CODIFICACION DE VIDEO DIGITAL BASADO EN LA ESTIMACION IMPLICITA DE MOVIMIENTO INTER-IMAGEN UTILIZANDO REDES DE NEURONAS

Por el Dr. Eduardo J. García García

Profesor Investigador del Departamento de Ciencias Computacionales
Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Estado de México
Carretera Lago de Guadalupe Km. 3.5
52926 Atizapán de Zaragoza, Edo. de México
e-mail : egarcia@campus.cem.itesm.mx

RESUMEN

En este trabajo se presenta la aplicación de las Redes de Neuronas Artificiales (RNA) en la codificación de video digital. Se aborda en particular la codificación inter-imagen con el objeto de eliminar en lo posible la información visual redundante en el tiempo. Dicha codificación inter-imagen consiste en un proceso de predicción donde se intenta estimar la imagen a codificar, a partir de la imagen precedente compensada en movimiento. En los sistemas normalizados de codificación de video, la predicción inter-imagen es realizada con la ayuda de métodos iterativos de estimación de movimiento como el *Block Matching*. En este artículo se muestra que es posible substituir el proceso de estimación/compensación de movimiento por una RNA. El predictor neuronal es capaz de obtener valores de predicción de las imágenes a codificar basándose en un aprendizaje sobre las imágenes precedentes. Los valores de predicción llevan implícitamente la evolución del movimiento en la escena, de donde se sugiere que la RNA es una forma de predictor con una estimación implícita de movimiento.
Palabras Clave : redes de neuronas artificiales, algoritmo de retropropagación, block matching, codificación inter-imagen, estimación de movimiento, compensación de movimiento.

I. Introducción

El desarrollo vertiginoso de la tecnología nos conduce poco a poco a abandonar los sistemas analógicos y remplazarlos por sistemas digitales. En este contexto, el procesamiento y la transmisión del video forman parte de esta evolución. Sin embargo, la representación digital del video nos confronta al problema de manipulación de grandes volúmenes de información. Es por esto que surge la necesidad de buscar métodos y algoritmos capaces de codificar y comprimir el video digital, sin introducir degradación en la información visual. Respondiendo a esta necesidad, los organismos internacionales de normalización proponen diversos esquemas de compresión de video digital. Dichos esquemas normalizados involucran a su vez dos procedimientos de codificación : la codificación espacial o *intra-imagen* y la codificación temporal, llamada también *inter-imagen*.

La codificación *intra-imagen* consiste en la eliminación de la información redundante en el espacio, proyectando la señal imagen sobre una base ortogonal. Dicha proyección permite pasar del dominio imagen al dominio de las frecuencias espaciales, donde se realiza la cuantificación de aquellas no reelevantes, desde el punto de vista de la percepción humana.
La codificación *inter-imagen*, en cambio, permite la eliminación de la información redundante en el tiempo mediante la predicción de imágenes. Dicha predicción consiste en generar una imagen muy cercana a la que se desea codificar, utilizando las imágenes precedentes y la estimación de movimiento de los objetos en la escena. De este modo, la transmisión o almacenamiento de la diferencia entre la imagen original y su predicción permite un ahorro importante en el volumen de información del video.
La Fig. 1 muestra la estructura de base de un codificador híbrido de video. Las diferencias básicas entre cada norma de codificación propuestas por los organismos internacionales residen en la elección de la cuantificación de frecuencias espaciales, el método de estimación de movimiento, la codificación entrópica de canal y la regulación del flujo de datos a la salida del codificador. Sin embargo, el esquema en todo caso es similar al de la Fig. 1.

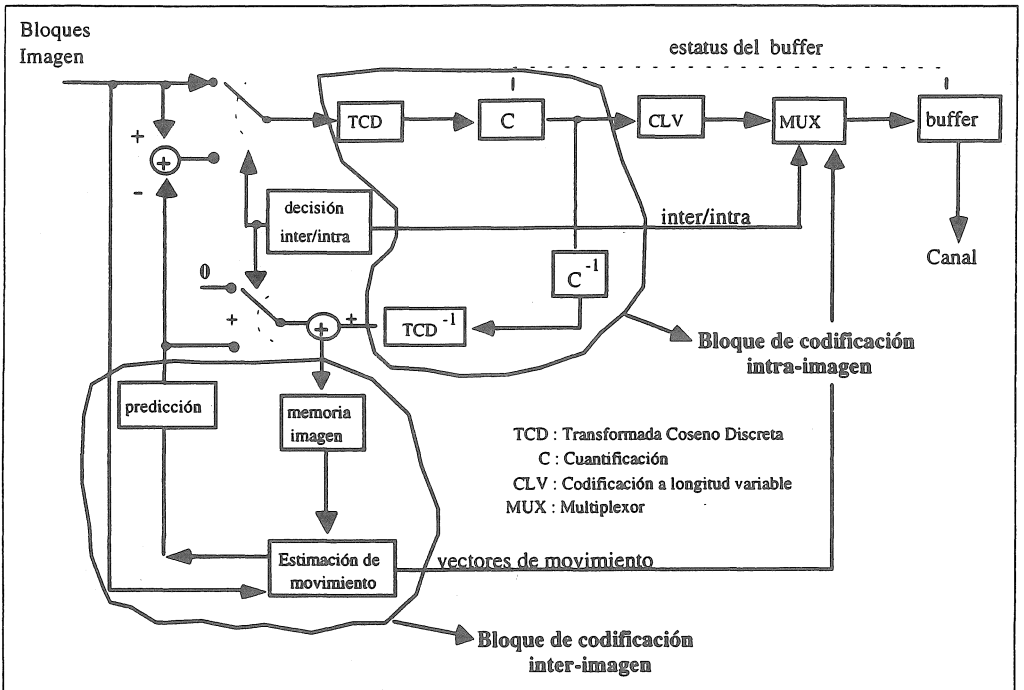


Fig. 1 Esquema Híbrido de Codificación de Video

En el presente artículo se abordará el problema de la codificación inter-imagen, la cual es decisiva en la adecuada compresión del video digital.

Primeramente, se hará una descripción breve del esquema de codificación utilizado en nuestra investigación. Se eligió trabajar sobre un esquema del European Telecommunications Standards Institute (*ETSI*)[1] para transmisión de video sobre canales inferiores a 17 Mbits/s . A continuación se detallará el método de codificación inter-imagen utilizado en dicho esquema, haciendo énfasis especial en la estimación de movimiento. Enseguida, explicaremos el principio de predicción propuesto y su integración al esquema de codificación. Finalmente, mostraremos los resultados obtenidos en la compresión de video digital formato TV según la Rec. 601 del Comité Consultivo Internacional de Radiocomunicaciones (CCIR)[2].

II. El Esquema de Codificación Video ETSI

En el marco de este estudio, se escogió un esquema de codificación video muy cercano a la *Recomendacion H.261*[3] de *CCITT* para video-conferencia : el esquema *ETSI* definido en el proyecto Europeo *RACE 1018*[1]. A diferencia de la *Rec. H.261*, usada en la transmisión de video-conferencia entre 64 y 384 Kbps, el esquema *ETSI* permite la transmisión de imágenes en formato televisión sobre canales de a lo más 17 Mbits/s. Este esquema es derivado de la norma de transmisión de televisión digital con calidad estudio a 34 Mbits/s[1].

Algunas de las características del esquema *ETSI* son las siguientes :

- Codificación de imágenes entrelazadas en formato *CCIR Rec. 601*[2]. Codificación trama por trama.
- Tres modos de codificación son posibles : codificación intra-imagen, codificación inter-imagen y codificación inter-imagen con compensación de movimiento. La elección del tipo de codificación se realiza sobre cada macro-bloque de 16x8 pixels.
- Estimación de movimiento usando el método de *Block Matching*
- Predicción a partir de la imagen precedente por medio de interpolación lineal, en caso de movimientos en fracción de pixel.
- Compensación de movimiento sobre macro-bloques de 16x8 pixels.
- Codificación por transformación : Uso de la *Transformación Coseno Discreta* (TCD) sobre bloques imagen 8x8 pixels. Dicha transformación se aplica sea sobre bloques de la imagen directamente, sea sobre un bloque de diferencias. Este último se obtiene de sustraer a un bloque de imagen original, un bloque constituído de valores de predicción (inter-imagen).
- Codificación entrópica de los coeficientes TCD resultantes.

La eficiencia de este esquema de codificación depende estrechamente de los módulos de estimación/compensación de movimiento y de la cuantificación de los coeficientes TCD. El esquema de codificación ETSI sigue la estructura mostrada en la Fig. 1.

III. La Estimación de Movimiento usando Block Matching

La estimación de movimiento es una técnica que nos permite evaluar el movimiento de los objetos en la escena. En el codificador, esta información es utilizada para generar imágenes de predicción, que siendo suficientemente acertadas, permitirán disminuir el volumen de información mediante la diferencia de cada imagen original y su predicción. Cabe notar que el decodificador del video debe estar en capacidad de reconstruir la predicción utilizada, para lo cual, necesita la información relativa al movimiento de los objetos en la escena. Esto provoca la transmisión o almacenamiento de información suplementaria.

Uno de los métodos de estimación de movimiento más comunes en los esquemas de codificación es el denominado "apareo de bloques" o *Block Matching*. El *block matching* consiste en la búsqueda iterativa de un conjunto de pixels (macro-bloque) en la imagen *t*, dentro de una ventana de pixels en la imagen *t-1*. Si la búsqueda da resultados positivos, el macro-bloque puede ser representado por su homónimo de la imagen precedente afectado por un vector de movimiento $(x,y)^T$, como se ve en la Fig. 2.

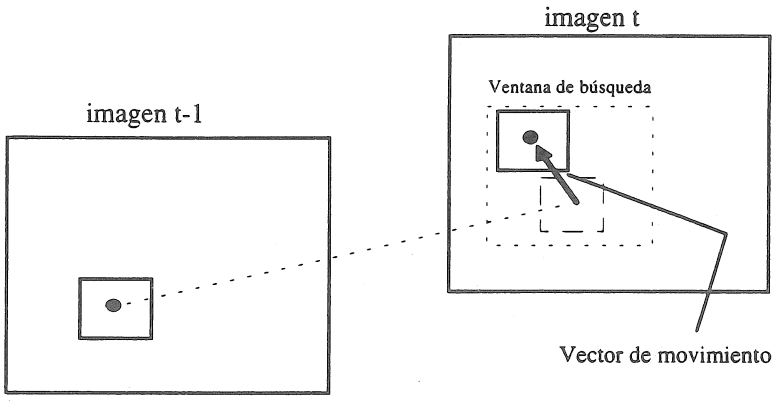


Fig. 2. El método de Block Matching

En el caso del esquema de codificación ETSI, la búsqueda del macro-bloque es de tipo *full-search*, sobre una ventana de ±16 pixels horizontalmente por ±8 pixels verticalmente. El criterio de "apareo" (match) utilizado es la minimización del error cuadrático medio (Mean Square Error) :

$$\mathbf{MSE}\ (x,y)^T = \frac{1}{(16)(8)} \sum_{(i,j)\ \in\ \text{ventana}} [p(i,j,k) - p(i+x, j+y, k+1)]^2$$

siendo $p(i,j,k)$ el valor de un pixel en la posición (i,j) de la imagen k y $(x,y)^T$ el vector de movimiento en consideración.

Como se podrá constatar, el hecho de efectuar una búsqueda exhaustiva del bloque involucra un número considerable de cálculos de MSE. Así pues, es necesario encontrar métodos menos costosos, en términos de consumo de tiempo y complejidad. Varias técnicas de búsqueda son propuestas en la literatura [4], pero los resultados no son del todo satisfactorios.

Una vez calculado el movimiento de los macro-bloques es posible, como ya se dijo, representarlos por sus homónimos en las imagenes precedentes pero con el desplazamiento indicado por los vectores de movimiento respectivos. A este proceso se le conoce como compensación de movimiento. De este modo, los bloques compensados en movimiento son utilizados como una predicción de los bloques originales en la imagen t.

IV. La Estimación Implícita de Movimiento Usando RNA

Como ya se mencionó, la estimación de movimiento usando la técnica de *block matching*, es una solución bien definida y adoptada en la mayor parte de los esquemas de codificación de video. Sin embargo, la complejidad del algoritmo depende estrechamente del tamaño de la ventana de búsqueda así como del criterio de semejanza, en este caso el MSE.

En este contexto, se efectuó un estudio que permite mostrar la factibilidad y el interés de utilizar RNA para realizar la predicción de imágenes, sin necesidad de una estimación/compensación explícita de movimiento.

Las RNA han sido ya utilizadas en aplicaciones, en las cuales, se busca predecir comportamientos a partir de una serie de observaciones en el tiempo. Tal es el caso de predicciones en los índices bursátiles, predicción en el consumo de agua, luz, etc. En nuestra aplicación, las RNA serán utilizadas como un predictor de pixels de una imagen. El conjunto de valores de entrada del predictor es la vecindad causal de cada pixel a predecir. Primeramente, la predicción se realizará pixel por pixel considerando sólo el entorno inmediato en la misma imagen y aquel de la imagen precedente.

El modelo de RNA escogido es el perceptron multi-capas, en el cual los pesos de conexión serán calculados mediante el algoritmo de retropropagación de gradiente con factor de modulación adaptativo[5].

Las entradas del perceptron multi-capas son constituídas por los pixels vecinos del valor a predecir, tanto en el dominio espacial (misma imagen), tanto como en el dominio temporal (imagen precedente).

Esto es debido a que el valor de un pixel *p* esta estrechamente ligado a los valores de sus vecinos (correlación espacial) y es probable que el pixel *p* se encuentre en la misma región de la imagen precedente (correlación temporal). Así pues, es posible distinguir dos ventanas causales como se muestra en la Fig. 3.

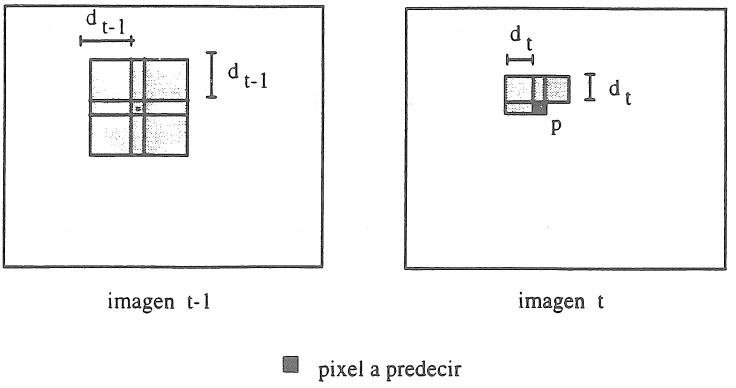


Fig. 3. Ventanas consideradas en la predicción de un pixel

Podemos ver que el número de pixels considerado para la predicción depende de los valores d_{t-1} y d_t que definen las ventanas causales en la imagen *t-1* y la imagen *t* respectivamente. Se escogieron los valores $d_{t-1}=4$ y $d_t=2$ considerando una distancia de movimiento inferior a 1/2 bloque de tamaño 8x8.

De este modo, el número de pixels considerados en la predicción es :

$$\{4(d_{t-1}^2 + d_{t-1}) + 1\} + 2(d_t^2 + d_t) = 81 + 12 = 93$$

Los experimentos llevados a cabo sobre secuencias de video de prueba, sugeridas por los organismos de normalización, mostraron que los valores escogidos para d_{t-1} y d_t son adecuados. La consideración de ventanas de mayor tamaño no introduce una mejora sustancial en la predicción. En cambio, se aumenta el costo en términos de complejidad de cálculo y aprendizaje de la RNA. Más aún, los estudios de movimiento efectuados por otras instancias mostraron que es poco frecuente encontrar desplazamientos superiores a ±4 pixels, en el caso de secuencias de video formato televisión[6].

En cuanto a la RNA utilizada, se trata de un perceptron multicapas con 93 neuronas en la capa de entrada, correspondientes a las vecindades causales, y una neurona en la capa de salida correspondiente a la predicción realizada. La estructura de la RNA se muestra en la Fig. 4.

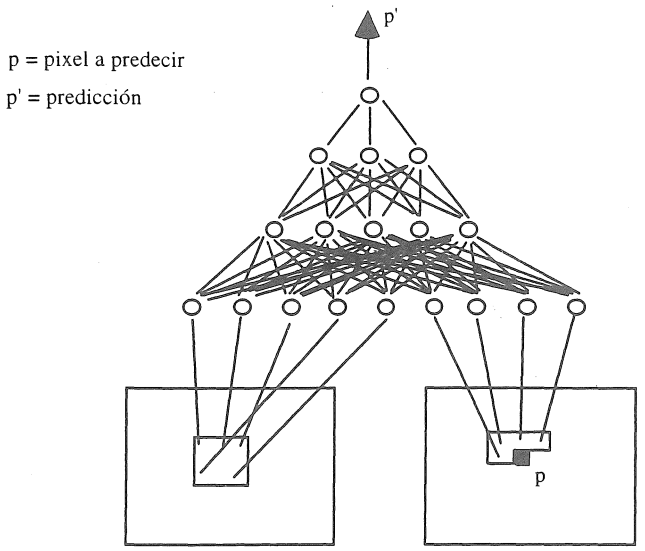
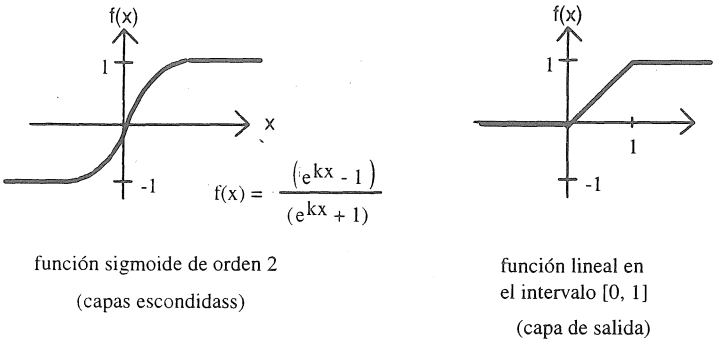


Fig. 4. Predicción utilizando una RNA multi-capas

Sin embargo, el número de capas escondidas así como el número de neuronas en cada una de ellas es menos evidente a obtener. Hasta ahora, no existe un método eficiente para determinar el número de neuronas en las capas escondidas de una RNA multi-capas, en función de sus entradas y salidas. Se probaron múltiples configuraciones, todas ellas de forma a generar una estructura en pirámide, es decir, cada capa escondida contiene un número menor de neuronas a medida que se encuentra más cerca de la capa de salida. Esto se basa en el hecho que en cada capa se efectua una proyección de las entradas sobre un espacio de menor dimensión, lo cual nos lleva a una síntesis de la información, que es lo que se busca. La elección de capas ocultas de mayor tamaño a cada vez, ocasiona una expansión del problema en espacios de mayor dimensión, lo que implica una descomposición en vez de una síntesis[**6**].

En cuanto a la función de activación de las neuronas, se escogió la tangente hiperbólica (***tanh***). Esta función permite un rango de valores de salida de las neuronas en el intervalo [-1, +1], como se ve en la Fig. 5. La función ***tanh*** es idéntica a la función sigmoide de orden 2 y es utilizada al interior de la RNA, es decir en las capas ocultas. El objetivo es generar una amplia dinámica en los valores de salida de dichas neuronas. Sólo la neurona de salida tiene una función lineal, para limitar su salida en el intervalo [0, 1].



función sigmoide de orden 2
(capas escondidass)

función lineal en
el intervalo [0, 1]
(capa de salida)

Fig. 5. Funciones de activación del perceptron multi-capas

Los valores de los pixels en la entrada del perceptron pueden tomar valores entre 1 y 255. En este caso, el uso directo de estos valores ocasionaría la saturación de la función de activación de las neuronas y ninguna solución será posible. Por este motivo, los valores en entrada son normalizados dividiéndolos entre 255. Con esto se logra obtener valores en el intervalo [0,1]. En cuanto a la salida de la RNA, los valores obtenidos se encuentran también en el rango de 0 a 1, por lo que la multiplicación por 255 permite encontrar un valor válido de predicción en el dominio imagen.

Finalmente, y como ya se mencionó, el algoritmo de aprendizaje de la RNA es la retropropagación de gradiente con factor de modulación adaptativo.

V. Resultados Obtenidos

La Tabla 1 presenta los resultados obtenidos para la secuencia de video *mobile_calendar*, utilizada en las pruebas de codificadores normalizados (cortesía del *Centro Común de Estudios en Teledifusión y Telecomunicaciones* de Francia). Esta secuencia consiste en un muro con tapiz multicolor en donde se desplaza un tren a escala de derecha a izquierda. Al mismo tiempo, un calendario sobre el muro se desplaza de arriba a abajo. La complejidad de la escena así como los movimientos múltiples la hacen difícil de procesar.

Se compararon los resultados con la salida del predictor original usado en el esquema *ETSI*, que involucra la estimación de movimiento por *block matching*. La secuencia de entrenamiento usada para el aprendizaje de la RNA consistió en las 2 primeras imágenes de la secuencia *mobile_calendar*. Se eligieron un total de 6000 muestras uniformemente espaciadas.

Los mejores resultados se obtuvieron con una RNA de 5 capas : 93 en la capa de entrada, 3 capas escondidas con 50, 25 y 10 neuronas respectivamente y una capa de salida de una neurona.

Método Usado	√EQM
Perceptron (93-40-1)	87.2
Perceptron (93-60-1)	85.4
Perceptron (93-50-25-10-1)	14.3
Block Matching	9.6

Tabla 1. Comparación de diferentes predictores para mobile_calendar
(Error de predicción de la imagen 2 de mobile_calendar)

Podemos notar que el hecho de aumentar el número de capas escondidas permite obtener mejores resultados. Sin embargo, este aumento genera también un aumento considerable en el tiempo de aprendizaje. Por supuesto, la implementación del algoritmo en una máquina paralela permitiría eliminar este problema. Cabe destacar que una vez entrenada la RNA, su funcionamiento como predictor es mucho más rápido que la búsqueda exhaustiva del block matching.

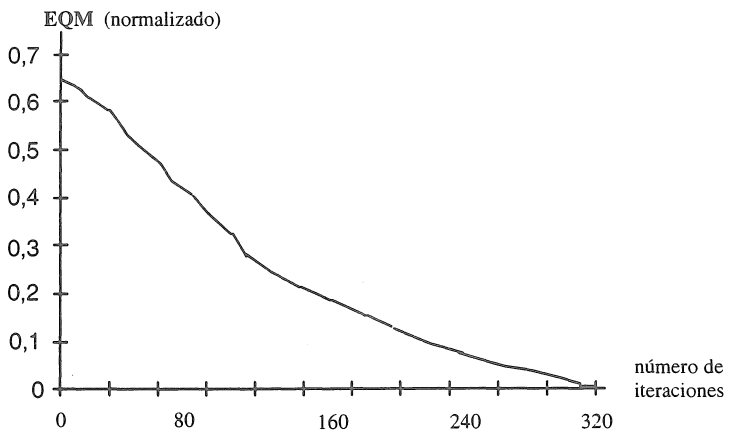


Fig. 6. Convergencia del algoritmo de retropropagación para la RNA de 5 capas.

Se puede observar en la Fig. 7 que los resultados, en términos de error cuadrático medio, son muy cercanos. Sin embargo, desde el punto de vista visual, la predicción por RNA nos da imágenes más claras y agradables.

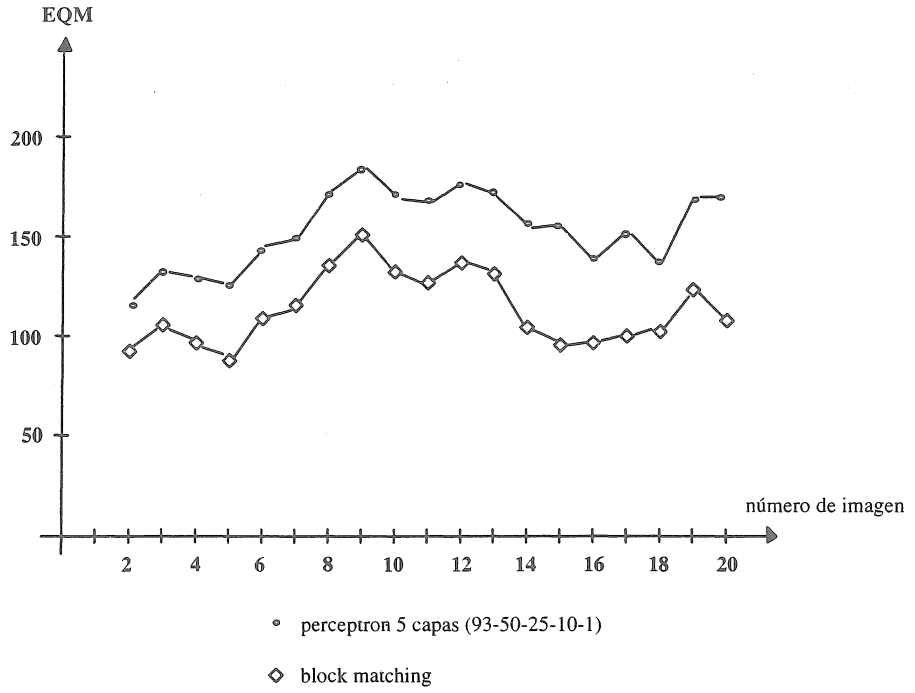


Fig.7. Comparación de los errores de predición sobre la secuencia mobile_calendar

Puesto que el movimiento de los objetos en la escena se encuentra implícito en la predicción dada por la red, no es necesario transmitir o almacenar ningún tipo de información adicional, como es el caso de los vectores de movimiento en el esquema convencional. Sin embargo, si existe la necesidad de hacer llegar los pesos de conexión de la RNA al decodificador. Así pues, la información relativa al movimiento, en el caso del esquema clásico, es remplazada por los pesos de conexión lo cual no afecta el volumen global de información video codificada. De cualquier forma, la simplicidad del predictor neuronal y su rapidez, comparado con el método de *block matching*, lo coloca todavía como una buena elección.

Finalmente, se integró el predictor neuronal en el esquema de codificación ETSI. En este caso, la parte de predicción y estimación de movimiento fue remplazada por el predictor neuronal. Los primeros experimentos se realizaron sobre el esquema de codificación en lazo abierto, es decir, los valores de la vecindad causal del pixel a predecir son tomados de las imágenes originales. Esto nos permitió evaluar el comportamiento del sistema sin el efecto de deriva. Hay que recordar que la deriva consiste en la acumulación del error de codificación debido al uso de informaciones codificadas precedentemente en vez de información original. Es esta deriva la que justifica el refrescado periódico de la secuencia de video.

Por otra parte, con el fin de eliminar problemas en la predicción de los pixels de los bordes de las imágenes, el entrenamiento de la RNA se efectuó tomando un marco de 4 pixels alrededor de cada imagen con valor igual al promedio de luminosidad. Esta elección se basa en el hecho que el sistema visual humano es más o menos sensible a las variaciones alrededor de la media de luminosidad de la escena que percibe.

El flujo de datos escogido para las simulaciones fue de 11 Mbits/s. Sólo la componente de luminosidad (imagen blanco/negro) fue considerada. Se codificaron las primeras 20 imágenes de la secuencia *mobile_calendar*. Se midió la relación señal/ruido (Signal to Noise Ratio SNR) y se comparó con el esquema clásico con estimación de movimiento por block matching. Se observó una ganancia de al menos 2% desde el punto de vista SNR. Se observó también que la calidad visual de las imágenes reconstruídas así como la rapidez de codificación son superiores al del esquema convencional *ETSI*. Los resultados se observan en la Fig. 8.

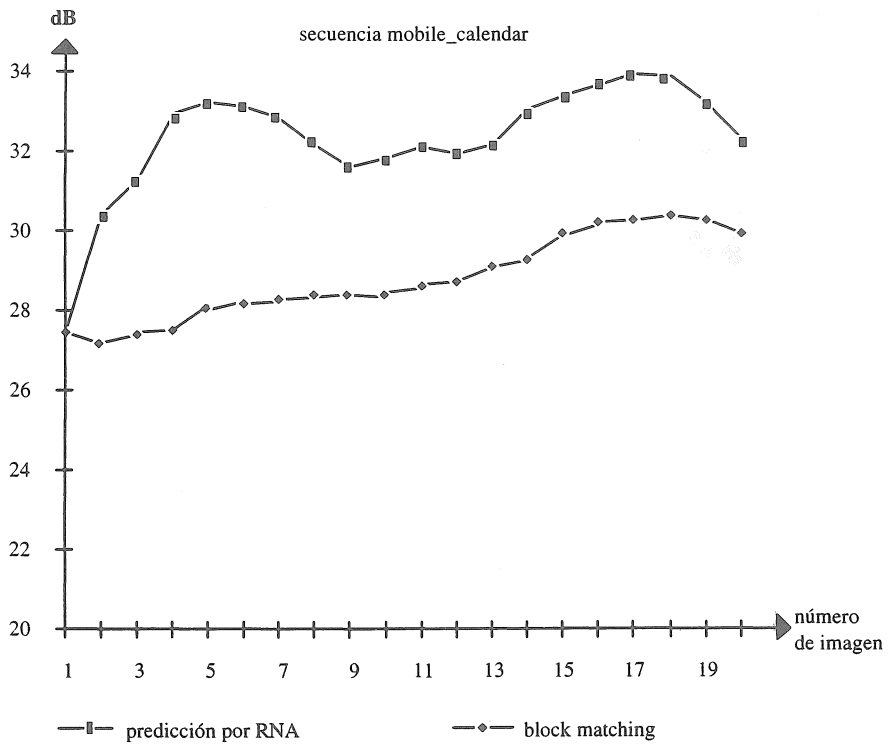


Fig. 8. Comparación de predictores en el esquema ETSI

VI. Conclusiones y Perspectivas

Hasta aquí, podemos concluir que el uso de RNA para la predicción de señal imagen es factible y se obtienen resultados, en cuanto a la calidad visual de las imágenes, mejores a los métodos convencionales. Dejando aparte el problema del aprendizaje, el uso de la RNA tiene sus ventajas en cuanto a la rapidez de ejecución y la simplicidad de su implementación. La RNA asegura un buen almacenamiento de las variaciones espacio-temporales de las imágenes.

Los experimentos realizados sobre un esquema de codificación real confirman la eficiencia del predictor neuronal, aunque hasta ahora sin introducir deriva en el sistema (lazo abierto). Nos queda verificar el comportamiento del predictor neuronal en lazo cerrado. En este caso y dada la estructura de codificación por bloque del esquema ETSI, tendremos que pensar en implementar un predictor neuronal que involucre varios perceptrones en paralelo. Esto se debe a que en algunos casos, no contaremos con la vecindad causal. Esto se ejemplifica en la Fig 9. Sin embargo, el nuevo paradigma neuronal parece prometedor y se espera obtener resultados satisfactorios a corto plazo.

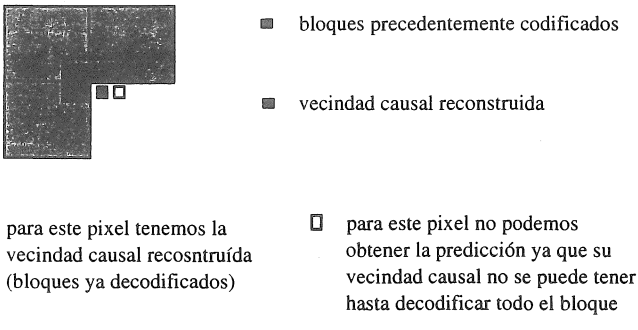


Fig. 9. Problema de vecindad causal en lazo cerrado.

BIBLIOGRAFIA

[1] ETSI, "Network Aspects Digital Coding of Component TV Signals for Contribution Quality Applications in the range 34-45 Mbits/s", Draft prETS 300 174, June 1991

[2] CCIR, "Recommendation 601-I Digital Methods of Transmittig TV Information", 1986

[3] CCITT, "Recommendation H.261, Video codec for audiovisual services at px64 kbps", 1990

[4] Choquet B., "Estimation de Mouvement et Segmentation Spatio Temporelle en Sequences d'images", Tesis de Doctorado, Univ. Rennes I, Sep. 1988

[5] Qiu G. et al, "Accelerated Training of Backpropagation Networks using Adaptive Momentum Step", Electronics Letters, Vol. 28, N. 4, Feb. 1992.

[6] García E., "Etude et Utilisation des Réseaux de Neurones pour le Codage d'images de TV", Tesis de Doctorado, Univ. Rennes I, Dic. 1994.