frames are captured simultaneously for both methods, of which 1345 effective frames are evaluated.

For each skeleton extracted from a effective frame, joint positions are normalized relative to the hip center, avoiding deviation introduced by different shot distances.

Six major joints are considered in evaluation, including head, spine, both keens and both feet. Figure 7 depicts the averaged precision evaluation result.

As presented in Figure 7, due to the larger scale of upper body shadow on the ground, relative high deviations appear at joints of the head and spine, where averaged deviations reach 14.5 cm and 12.1 cm, respectively. For the remaining joints, the averaged deviations are around 4 cm and the highest deviation remains below 8 cm. In summary, SSSE extracts joint positions in a reasonable precision at 9 m away from the target human, compared with the ground truth Kinect skeletonization result obtained at a position 6 m closer to the subject human.

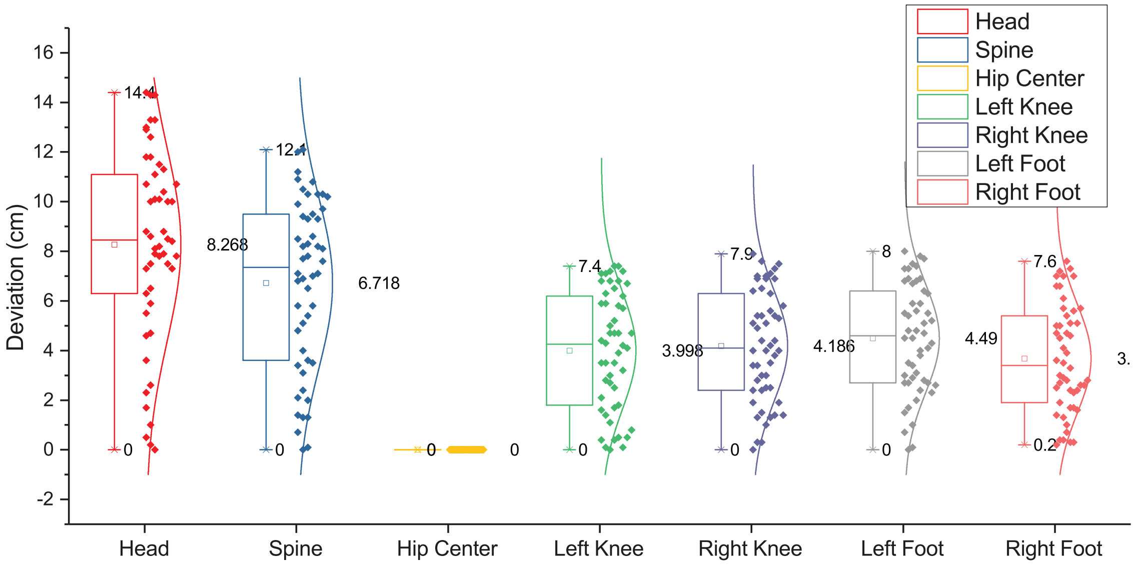


Figure 7. Deviation between SSSE-extracted joints and ground truth.

5. Discussion

Based on the experimental results in Section 4.2, an interesting phenomenon can be observed in that the effective ranges of the proposed SSSE and traditional RGB-D method are highly complementary. Thus, the fusion application of SSSE and traditional RGB-D method can provide wide range human skeletonization for indoor and outdoor scenarios. In the fusion method, the traditional RGB-D method and SSSE are deployed under different scenarios. For humans inside the effective range of RGB-D cameras, the traditional RGB-D based skeletonization method can provide solid human skeleton extraction method. For humans outside the effective range of RGB-D cameras, SSSE method can redress the unreliable 3D joint positions appears in RGB-D skeletonization result. In order to evaluate the fusion application effectiveness, a comparison between the reliable joint percentage of original skeletons extracted by Kinect and redressed skeletons processed by SSSE is carried out in this section. Reliable joints are defined as joints generated by SSSE not affected by sheltering, and joints generated by Kinect with a confidence index above 0.7. On the contrary, unreliable joints are unavailable joints affected by sheltering in SSSE methods, or joints generated by Kinect with confidence index under 0.7. For better evaluation of the fusion application, Kinect is set up to skeletonize a human subject outside its effective range.

The unreliable 3D joint positions in Kinect skeletonization result is redressed by SSSE simultaneously. In total, 20 sets of experiments have been launched to evaluate the reliable joint percentage enhancement.

5.1. Reliable Joint Percentage Enhancement

The enhancement of the reliable joint percentage is evaluated by determining the precisely recovered joint rate J_R and precisely recovered frame rate F_R. As shown in Equation (6a)–(6c), N_{Ej} and N_{Ef} are the unreliable joint number and relevant affected frame number, respectively. N_{rj} is the number of total recovered unreliable joint positions after deploying the SSSE procedure, while E_{rj} is the number of inaccurately recovered joints. From the aspect of frame statistics, N_{rf} is the total number of recovered frames and E_{rf} is the number of frames containing inaccurately recovered joints.

$$\begin{aligned} J_R &= \frac{N_{rj} - E_{rj}}{N_{Ej}} \\ F_R &= \frac{N_{rf} - E_{rf}}{N_{Ef}} \end{aligned} \tag{47}$$

The 20 test sets presented in Table 1 indicate that more than four-fifths of all unreliable joints are successfully redressed based on the proposed SSSE method, and more than three-quarters of all frames containing unreliable joint skeletonization results are accurately fixed. Based on the experimental results above, the fusion application of SSSE and traditional RGB-D method proved effective in reliable joint percentage enhancement.

Table 1. Result of unreliable joint position redress. J_R is the precisely recovered joint rate. F_R is the precisely recovered frame rate. N_{Ej} is the number of unreliable joints. N_{Ef} is the number of frames affected by unreliable joints. N_{rj} is the number of total recovered unreliable joint. E_{rj} is the number of inaccurately recovered joints. N_{rf} is the total number of recovered frames. E_{rf} is the number of frames containing inaccurately recovered joints.

Test Set	**Joints**				**Frames**			
Set No.	J_R	N_{Ej}	N_{rj}	E_{rj}	F_R	N_{Ef}	N_{rf}	E_{rf}
1	0.8644	1221	1111	56	0.7923	1385	1291	194
2	0.8554	1666	1516	91	0.7943	1218	1125	158
3	0.8357	1654	1519	137	0.8603	1293	1236	124
4	0.8636	1632	1532	123	0.8410	1377	1316	158
5	0.8391	1913	1764	159	0.7705	1176	1066	160
6	0.8929	1404	1348	94	0.7606	1198	1072	161
7	0.8816	1985	1842	92	0.8170	1048	973	117
8	0.8303	1254	1096	55	0.8189	1383	1287	154
9	0.8476	1907	1796	180	0.8264	1346	1236	124
10	0.8374	1516	1395	126	0.7877	1132	1049	157
11	0.7664	1944	1817	327	0.7190	1228	1132	249
12	0.7254	1135	992	169	0.7209	1063	970	204
13	0.7480	1446	1319	237	0.7092	1242	1159	278
14	0.7234	1129	996	179	0.6987	1037	941	216
15	0.7282	1076	944	160	0.7419	1247	1171	246
16	0.8029	1561	1492	239	0.7348	1269	1211	279
17	0.8113	1959	1892	303	0.7542	1414	1333	267
18	0.7999	1562	1470	221	0.6861	1138	1041	260
19	0.7806	1599	1486	238	0.6800	1039	942	236
20	0.7849	1589	1521	274	0.7445	1182	1100	220
Total	0.8149	31152	28848	3460	0.7655	24415	22651	3962

5.2. Computational Cost Evaluation

The simultaneous collaboration between the RGB-D skeletonization method and proposed SSSE method is crucial for the real-time deployment of the fusion application. Thus, limiting the computational cost is essential for the effectiveness of the fusion method. The test platform is a mainstream personal laptop connected with the first generation Kinect, equipped with one Intel Core i7 central processing unit (CPU) and 16 Gigabyte of random access memory (RAM). Two indicators, i.e., maximum process capability per second and single frame delay are concerned in order to evaluate the computational cost. This evaluation test aims to process as many frames as the computational capability allows based on the proposed method. The computation cost efficiency of the fusion application is determined by the number of frames processed per second. As shown in Figure 8, the stable maximum process capability remains around 25 frames per second after the initial stage where less than 10 frames are processed per second. The experimental result indicates that the fusion application is feasible for real-time deployment based on its stable maximum process capability.

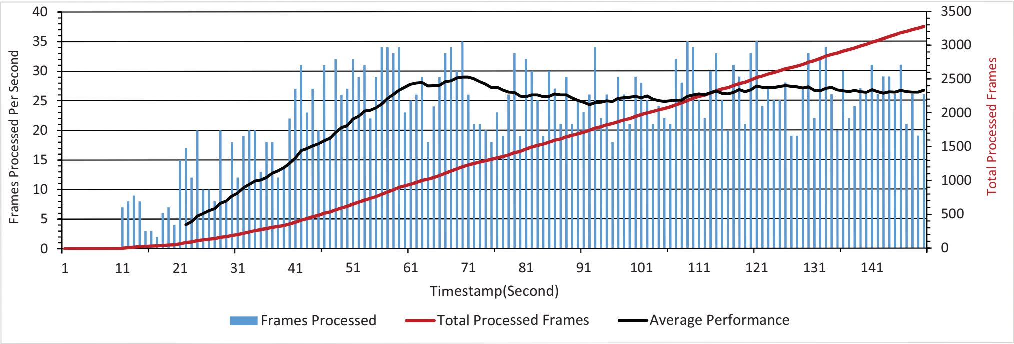


Figure 8. Maximum process capability test.

6. Conclusions

In this paper, we proposed a shadow silhouette-based skeleton extraction (SSSE) method. SSSE extracts three-dimensional human skeleton based on the human shadow information on the ground. Specifically, the proposed SSSE method comprises the following:

(1) A block matrix-based projection transformation is proposed, allowing the reconstruction of precise shadow silhouette information from human shadow captured by monocular camera.
(2) A silhouette shadow-based human skeleton extraction method is proposed. The proposed SSSE method extracts 3D positions of seven major joints in the human skeleton based on the reconstructed human shadow silhouette information and light source position.
(3) A temporal–spatial integration algorithm for discrete shadow silhouette information is proposed, empowering the SSSE-based human skeletonization in single light source scenario.

As shown in Table 2, compared with the traditional RGB-D human skeletonization method and other mono-RGB method, the proposed SSSE method has the following advantages:

(1) The SSSE method can be deployed in large-scale outdoor scenarios where traditional 3D human skeletonization algorithms are not effective.
(2) the SSSE method is capable of extracting human skeleton from frames shot by any normal monocular camera.
(3) The SSSE method can be deployed in stretching the effective range of traditional RGB-D skeletonization method in the fusion application.

Table 2. Comparison between external sensor information-based quadcopter monitoring methods.

Methods	**Device Requirement**	**Effective Range R_e**	**Output Format**	**Joint Numbers**
SSSE	Single RGB Camera	7.0 m <R_e < 10 m	Human Skeleton	7
Traditional RGB-D Method [20]	RGB-D Camera	0.8 m < R_e < 3.5 m	Human Skeleton	20
SSSE and RGB-D Fusion	RGB-D Camera	0.8 m <R_e < 10 m	Human Skeleton	7 to 20
Jafari's RGB-D method [16]	RGB-D Camera	Not Available (N/A)	Human Voxel	0
Yang's mono-RGB method [6]	Multiple RGB Cameras	N/A	Partial Voxels	0

For traditional outdoor surveillance systems, the limited 8-Bit color depth in the analogy transmission system restricts the precision of depth information. Based on the proposed SSSE method, precise 3D human skeleton activities can be extracted at any monitoring terminal. The extracted 3D human skeleton activities will enrich the information for surveillance video analyses, empowering convenient 3D scenario reproduction. Because of the simplicity in device requirement and the compatibility with the traditional surveillance network, the proposed SSSE is an ideal upgrade solution for a traditional surveillance system without extra hardware expenditure.

In conclusion, SSSE offers an extra choice for 3D human skeletonization other than depth camera, wearable sensors, or illuminator array, laying down a milestone to deploy in-lab human skeleton-related methods [6,16,20] in outdoor scenarios with normal photographic devices. Based on the unique outdoor merits provided by SSSE, we will focus our future research on applications of SSSE on outdoor surveillance and unmanned aerial vehicle navigation.

Acknowledgments: We appreciate the detailed comments and kind suggestions provided by reviewers. This research was supported by the National Natural Science Foundation of China under Grants No. 61571346 and'61305040.

Author Contributions: All authors contributed to the research work. Jie Hou conceived the new SSSE method and designed the experiments. Baolong Guo and Wangpeng He reviewed the research work. Jinfu Wu participated in the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimedia* **2012**, *19*, 4–10.
2. Xia, L.; Aggarwal, J. Spatio–temporal depth cuboid similarity feature for activity recognition using depth camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2834–2841.
3. Zhang, X.; Gao, Y. Heterogeneous specular and diffuse 3-D surface approximation for face recognition across pose. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 506–517.
4. Zhang, Y.; Mu, Z.; Yuan, L.; Zeng, H.; Chen, L. 3D Ear Normalization and Recognition Based on Local Surface Variation. *Appl. Sci.* **2017**, *7*, 104.
5. Lay, Y.L.; Yang, H.J.; Lin, C.S.; Chen, W.Y. 3D face recognition by shadow moiré. *Opt. Laser Technol.* **2012**, *44*, 148–152.
6. Yang, T.; Zhang, Y.; Li, M.; Shao, D.; Zhang, X. A multi-camera network system for markerless 3d human body voxel reconstruction. In Proceedings of the 2009 IEEE Fifth International Conference on Image and Graphics, Xi'an, China, 20–23 September 2009; pp. 706–711.
7. Chen, L.C.; Hoang, D.C.; Lin, H.I.; Nguyen, T.H. Innovative methodology for multi-view point cloud registration in robotic 3D object scanning and reconstruction. *Appl. Sci.* **2016**, *6*, 132.
8. Gouiaa, R.; Meunier, J. 3D reconstruction by fusioning shadow and silhouette information. In Proceedings of the IEEE 2014 Canadian Conference on Computer and Robot Vision (CRV), Montreal, QC, Canada , 6–9 May 2014; pp. 378–384.
9. Wang, Y.; Huang, K.; Tan, T. Human activity recognition based on r transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
10. Chen, C.C.; Aggarwal, J. Recognizing human action from a far field of view. In Proceedings of the 2009 IEEE Workshop on Motion and Video Computing, Snowbird, UT, USA, 8–9 December 2009; pp. 1–7.
11. Jin, X.; Kim, J. A 3D Skeletonization Algorithm for 3D Mesh Models Using a Partial Parallel 3D Thinning Algorithm and 3D Skeleton Correcting Algorithm. *Appl. Sci.* **2017**, *7*, 139.
12. Shotton, J.; Girshick, R.; Fitzgibbon, A.; Sharp, T.; Cook, M.; Finocchio, M.; Moore, R.; Kohli, P.; Criminisi, A.; Kipman, A.; et al. Efficient human pose estimation from single depth images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2821–2840.
13. Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; Moore, R. Real-time human pose recognition in parts from single depth images. *Commun. ACM* **2013**, *56*, 116–124.
14. Song, Y.; Liu, S.; Tang, J. Describing trajectory of surface patch for human action recognition on RGB and depth videos. *IEEE Signal Proc. Lett.* **2015**, *22*, 426–429.
15. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334.
16. Jafari, O.H.; Mitzel, D.; Leibe, B. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 5636–5643.
17. Juang, C.F.; Chang, C.M.; Wu, J.R.; Lee, D. Computer vision-based human body segmentation and posture estimation. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* **2009**, *39*, 119–133.

18. Hsieh, J.W.; Hsu, Y.T.; Liao, H.Y.M.; Chen, C.C. Video-based human movement analysis and its application to surveillance systems. *IEEE Trans. Multimedia* **2008**, *10*, 372–384.
19. Yuan, X.; Yang, X. A robust human action recognition system using single camera. In Proceedings of the 2009 International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 11–13 December 2009; pp. 1–4.
20. Hu, M.C.; Chen, C.W.; Cheng, W.H.; Chang, C.H.; Lai, J.H.; Wu, J.L. Real-time human movement retrieval and assessment with Kinect sensor. *IEEE Trans. Cybern.* **2015**, *45*, 742–753.