IDENTIFICACIÓN DE POSES DEL CUERPO HUMANO PARA CARACTERIZAR Y OBTENER INFORMACIÓN DE MOVIMIENTO.

BELTRÁN PERÉZ CRISTIAN CAMILO

Trabajo de grado presentado como requisito para optar al título de:

INGENIERO EN MULTIMEDIA

Director:

Ing. Wilman Helioth Sánchez R., Esp. M.Ed.

UNIVERSIDAD MILITAR NUEVA GRANADA

FACULTAD DE INGENIERÍA

PROGRAMA INGENIERIA EN MULTIMEDIA

BOGOTÁ D.C

ENERO 2019

Identification of Human Body Poses to Characterize and Obtain Movement Information.

C. C. Beltrán, W. H. Sanchez and E. L. Sierra

***Abstract*— This paper, studies human poses classification and human body movement. It was built a convolutional neural network to approach the problem of classification, who they were defined 6 ideal poses to classify. The approach has the advantage of reasoning about pose in a controlled, with an 84.78% of precision. For each pose, a descriptor was located in an area of the human body, were the movement is high. The different descriptors was located in head, torso, arms and legs, were the activity recognition approach an 82.35% of precision.**

***Keywords*— Feature extraction, image classification, image capture, image processing, neural networks.**

I. INTRODUCCION

LA CANTIDAD de videos que son realizados en la actualidad está creciendo exponencialmente, lo cual implica mayor consumo de recursos al momento de analizar información que pueda ser de interés [1]. Existe diferente información que puede ser estudiada en videos como identificación de objetos o de personas que realizan acciones; Este último implica un análisis más profundo del comportamiento humano, lo que lleva a caracterizar el movimiento para su posterior estudio.

El análisis del movimiento humano actualmente ha sido de gran interés en áreas de la multimedia como la visión por computador. En particular, el reconocimiento de la actividad humana puede verse reflejado en múltiples aplicaciones como la Animación 3D, captura de movimiento y efectos especiales, psicología, seguridad, cinestesia entre otros [2]. Los cuales requieren de bases de conocimiento de video o imágenes que permitan el análisis morfológico de personas que realizan acciones.

El procesamiento de repositorios de videos estandarizados puede resultar beneficiosos para los proyectos, sin embargo, pueden sesgar el análisis de la información debido elementos culturales, de contexto y temporales. Por lo tanto, es necesario contar con un repositorio local con amplia cantidad información en cuanto a las necesidades del proyecto. Dado lo anterior, mediante la clasificación de parámetros como la constitución física, peso corporal, altura, vestimenta y otros [3], se puede definir diferentes tipos de acciones que pueden ser vistas como una secuencia de poses que realiza una persona en el espacio- tiempo.

C. C. Beltrán, Universidad Militar Nueva Granada, Bogotá, Colombia, u1201539@unimilitar.edu.co

W. H. Sanchez, Universidad Militar Nueva Granada, Bogotá, Colombia, wilman.sanchez@unimilitar.edu.co

E. L. Sierra, Universidad Militar Nueva Granada, Cajicá, Colombia, edward.sierra@unimilitar.edu.co

Corresponding author: Eduard Leonardo Sierra Ballén

Diferentes estudios han proporcionado avances en cuanto a características espacio-temporales que puedan ser usadas en el estudio del cuerpo humano como: la segmentación y clasificación de acciones humanas en uniones [4], que pueden representar dificultad debido a la falta de información puntual del movimiento; El análisis de acciones a distancia [5], que implica tener un mayor control en la medición y extracción de la información de los múltiples puntos de interés que se pueden ubicar sobre una imagen, los son de poca confiabilidad al tratarse de superficies muy curvas o en videos con muy baja resolución; y por último el estudio de siluetas humanas [1] [6], que requiere un mayor control de la escena de estudio, en cuanto al fondo, la iluminación y objetos dentro de la escena. Por lo anterior, se requiere de sistemas que permitan mitigar las anteriores falencias.

Por otro lado, los diferentes métodos para capturar la información requieren de múltiples sensores ubicados en diferentes partes del cuerpo, que proporcionan la información directa de las acciones realizadas por un sujeto, lo que implica un alto costo computacional. Asimismo, existen otros métodos que requieren de amplios espacios controlados, etc.

El aporte del presente artículo es la realización de un método que permita identificar y recrear movimientos humanos lo más cercanos a la realidad, sin el uso de sensores o sistemas en adición o de amplios espacios.

El presente artículo está organizado de la siguiente forma. La sección 2 muestra los materiales y métodos utilizados durante el proceso. La sección 3 presenta los resultados obtenidos. En la sección 4 se tiene el análisis de resultados y por último las conclusiones del trabajo.

II. MATERIALES Y METODOS

Esta sección describe el marco de trabajo empleado para identificar y clasificar las diferentes poses mediante la ubicación de descriptores en secuencias del cuerpo humano. Los puntos de interés se ubican mediante el uso de histogramas de gradientes orientados para un posterior entrenamiento y clasificación. Los pasos empleados en el proceso son 1. Captura de acciones, 2. Segmentación de siluetas 3. Ubicación de descriptores, 4. Clasificación de poses (Fig 1).

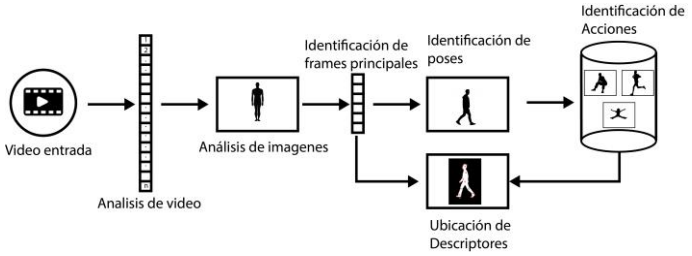

Figura 1. Diagrama del sistema principal.

Captura de acciones

Se realizó la captura de las diferentes secuencias en el laboratorio de hipermedia del campus de la Universidad, ya que cuenta con un Chroma key de gran tamaño, ideal para la captura de las acciones. Cada video cuenta con una resolución de 1920x1080 y una duración de alrededor 5 segundos por video. Inicialmente se definieron 12 diferentes poses a analizar, de las cuales se estudiaron 6 distintas poses ideales a clasificar debido a la gran cantidad de movimiento que estas tienen, dichas poses o categorías son: persona agachada, persona caminando, persona preguntando, saludo de manos, saludo militar, persona tocando el suelo.

Segmentación de poses

Para segmentar las diferentes poses, se hizo uso de la técnica del Chroma Key llevar a cabo el proceso y separación de la silueta humana de un fondo de cada imagen. Inicialmente se usa un espacio de color denominado Hue, Saturation, Value (HSV), en donde se varían estos parámetros para determinar un estándar para todas las imágenes, dichos valores para cada canal son los siguientes: El matiz identifica colores que estén por encima de los 165 niveles y debajo de los 65 niveles, la saturación maneja valores mayores a 35 niveles, y el valor de estos trabaja valores por encima de 45 niveles presentada em la Fig. 2. El resultado es una imagen binaria para cada pose, con la información representada mediante una silueta.

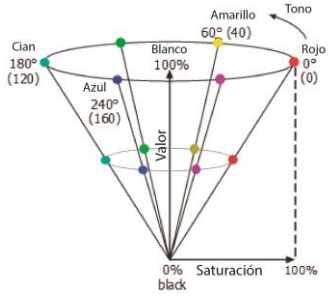

Figura 2. Espacio de color HSV.

Debido a la gran cantidad de información y a la diferente variación en cuanto al tamaño de las imágenes, se estableció un método para normalizar todas las entradas al sistema. Dicho procedimiento consiste en identificar el centro de la imagen (Centroide) con mayor tamaño [3], para establecer un factor de escala, el cual compara con cada una de las imágenes con tamaño mayor (Fig 3). Una vez se ha comparado la diferencia de tamaños para todas las imágenes, estas se escalan de acuerdo a la imagen más grande, con el fin de tener un tamaño igual para todas las imágenes.

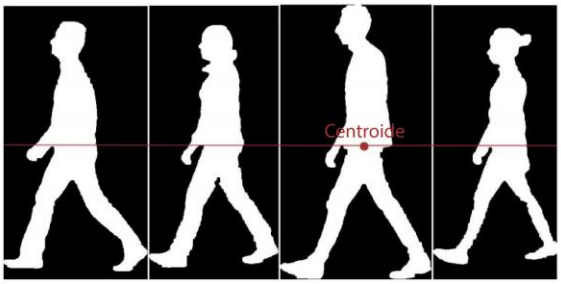

Figura 3. Normalización de imágenes, Todas las imágenes se re escalan de acuerdo al centroide de la imagen de mayor tamaño.

Ubicación de descriptores

Para la ubicación de descriptores sobre las diferentes siluetas humanas, se calculan histogramas locales dentro de una imagen dividida por celdas, identificando a su vez la dirección del gradiente en cada casilla de la misma [1]. La apariencia y forma local de un objeto dentro de una imagen, pueden ser descritas por la distribución de la intensidad del gradiente o la dirección de los bordes. El cálculo del gradiente tiene como antecedente el cálculo de la pendiente, para esto se aplica un filtro lineal centrado en las direcciones vertical y horizontal de cada celda (Fig 4). Existen diferentes filtros que se puede aplicar dentro de este apartado, en el proceso se empleó un filtro lineal 3x3, ya que el uso de máscaras más grandes, disminuye el rendimiento y crea un suavizado en las imágenes que resulta significativo.

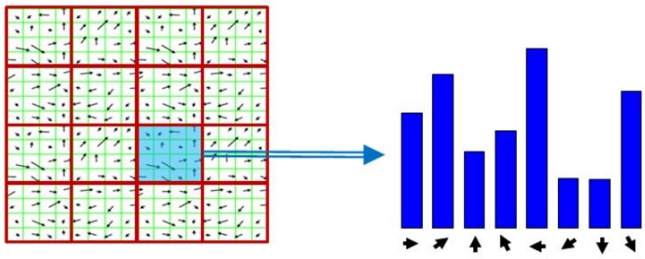

Figura 4. Histograma de gradientes orientados.

El siguiente paso consiste en reunir la información de cada pixel que ha calculado un histograma de orientación basado en la dirección del gradiente ubicado en cada uno, los cuales se van acumulando en contenedores de orientación denominados celdas, cada celda maneja un tamaño de 8x8 pixeles. Cada contenedor maneja diferentes direcciones, las cuales están definidas uniformemente entre 0°-360°, lo que significa que es un gradiente identificado.

Sin embargo, cada gradiente se ve afectado por la iluminación, por lo tanto, al igual que el trabajo de las siluetas, es necesario normalizar cada celda. La normalización toma los valores de los pixeles y los escala para ocupar el mayor rango posible. El rango de cada celda toma un valor mínimo de 50 pixeles, y un valor máximo de 180 pixeles. El descriptor final es entonces un vector de todos los componentes de cada celda normalizada. [1]

Por último se estableció un rango de precisión para la ubicación correcta de 4 descriptores primordiales, los cuales están ubicados en la cabeza (Cabeza y cuello), torso (Hombros, cadera), brazos (Muñeca, codos) y piernas (Tobillo, rodillas).

Clasificación de poses

La clasificación de poses puede verse como un problema de clasificación múltiple al tener información variada de personas que ejecutan acciones, el cual puede ser modelado por una red neuronal convolucional. El sistema toma como entrada una imagen completa de 512 x 512 pixeles, el cual genera un vector que en cada una de sus posiciones representa una probabilidad de cada una de las etiquetas de actividad para las 6 categorías a analizar.

Esta red neuronal se implementó mediante Caffe[3], el cual consta de 5 capas de convolución, y 3 capas totalmente conectadas intercaladas no lineales. La primera capa convolucional tiene una profundidad de 8 × 8, con un paso de 4. La segunda capa convolucional usa filtros de 4 × 4 con un paso de 2, y las tres capas convolucionales restantes usan filtros 3 × 3 con un paso de 1. [4]

Información de entrenamiento

Se entrenó el sistema usando 150 imágenes del MPIIdataset el cual es de uso libre, este conjunto de datos contiene aproximadamente 25,000 imágenes con más de 40,000 personas. Cada imagen es extraída de un video de YouTube, y todas las imágenes tienen un tamaño de 1280 × 720 píxeles. El conjunto de datos cubre 410 categorías de actividad humana y 20 categorías generales. Cada imagen usada está etiquetada con un descriptor de actividad. Además, cada imagen está anotada con las secciones delimitadores del cuerpo y cabeza, las coordenadas x-y de cada articulación donde existe un punto clave, y una indicación de si la articulación es visible o no. Las anteriores imágenes se validaron contra 90 imágenes del dataset propio, el cual cuenta con información de la morfología local, para establecer un mayor rango de clasificación del cuerpo humano. Además, el entrenamiento para 50 iteraciones usa un tamaño para cada lote de 24. La taza de aprendizaje es del 0.001, la cual va disminuyendo un 0.1 cada 10 iteraciones.

Medidas de Evaluación

La estimación de poses se evalúa mediante la métrica de Porcentaje de partes correctas (PCP) [5]. La cual mide la tasa de detección de extremidades. Una extremidad se considera correctamente detectada si la distancia entre dos puntos de interés y sus ubicaciones están en la mitad de la longitud de la extremidad, Sin embargo, este método no es eficiente en las extremidades cortas, como la parte inferior de los brazos, que son más difíciles de detectar.

III. RESULTADOS

Se probó la clasificación de la red neuronal en un conjunto de validación de aproximadamente 300 imágenes distintas capturadas desde una vista lateral. La clasificación de este conjunto de validación en las 6 categorías seleccionadas de actividad, logra una precisión máxima de 41.64% para las acciones de persona preguntando, saludo y saludo militar. Al analizar las posibles causas se evidencio que las actividades de personas realizando una pregunta, saludo militar y saludo de manos, son similares en su modo de ejecución (Fig 5). Por lo tanto, al ubicar un descriptor e identificar cada una, el sistema tiende a tomar las acciones como una sola, la cual es la de saludo.

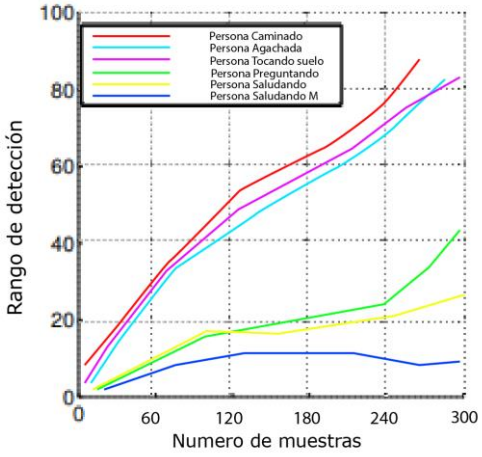

Figura 5. Detección de las diferentes acciones, desde una vista lateral.

Por tanto, el sistema se probó con un conjunto de imágenes capturadas, desde una vista frontal, en donde las acciones descritas anteriormente se pueden estudiar con mayor precisión. Se evidenció que las actividades de persona realizando una pregunta, saludo militar y saludo de manos, son detectadas con un máximo de 70.54% de precisión. Por otro lado, el desempeño de las actividades restantes como el caminado, presentan una caída de desempeño con un máximo de precisión de 81.23 (Fig 6).

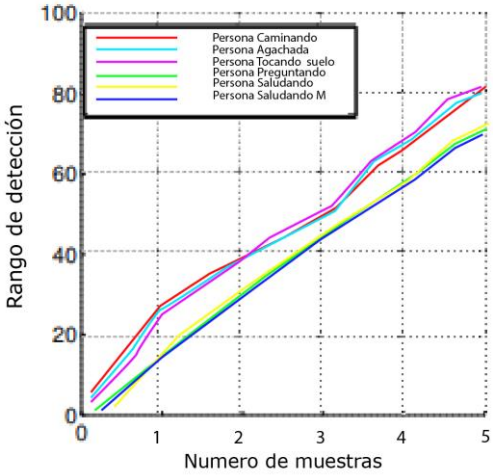

Figura 6. Evaluación de ubicación de descriptores, en un tiempo de 5 segundos.

Se estableció una comparación entre el conjunto de datos proporcionado por MPIdataset y el conjunto de datos propio, en donde se puede observar que el método propuesto, logra identificar mejor las acciones cuando se tiene un entorno controlado, ya que se omiten errores que puedan estar relacionados a la interferencia de la luz o sombras. Las acciones identificadas en un entorno controlado presentan un error máximo en la acción de saludo de 32.15%, mientras que cuando no se tiene un entorno controlado, el error máximo tiende a aumentar a un 55.78% como se muestra en la Tabla I.

TABLA I
Error de clasificación en Datasets.

Clasificación	UMNG Dataset Entorno controlado	MPII Dataset Entorno No Controlado
Persona Caminando	22,31%	32,51%
Persona Agachada	29,40%	44,87%
Persona Tocando Suelo	26,65%	39,55%
Persona Preguntando	30,87%	46,21%
Persona Saludando	28,62%	41,34%
Persona Saludando M	32.15%	55.78%

En los dos Dataset, la acción de caminado es la que mayor precisión presentó. Al identificar y analizar los videos se pudo observar que dicho movimiento es más fácil de reconocer debido a la gran cantidad de descriptores que se recrean en las extremidades inferiores del cuerpo humano, por lo tanto, en dichas secciones donde se encuentra mayor movimiento es más fácil ubicar un descriptor.

Se evaluó la ubicación de los descriptores a lo largo del tiempo, para de esta forma reunir información del movimiento humano. Se evaluó el tiempo que un descriptor permanecía ubicado en las áreas de movimiento, para establecer una comparación con cada una de las categorías, y verificar si el posicionamiento de cada descriptor se mantiene y si es el correcto.

Se pudo evidenciar que la ubicación de los cuatro descriptores primordiales alcanza un rango aceptable de precisión cerca del 83.78%, en donde los descriptores ubicados en la sección de la cabeza, se posicionan correctamente para la mayoría de las muestras. Además, se estableció un rango máximo de distancia de 0.5 pixeles de acuerdo con el diámetro del torso, para la ubicación de cada uno de los puntos de interés. De este modo se evita que descriptores sean ubicados en áreas completamente alejadas del rango morfológico establecido (Fig 7).

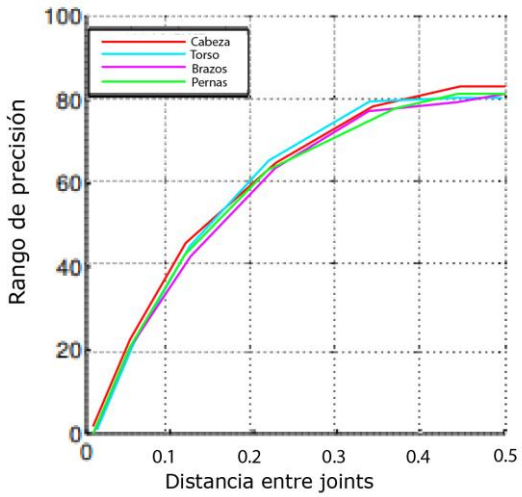

Figura 7. Resultado del posicionamiento de los descriptores. Se grafica en un rango entre [0, 0.5] de acuerdo al diámetro del torso.

La ubicación de cada descriptor alcanza una precisión máxima del 82.35% cuando se tiene una imagen fija, sin embargo, al realizar el seguimiento a lo largo del tiempo, se puede observar una variación en su posición. Algunos descriptores como el de brazos y piernas, eventualmente se pierden mediante el seguimiento, pero vuelven a su posición pocos segundos después. Por otro lado, los descriptores ubicados en la cabeza y torso presentan mayor estabilidad en medio del movimiento, esto debido a que son áreas que permanecen a lo largo del video (Fig 8).

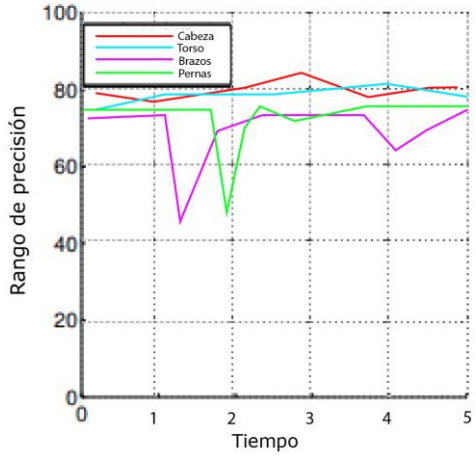

Figura 7. Resultado del posicionamiento de los descriptores. Se grafica en un rango entre [0, 0.5] de acuerdo al diámetro del torso.

IV. CONCLUSIONES

Para identificar poses en videos o secuencias de imágenes, se requiere utilizar dos perspectivas o puntos de vista perpendiculares, con el fin de evitar superposición de elementos que interfieran con la identificación y ubicación de descriptores, lo que conlleva a errores en el análisis.

Es posible la identificación de poses mediante métodos computacionales, de análisis de video, evitando el uso de sensores y trajes especiales, lo que reduce los costos de implementación de un sistema de reconocimiento.

El uso de descriptores mediante histogramas de gradientes orientados, establecidos como primordiales por áreas permite identificar los elementos ya que se mantienen en el tiempo y en mayor medida si es en un entorno controlado.

AGRADECIMENTOS

Agradecemos a nuestra colega Yessica Tatiana Leguizamon por formar parte de este proceso y contribuir con la creación del dataset propio usado en el estudio, y al equipo de estudiantes que contribuyeron con el muestreo de videos capturados en el laboratorio.

REFERENCIAS

[1] N. Dalal and B. Triggs, (2005). Histograms of oriented gradients for human detection, (pp. 886-893 vol. 1) IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA.

[2] L. Wang and D. Suter. (2007) Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model, IEEE Conference on Computer Vision and Pattern Recognition.(pp. 1-8), Minneapolis, MN.

[3] T. Alexander and S. Christian. Human Pose Estimation via Deep Neural Networks. Google.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell 2014. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093.

[5] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d 2012. articulated human pose estimation and retrieval in (almost) unconstrained still images. International Journal of Computer Vision, 99(2):190–214.

[6] Collins, R. T. Gross,R. & Shi, J.(2002). Silhouette-based Human Identification from Body Shape and Gait. "Automatic Face and Gesture Recongnition.

[7] G. Mori, X. Ren, A. Alexei and J. (2004). Malik. Recovering Human Body Configurations: Combining Segmentation and Recognition, University of Oxford.

[8] M. Hoai, Z. Z. Lan and F. De la Torre. (2011). Joint segmentation and classification of human actions in video, Computer Vision and Pattern Recognition (CVPR), (pp. 3265-3272), IEEE Conference on, Providence.

[9] L. Wang and D. Suter. (2007) Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model, IEEE Conference on Computer Vision and Pattern Recognition.(pp. 1-8), Minneapolis, MN.

[10] M. Dantone, J. Gall, C. Leistner and L. Van Gool.(2013). Human Pose Estimation Using Body Parts Dependent Joint Regressors, Computer Vision and Pattern Recognition (CVPR), (pp. 3041-3048) IEEE Conference on, Portland, OR.

[11] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal and R. Bajcsy,(2012). Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition, (pp. 8-13) IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI.

Cristian Camilo Beltrán Pérez Estudiante de ingeniería en Multimedia de la Universidad Militar Nueva Granada. Sus temas de investigación se basan en áreas como la animación, modelado 3D, procesamiento de imágenes y visión por computador.

Wilman Helioth Sánchez Rodríguez Ingeniero de Sistemas de la Universidad Nacional de Colombia. Magister en Educación y Especialista en Docencia Universitaria de la Universidad Militar Nueva Granada. Igualmente ha realizado educación no formal como Auditor Interno de Calidad y posee formación como coordinador de autoevaluación y par académico en procesos de calidad.

Docente asociado en la Facultad de Ingeniería en la UMNG sede calle 100, en las áreas de Ingeniería aplicada y básicas de Ingeniería. Maestro con más de 13 años de experiencia en asignaturas como Introducción a la Ingeniería, Programación Estructurada y Orientada a Objetos, Estructuras de Datos, Ingeniería del Software, Inteligencia Artificial, Procesamiento de Imágenes, Seminario de Investigación y Sistemas Multimedia. Director, coordinador académico, coordinador de autoevaluación y acreditación del Programa Ingeniería en Multimedia.

Eduard Leonardo Sierra Ballen Ingeniero de Sistemas de la Universidad Nacional de Colombia, Magister en Educación de la Universidad Militar Nueva Granada. Se encuentra vinculado como profesor asistente del Programa de Ingeniería en Multimedia, Universidad Militar Nueva Granada, Sede Campus Nueva Granada, Cajicá. Cofundador del Grupo de Investigación en Multimedia GIM de la misma universidad.

Docente en la Facultad de Ingeniería en la UMNG sede cajica, en las áreas de Ingeniería aplicada. Maestro en asignaturas, Programación Estructurada y Orientada a Objetos, Estructuras de Datos ,Ingenieria de Software, Procesamiento de Imágenes, Seminario de Investigación y Sistemas Multimedia. Director del Programa Ingeniería en Multimedia.