Extracción de Características de Imagen para Recuperación 3D

M. Guerrero[1], J. M. Santos[1], J. Gambini[1,2,3]

[1]Depto. de Ingeniería en Informática, Instituto Tecnológico de Buenos Aires, Buenos Aires, Argentina
[2]Depto. de Ingeniería en Computación, Universidad Nacional de Tres de Febrero, Caseros, Pcia. de Buenos Aires, Argentina
[3]Centro de Procesamiento de Señales e Imágenes, Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Buenos Aires, Argentina.
mguerrero@itba.edu.ar, mgambini@itba.edu.ar, jsantos@itba.edu.ar

Resumen La reconstrucción 3D a partir de imágenes 2D es un desafío en el campo de imágenes y visión, con múltiples aplicaciones en áreas muy diversas. Algunos métodos se basan en marcadores, los cuales son puntos estratégicamente ubicados en el objeto de interés o sobre el traje de un ser humano, diseñado para este propósito. A diferencia de estos métodos, este trabajo se enfoca en encontrar características en imágenes 2D las cuales pueden ser utilizadas como marcadores, permitiendo la reconstrucción 3D automáticamente. Utilizamos el método SIFT (Scale Invariant Feature Transform) para asociar puntos característicos en imágenes de la misma escena provenientes de diferentes puntos de vista. Nos encontramos mejorando el proceso por medio del reconocimiento de *skeletons*. El objetivo de este trabajo es que los puntos encontrados se utilicen para estimar estructuras 3D. Los resultados obtenidos hasta el momento son alentadores.
Keywords: Recuperación 3D, Recuperación de Pose, método SIFT, *Skeleton*

1. Introducción

La reconstrucción de objetos 3D a partir de imágenes 2D posee múltiples aplicaciones como la recuperación topográfica, la reconstrucción de huesos para la creación de prótesis, el reconocimiento de pose, la industria del cine y de la robótica, entre otras [1]. Este proceso presenta varias dificultades, como la representación de la forma del objeto, las medidas de similitud entre dos formas distintas, la robustez bajo transformaciones afines y ruido o el alto costo computacional [2]. Uno de los enfoques para abordar este problema está basado en múltiples vistas, el cual recupera el objeto 3D utilizando un conjunto de imágenes capturadas a partir de distintos puntos de vista de la misma escena [1,3]. Éste es el objetivo del presente artículo.

Algunos programas de captura de movimiento recuperan la información 3D de la escena utilizando marcadores, los cuales se ubican en lugares estratégicos del cuerpo, como por ejemplo las articulaciones, la cabeza, los pies, etc. Este

enfoque se basa en modelos del cuerpo humano previamente definidos [4]. Los algoritmos más recientes combinan estos modelos con el proceso de reconstrucción 3D aplicando técnicas de *deep learning* [5,6].

Otro tipo de métodos son aquellos que estiman la pose del cuerpo reconociendo características de las imágenes sin utilizar información a priori [7]. Inspirados en ese tipo de metodología, asociamos dos imágenes por medio del reconocimiento de correspondencias en esas características, las cuales pueden ser utilizadas como marcadores. Para esto, aplicamos el método SIFT, el cual nos permite detectar características de las imágenes [8]. Las mismas tienen dos atributos: la ubicación centrada en un pixel y un descriptor. El primero es el centro de un parche en la imagen y el segundo es un vector de 128 elementos que contiene información sobre el histograma de orientaciones dentro del parche.

Otro tipo de método que resulta de nuestro interés son aquellos basados en *skeleton* [9], los cuales encuentran una simplificación del cuerpo humano o esqueleto que se utiliza para capturar movimientos, posturas y pose.

Finalmente, queremos reconstruir el objeto en 3D aplicando los métodos mencionados a partir de imágenes capturadas por distintos puntos de vista dentro de un laboratorio. En este artículo presentamos la idea principal del proyecto.

2. Sistema de Captura de Múltiples Vistas

Las imágenes fueron tomadas en el Laboratorio de Captura de Movimiento, situado en el Instituto Tecnológico de Buenos Aires. Corresponden a ocho cámaras Flex 3 Optitrack, con una resolución de 0,3Mp y con imágenes de (480×640)px. Las cámaras está ubicadas en trípodes alrededor de un círculo y en el centro del mismo está el objeto de interés. La Figura 1 muestra un esquema del sistema. Las ubicaciones y orientaciones de las cámaras se muestran con flechas y el centro es la ubicación del objetivo. De esta forma, es posible tomar imágenes desde ocho puntos de vista diferentes.

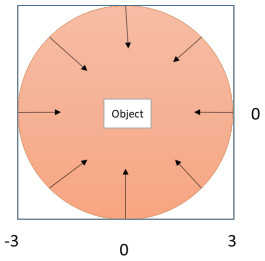

Figura 1. Esquema de posiciones y orientaciones de las cámaras (indicadas con flechas) dentro de la sala del Laboratorio de Captura de Movimiento.

En primer lugar, se calibra el sistema y se calculan los parámetros intrínsecos y extrínsecos de las cámaras. Los primeros se refieren a las características de la

cámara como la distancia focal y la ubicación del punto principal, mientras que los segundos son utilizados para describir la transformación entre la cámara y el sistema de referencia del mundo [10].

El siguiente paso es capturar los movimientos de la persona en la escena, sin utilizar marcadores. La Figura 2 muestra dos imágenes de la misma escena capturadas por dos cámaras consecutivas.

Figura 2. Dos imágenes de la misma escena tomadas por dos cámaras consecutivas.

La Figura 3 muestra los resultados que se generan al aplicar el método SIFT a las imágenes de la Figura 2.

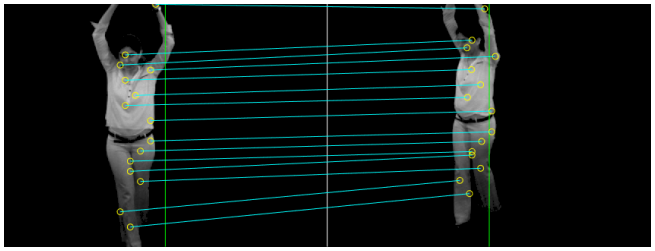

Figura 3. Correspondencias entre dos imágenes provenientes de distintos puntos de vista.

Si bien los resultados hallados son alentadores, observamos que los puntos característicos encontrados no son suficientes para describir la pose o la forma 3D, por ejemplo puede ocurrir que no estén ubicados en todas las extremidades del cuerpo o en las articulaciones, como manos, rodillas, codos, etc., los cuales son elementos claves para describir la estructura 3D.

Para resolver este problema, recurrimos al estimador corporal conocido como *skeleton*. La ventaja de utilizar *skeleton* es que puede restringirse la búsqueda de puntos asociados entre aquellos que se encuentren en regiones correspondientes, además de reducir el costo computacional. El método utilizado para hallar

el *skeleton*, consiste en encontrar el contorno del cuerpo humano utilizando el método presentado en [11], para luego detectar las extremidades y posteriormente utilizar las métricas corporales para detectar partes del cuerpo, por ejemplo, si h es la altura del sujeto, la *cabeza* mide $\frac{h}{8}$ y el cuello tiene una longitud $0{,}37*$ *cabeza*. Los *skeletons* permiten buscar puntos correspondientes entre dos imágenes, utilizando como referencia los puntos principales de ellos en cada una de las imágenes.

El Algoritmo 1 describe el pseudocódigo del método utilizado. Sean I_j, C_j y LI_{C_j}, $j = 1, \ldots, 8$ las imágenes capturadas, el contorno del objeto de interés de cada una y la lista de píxeles en el interior del contorno, respectivamente.

Algoritmo 1 Algoritmo utilizando *skeletons*.

Dado C_1 y las métricas corporales, obtener un *skeleton* S_1 de 14 puntos principales.
Obtener los *skeletons* S_i,$i = 2, \ldots, 8$ de acuerdo a los parámetros extrínsecos calculados en la calibración.
Sobreimprimir los *skeletons* S_i sobre las imágenes I_i, $i = 1, \ldots, 8$.
Inicializar la lista L vacía.
for $i = 1$ **to** 8 **do**
  for $p_i \in LI_{C_i}$ **do**
    Obtener su descriptor d_{p_i} usando SIFT y encontrar la correspondencia p_{i+1} en $LI_{C_{i+1}}$, usando como referencia los puntos principales de los *skeletons* S_i y S_{i+1}.
    Obtener las coordenadas (x, y, z) respecto al centro de referencia del sistema.
    Agregar (x, y, z) a L.
  end for
end for
return L

La Figura 4 muestra el *skeleton* encontrado en una imagen.

En el paso 8 del Algoritmo 1, se obtiene la profundidad del punto en la escena a partir de los píxeles correspondientes y de los parámetros extrínsecos e intrínsecos del sistema, utilizando triangulación. En el paso 12, el algoritmo devuelve una lista de puntos (x, y, z) con la cual puede calcularse un modelo 3D del sujeto usando, por ejemplo, triángulos o trapecios.

3. Conclusiones

A partir de las pruebas realizadas, es posible encontrar correspondencias entre dos imágenes del cuerpo humano, tomadas desde distintos puntos de vista. Los píxeles que pueden ser asociados, son utilizados como marcadores para la reconstrucción 3D. Estudiamos el uso de *skeleton* y del método de triangulación para reconstruir su superficie. Sabemos que tenemos trabajo por delante pero los resultados obtenidos hasta el momento son prometedores.

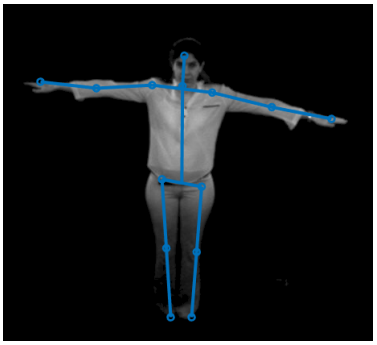

Figura 4. *Skeleton* del cuerpo humano. Pueden identificarse los 14 puntos principales.

Referencias

1. Liu, A., Hu, N., Song, D., Guo, S., Zhou, H., Hao, T.: Multi-view hierarchical fusion network for 3D object retrieval and classification. IEEE Access **7** (2019) 153021–153030
2. Bronstein, A., Bronstein, M., Guibas, L., Ovsjaniko, M.: Shape google: Geometric words and expressions for invariant shape retrieval. ACM Transactions on Graphics **30**(1) (2011) 11–20
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
4. Zhang, D., Miao, Z., Chen, S.: Human model adaptation for multiview markerless motion captures. Mathematical Problems in Engineering **2013**(1) (2013) 1–7
5. Yasin, H., Iqbal, U., Krugerand, B., Weber, A., Gall, J.: A dual-source approach for 3D pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 4948–4956
6. Mathis, A., Schneider, S., Lauer, J., Mathis, M.: A primer on motion capture with deep learning: Principles, pitfalls, and perspectives. Neuron **108**(1) (2020) 44–65
7. Park, J., Park, K., Baeg, S., Baeg, M.: Reliable feature point detection and object pose estimation using photometric quasi-invariant SIFT. In: International Conference on Control, Automation and Systems. (2008) 2142–2147
8. Guerrero, M., Santos, J., J.Gambini: 3d features recovery from images of a multi camera system. In: International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPC). (2021) 1–5
9. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H., Rhodin, H., Pons-Moll, G., Theobalt, C.: XNect: Real-time multi-person 3D motion capture with a single RGB camera. In: ACM Transactions on Graphics. Volume 39. (July 2020)
10. Zhang, Z.: Camera Parameters (Intrinsic, Extrinsic). Springer (2014)
11. Gambini, J., Rozichner, D., Buemi, M.E., Mejail, M., Berllés, J.: Occlusion handling for object tracking using a fast level set method. In: XXI Brazilian Symposium on Computer Graphics and Image Processing. (2008) 61–68