Reconstrucción y animación 3D

Dana Urribarri[1], Juan Ignacio Larregui[1,2], Martín Larrea[1,2], Silvia Castro[1,2]

[1]Laboratorio de I+D en Visualización y Computación Gráfica (UNS–CIC Prov. Buenos Aires)
Departamento de Ciencias e Ingeniería en Ciencias de la Computación
Universidad Nacional del Sur, Av. Alem 1253, Bahía Blanca

[2]Instituto de Ciencias e Ingeniería de la Computación (UNS–CONICET)
Departamento de Ciencias e Ingeniería en Ciencias de la Computación
Universidad Nacional del Sur, Av. Alem 1253, Bahía Blanca

{dku, juan.larregui, mll, smc}@cs.uns.edu.ar

RESUMEN

En los campos de la Computación Gráfica y de la Visión por Computadora nos encontramos con dos desafíos importantes. En primer lugar, dentro del área de la Reconstrucción 3D, la recuperación de información de profundidad a partir de imágenes es una tarea laboriosa que requiere no sólo el análisis de características en las imágenes, sino también la correcta aplicación de las propiedades de la geometría de la perspectiva.

Por otro lado, dentro del área de animación, conseguir una animación realista de humanos virtuales no es sólo una tarea sumamente compleja, sino que además, cualquier imperfección es altamente perceptible y produce el rechazo de quien lo observa. Es por esto que muchas aplicaciones utilizan capturas de movimientos para animar sus humanos virtuales.

El objetivo general de esta línea de investigación consiste tanto en el estudio y análisis de técnicas de Reconstrucción 3D a partir de imágenes, como en el análisis de capturas de movimientos para identificar las principales características de los movimientos reales y modelar estos movimientos de manera que permitan ser reproducidos en la animaciones.

Los trabajos se realizan dentro del VyGLab entre becarios y docentes investigadores de la Universidad Nacional del Sur.

Palabras claves: *Computación gráfica, Visión por Computadora, Reconstrucción 3D, Animación 3D, Mo-Cap.*

CONTEXTO

Este trabajo se lleva a cabo en el Laboratorio de Investigación y Desarrollo en Visualización y Computación Gráfica (VyGLab) del Departamento de Ciencias e Ingeniería de la Computación, de la Universidad Nacional del Sur. Los trabajos realizados bajo esta línea involucran a docentes investigadores y becarios.

La línea de Investigación presentada está inserta en el proyecto acreditado *Análisis Visual de Grandes Conjuntos de Datos* (24/N037), dirigido por la Dra. Silvia Castro y en el proyecto *Análisis de Capturas de Movimientos para la Animación de Humanos Virtuales* (24/ZN33) dirigido por la Dra. Dana Urribarri; ambos financiados por la Secretaría General de Ciencia y Tecnología de la Universidad Nacional del Sur.

1. INTRODUCCIÓN

Dentro de esta línea de Investigación se está trabajando en la reconstrucción 3D a partir de fotografías y en el análisis de Capturas de movimientos (*Mo-Caps)* para la animación de humanos virtuales

1.1 Reconstrucción 3D

La Reconstrucción 3D a partir de imágenes consiste en recuperar la información de la profundidad de la escena en cada uno de los puntos proyectados en el plano de la imagen (píxeles). Se trata de un área que ha recibido gran atención durante el siglo pasado y el

actual, en la cual se ha desarrollado el estudio de las relaciones geométricas entre las proyecciones de una escena a diferentes puntos de vista [17], pero que aún dista de ser un problema solucionado.

En los últimos años, el campo de la Visión por Computadora se ha visto ampliamente afectado por los avances en las técnicas de Deep Learning (DL), especialmente por la aplicación de Redes Neuronales Convolucionales que han obtenido rendimientos superiores a las técnicas tradicionales en la amplia mayoría de los tópicos de interés del campo, como Clasificación de Imágenes, Detección de Objetos, Segmentación Semántica y Reconstrucción 3D, entre otros. La capacidad de estas redes para capturar características de las imágenes de diferentes complejidades sin la necesidad del diseño explícito por parte del humano las hace atractivas para la Reconstrucción 3D, donde la gran variabilidad de los datos visuales y su alta dimensionalidad (potencialmente millones de píxeles) dificulta dicho diseño. Varios trabajos relacionados han sido publicados en los que se entrena las redes para traducir una imagen a un mapa de profundidades [12, 13].

A pesar de su atractivo, una de las principales dificultades en la aplicación del DL a la Reconstrucción 3D se encuentra en la necesidad de contar con un gran volumen de datos etiquetados con información de profundidad para supervisar el entrenamiento de estas redes. La obtención de estas etiquetas debe realizarse mediante sensores y/o en entornos controlados o sintéticos, y en muchos casos requiere un procesamiento manual antes de poder ser utilizadas por las técnicas de DL, haciendo que el volumen de datos no sea fácilmente escalable.

De esto se desprende la necesidad de nuevas técnicas que permitan entrenar redes neuronales convolucionales con menor cantidad de datos, aprovechando el conocimiento geométrico desarrollado durante las últimas décadas, evitando el acercamiento *naive* que deja la totalidad de la tarea de la reconstrucción a las redes neuronales, esperando que no sólo aprendan características de las imágenes, sino también relaciones propias de la geometría de la perspectiva.

1.2 Animación de Humanos Virtuales

El análisis del movimiento humano (*Human Movement Analisys*, *HMA*) se refiere al análisis e interpretación de los movimientos humanos en el tiempo [1,3]. Durante décadas, fue un campo de investigación que atravesaba varias áreas: biología, psicología, multimedia, etc. En el campo de la visión por computadora, el *HMA* emergió gracias al video y a la aparición de sofisticados algoritmos de dominio público. Las tecnologías de *Mo-Cap* han agregado al *HMA* la posibilidad de analizar el movimiento a partir de una representación en 3D del esqueleto [14].Por otro lado, hoy en día, los ambientes sintéticos habitados por humanos virtuales (HHVV) son habituales en un sinnúmero de aplicaciones [6,15,16,18]. Sin embargo, crear un humano virtual (HV) es una tarea sumamente compleja. Dado que estamos acostumbrados a cómo luce hasta el último detalle de un humano, cualquier imperfección en el HV es altamente perceptible y produce el rechazo de quien lo observa [2,7,9]. La teoría del valle inquietante sostiene que cuanto más cerca se está de lograr algo artificialmente humano, mayor es el nivel de rechazo que hay en los observadores humanos [19]. Actualmente existen diversas técnicas para realizar animaciones interactivas en tiempo real [20]; éstas técnicas difieren en el *trade–off* que ofrecen entre la cantidad de control sobre el movimiento, la exactitud y naturalidad del movimiento resultante y el tiempo de cálculo requerido. Elegir la técnica adecuada depende de las necesidades de la aplicación.

Es por esto que en muchas aplicaciones se utilizan *Mo-Caps* almacenados en bases de datos que posteriormente se trasladan a los modelos de HHVV para animarlos. Teniendo en cuenta que se debe almacenar una gran cantidad de *Mo-Caps* para obtener diversidad de movimientos y que estos pueden aplicarse solo en escenarios previamente planeados, surge la necesidad de contar con métodos alternativos para sintetizar humanos que se comporten naturalmente. Una estrategia

tradicionalmente empleada es la utilización de las capturas de movimiento conjuntamente con métodos algorítmicos; sin embargo estos últimos aproximan burdamente las restricciones físicas del cuerpo y del entorno y por lo tanto generan artefactos visuales e intersecciones entre los objetos.El movimiento del cuerpo humano se puede describir desde varios puntos de vista, por ejemplo el mecanismo del movimiento en el espacio y el tiempo, la expresividad cualitativa del movimiento, la trayectoria del movimiento en el espacio, el ritmo y la coordinación del movimiento, entre otras características.Lograr que los HHVV se muevan de manera aceptable es un desafío que requiere identificar las principales características de los movimientos reales y modelar estos movimientos de manera que permitan ser reproducidos en la animación de HHVV.

Un mejor entendimiento de los factores que hacen al movimiento humano reconocible y aceptable es de gran valor en las aplicaciones que requieren realismo en los movimientos de los personajes virtuales [4,7]. La animación realista de un HV es un problema desafiante. Los procesos biomecánicos y fisiológicos que ocasionan el movimiento son difíciles de entender y replicar.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

En el contexto de esta línea de investigación se están realizando paralelamente los siguientes trabajos:

- Reconstrucción 3D a partir de imágenes en tiempo real.
- Análisis de *Mo-Caps* para la animación de humanos virtuales.

1.1 Reconstrucción 3D

Los desafíos más significativos que se plantean en el proceso de Reconstrucción 3D mediante técnicas de Deep Learning los constituyen:

- La obtención tanto de la secuencia de imágenes como de la información de profundidad de cada píxel.
- La variabilidad de iluminación/color de puntos correspondientes en imágenes desde distintos puntos de vista.
- El entrenamiento de redes neuronales con volúmenes de datos limitados, que tengan la capacidad de predecir correctamente la profundidad de una escena nunca vista (generalización).
- La asignación de información de color a los puntos reconstruidos, pese a cambios de iluminación en las distintas imágenes.
- El diseño y desarrollo eficiente y usable del proceso de reconstrucción 3D en tiempo real.

1.2 Animación de Humanos Virtuales

En el contexto de animación de humanos virtuales, hay varias líneas de trabajo que es necesario atacar para enfrentar este desafío:En primera instancia es necesario contar con herramientas que permitan analizar comparativamente diferentes repeticiones de una secuencia de movimientos. Este análisis puede llevar a identificar secuencias correctamente ejecutadas, medir la experiencia de una persona realizando un movimiento, identificar cuáles son las falencias en la realización de una rutina, etc. Por otro lado, es necesario identificar las propiedades que, además de la trayectoria, hacen al movimiento humano. ¿Qué hace que dos rutinas que ejecutan la misma secuencia de movimientos se perciban de forma diferente?En cuanto a la animación de HHVV, contar con herramientas de comparación permite identificar cuáles son los puntos débiles de los movimientos sintéticos y tomar medidas para corregirlos.

3. RESULTADOS OBTENIDOS/ESPERADOS

1.1 Reconstrucción 3D

Esta línea de investigación explora la reconstrucción 3D de escenas continuas, donde se requiere la estimación de profundidad de cada uno de los cuadros que la constituyen.

El problema puede clasificarse como una regresión densa donde se estima, para cada píxel, su distancia respecto a la cámara. Para ello se han diseñado diferentes arquitecturas de redes neuronales convolucionales, partiendo de redes probadas en tareas de regresión por píxel [10, 11] y se ha supervisado su entrenamiento con diferentes datasets públicamente disponibles [5, 8].

Hasta el momento, la estimación se ha realizado de manera independiente para cada cuadro, sin tener en cuenta la relación de los mismos en el tiempo. Se investigará la aplicación de técnicas de Deep Learning que permitan incorporar esta relación temporal, como la utilización de Redes Neuronales Recurrentes. El objetivo es aprovechar predicciones de cuadros previos en la secuencia de manera de predecir la profundidad de cada cuadro no sólo por la información presente en el mismo, sino también en su contexto.

La correcta estimación de la profundidad en los cuadros de la secuencia representa un paso fundamental en la reconstrucción, modelado y posterior animación de la misma.

1.2 Animación de Humanos Virtuales

Esta línea de investigación se centra en el análisis comparativo de capturas de movimientos en el dominio específico del karate.

En colaboración con el "*Geometry and Graphics Group*" del Departamento de Informática, Bioingeniería, Robótica e Ingeniería en Sistemas (DIBRIS) de la Universidad de Génova (www.unige.it), Italia, se han conseguido *Mo-Caps* de varios karatekas, tanto expertos como novatos, realizando la misma rutina de entrenamiento. Actualmente se logró acondicionar los datos para comenzar con los análisis comparativos:

- Inicialmente, se han completado las capturas de movimiento.
- Luego, las capturas se alinearon y normalizaron en el tiempo. De esta forma el análisis posterior es independiente de la altura de las personas y de la orientación con la se realiza la rutina.
- Finalmente, se han alineado en el tiempo para evitar que pequeños desajustes en la velocidad de ejecución de la rutina incidan negativamente en la comparación.

A partir de ahora se continuará analizando técnicas para, a partir de parámetros estadísticos, *clusterización* y técnicas de *Deep Learning* como *Recurrent Neural Networks* y *Convolutional Neural Networks* entre otras, comparar las secuencias y distinguir automáticamente las secuencias realizadas por atletas expertos de atletas con niveles de experiencia menores. De esta forma se espera lograr un análisis comparativo de movimientos realizados por atletas expertos, intermedios y novatos.

4. FORMACIÓN DE RECURSOS HUMANOS

En lo que concierne a la formación de recursos humanos se incentiva la incorporación de alumnos que deseen realizar su tesina o trabajo final de carrera en alguno de estos temas. Por otro lado, el Ing. Larregui está realizando su trabajo de tesis doctoral bajo la dirección de la Dra. Castro en el tema de Estimación de Profundidad en tiempo real mediante la utilización de técnicas de Deep Learning.

5. BIBLIOGRAFÍA

[1] Jasbir Arora and Karim Abdel-Malek. Human Motion Simulation: Predictive Dynamics. Academic Press, 1st edition, 2013.

[2] James E. Cutting and Lynn T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. Bulletin of the Psychonomic Society, 9:353–356, 1977.

[3] Yu Ding, Ken Prepin, Jing Huang, Catherine Pelachaud, and Thierry Artières. Laughter animation synthesis. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, AAMAS '14, pages 773–780, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.

[4] Rukun Fan, Songhua Xu, and Weidong Geng. Example-based automatic music-driven conventional dance motion synthesis. IEEE Transactions on Visualization and Computer Graphics, 18(3):501–515, 2012.

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1. 2017..

[6] S. Hagler, D. Austin, T.L. Hayes, J. Kaye, and M. Pavel. Unobtrusive and ubiquitous inhome monitoring: A methodology for continuous assessment of gait velocity in elders. IEEE Transactions on Biomedical Engineering, 57(4):813–820, 2010.

[7] Ludovic Hoyet, Kenneth Ryall, Katja Zibrek, Hwangpil Park, Jehee Lee, Jessica Hodgins, and Carol O'Sullivan. Evaluating the distinctiveness and attractiveness of human motions on realistic virtual bodies. ACM Transactions on Graphics, 32(6):204:1–204:11, November 2013.

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research 32, no. 11 (2013): 1231-1237.

[9] K. L. Johnson and L. G. Tassinary. Perceiving sex directly and indirectly: Meaning in motion and morphology. Psychological Science, 16(11):890–897, 2005.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017).

[11] Olaf, Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[12] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. arXiv preprint arXiv:1704.07804 (2017).

[13] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. CoRR, vol. abs/1703.04309 (2017).

[14] Liliana Lo Presti and Marco La Cascia. 3D skeleton-based human action classification. Pattern Recognition. 53, C (May 2016), 130-147. 2016.

[15] Nadia Magnenat-Thalmann and Zerrin Kasap. Virtual humans in serious games. In Proceedings of the 2009 International Conference on CyberWorlds, CW '09, pages 71–79, Washington, DC, USA. IEEE Computer Society. 2009.

[16] J. Music, M. Cecic, and M. Bonkovic. Testing inertial sensor performance as hands-free human-computer interface. WSEAS Transactions on Computers, 8:715–724, April 2009.

[17] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003.

[18] V. B. Zordan, A. Majkowska, B. Chiu, and M. Fast. Dynamic response for motion capture animation. ACM Transactions on Graphics, 24:697–701, 2005.

[19] Katsu Yamane, Yuka Ariki, and Jessica K. Hodgins. Animating non-humanoid characters with human motion data. In Zoran Popovic and Miguel A. Otaduy, editors, Symposium on Computer Animation, pages 169–178. Eurographics Association, 2010.

[20] H. van Welbergen, B.J.H. van Basten, A. Egges, Z.M. Ruttkay, and M.H. Overmars. Real time character animation: A trade-off between naturalness and control. In M. Pauly and G. Greiner, editors, Eurographics - State-of-the-Art-Report, pages 45–72, Munich, 2009. Eurographics Association. ISSN: 1017-4656.