DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico Nº 2

Segundo cuatrimestre de 2015 — Organización del Computador II

Grupo: Smelly Cat

Integrante	LU	Correo electrónico
Frizzo, Franco	013/14	francofrizzo@gmail.com
Martínez, Manuela	160/14	martinez.manuela.22@gmail.com
Rabinowicz, Lucía	105/14	lu.rabinowicz@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

1. Introducción

En el presente trabajo, se aplica el modelo de programación vectorial SIMD (*Single Instruction, Multiple Data*) para la implementación de filtros para el procesamiento de imágenes. Más precisamente, se lleva a cabo la implementación de los siguientes dos filtros:

- *Diferencia* (*diff*), que recibe como entrada dos imágenes y devuelve como resultado otra imagen que indica dónde difieren las dos primeras.
- *Blur gaussiano* (*blur*), que suaviza la imagen reemplazando cada píxel por un promedio de los píxeles circundantes, ponderado según una función gaussiana.

La elaboración del trabajo se dividió en dos etapas. En primer lugar, se implementaron ambos filtros tanto en lenguaje C como en lenguaje ensamblador para la arquitectura x86 de Intel. En este último caso, se utilizaron las instrucciones SSE de dicha arquitectura, que aprovechan el ya mencionado modelo SIMD para procesar datos en forma paralela.

Una vez realizadas estas implementaciones, fueron sometidas a un proceso de comparación para extraer conclusiones acerca de su rendimiento. Con este fin, se experimentó con variaciones tanto en los datos de entrada como en detalles implementativos de los mismos algoritmos. De esta manera, se pudo recopilar datos sobre el comportamiento de cada implementación, y contrastar estos resultados con diversas hipótesis previamente elaboradas.

2. Desarrollo

2.1. Consideraciones generales

Las imagenes que se utilizan como entrada y salida de los algoritmos a implementar son matrices de píxeles. Cada uno de estos píxeles está representado por cuatro enteros sin signo de 8 bits de profundidad (es decir, en el rango $[0, 256)$), que contienen, respectivamente, los valores de los colores azul (b), verde (g) y rojo (r), y la transparencia (a).

Se usará la notación $I_{x,y}$ para referirse al píxel ubicado en la fila x y la columna y de la imagen I, y la notación $I^k_{x,y}$ para hacer referencia al valor de la componente k de este píxel, donde $k \in \{\mathsf{b}, \mathsf{g}, \mathsf{r}, \mathsf{a}\}$.

2.2. Diferencia de imágenes

Este filtro recibe dos imágenes como entrada y devuelve como salida una tercera imagen que muestra, en cada píxel, la diferencia entre los píxeles correspondientes de las imágenes de entrada, ignorando la componente a. Más especificamente, si I_1 e I_2 son las imágenes de entrada y O es la imagen de salida, entonces:

$$O^k_{x,y} = \begin{cases} \max\limits_{k \in \{\mathsf{b},\mathsf{g},\mathsf{r}\}} \left(\left| {I_1}^k_{x,y} - {I_2}^k_{x,y} \right| \right) & \text{si } k \in \{\mathsf{b}, \mathsf{g}, \mathsf{r}\} \\ 255 & \text{si } k = \mathsf{a} \end{cases}$$

2.2.1. Implementación en lenguaje C

La implementación de este filtro en lenguaje C es sumamente sencilla. Ambas imágenes de entrada se recorren simultáneamente mediante dos ciclos anidados, que iteran sobre sus filas y sus columnas, respectivamente. Para las componentes `b`, `g` y `r` de cada píxel, se calcula el valor absoluto de la diferencia entre los valores correspondientes a ambas imágenes. Luego, se computa el máximo entre estos tres valores, que se guarda como el valor de las componentes `b`, `g` y `r` para el píxel correspondiente en la imagen de salida. Por último, la componente `a` de dicho píxel se define como 255.

2.2.2. Implementación en lenguaje ensamblador

Al implementar el filtro en lenguaje ensamblador, es posible aprovechar las ventajas que brinda el modelo SIMD. En particular, dado que los registros `XMM` son de 16 bytes, se los puede utilizar para procesar 4 píxeles de las imágenes en paralelo, reduciendo la cantidad de iteraciones del algoritmo y, particularmente, de accesos a memoria necesarios para completar el algoritmo.

La implementación en este lenguaje del filtro consiste principalmente de un ciclo que itera sobre la imagen. Al comienzo de cada ejecución, se copian 4 píxeles de I_1 al registro `XMM0`, y los correspondientes 4 píxeles de I_2 a `XMM1`.

`XMM0:`	A_4^{b}	A_4^{g}	A_4^{r}	A_4^{a}	A_3^{b}	$\cdots$	A_1^{a}
`XMM1:`	B_4^{b}	B_4^{g}	B_4^{r}	B_4^{a}	B_3^{b}	$\cdots$	B_1^{a}

El paso siguiente consiste en calcular, para cada una de las componentes de estos píxeles, el valor absoluto de la diferencia entre ambas imágenes. Para realizar esto, se realiza la resta de las dos maneras posibles, obtenie `XMM0` $=$ `XMM0` $-$ `XMM1` y `XMM1` $=$ `XMM1` $-$ `XMM0`. En las posiciones donde el valor contenido en `XMM0` sea mayor que el de `XMM1`, será válido el resultado de la primera operación, mientras que en las demás posiciones se deberá tener en cuenta el segundo resultado.

Para seleccionar cuál de los dos resultados es el correcto, se utiliza una máscara que se obtiene comparando los valores de `XMM0` y `XMM1`. Aquí aparece un problema, ya que debemos comparar enteros sin signo, y SSE no brinda instrucciones para hacer esto. Es por eso que se recurre a desempaquetar los números y considerarlos como enteros con signo de dos bytes, que sí se pueden comparar. Empaquetando luego el resultado obtenido, se logra la máscara buscada. Esta se aplica mediante `PAND` al valor contenido en `XMM0` y mediante `PANDN` al valor presente en `XMM1`. Posteriormente se computa un `POR` entre los valores recién calculados, llegando al valor deseado, que se almacena en `XMM0`.

Por último, es necesario calcular la norma infinito de las componentes `b`, `g` y `r` obtenidas. Para hacer esto, se copia en `XMM1` el último resultado y se lo desplaza un byte hacia la izquierda.

`XMM0:`	C_4^{b}	C_4^{g}	C_4^{r}	C_4^{a}	C_3^{b}	C_3^{g}	$\cdots$	C_1^{a}
`XMM1:`	0	C_4^{b}	C_4^{g}	C_4^{r}	C_4^{a}	C_3^{b}	$\cdots$	C_1^{a}

A continuación puede usarse la instrucción `PMAXUB XMM1, XMM0` para calcular el máximo entre estos dos registros, obteniendo

XMM1:	*	máx($C_4^{\mathsf{b}}, C_4^{\mathsf{g}}$)	*	*	⋯	*	máx($C_1^{\mathsf{b}}, C_1^{\mathsf{g}}$)	*	*

Repitiendo el proceso anterior, pero esta vez almacenando el resultado en `XMM0`, se obtiene

XMM0:	*	*	máx($C_4^{\mathsf{b}}, C_4^{\mathsf{g}}, C_4^{\mathsf{r}}$)	*	⋯	*	*	máx($C_4^{\mathsf{b}}, C_4^{\mathsf{g}}, C_4^{\mathsf{r}}$)	*

A partir de aquí, utilizando la instrucción `PSHUFB` para replicar el valor calculado, se almacena este máximo en las componentes r, g y b de los píxeles correspondientes de la imagen de destino, y utilizando una máscara adecuada, se define la componente a de todos ellos como 255.

2.3. Blur gaussiano

Este filtro recibe una imagen como entrada y devuelve como salida el resultado de aplicarle una convolución[1] con una función gaussiana, que dependerá de un parámetro σ que podrá ser modificado. Dada la naturaleza del problema, trabajaremos con una convolución discreta en dos dimensiones, y como nuestro poder de cómputo es limitado, procesaremos solo una vecindad acotada de cada píxel, cuyo radio quedará determinado por un parámetro configurable r. En definitiva, el resultado del filtro será

$$O_{x,y}^k = \begin{cases} \displaystyle\sum_{i=-r}^{r} \sum_{j=-r}^{r} O_{x+i,y+j}^k K_{r-i,r-j} & \text{si } k \in \{\mathsf{b}, \mathsf{g}, \mathsf{r}\} \\ 255 & \text{si } k = \mathsf{a} \end{cases}$$

donde K es la matriz o *kernel* de la convolución, con

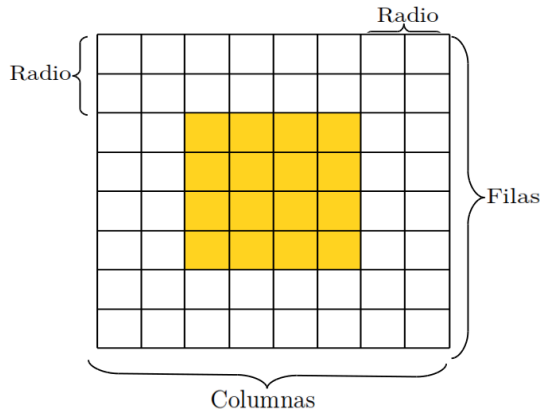
$$K_{i,j} = \frac{1}{2\pi\sigma^2} e^{-\frac{(r-i)^2+(r-j)^2}{2\sigma^2}} \qquad \text{para todo } 0 \leq i, j \leq 2r$$

2.3.1. Implementación en lenguaje C

Para aplicar el filtro, es necesario que el parámetro r sea menor que la mitad de la cantidad de filas y menor que la mitad de la cantidad de columnas. Por esto, verificamos que el parámetro cumpla con estas condiciones. El siguiente paso consiste en crear la matriz de convolución del filtro llamando a una función auxiliar, explicada más adelante.

Luego, mediante dos ciclos, se recorre toda la parte de la imagen original a la que es posible aplicarle el filtro. Esta parte es la que contempla las filas desde el valor del radio hasta $filas - (r+1)$, y las columnas desde el valor del radio hasta $columnas - (r+1)$. El resto de la imagen no será afectada.

[1] Dadas dos funciones f y g, una *convolución* $f * g$ es una operación que las transforma en una tercera función: $(f*g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t-\tau)\,d\tau$ en el caso continuo, $(f*g)_n = \sum_{k=-\infty}^{+\infty} f_k g_{n-k}$ $(n, k \in \mathbb{Z})$ en el caso discreto.

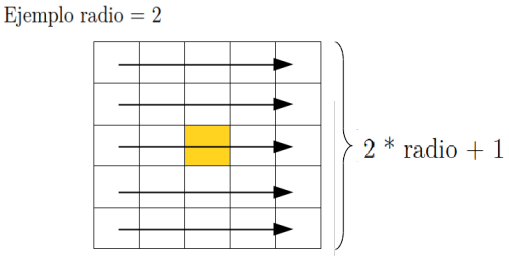


En cada paso de esta iteración, se llama a la función `afectarPixel`, que se ocupa de modificar cada píxel correctamente, utilizando la matriz de convolución creada anteriormente.

Función `matrizDeConvolucion`

En esta parte del algoritmo se calcula la matriz de la convolución. Para ello, en primer lugar, se piden $4 \times (2r+1)^2$ bytes de memoria, lugar que va a ocupar la matriz ya que su altura es $2r+1$ y su ancho, $4 \times (2r1)$ (porque cada píxel ocupa 4 bytes).

Se utilizan dos ciclos; ambos van desde 0 hasta $2r+1$.



En cada paso, se calcula la función gaussiana con el σ, r, i y j correspondientes, donde i representa la fila y j la columna. Luego, se coloca el resultado en la fila i, columna j de la matriz. Finalmente, se devuelve un puntero a la primera posición de la matriz.

Función `afectarPixel`

Primero se inicializan 3 variables donde luego se van a almacenar las sumas que corresponden a cada componente del píxel a afectar (`b`, `g` y `r`).

Seguidamente, se utilizan dos ciclos para recorrer la matriz de convolución y la parte de la imagen correspondiente a los vecinos del píxel a afectar.

Estos van desde 0 hasta $(2r)$, para recorrer las filas, y desde 0 hasta $(2r \times 4)$ para recorrer las columnas. En cada paso se multiplica el valor de las componentes del píxel observado por el valor de la matriz de convolución. El resultado de esta multiplicación se suma en las 3 variables creadas en el principio.

Finalmente, se copia el valor de cada variable en cada componente del píxel en la imagen destino. Luego, en la componente `a` se coloca el valor 255.

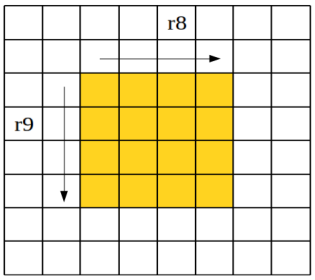
2.3.2. Implementación en lenguaje ensamblador

Este algoritmo se ocupa de recorrer toda la porción de la imagen a la que es posible aplicarle el filtro.

Al igual que en la implementación en C, primero se hace una comparación para revisar si el radio es válido.

Luego, utilizando la instrucción `CALL`, se hace un llamado a la función `matrizDeConvolución` (implementada en C), la cual devuelve un puntero a la matriz de convolución creada.

Posteriormente, se utilizan dos ciclos para recorrer la porción a modificar de la imagen. Estos utilizan los registros `R8` para recorrer las columnas y `R9` para recorrer las filas.

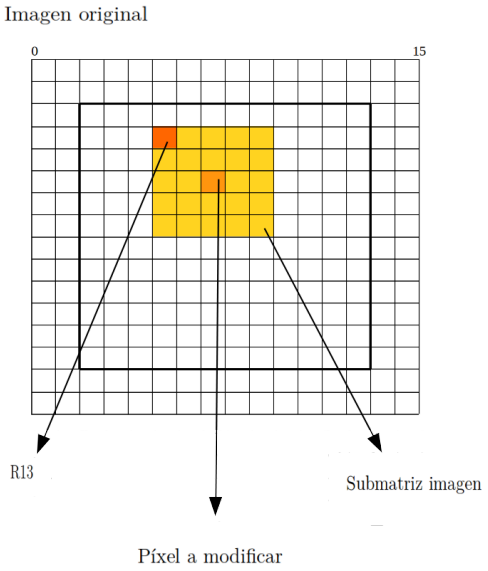


Radio = 2

En cada paso del ciclo, utilizando nuevamente la instrucción `CALL`, se realiza un llamado a la función afectarPixel (implementada en assembler) que se ocupa de modificar el píxel correspondiente.

Función `afectarPixel` El primer paso de este algoritmo es encontrar el puntero al píxel que se debe afectar en la imagen original y otro puntero al mismo píxel pero en la imagen destino. Estos punteros son guardados en los registros `R12` y `R14`, respectivamente.

Llamaremos *submatriz imagen* a la porción de la imagen original que se debe utilizar para que, junto con la matriz de convolución, se obtenga el nuevo valor del píxel. Esta submatriz es la que contiene al píxel a modificar en el centro y su tamaño es el mismo que el de la matriz de convolución.



Ejemplo Radio = 2

Luego, se debe encontrar el puntero al primer píxel de la submatriz imagen. Esto se calcula de la siguiente manera: `R12` $- 4 \times (r - r \times columnas)$. Este puntero se encuentra en el registro r13 y es utilizado para recorrer la submatriz.

En el registro `R11` se guarda la cantidad total de píxeles que componen la submatriz imagen.

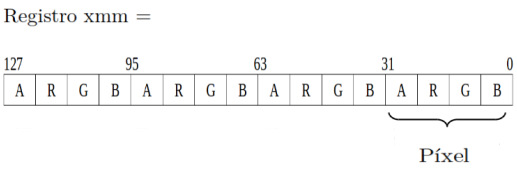
`R11` $= (2 \times r + 1)^2$

El registro `R15` contiene el tamaño de las filas.

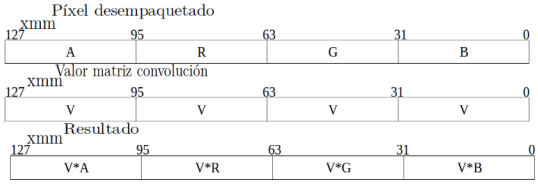
`R15` $= (2 \times r + 1)$

Estos últimos dos se utilizan como registros contadores.

A continuación, se recorren por filas la submatriz imagen y la matriz de convolución a la vez. `R11` se utiliza para saber cuando finalizar el bucle, es decir, cuando ya se observaron todos los píxeles. Estos son procesados utilizando instrucciones SSE y registros del tipo `XMM`. En cada uno de estos registros es posible guardar 4 píxeles, ya que cada píxel ocupa 4 bytes.



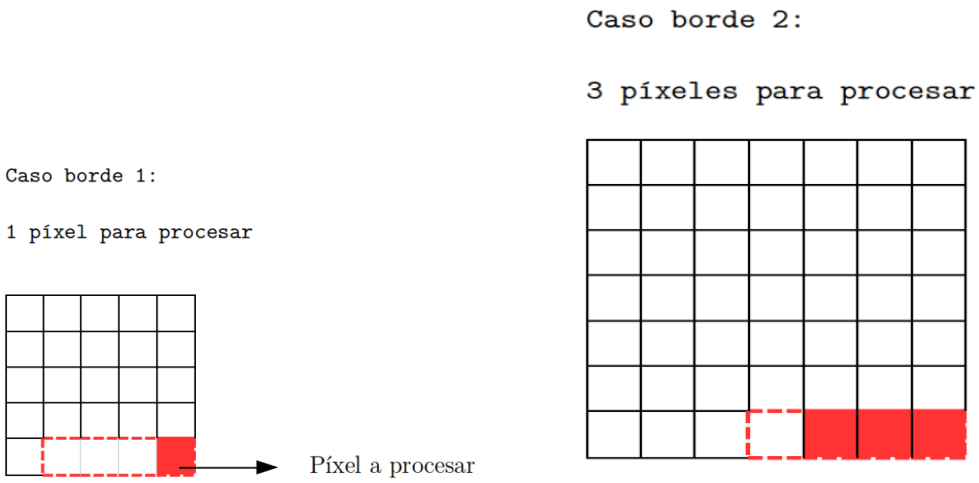
En cada paso de este ciclo se considera una fila, utilizando `R15` para saber cuándo termina. Se toman 4 píxeles de la submatriz imagen y los correspondientes 4 de la matriz de convolución. Luego de desempaquetar los píxeles se realiza la multiplicación de cada componente (`b`, `g` o `r`) con el valor de la matriz de convolución que le corresponde. El resultado de cada producto se suma al registro `XMM6`.



Se puede notar que el tamaño de las filas es congruente a 1 o 3 módulo 4. Por lo tanto, se procesan de a 4 píxeles hasta llegar a alguno de los dos casos borde posibles: cuando queda 1 píxel por computar o cuando quedan 3.

La solución a este problema es tomar y desempaquetar los píxeles de la submatriz imagen que todavía no fueron procesados junto con sus siguientes, hasta completar 4. Esto es así en todos los casos, excepto en el que los píxeles restantes son los últimos de la submatriz imagen (es decir, los que se corresponden con la última fila). En este caso, se toman los píxeles anteriores, ya que de otra forma, si el píxel a modificar se encontrase en alguna de las 3 últimas posiciones se accedería a posiciones de memoria fuera de la imagen. Estos píxeles no se tienen en cuenta a la hora de realizar el cálculo.

Esto mismo se realiza para la matriz de convolución.



Al finalizar el ciclo, se tiene en `XMM6` el valor esperado para cada componente (`b`, `g` y `r`) del píxel a afectar, en punto flotante de 32 bits. Primero se pasan estos valores a enteros; después, en la componente `a` se coloca un 255, y luego se empaqueta `XMM6` para copiarlo en la posición del píxel que se quiere modificar en la imagen destino.

3. Experimentación

A continuación, se presentan los cuatro experimentos que se realizaron. Todas las instancias de experimentación se ejecutaron 20 veces, calculando luego el promedio de los valores obtenidos. Con el código del trabajo práctico se incluye una serie de scripts de *bash* que permiten recrear los experimentos realizados, como así también los gráficos que se incluyen en este informe; esto puede hacerse ingresando al directorio `exp` dentro de la raíz, y ejecutando el comando `./exp -n 20`.

3.1. Experimento 1

En el primer experimento se genera una serie de imágenes de diferentes tamaños, tomando una imagen grande y disminuyendo progresivamente sus dimensiones. Luego, se ejecuta el filtro *blur* con cada una de las imágenes generadas y se compara el tiempo de ejecución de las implementaciones en C y lenguaje ensamblador. Esto se repite para el filtro *diff*, con la diferencia de que para cada tamaño de imagen se genera un par de imágenes con ciertas diferencias entre ellas, para poder verificar el buen funcionamiento del mismo.

Hipótesis

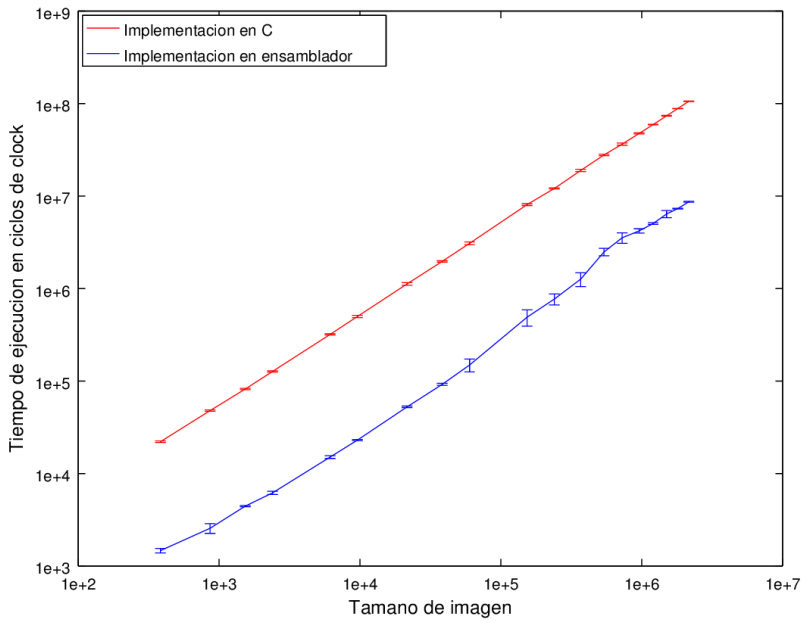
Se espera observar que la implementación en lenguaje ensamblador de ambos filtros sea más eficiente, independientemente del tamaño de la imagen. Esto se debe a que hacen uso del modelo SIMD, con todas las ventajas ya mencionadas que esto tiene sobre el rendimiento del código, a diferencia de las implementaciones de los algoritmos en C, que procesan cada píxel de manera independiente. Esto último puede inferirse no solo de la estructura propia del código, cuyos ciclos iteran sobre un único píxel a la vez, sino también de la ausencia de instrucciones SEE para procesamiento de valores empaquetados que se observa al desensamblar los objetos obtenidos a partir de este código.

Valores utilizados como parámetros

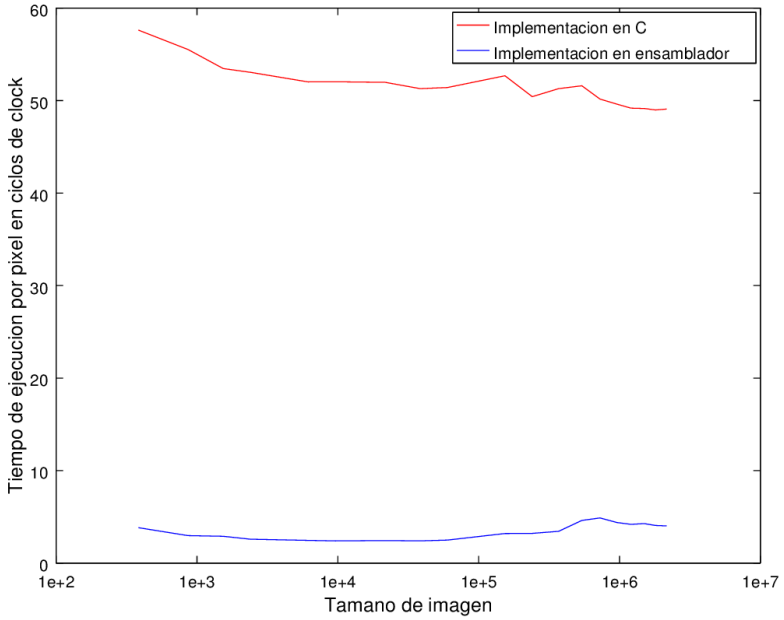
En este experimento el ancho de las imágenes utilizadas como parámetro se encuentran en un rango entre 24 y 1800 píxeles.Además, para el filtro *blur* se utilizó $r = 15$ y $\sigma = 5$.

Resultados

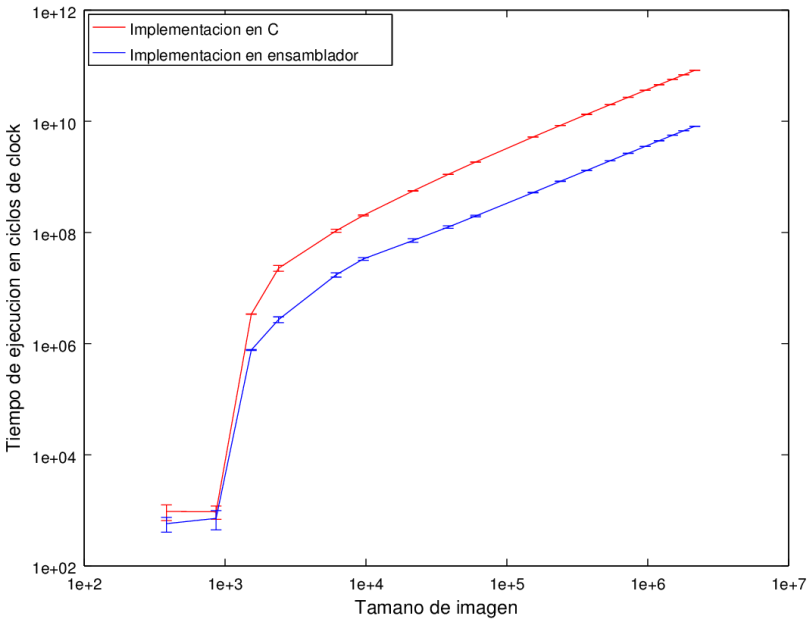
Filtro *diff*



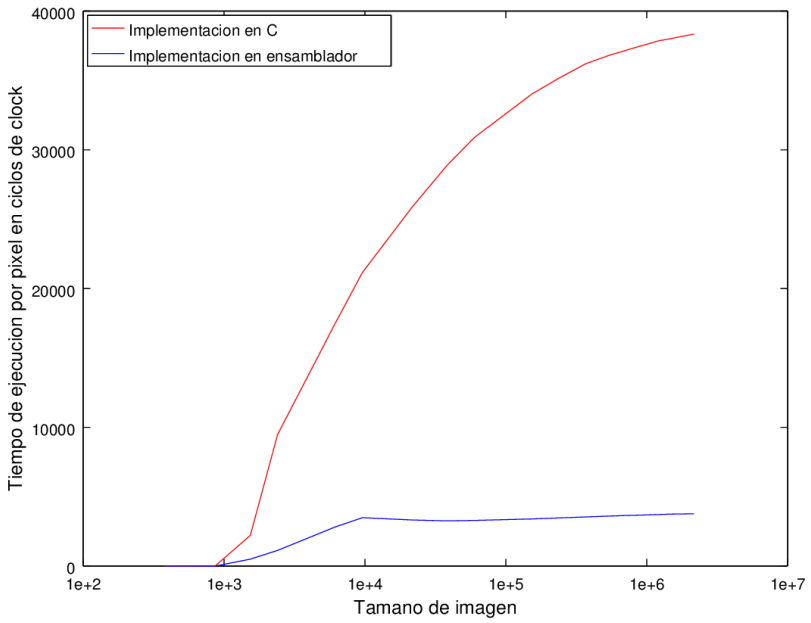
Filtro *diff* - Tiempo de ejecución normalizado por píxel



Filtro *blur*



Filtro *blur* - Tiempo de ejecución normalizado por píxel



Conclusiones y observaciones

Como se observa en los resultados, se pudo confirmar la hipótesis planteada: la implementación en lenguaje ensamblador resultó más rápida que la implementación en C para todos los tamaños de imagen. En *blur*, cuando llamamos a la función con un valor de r mayor a la mitad de la altura o a la mitad del ancho de la imagen, no se producen cambios. Dado que en el experimento el valor de r se mantiene constante, las dos imágenes más pequeñas no se ven afectadas por el filtro, lo cual se ve reflejado en los resultados, ya que para estas dos imágenes el tiempo de ejecución es notablemente menor.

3.2. Experimento 2

El objetivo de este experimento es observar como se ve afectada la eficiencia del algoritmo *blur*, en ambas implementaciones, para diferentes valores del parámetro r manteniendo constante la imagen de entrada.

Hipótesis

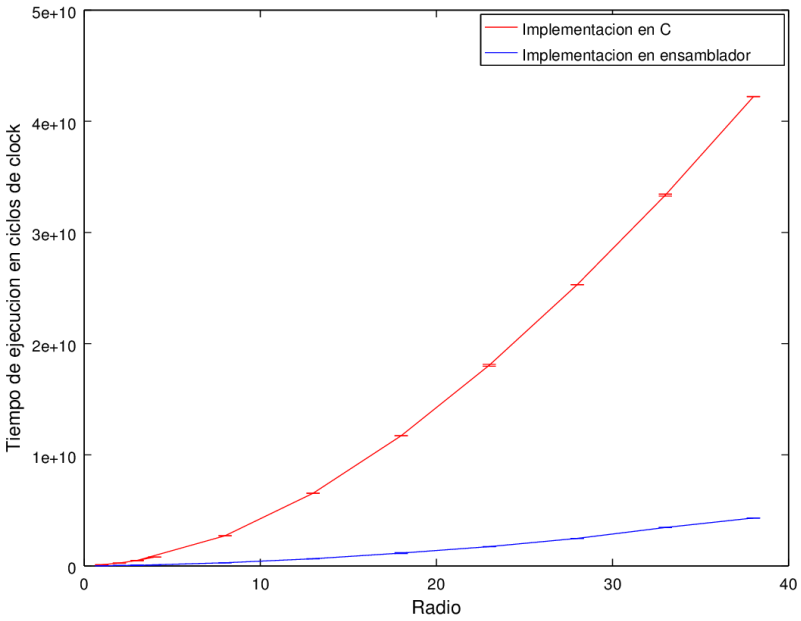
Se conjetura que, a medida que el valor del radio r se incrementa, el tiempo de ejecución en las dos implementaciones aumentará, y que lo hará de manera cuadrática con respecto al incremento en r. Esto se debe a que la complejidad temporal de cada ejecución del ciclo principal del algoritmo depende del tamaño de la matriz de convolución, que es $(2r + 1) \times (2r + 1) \times 4$, es decir, es cuadrático en el valor de r.

Valores utilizados como parámetros

La dimensión de la imagen utilizada es 400 filas y 600 columnas. El valor del sigma es 5 y los radios toman valores entre 1 y 40.

Resultados

Filtro *blur* - Tiempo de ejecución según radio



Conclusiones y observaciones

Se puede observar en los gráficos que a medida que los r aumenta, también lo hace el tiempo de ejecución. En este sentido, se pudo confirmar la hipótesis. Sin embargo, si se dividen los valores del tiempo de ejecución por su correspondiente r^2, se comprueba que la relación no es lineal; es decir, el tiempo de ejecución no varía cuadráticamente con el valor de r.

3.3. Experimento 3

Este experimento es similar al anterior; también se realiza sobre las dos implementaciones del filtro *blur* y se considera siempre la misma imagen. En este caso el valor de r se mantiene constante pero el de σ se modifica.

Hipótesis

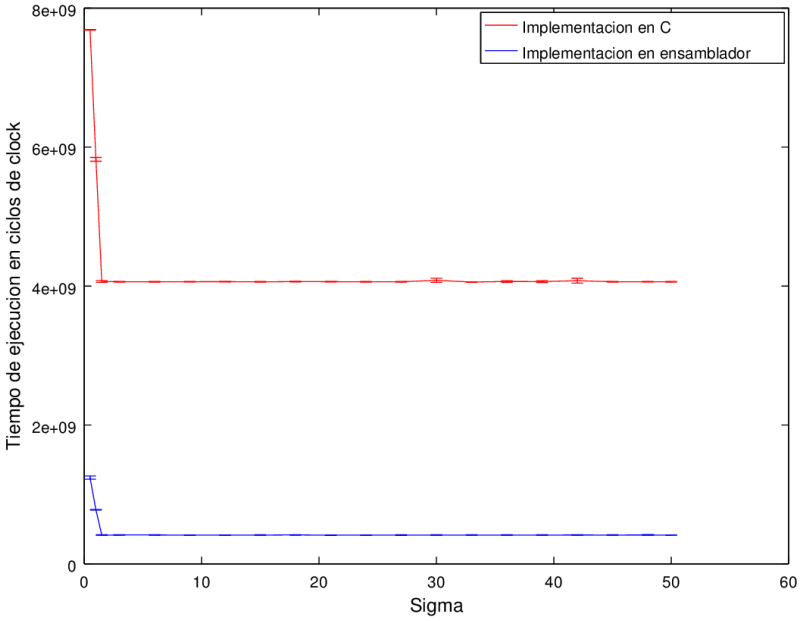
Debido a que el valor del sigma es utilizado solamente para realizar un cálculo por cada posición de la matriz de convolución, se estima que modificar este valor no alterará el tiempo de ejecución.

Valores utilizados como parámetros

La dimensión de la imagen utilizada es de 400 filas y 600 columnas. El valor del r es 10, y σ toma valores entre 0.5 y 50.

Resultados

Filtro *blur* - Tiempo de ejecución según sigma



Conclusiones y observaciones

Como se había previsto en la hipótesis, la variación del σ no afecta el tiempo de ejecución del algoritmo, tanto en lenguaje ensamblador como en C. Puede observarse que, para valores de σ menores que 1, el tiempo de ejecución es notoriamente mayor. Este es un comportamiento inesperado, que sería interesante estudiar posteriormente.

3.4. Experimento 4

Otras de las pruebas consiste en comparar los tiempos de ejecución de diferentes implementaciones de los filtros en lenguaje ensamblador. Nos interesa medir el peso que tienen en el tiempo de ejecución los llamados a funciones auxiliares. Para esto, queremos comparar el rendimiento de una implementación que utiliza llamados a estas funciones, con el de otra que tiene todas las instrucciones necesarias en el mismo bloque de código (sin utilizar esas funciones auxiliares). En particular, se consideraron la versión en lenguaje C de *diff*, a la que se le reemplazaron macros de preprocesador utilizadas para realizar operaciones aritméticas por llamados a función, y la implementación en ensamblador de *blur*, con la que se hizo el proceso opuesto, eliminando los llamados a funciones auxiliares y colocando todo el código directamente en el cuerpo de la función principal.

Este experimento se realiza una determinada cantidad de veces con distintos tamaños de imagen.

Hipótesis

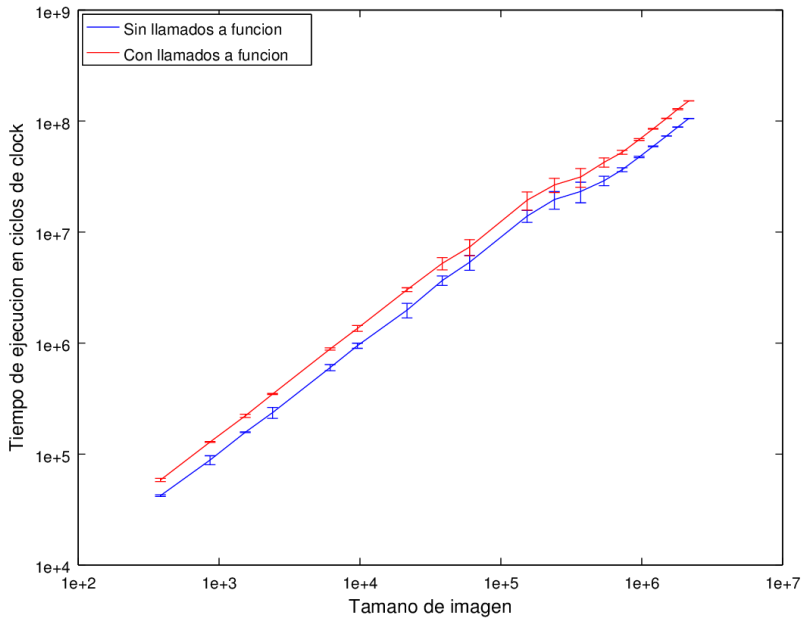
Creemos que la versión del código implementada en lenguaje ensamblador que no realiza llamados a funciones va a tener un mejor rendimiento, ya que se evita el overhead que producen estos llamados.

Valores utilizados como parámetros

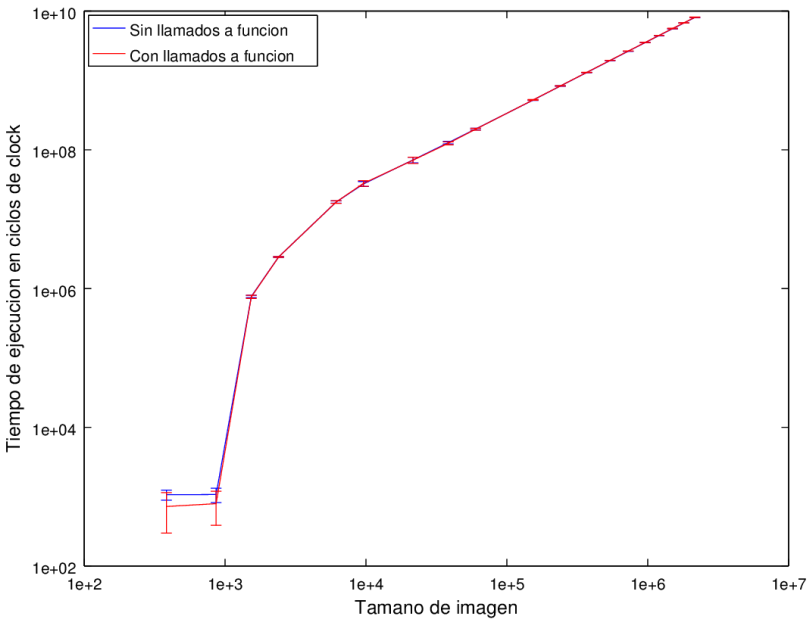
En este experimento el ancho de las imágenes utilizadas como parámetro se encuentran en un rango entre 24 y 1800 píxeles. Además, para el filtro *blur*, se utilizó $r = 15$ y $\sigma = 5$.

Resultados

Filtro *diff*



Filtro *blur*



Conclusiones y observaciones

Como muestran los gráficos de la implementación en C, los algoritmos que hacen llamados a otras funciones tienen un mayor tiempo de ejecución que las que no los hacen. Esto se debe a que cada vez que se hace un llamado a función, en necesario modificar la pila, manteniéndola alineada y guardando los registros que se deben preservar según la convención C y fueron utilizados a lo largo de la función.

En lenguaje ensamblador, cuando implementamos el algoritmo sin el llamado a la función auxiliar, sigue siendo necesario acceder a la pila ya que hay que reutilizar registros que se tienen que mantener para la convención C. Por esto, seguimos haciendo accesos a memoria. Una vez que la accedemos es muy posible que en la caché se encuentren los siguientes accesos a realizar. Entonces, la diferencia entre accederla pocas veces o algunas más es muy pequeña, ya que el acceso a memoria caché no es muy caro.

Posiblemente, si se reformulara por completo la estructura del algoritmo y la manera en que se utilizan los registros, podría lograrse una implementación en lenguaje ensamblador que saque una ventaja más considerable. Sin embargo, esto representaría una labor muy costosa, y hay que tener en cuenta que un código que no realiza llamadas a funciones auxiliares es más difícil de mantener y menos legible, por lo que la ganancia obtenida en rendimiento sería probablemente muy pequeña en relación con las desventajas que se ocasionarían.