SISTEMA DE VISIÓN ARTIFICIAL PARA EL REGISTRO DE DENSIDAD PEATONAL EN TIEMPO REAL

Computer Vision System for Real-time Registry of Pedestrian Density

RESUMEN

En este trabajo se presenta el desarrollo, implementación y comprobación de un sistema de visión artificial para la detección de transeúntes bajo un ambiente controlado durante una ventana de tiempo determinada. El sistema presentado es libre de parámetros y calibración y utiliza una sola cámara para localizar y seguir en tiempo real a los transeúntes en la zona de interés, determinar su dirección y mantener un registro discriminado. En el artículo se resumen las técnicas de procesamiento de imágenes utilizadas en el desarrollo, los mecanismos de implementación, las pruebas experimentales, las limitaciones del sistema y se discuten los resultados obtenidos.

PALABRAS CLAVES: Máquina de visión, procesamiento de imágenes, densidad peatonal, conteo de personas, seguimiento visual.

ABSTRACT

In this work, we describe development, implementation, and testing of a computer vision system for passer-by detection in a controlled environment during a specified time window. Proposed system is parameter-free, calibration-free, and single-camera. System localizes and tracks in real-time a passer-by within the interest zone, determines his or her direction, and keeps a detailed record. This paper also summarize image processing techniques used in this process, implementation tools, experimental tests, system's constraints, and discuss the obtained results.

KEYWORDS: *Computer vision, digital image processing, pedestrian density, counting people, visual tracking.*

MAURICIO ABRIL CAÑAS
Ingeniero Electricista
Universidad Tecnológica de Pereira
abrilmauricio@hotmail.com

MAURICIO VALENCIA L
Ingeniero Electricista
Universidad Tecnológica de Pereira
zagat@utp.edu.co

BONIE JOHANA RESTREPO
Ingeniero Electricista
Profesor Catedrático
Universidad Tecnológica de Pereira
bonie@ohm.utp.edu.co

GERMÁN ANDRÉS HOLGUÍN
Ingeniero Electricista, M.Sc.
Profesor Asistente
Universidad Tecnológica de Pereira
gahol@ohm.utp.edu.co

1. INTRODUCCIÓN

El registro de objetos en movimiento es una herramienta importante para el diseño de aplicaciones de seguridad, control de acceso restringido, control de multitudes en espacios abiertos o cerrados, control de inventarios, mercadeo en centros comerciales y urbanos, entre otros. Teniendo en cuenta que la presencia de cámaras de seguridad en cualquier entorno urbano es un hecho [1], es natural pensar que en el mediano plazo, tengan que convertirse en parte de sistemas automáticos que permitan el procesamiento de video en tiempo real, con la finalidad de mantener un conteo discriminado de las personas que circulan en una zona de interés.

La detección específica de transeúntes es un problema de visión artificial complejo, con muchas dependencias y restricciones. El solo hecho de segmentar personas en movimiento es un reto para el que Shio y Sklansky [2] presentaron una metodología, empleando la correlación entre frames en la secuencia de video, con cierto éxito en el tratamiento de oclusiones parciales, pero con un costo computacional muy elevado como para ser viable en tiempo real. Posteriormente, se presentaron algunos trabajos, como el de Rossi y Bezzoli [3] donde se restringe el problema a una sola dirección en la escena y para una posición específica de cámara, o como el trabajo de Segen y Pingali [4], donde se utiliza análisis de contornos y siluetas, lo que mejora la detección y el seguimiento, pero a costa de la sensibilidad a las oclusiones. Algunas técnicas multicámara han resultado mejores, como en Cai y Aggarwal [5] o en Alboil, Naranjo y Mora [6], pero aún requieren una costosa infraestructura y altos niveles de parametrización. Más tarde, Choi [7], simplifica la segmentación y el seguimiento en la secuencia de video mediante predicción de movimiento, lo que permite una implementación de tiempo real.

En este artículo se propone una metodología que permita realizar en tiempo real el registro discriminado de transeúntes en un ambiente controlado, sin considerar el

Fecha de Recepción: 31 Mayo de 2007
Fecha de Aceptación: 19 Julio de 2007

problema de oclusiones, que apunte a minimizar el costo computacional y a maximizar la precisión del sistema. Se requiere además que la aplicación sea económicamente viable, por lo que se restringirá el estudio a la utilización de una sola cámara y a la arquitectura PC. Se desea que la aplicación tenga un funcionamiento autónomo, libre de parámetros de ajuste, es decir el usuario no tenga que manipular variables durante su ejecución. La estructura de este trabajo se describe a continuación: inicialmente se muestra la metodología planteada para el registro de densidad peatonal, luego se describe su implementación para finalizar con conclusiones y recomendaciones para trabajos futuros a partir de los resultados obtenidos.

2. METODOLOGÍA

La metodología propuesta está representada en el diagrama de bloques de la figura 1. La función de cada una de las etapas es detallada a continuación.

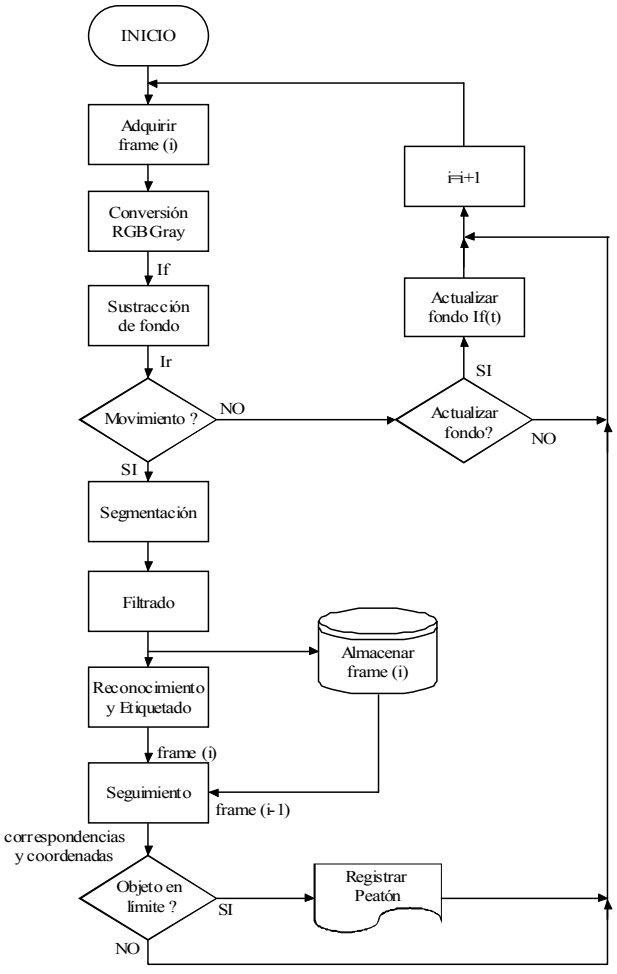

Figura 1. Diagrama de bloques.

1) Adquisición

Esta es la etapa encargada de controlar el hardware. Debe garantizar que cada frame sea adquirido a una velocidad predeterminada y constante, por lo que requiere de alta prioridad de ejecución. Todo el procesamiento posterior de cada frame, deberá ser realizado en la ventana de tiempo entre muestras, dejada por este módulo, lo que garantizará la operación del sistema en tiempo real. La salida de este módulo es una imagen en formato RGB.

Para fines experimentales, se desarrollaron dos módulos de adquisición, uno para operar directamente de la cámara y otro para recibir imagen de videos pregrabados en algún formato estándar.

2) Pre-procesamiento

Esta etapa realiza dos tareas. En primer lugar convierte la imagen de entrada a escala de grises utilizando la norma euclidiana [8] con un factor de escala de √ (1/3) que asegura que el rango de valores de cada píxel en la imagen de salida esté limitado por el rango de los valores de entrada. La ecuación (1) ilustra este procedimiento

$$f(x,y)=\sqrt{\frac{1}{3}\left(R^2+G^2+B^2\right)} \qquad (1)$$

Donde *R*, *G* y *B* son los valores de los píxeles para cada uno de los planos de color rojo, verde y azul respectivamente.

En segundo lugar, es necesario que esta etapa detecte el movimiento de transeúntes en la zona interés. Para ello, una sustracción de fondo es suficiente, ya que permite completar la tarea con muy bajo costo computacional y cierta robustez a los cambios de iluminación. La sustracción consiste en obtener la diferencia, píxel a píxel, entre el frame actual $I(t)$ y el fondo $If(t)$ escogido como referencia [8], y que debe haber sido capturado previamente al menos una vez. La captura del fondo en este experimento se realiza solo a solicitud del usuario. El resultado de la sustracción, ver ecuación (2), es una nueva imagen donde los transeúntes están representados por píxeles conectados con valores mucho mayores que cero.

$$Ir(m,n)=\left|i(m,n)-If(m,n)\right| \qquad (2)$$

3) Segmentación.

En esta etapa se busca separar a los transeúntes del fondo. Aquí, se empleó una binarización de la imagen por umbral que aprovecha bastante bien la característica bimodal de la imagen [9]. Como el objetivo es obtener un sistema libre de parámetros, dicho umbral *k*, tendrá que ser obtenido automáticamente. Para ello *k*, se obtiene mediante el agrupamiento (clustering) de los píxeles en dos grandes grupos que cumpla con la siguiente condición:

$$k = \frac{u_1 + u_2}{2} \quad (3)$$

Donde u_1 y u_2 son las máximas frecuencias de nivel de gris comprendidas entre 0 y k; y entre $k+1$ y el valor máximo (255 para 8 bits).

Para refinar el resultado de la segmentación y eliminar píxeles aislados en la imagen, se utiliza una etapa de filtrado espacial no lineal, implementada mediante la operación morfológica conocida como closing. El closing consiste básicamente en una dilatación seguida de una operación de erosión [10].

4) Seguimiento en la secuencia de imágenes

En la imagen resultado de la etapa anterior, el fondo está representado por píxeles de valor cero. Por tanto, toda área de píxeles conectados con valor diferente de cero y que esté comprendida entre 4000 y 7000 píxeles ^{2} será considerada transeúnte. Este rango fue determinado únicamente en función de la altura de ubicación de la cámara, y es estático, por lo que no representa un parámetro configurable.

A todo objeto reconocido como transeúnte se le asigna una etiqueta en el frame actual. Para cada transeúnte, se calcula el momento de inercia [11], utilizando la expresión:

$$G_{KO} = \sum_{i=1}^{m}\sum_{j=1}^{n} G_K(i,j) \quad (4)$$

Donde $G_k(i,j)$ es el conjunto de valores umbralizados, m y n son el número de filas y columnas. Con el momento de inercia es posible determinar las coordenadas de los centroides [11], utilizando:

$$\hat{\delta}_x(k) = \frac{1}{G_{KO}}\sum_{i=1}^{m}\sum_{j=1}^{n} x_j G_K(i,j)$$

$$\hat{\delta}_y(k) = \frac{1}{G_{KO}}\sum_{i=1}^{m}\sum_{j=1}^{n} y_j G_K(i,j) \quad (5)$$

Este cálculo se realiza para cada transeúnte en cada frame. Luego, la tarea de seguimiento consiste en determinar la correspondencia de etiquetas a un mismo transeúnte en los frames *i* e *i-1*. Para esto, es válido suponer que la posición estimada de transeúnte en el nuevo frame depende de la velocidad y dirección con que se mueve, y sin pérdida de generalidad se puede suponer que dicha posición estimada es un valor viable dentro de la imagen [7]. Por lo tanto, las velocidades actual y estimada hacen parte de un conjunto convexo que puede ser representado por:

$$\hat{v}(t+1) = (1-a)v(t) + a\hat{v}(t) \quad (6)$$

Donde *a* es un coeficiente de amortiguamiento que depende de la altura de ubicación de la cámara. La velocidad *v* y el desplazamiento *d*, tanto actual como estimada, están relacionadas mediante la tasa de muestreo dada en frames por segundo. El siguiente diagrama ilustra la predicción de movimiento.

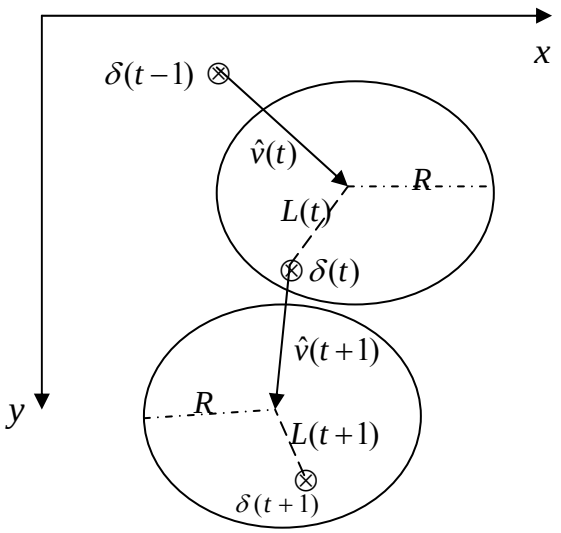

Figura 2. Diagrama del seguimiento mediante predicción de movimiento

Donde δ representa las coordenadas del centroide y $\hat{v}(t+1)$ es el vector velocidad estimado. R es la distancia en píxeles mínima necesaria para identificar a un mismo transeúnte en dos frames consecutivos, tal que $\delta(t)$ y $\delta(t+1)$ corresponden al mismo transeúnte si y solo si $R > L$.

Se considera que un transeúnte debe ser contabilizado, cuando ha atravesado el área de conteo de arriba abajo o viceversa.

3. IMPLEMENTACIÓN Y RESULTADOS

Para la implementación del sistema se utilizó un PC con un procesador Intel Pentium IV de 2.8 GHz y 512 MB de RAM corriendo Windows XP de Microsoft. La aplicación se desarrolló en LabVIEW 8.0 de National Instruments. La figura 3, muestra el panel frontal de la aplicación.

La cámara utilizada fue una cámara web convencional, Video BLASTER WEBCAM Plus de Creative, de 24 bits con salida RGB de 320x240 y una frecuencia de muestreo máxima de 30 fps. La cámara fue ubicada a 3.5

metros del piso en posición acimutal, obteniendo un área de conteo de 9 m^2.

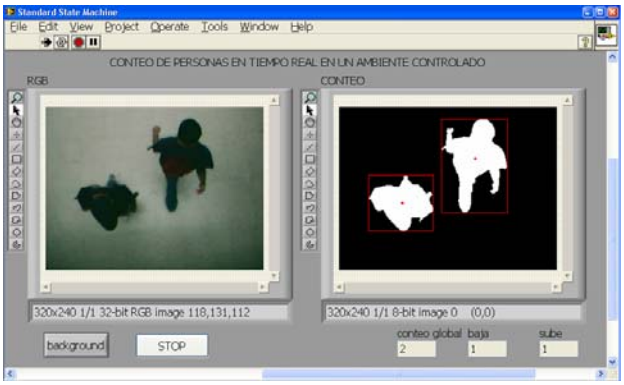

Figura 3. Panel Frontal de la aplicación

La posición y disposición de cámaras y las características del área de conteo varían según la aplicación, esto lleva a que no exista una base de datos estándar diseñada para la comprobación de sistemas de visión artificial para el conteo de personas; por tanto se creó una propia. Esta base de datos está conformada por 50 videos etiquetados según el número de transeúntes presentes en la escena, según la dirección de cruce del área de conteo y algunos casos especiales, tales como: oclusiones leves, oclusiones severas, peatones detenidos, cambios de iluminación, entre otros.

En el momento de cuantificar el rendimiento y confiabilidad del sistema se optó por utilizar una matriz de confusión donde se definen las siguientes salidas del sistema.

- FP: Falso Positivo. Grupo de píxeles conectados que son detectados como movimiento, pero que no corresponden a ningún transeúnte.
- FN: Falso Negativo. Grupo de píxeles conectados que son detectados como movimiento, pero no son identificados por el sistema como transeúntes.
- VP: Verdadero Positivo. Grupo de píxeles conectados que son detectados como movimiento, que son identificados como transeúntes por el sistema.
- VN: Verdadero Negativo. Grupo de píxeles no conectados, que no son detectados por el sistema como Transeúntes.

Se realizaron las mismas pruebas para diferentes velocidades de muestreo: 7.5, 15 y 30 fps. Para 7.5 fps no fue posible garantizar la operación en tiempo real con el hardware utilizado. Para 15 y 30 fps se obtuvieron resultados estadísticamente equivalentes, por lo que se presentan los resultados de 15 fps.

Dirección	VP	FP	FN	Real
Arriba	39	1	0	39
Abajo	40	0	2	42

Tabla 1. Matriz de confusión para 15 fps.

Estos resultados, fueron obtenidos para una altura de cámara de 3.5 metros, por lo que los valores de a y R fueron 0.3 y 60 respectivamente. El número máximo de peatones en el área de conteo es 4.

4. CONCLUSIONES

Este artículo presenta la metodología utilizada en el diseño, implementación y prueba de un sistema de visión artificial concebido para el conteo de transeúntes en una región de interés determinada sin tener en cuenta oclusiones ni traslapes.

El sistema propuesto es libre de parámetros en tiempo de ejecución. Es decir, el usuario no requiere estimar ni calcular ningún parámetro de funcionamiento una vez se ha completado la instalación del sistema.

Se concluye que 15 fps es una velocidad de muestreo óptima para el sistema, ya que los resultados obtenidos con velocidades mayores son estadísticamente equivalentes con mayores costos computacionales y con velocidades de muestreo menores se presentan pérdidas de información.

El seguimiento de los objetivos reconocidos en la secuencia de vídeo (Tracking), por medio de la predicción de movimiento de sus centroides es útil dado que, los transeúntes no presentan un patrón igual entre frame y frame, que permita reconocerlo y asociarlo en cada una de las imágenes el mismo objeto a una etiqueta.

La metodología fue probada en recintos cerrados donde el fondo es estático y su mayor limitación en su utilización práctica radica en la identificación de transeúntes ante oclusiones severas. Este es un problema de investigación que sigue abierto como continuación de este trabajo; para darle solución se debe pensar en una metodología de solución que involucre técnicas más robustas pero computacionalmente más costosas, tanto en la sustracción del fondo, utilizando un modelamiento dinámico; como en la segmentación, con técnicas que involucren jerarquía en color ó visión estereo; y en la etapa de seguimiento, con técnicas basadas en colores o en modelos interframes.

5. BIBLIOGRAFÍA

[1] C. Sandoval. “Seguimiento de objetos en secuencias de vídeo”. Dept. Comunicaciones. Univ. Politécnica de Valencia, 2003.

[2] A. Shio and J. Sklansky, "Segmentation of people in motion," Proc. IEEE Workshop Visual Motion, pp. 325–332, Oct. 1991.
[3] M. Rossi and A. Bozzoli, "Tracking and counting moving people," Proc.2nd IEEE Int. Conf. Image Processing, pp. 212–216, Nov. 1994.
[4] J. Segen and S. Pingali, "A camera-based system for tracking people in real time," in Proc. IEEE 13th Int. Conf. Pattern Recognition, Los Alamitos, CA, pp. 63–67 Aug. 1996.
[5] Q. Cai and J. K. Aggarwal, "Tracking human motion using multiple cameras," in Proc. IEEE 13th Int. Conf. Pattern Recognition, Los Alamitos, CA, pp. 68–72, Aug. 1996
[6] A. Albiol, V. Naranjo. I, Mora,. "Real-Time Hight Density People Counter using Morphological Tools". IEEE Trans. Intelligent Transportation Systems, vol. 2, pp. 201-218, Dic. 2001.
[7] J. Kim, K. Choi "Real-Time vision-based people counting system for the security door". Department of Electronics Engineering, Korea University, 2002.
[8] N. Siebel, "Design and implementation of people tracking algorithms for visual suveirllance aplications," Marzo 2003.
[9] National Instrument. "NI vision concepts manual". November 2005.
[10] G. Pajares, J. D. la Cruz, Visión por computador Imágenes Digitales y Aplicaciones. España: Rama, 2001.
[11] D. C. P. Albores, "Seguimiento de objetos por medio de visión activa," Ph.D. dissertation, Instituto Nacional de Astrofísica, Óptica y Electrónica, Mayo 2002.
[12] A. Bovik, Handbook of Image and Video Processing. Austin, Texas: University of Texas, 2000.
[13] M. Petrou and P. Bosdogianni, Image Processing The Fundamentals. England: John Whiley and Sons, 1999.
[14] G. A. Holguín, S. Pérez, A. Orozco, Curso Básico LabVIEW 6i, Universidad Tecnológica de Pereira, Facultad de Ingeniería Eléctrica, 2002.
[15] G. Betancourt, G. Holguín, A. Orozco, Libro electrónico de Medidas, Universidad Tecnológica de Pereira, Facultad de Ingeniería Eléctrica, 2005.