Sequence matching using spatiotemporal wavelet decomposition (Proceedings Paper)

Author(s): **A. Corghi; Riccardo Leonardi**

Date: **10 January 1997**

ISBN: **9780819424358**

Proceedings Vol. 3024

Visual Communications and Image Processing '97, Jan Biemond; Edward J. Delp III, Editors, pp.938-2423

Date: **10 January 1997**
ISBN: **9780819424358**

Paper Abstract

Indexing and retrieval of image sequences are fundamental steps in video editing and film analysis. Correlation-based matching methods are known to be very expensive when used with large amounts of data. As the size of sequence database grows, traditional retrieval methods fail. Exhaustive search quickly breaks down as an efficient strategy for sequence databases. Moreover, traditional indexing with labels has a lot of drawbacks since it requires a human intervention. New advanced correlation filters are being proposed so as to decrease the computational load of the task. A new method for retrieval of images sequences in large database based on a spatio-temporal wavelet decomposition is proposed here. It will be shown how the use of the multiresolution approach can lead to good results in terms of computationally efficiency and robustness to noise. We will assume that the query sequence may not be contained in the database for different reasons: the presence of a noise signal on the query, or different digitation process, or the query is only similar to sequences in the database. As a

consequence we are providing have developed a new efficient retrieval strategy that analyses the database in order to extract the most similar sequences to a given query. The wavelet transform has been chose as the framework to implement the multiresolution formalism, because of its good compression capabilities, especially for embedded schemes. And the good features it provides for signal analysis. This paper describes the principles of a multiresolution sequence matching strategy and outlines its performance through a series of experimental simulations.

Out of print 11/14/07.

DOI: 10.1117/12.263306

SEQUENCE MATCHING USING A SPATIO-TEMPORAL WAVELET DECOMPOSITION

A. Corghi & R. Leonardi,

Signals & Communications Lab., Dept. of Electronics for Automation, *
University of Brescia, I-25123, E-mail: leon@bsing.ing.unibs.it

ABSTRACT

Indexing and retrieval of image sequences are fundamental steps in video editing and film analysis. Correlation-based matching methods are known to be very expensive when used with large amounts of data. As the size of sequence database grows, traditional retrieval methods fail. Exhaustive search quickly breaks down as an efficient strategy for sequence databases. Moreover, traditional indexing with labels has a lot of drawbacks since it requires a human intervention. New advanced correlation filters are being proposed so as to decrease the computational load of the task.[10,11]

A new method for retrieval of images sequences in large database based on a spatio-temporal wavelet decomposition is proposed here. It will be shown how the use of the multiresolution approach can lead to good results in terms of computationally efficiency and robustness to noise. We will assume that the query sequence may not be contained in the database for different reasons: the presence of a noise signal on the query (e.g. due to a lossy compression scheme), or different digitization process (e.g. a different sampling rate), or the query is only similar to sequences in the database. As a consequence we are providing have developed a new efficient retrieval strategy that analyzes the database in order to extract the most similar sequences to a given query. The wavelet transform has been chosen as the framework to implement the multiresolution formalism, because of its good compression capabilities, especially for embedded schemes. and the good features it provides for signal analysis. This paper describes the principles of a multiresolution sequence matching strategy and outlines its performance through a series of experimental simulations.

Keywords: Database, Matching, Multiresolution, Querying, Synchronization, Video Indexing, Video Retrieval, Video Segmentation, Wavelets.

1. INTRODUCTION

An important functionality of current and future image databases is automatic content-based retrieval. This corresponds to the capability of retrieving images or part of them, solely on the basis of visual properties, without using any external textual label. This feature is even more essential when dealing with image sequence databases, because of the motion component. Sequence matching in very large databases poses challenges in addition to the limitations of the particular matching method used. A brute-force technique which matches a query pattern exhaustively against all possible locations does not scale very well on large database.
One way to to achieve scalability is to apply more elaborated matching methods, combining multiple patterns into one filter. Matching is then done *in parallel*, using the composite filter instead of the individual pattern.

This research has been supported in part by the Italian Ministery of the University and of the Scientific and Technological Research (MURST) and by the Italian National Council for Research (CNR).

SPIE Vol. 3024 • 0277-786X/97/$10.00

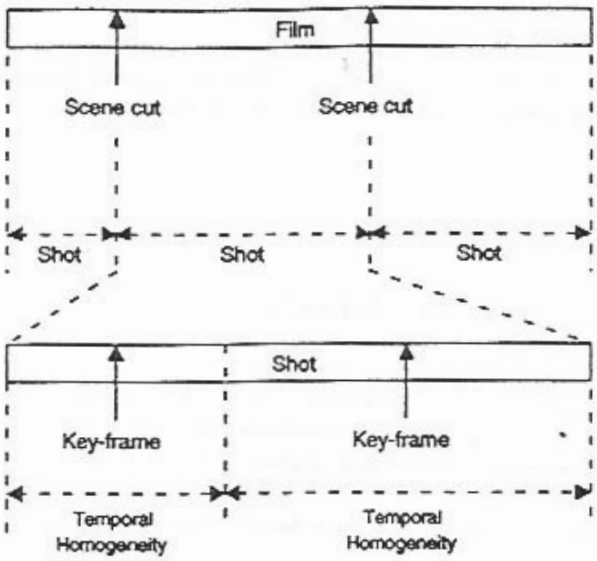

Figure 1: Database structure: a film is composed by one or more shots, each of them being identified by one or more key-frames.

Another way is to use simple matching methods, simultaneously, to approach the problem in a structured way. We proposed a multiresolution approach based on the wavelet transform. As discussed in the following, this data organization allows to decrease the computational load of the task, so as to move in the scalability direction, and to improve the robustness to noise or other artifacts due to the digitalization processes.
Our technique is based on an alignment algorithm that compares, with a multiresolution strategy, the wavelet transforms of two generic video sequences. This allows to determine the exact position of a sequence $s_q[x, y, n]$ in the sequence $S[x, y, n]$ if $s_q[x, y, n]$ is temporally contained in $S[x, y, n]$. If the sequence $s_q[x, y, n]$ is instead similar to a subset of $S[x, y, n]$, then the algorithm solves a similarity problem, synchronizing the two sequences in the best relative position. Since each video sequence is however represented by a big amount of data, this approach by itself is not efficient enough to solve the scalability of the task. A previous video indexing step is needed. As a consequence we have designed a database structure based on the concept of homogeneous sequence. In other words, when a video sequence is added to the database, it is preprocessed, so as to segment it in a group of homogeneous subsets each of them being identified by a visual index entry. So as to have a further structure of the sequence database, we have introduced the concept of shot.[3] A shot is a video sub-sequence which contains no scene cuts. Therefore, there is an embedded relationship between film and shot, and, in turn, between shot and homogeneous sequence; each homogeneous sequence identifies a visual index entry, that we call *key frame* of the corresponding sequence. In Figure 1, the database structure is shown. We have chosen the shot as the fundamental element of the database structure. From there, the retrieval phase can be organized in two consecutive steps: in the first, the visual index is used so as to extract from the vast number of shots a subset which may be identified by similar key-frames. In the second, using this candidate set, the multiresolution matching technique is used to find the best candidate shot, and the position of the query with respect to it. The first step step improves the algorithm efficiency, because it prevents from applying the sequence matching process to all the sequences contained in the database. Only the ones in the set of candidate shots will be correlated to the query. Scanning the index is much less critical, from a performance point of view, than the whole database scanning.[5]
In section 2 a brief introduction to the wavelet theory is presented; section 3 describes then the key-frame identification step characterizing the database generation step; sections 4, 5 and 6 discuss the retrieval phase and finally, in section 7, experimental simulations are reported.

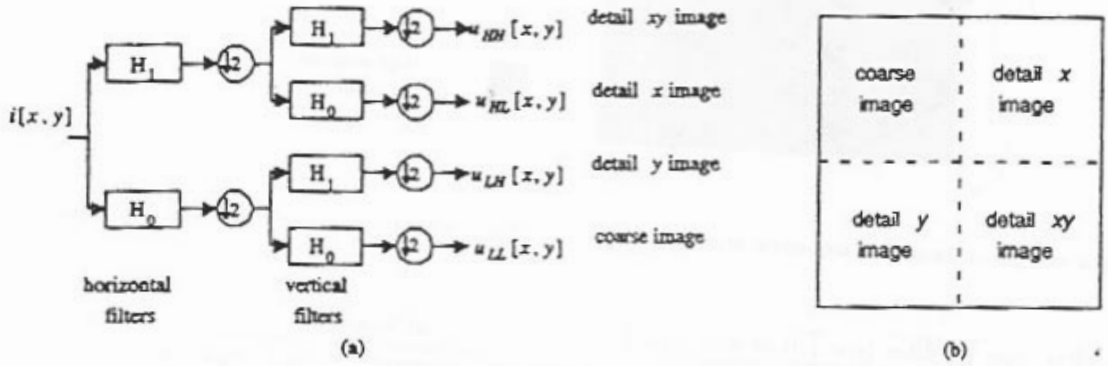

Figure 2: Wavelet spatial decomposition filterbank (a); the first sub-stage operates along raws, while the second sub-stage operates along columns. A possible data reorganization is shown in (b).

2. REPRESENTATION OF THE INFORMATION: A MULTIDIMENSIONAL WAVELET TRANSFORM

Before describing the proposed technique for sequence matching, it is essential to summarize briefly the wavelet transform theory.
In what follows, we assume without loss of generality, that a video sequence is simply characterized by its luminance component. A video sequence can be described by a multiresolution pyramid obtained from a spatio-temporal wavelet representation of the luminance signal. The signal domain space is decomposed using a dyadic wavelet transform so as to segment the information in roughly octave frequency bands over space and time.[1,4,2]
A spatio-temporal wavelet representation of a video-sequence can be obtained by constructing a separable multiresolution description of the luminance function along every dimension (space/time).The decomposition at each stage is performed in two steps: a two dimensional spatial decomposition followed by a one-dimensional temporal decomposition. Figure 2 shows the step of a single stage spatial decomposition process (a) and a possible reorganization of the output coefficients (b). A multiresolution of each frame takes place applying the filterbank in cascade on the coarse signal. In order to distinguish different resolution levels we use a specific index to identify them. It is important to note that the output coefficients do not belong solely to the frequency domain, but also to the space-time domain. The next time stage operates on one-dimensional signals obtained by considering the pixels at the same spatial location in re-organized successive frames. This step splits the corresponding sequence in different sub-sequences, each of them with different lengths. These sub-sequences describe the signal as a function of the time-frequency information. The final coefficient reorganization after a wavelet transform with one spatial stage and one temporal stage is shown in Figure 3. Table 1 identifies the different type of information represented by each sub-sequence. Since we select a separable wavelet transform, a different number of spatial stages and temporal stages can be performed. We however assume that is not possible to use a different number of stages in the x direction and in the y direction.

3. KEY-FRAMES IDENTIFICATION

The generation of the sequence database consists in the indexing of all the shots that compose the database, in order to obtain better performance in the next retrieval phase.[7,8]
Let us introduce the following notation to identify all elements of interest. We assume that a film is generally made of a set of non-overlapped shots. We identify the film with $F[x, y, n]$ where x, y and n represent respectively

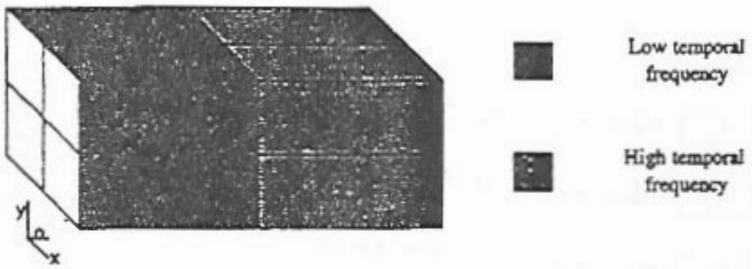

Figure 3: Data reorganization of a sequence after a wavelet transform with one spatial stage and one temporal stage.

x filter type	y filter type	time filtr type	nickname
low	low	low	coarse low temporal frequency sequence
high	low	low	detail-x low temporal frequency sequence
low	high	low	detail-y low temporal frequency sequence
high	high	low	detail-xy low temporal frequency sequence
low	low	high	coarse high temporal frequency sequence
high	low	high	detail-x high temporal frequency sequence
low	high	high	detail-y high temporal frequency sequence
high	high	high	detail-xy high temporal frequency sequence

Table 1: Wavelet decomposition of a video sequence: renaming the information channels.

:he space and time dimensions; the shots are instead identified by $S[x, y, n]$. With the word *sequence* we denote :ither a shot or a part of it (from a temporal point of view) and it is referred to $s[x, y, n]$. Frames will be identified ›y $f[x, y]$ or $s[x, y, \overline{n}]$ where $\overline{n}$ is fixed. Finally, at times, we will use $\vec{x}$ to indicate the spatial coordinates (x, y).
A usual technique for video indexing is sequence labeling with *key-words*. However this approach has poor interest ›ecause it is often a non automatic procedure in the sense that it requires human intervention. Moreover the iterature reports that different visual cues can be labelled in different ways and the correlation of the key-words dentifying the visual cues could be difficult.[6]
A visual index could be used instead. The entries in that index are images, more exactly frames (called *key-frame*), :xtracted from the database video sequences. In this case it is possible to define an automatic strategy in order ›o build the index using a simple mathematical framework.
As described before we assume that the database is composed of shots. Therefore, we could use the first frame ›f each shot as an entry in the index table. Unfortunately such an index is too sparse and would cause many 'etrieval errors. A possible solution is to increase the index size by using more than one entry for each shot, e.g. ısing as an index one frame every T frames. This technique is however quite computationally intensive, and is ›ften not necessary for most shots assuming that a rather small value of T has been selected. Besides selecting too mall a value for T would require a large amount of memory for the index and may slow down the retrieval step. n this paper we propose a key-frame identification strategy by introducing a sequence homogeneity measure.

3.1. Definition of a homogeneity measure

A sequence homogeneity measure should indicate how much a sequence varies at a macroscopic level over ime. To be more explicit, let us consider the following example: in a first shot two men are speaking, while in a econd one a camera zooms out then pans a given scene. Obviously the visual content of the first shot varies less

than the visual content of the second one. We can therefore state that the first shot is more homogeneous than the second. It should hence be possible to characterize each sequence using a homogeneity measure. In order to mathematically define such a measure, we refer first to the usual definition for image similarity.
In the literature different complex image similarity measures have been proposed. Typically, one has:

$$\vec{c} = \psi(f_1[\vec{x}], f_2[\vec{x}]), \tag{1}$$

where $f_1[\vec{x}]$ and $f_2[\vec{x}]$ are two arbitrary images, while $\vec{c}$ is represents an output coefficient vector that summarizes the correspondence between $f_1[\vec{x}]$ and $f_2[\vec{x}]$. A good choice for ψ is the normalized cross-correlation measure evaluated at the origin (i.e. without any spatial shifts):

$$\varphi(f_1[\vec{x}], f_2[\vec{x}]) = \frac{2\langle f_1[\vec{x}], f_2[\vec{x}]\rangle}{\|f_1[\vec{x}]\|^2 + \|f_2[\vec{x}]\|^2} = \frac{2\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1}(f_1[x,y]\cdot f_2[x,y])}{\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1}|f_1[x,y]|^2 + |f_2[x,y]|^2} \tag{2}$$

Previously we suggested to use a sequence homogeneity measure at a macroscopic level. This can be efficiently achieved by considering the lowest spatial band of the wavelet decomposition of the video sequence using N spatial stages and no temporal stages, i.e. the coarse low temporal frequency (N,0) sequence at any instant in time. Because we use a normalized cross-correlation measure, if $f_1[x,y] \geq 0$ and $f_2[x,y] \geq 0$ for every x and y, then $0 \leq \varphi(f_1, f_2) \leq 1$.
Let us define the *representative frame* of a given generic sequence $s[\vec{x}, n]$ as the middle frame of the sequence $s[\vec{x}, n]$; we denote it with $m_s[\vec{x}]$.
The sequence $s[\vec{x}, n]$ has a degree α of homogeneity if

$$\varphi(f[\vec{x}], m_s[\vec{x}]) \geq \alpha \qquad \forall f[\vec{x}] \in s[\vec{x}, n]. \tag{3}$$

This means that the cross-correlation between any frame of the sequence and the representative frame has to be higher than the degree of homogeneity of that sequence. Thus, to obtain the homogeneity degree of a sequence it is sufficient to find the lowest cross-correlation value with respect to its middle frame. Unfortunately, for long sequences, this definition is very computationally intensive as it is necessary to estimate N cross-correlation measures, where N identifies the sequence length. In order to draw a simple criterion, let us define the *sequence cross-correlation function* as:

$$\xi_h[n] = \varphi(s[\vec{x}, n], s[\vec{x}, h]). \tag{4}$$

The sequence cross-correlation function expresses how the cross-correlation evolves with respect to any given frame of that sequence. It is possible to prove that, under certain conditions, the cross-correlation function is monotonically decreasing from $n = h$. Such conditions, denoted as *homogeneity assumption*, are the following:

1. the sequence has to be smooth enough temporally and last a few frames ($\sim$ 200 frames at a rate of 25 fps);
2. the sequence contains neither scene cuts nor periodic motion.

In Figure 4 a cross-correlation function is shown.
Let $s[x, y, n]$ be a sequence with $n = 0, 1, \ldots, N$ and $m_s[\vec{x}]$ its representative frame. If the cross-correlation with respect to $m_s[\vec{x}]$ is strictly decreasing, then $s[\vec{x}, n]$ has degree α of homogeneity, where:

$$\alpha = \min\{\varphi(m_s[\vec{x}], s[\vec{x}, 0]), \varphi(m_s[\vec{x}], s[\vec{x}, N])\} = \min\{\xi_{N/2}[0], \xi_{N/2}[N]\}. \tag{5}$$

Thus, the homogeneity assumption allows to estimate the homogeneity degree of a sequence with only two cross-correlation measures instead of N.

3.2. Optimal decomposition of a shot into homogeneous sequences

So as to design a good video indexing strategy, we further introduce the concept of homogeneous sequence to identify the key-frames of the index. The main objective becomes how to determine the best decomposition of

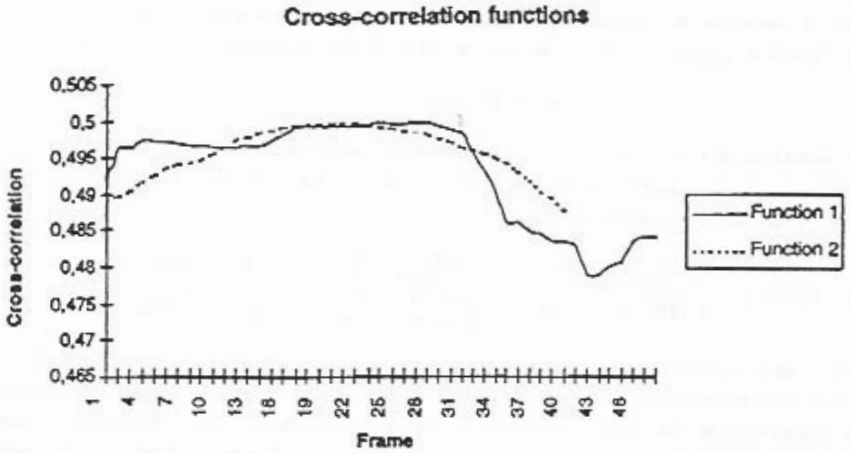

Figure 4: Cross-correlation function behavior for two sequences: in the first the homogeneity assumption is satisfied, while in the second it is not.

a shot into homogeneous sequences so that the extracted key-frames represent a good video index. In order to analyze the problem we make the following three considerations:

1. for any shot $S[x, y, n]$, there are two trivial decompositions:
   (a) every frame of $S[x, y, n]$ is considered an homogeneous sequence with unitary length; the corresponding decomposition D_S^{inf} is made up of N homogeneous sequences each one having homogeneity degree equal to 1; it will be named the *lower trivial decomposition*;
   (b) the whole shot $S[x, y, n]$ is considered a homogeneous sequence; the corresponding decomposition D_S^{sup} is made of a single sequence; it will be named the *upper trivial decomposition*;
2. in order to minimize the index size, the number of sequences contained in a decomposition must be minimized;
3. each homogeneous sequence contained in a decomposition D_S must have the highest possible homogeneity degree, so as to decrease the error probability in the retrieval phase.

The last two considerations lead to the following trade-off: the need to minimize the number of sequences brings a decrease of the homogeneity degree. On the other hand, maximizing the homogeneity degree brings an increase of the number of homogeneous sequences (that is the number of key-frames). The segmentation task can be formulated as an optimum problem so as to identify the optimal decomposition of a shot.
The optimal decomposition $D_S^{opt} = \{s_j\}_{j=0,1,\ldots,J-1}$ of a generic shot $S[x, y, n]$ can be obtained by maximizing an objective function that depends on the mean homogeneity degree of the sequences s_j and on the number of these sequences. In order to define a relative weight between these two terms it is possible to consider J, the number of sequences, to the power of β, where $0 \leq \beta \leq 1$. The objective function to maximize is given by:

$$\mathcal{L} = \frac{\sum_{j=0}^{J-1} \alpha_j}{J} \cdot \frac{1}{J^\beta} = \frac{\sum_{j=0}^{J-1} \alpha_j}{J^{\beta+1}} \tag{6}$$

where α_j represents the homogeneity degree of the sequence s_j.
The optimum problem is constrained by the two trivial decompositions. Instead of using a normalization factor

by $J^{\beta+1}$ it is also possible to add others contraints to the objective function $\mathcal{L}$ such as the maximum number of homogeneous sequences for each shot or such as the minimum homogeneity degree of each sequence.

3.3. A greedy segmentation algorithm

The correct algorithm for the solution of the non-linear optimum problem cannot be simply defined without an exhaustive search. Therefore, a technique that approximates the optimal solution has been studied. In the following, an iterative strategy is proposed. There are two main problems: first, the initial solution must not constrain the final solution; second, the used technique has to be computationally efficient. Generally, the definition of an iterative strategy for an algorithm involves three aspects: the identification of the possible operations that the algorithm can perform, the choice of the initial solution and finally the definition of the algorithm strategy.
Before describing these aspects for the specific optimization procedure, a consideration about the space of the possible solutions is required. Hence, it can be represented using a time-variant graph with a finite number of nodes. Each node represent a group of consecutive frames. Two adjacent groups are connected by a branch. So the number of branches connected to each node is equal to 2 (except for the extremity nodes). The actual decomposition D changes into a new decomposition D' where either new nodes are created or some nodes are merged together. In the lower trivial decomposition, the graph is composed by N nodes, each of them is linked to the previous and to the next frame. In the upper trivial decomposition there is just one node forming the graph. The proposed segmentation optimization consists in a two stage process: a split and merge procedure. Splitting a sequence $s_j[\vec{x}, n]$ means to generate two new sequences $s'_j[\vec{x}, n]$ and $s''_j[\vec{x}, n]$ so that:

$$\begin{aligned} s'_j[\vec{x}, m] &= s_j[\vec{x}, m] \\ s''_j[\vec{x}, m] &= s_j[\vec{x}, m + \tfrac{N_j}{2}] \end{aligned} \tag{7}$$

where $m = 0, 1, \ldots, N_j/2 - 1$ and N_j is the sequence length. Merging of two contiguous sequences $s_j[\vec{x}, n]$ and $s_{j+1}[\vec{x}, n]$ (whose lengths are respectively N_j and N_{j+1}), produce the sequence $s'_j[\vec{x}, n]$ such that

$$s'_j[\vec{x}, n] = \begin{cases} s_j[\vec{x}, m] & \text{for } m = 0, 1, \ldots, N_j - 1 \\ s_{j+1}[\vec{x}, m - N_j] & \text{for } m = N_j, N_j + 1, \ldots, N_j + N_{j+1} - 1 \end{cases}$$

The initial state has not to constrain the navigation in the space of solutions. A good decision is to start from one of the two trivial solutions. If the upper trivial decomposition is chosen, the algorithm may however result ill-posed, because of the definition of the homogeneity measure according to 5. Let consider a very static sequence with 100 frames. Between the 60-th frame and the 80-th frame a big object appears and disappears. Starting from the upper trivial decomposition and using the homogeneity assumption the sequence seems to have a high homogeneity degree, since the cross-correlations between the middle frame and the bound frames are calculated. This is a wrong conclusion as there is a set of non-homogeneous frames (from 60-th to 80-th frame). On the other hand, starting from the lower trivial decomposition and using a good navigation strategy, it is possible to build the decomposition around groups of homogeneous content.
The greedy optimization strategy is therefore based on iteratively applying a merge to the graph followed by a split step. More specifically, in the former the algorithm calculates whether it is possible to merge two contiguous sequences so as to increase the value of the objective function; for each possible fusion, the new objective function value is first estimated and assigned to each branch of the graph; then the merge where the highest increase of the objective function takes place is chosen. When no more merging is possible, the merging phase stops. This can be originated from two different situations: all possible merge do not increase the value of objective function or the upper trivial decomposition has been reached. In practice at each iteration it is not necessary to compute all the increases of the objective function but only those contiguous to the last nodes of the graph that were merged. If the maximum increase is not positive or the upper trivial decomposition is the actual state, the merging procedure is stop. Equivalently the split phase searches if any sequence of the actual decomposition can be split so as to improve the solution. The splitting stage is totally equivalent to the previous one: at each step

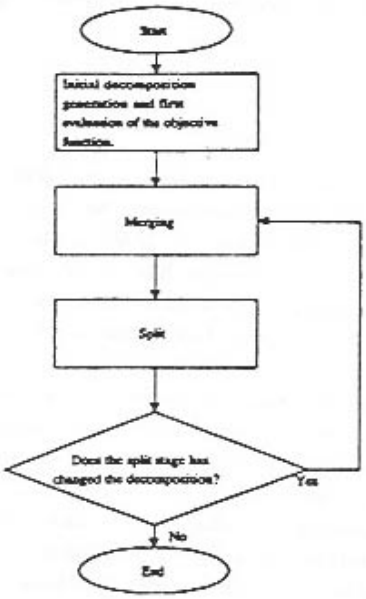

Figure 5: Block diagram of the video indexing algorithm.

the objective function variations are calculated so as to identify the best action. The splitting process stops either if all subdivision do not lead to an increase of objective function or if the lower trivial decomposition has been reached. If during the split process there is not modification of the graph structure, the segmentation is stopped, else a new iteration starts with a merge followed by a split. Figure 5 shows a block diagram of the proposed algorithm. The major drawback of the algorithm presented here lies upon the impossibility to estimate the error between the optimal solution and the one obtained by the greedy segmentation strategy. The only certainty is that the last state is the best among all the reached states by the algorithm.

Using the homogeneous sequence approach and the described segmentation strategy, it is possible to create a visual index from a sequence database. The index contains the representative frames of each identified sub-sequence in the shot decomposition. Therefore, when a shot is added to the database, its segmentation in homogeneous sequences is performed so as to include its representative frames in the index. The index includes also an homogeneity information associated to each key-frame.

It is important to note that the split and merge procedure neither does actually solve the optimum problem nor guarantees that the homogeneity assumption is verified for the identified sub-sequences. This last observation will have very important consequences on the next phase: the key-frame matching. The following tho phases, key-frame matching and sequence synchronization, are concerned with the retrieval task.

4. KEY-FRAMES MATCHING

The key-frame matching step is used for the extraction of a set of candidate shots. The shots contained in this set represents a selection of all the shots archived in the database that exhibit a rough similarity with the query sequence. Under some assumptions about the structure of this set, it is possible to decrease the computational load of the synchronization task without affecting the performance of the retrieval. In fact the synchronization task will operate on the set of candidate shots and not on the whole database.

The main objectives of the key-frame matching process are two folds: on one hand it is to extract the shot which actually contains the query sequence (i.e. avoiding a missed error); on the other hand is to extract the lowest number of wrong shots (i.e. leading to a false detection error). The missed error must be avoided because the

synchronization task is conditioned by the presence of the right shot in the candidate set. It is not possible to recover from this type of error. Contrarily the false detection error should be kept to a minimum as not to slow down the synchronization task. It will however not affect the result of the retieval itself.
The key-frames matching task uses the visual index in order to extract the candidate set. On the basis of what described in the previous section, each shot in the database has been segmented in an arbitrary number of sequences so as to maximize a given objective function. Each of these sequences is described by its representative frame (called key-frame) and by its homogeneity degree. Any sequence will be well represented by just one key-frame. So each key-frame and each homogeneity degree is linked to just a single shot and define the index entries. Using this information, it is possible to identify the candidate set. The key-frame matching algorithm compares an arbitrarily chosen frame of the query sequence, with all the key-frames contained in the visual index. The query frame will be denoted with $f_q[\vec{x}]$, while the index key-frames are denoted with $f_k[\vec{x}]$. The comparison is made using the same similarity measure used in the previous section. If

$$\varphi(f_q[\vec{x}], f_k[\vec{x}]) \geq \alpha_k \qquad (8)$$

for the key-frame $f_k[\vec{x}]$, then the linked shot is included in the candidate set, It is possible to show that, under some assumptions, the homogeneous sequence approach guarantees that no missed error is introduced. Let us assume that each stored sequence respects the homogeneity assumption and that the query is

$$s_q[\vec{x}, n] = S[\vec{x}, n + N] \qquad (9)$$

where $S[\vec{x}, n]$ is a shot contained in the database. The homogeneity assumption implies that, for each homogeneous sequence $s_j[\vec{x}, n]$ included in $S[\vec{x}, n]$, the following relation is verified:

$$\varphi(f[\vec{x}], f_j[\vec{x}]) \geq \alpha_j \qquad (10)$$

for each frame $f[\vec{x}] = s_j[\vec{x}, \overline{n}]$. In equation (10), α_j and $f_j[\vec{x}]$ represent respectively the homogeneity degree and the representative frame of $s_j[\vec{x}, n]$. If the query frame $f_q[\vec{x}]$ belongs to the homogeneous sequence $s_j[\vec{x}, n]$, then the equation 10 is verified and no missed error is possible. However since the split and merge segmentation strategy does not lead necessary to the best optimum of the objective function (6), as discussed in the previous section, the homogeneity criterion cannot be guaranteed, and so a missed error may occur.
No hypothesis are possible about the false detection error. The query frame can satisfy equation (10) with all the key-frames that identify different sub-sequences. The false detection error can occur in two situations: if there is a matching with a key-frame linked to the wrong shot (Type I False Detection Error) or with a key-frame linked to the right shot but associated with a wrong sub-sequence (Type II False Detection Error). The latter error is irrelevant since the algorithm extracts a set of shots and not a set of homogeneous sequences. The use of the cross-correlation implies that a global information measure is being used, but it does not contrain the spatial resolution level to which the comparison takes place. It is possible to compare images at different resolution levels. Low resolutions are concerned with an averaged information and so the probability of false detection increases. So as to avoid this effect, it is possible to use more than one resolution level of the query frame and more than query frame considering, in this case their relative position in time. The key-frame matching algorithm must take into consideration this relationship so as to decrease the probability of false detection decreases. As high spatial resolutions do not identify as well the macroscopic content, there is a higher probability of missed error if not conducted properly. Accordingly, a statistical evaluation of the missed error and the false detection error as functions of the spatial resolution levels will be presented in the last section.

5. SEQUENCE SYNCHRONIZATION

The sequence synchronization task must find the right position of the query sequence from the video database, by comparing the query sequence $s_q[\vec{x}, n]$ with the set of candidate shots $S_i[\vec{x}, n]$. So the main goal of the sequence synchronization task is to align temporally two sequences ($s_q[\vec{x}, n]$ and $S[\vec{x}, n]$), assuming that the query sequence $s_q[\vec{x}, n]$ is temporally contained in the shot $S[\vec{x}, n]$. The alignment may be done by matching any frame of $s_q[\vec{x}, n]$

with the corresponding frame of the shot. Unfortunately this technique is computationally intensive, because it could require on the average $N/2$ cross-correlation evaluation, and, moreover, it does not consider the temporal correlation between frames increasing the probability of bad alignment.

An alternative strategy leading to the same objective is based on the multiresolution approach. The basic idea is to work by successive approximations: the task is performed by starting from a coarse representation of the information to produce an initial guess for focusing then to a full resolution result. This is a greedy approach that is sometimes suboptimal. It is therefore necessary to provide a recovery mechanism so as to modify the decision made at the coarse resolution levels.

The multiresolution approach is usually applied to a multiresolution description of the information. There are different techniques to describe a signal using a multiresolution structure. The wavelet transform has been selected here. For simplicity let us suppose that no spatial decomposition stages are applied, that is no filters are used along the spatial dimensions x and y. Note that, additional spatial stages would have to be used so as to improve performance and robustness to noise, but in the description below, no spatial stages are supposed for simplicity purposes. Thus, a wavelet transform with K temporal stages, splits the original sequence into one low temporal frequency sub-sequence and K high temporal frequency sub-sequences. The former represents a coarse version of the original sequence downsampled by a factor of 2^K in time. The others represent different detail version downsampled by different factors from resolution -1 to resolution $-K$. The multiresolution tree obtained from the wavelet transform is shown in Figure 7. Each couple of sequences at a given resolution (i.e. $-M$) has the same parent, the coarse sequence at the resolution layer immediately above (i.e. resolution $-M+1$). Moreover, the generic frame $f[\vec{x}]$ at resolution $-M$ has $L_s + L_w$ parents in the coarse sequence at resolution $-M+1$ where L_s and L_w are respectively the smoothing and the wavelet analysis filter lengths. In order to rebuild the parents, it is necessary to use the information around $f[\vec{x}]$ and its corresponding frame in the detail sequence at the same resolution. Using the Haar basis, two consecutive frames are merged so as to build an average frame and a normalized difference frame, which belong respectively to the low and high temporal frequency sequences of the lower resolution layer. So any couple of frames at resolution $-M$ (the first extracted from the coarse sequence and the second from the detail sequence) can be used to rebuild the parent signal at the higher resolution level, which in the case of the Haar transform corresponds to the two parents. This property is fundamental when we want to rebuild the parents of a parent of a give frame at resolution $-M$. Generally, the single parent at resolution $-M+1$ is not sufficient, but some other frames in its neighborhood are needed. On the contrary, in the case of the Haar transform, the parent at resolution $-M+1$ contains all the information, together with its coupled frame in the high temporal frequency at resolution $-M+1$, to rebuild the grandparents at resolution $-M+2$.

A block based representation of the multiresolution algorithm is presented in Figure 6. It is composed of three main processing elements. The first step consists of finding the alignment between the query and the tested sequences at the lowest resolution, using the low temporal frequency sequences obtained for each one by the wavelet transform. Once the coarse sequences are aligned, the synchronization is improved using the detail information. When the full resolution has been reached, if the alignment is not correct, a backtracking procedure is started in order to re-route the synchronization along another branch of the projection tree used in the refinement of the synchronization procedure. In what follows a detailed description of each block is presented. The coarse version of the query sequence and the candidate shot are denoted respectively with $s_{ql}^{(-M)}[\vec{x},n]$ and $S_l^{(-M)}[\vec{x},n]$, while detail sequences are denoted with $s_{qh}^{(-M)}[\vec{x},n]$ and $S_h^{(-M)}[\vec{x},n]$ where the index between the round brackets refers to the resolution level. The alignment task of the coarse sequences is obtained by matching the first frame of the query sequence, using the cross-correlation as similarity degree. For each frame of $S_l^{(-M)}[\vec{x},n]$ the value

$$\phi_k = \varphi(s_{ql}^{(-M)}[\vec{x},0], S_l^{(-M)}[\vec{x},k]) \tag{11}$$

is calculated as a function of k and the the maximum is selected; this identifies the proper alignment.

Once the alignment at the coarse resolution has been found, it is necessary to increase the resolution to improve it. The successive approximation strategy uses the result of the lower resolution as an initial guess for the current resolution level; the new solution serves then for a further approximation stage until the full resolution has been reached. Suppose that the alignment at resolution $-m$ (where $0 \geq -m \geq -M$) has been found and that the selected frame has position h.The corresponding position $2h$, at resolution $-m-1$, is used as an initial guess. Since the cross-correlation is used as a measure, coarse information at resolution $-m+1$ is essential, both for

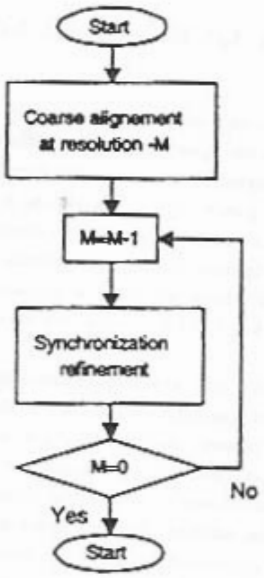

Figure 6: Block diagram of the sequence synchronization algorithm.

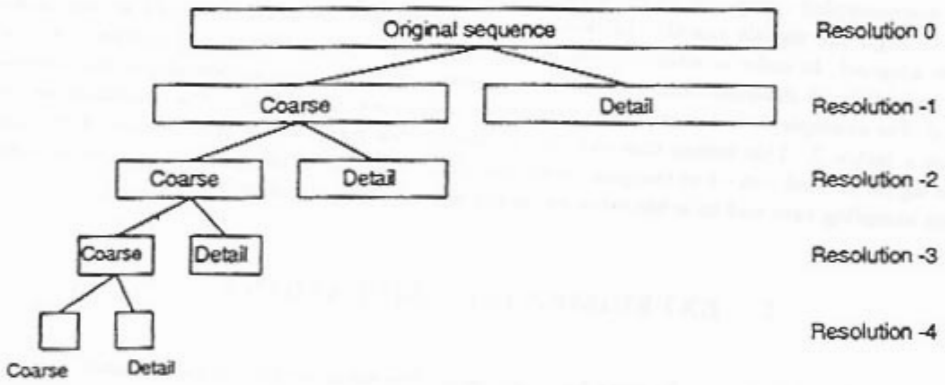

Figure 7: Multiresolution organization of a wavelet decomposition of a video sequence in the time domain.

the query sequence and for the candidate sequence. For the query the first frame of the sequence at resolution $-m+1$, $s_{ql}^{(-m+1)}[\vec{x},0]$ must be rebuilt. Similarly, the candidate sequence must use the frames around position $2h$ ($S_l^{(-m+1)}[\vec{x},2h\pm n]$ with n varying in the range of the wavelet analysis smoothing filter length). The comparison between the query frame and the other frames is calculated so as to extract the position identified by the highest cross-correlation value. As explained above, this new solution is used as an initial guess for the next alignment stages.

At resolution 0, if equation (9) is satisfied and the algorithm has found the proper alignment, the cross-correlation has reached its maximum:

$$\varphi(s_{ql}^{(0)}[\vec{x},0],S_l^{(0)}[\vec{x},\tilde{h}])=1. \tag{12}$$

Otherwise, a backtracking procedure is started in order to find another branch to cover. It consists of searching the last refinement decision, choosing another alignment and, consequently, another branch towards higher resolution levels.

6. NOISY QUERY SEQUENCE

If equation (9) is not satisfied, because the query sequence is noisy or slightly different from the test sequence (e.g., generated from a different digitization process, perhaps with a different sampling rate in time), some considerations of the previous section are not valid anymore, but the multiresolution approach may still be preserved. Three situations have been considered: a noisy query, $s_q[\vec{x}, n] = S[\vec{x}, n] + \epsilon[\vec{x}, n]$, a downsampled query, $s_q[\vec{x}, n] = S[\vec{x}, g \cdot n]$ and a query not contained in the database, but similar to some of the database shots.[12]
Two parts of the described algorithm will be affected in these situations. First, during the key-frames matching, equation (10) is not valid, because equation (9) does not hold anymore. Second, during the sequence synchronization task, the final condition at full resolution (12) will never be verified, because the cross correlation is not calculated on the same signals.
If the query sequence is noisy, the frame used for the key-frame matching task is not contained in the database. If so, a missed error can occur even if the homogeneity assuption on the sequence is verified. The missed error probability is therefore a function of the noise power. As we will show in the next section, describing the simulation results, the noise alteration can be decreased using spatial stages of thw wavelet transform jointly with those in time. The low resolution spatial information allows to access the data at a macroscopic level and, at the same time, to smooth the noise effect. The noise also affects the multiresolution synchronization strategy, because no exact alignment can be reached. Condition (12) cannot be used any longer. In this case the only possibility is to use the first detected solution as an approximation of the right alignment. No backtracking procedure is carried out.
The case of a downsampled query can be reconduced to the noisy situation because at all resolutions, there is no possibility to align the signals exactly. In this case the condition (12) is satisfied and so the backtracking strategy can be adopted. In order to eliminate the downsampling effect, a proposal is to compare the query and the candidate sequences at different resolution levels. Obviously, a-priori information about the downsampling factor is needed. For example, let us suppose that the query sequence is contained in the candidate sequence but is decimated by a factor 2. This means that the query contains one frame every other frame of the candidate. Comparing the signal at level $-m-1$ of the query with the one at level $-m$ of the candidate allows to compensate for the different sampling rate and to achieve better results in the synchronization task.

7. EXPERIMENTAL SIMULATIONS

In the tests a database containing 30 shots has been used. The mean length of the sequence is 42 frames at a sampling rate of 6 frame per second. Some significant parameters defined below are measured as a function of the SNR of the query sequence: the probability of missed error, the statistics of the false detection error and the statistics of the synchronization error, that is the difference between the correct position of the query and the one found by the algorithm. A sample of 8 frame query sequences has been used. To establish the robustness of the algorithm in presence of noise, a white gaussian noise has been added to the query sequence. This test has been computed on two different databases: one created using a wavelet decomposition with no spatial stages and 3 temporal stages, the other with 4 spatial stages and 3 temporal stages.
The missed error probability represents the probability that the correct key-frame is not extracted during the key-frame matching process. In Figure 8 the behaviors of this parameter are plotted in the two considered cases. A higher number of spatial wavelet stages improves the robustness to noise of this technique. The error probability, using no spatial stages, reaches 100% when the SNR is zero dB. A false detection error occurs when some wrong shots have been chosen during the key frames matching. Experimentally it has been verified that the false detection error causes only to decrease the efficiency of the retrieval algorithm. It does not involve the efficiency of the synchronization task if the correct shot is in the set extracted by the key-frame matching. The sensitivity to noise in the case with no spatial stages of the wavelet decomposition is a positive factor versus the false detection error: when the noise power is comparable with the signal power, the key-frame matching task enfavors the correct matching and so the false detection error decreases. Figure 9 shows the statistical behavior of this error. The synchronization error (see Figure 10) is the position difference between the real position of the

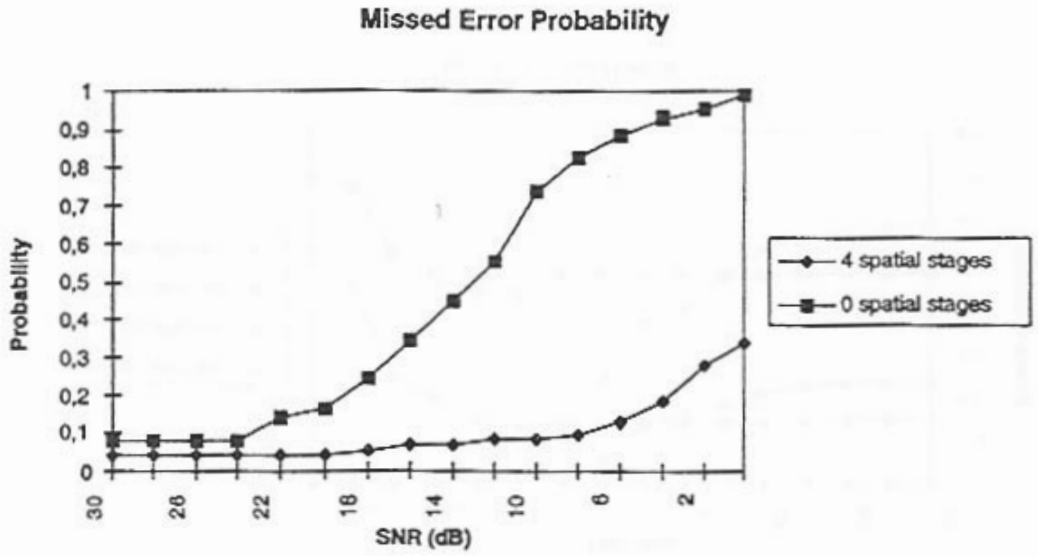

Figure 8: Missed error probability.

query sequence and the position found in the retrieval. This parameter can be evaluated only if no missed error occurs. This explains the low values of synchronization error in the case one does not use any spatial stages.

8. CONCLUSIONS AND FUTURE WORK

In this paper a new strategy for image sequence retrieval in large database has been described. We have assumed that the information for compression purposes were represented in a multiresolution structure using a wavelet transform. The approach is mainly based on two steps: an indexing phase and a sequence synchronization phase. The former is used during the database generation so as to extract a set of frames that could be considered representative of all the stored sequences. This visual index is then used in the first stage of the retrieval phase to identify a subset of sequences among which to perform the matching task. The second stage is concerned on the temporal alignment of two sequences using a multiresolution strategy based on the wavelet coefficients.
We have tried to perform a number of different tests to characterize the algorithm performance. Other tests will be done increasing the database size and introducing new matching criteria.
There are different areas for future research: first of all the introduction of a motion parameter that allows to perform a similarity retrieval based on motion models. Then it is reasonable to add a meta-index so as to improve the algorithm efficiency during the index scanning. Finally, it will be interesting to study a new methodology for the sequence alignment that does not require the reconstruction of the wavelet coefficients.

9. REFERENCES

[1] M.Vetterli and J.Kovacevic', *Wavelet and subband coding*, Prentice Hall, 1995.

[2] M.L.Hilton, B.D.Jawerth and A.Sengupta *Compressing Still and moving Images with Wavelets*, Multimedia Systems, 3(3).

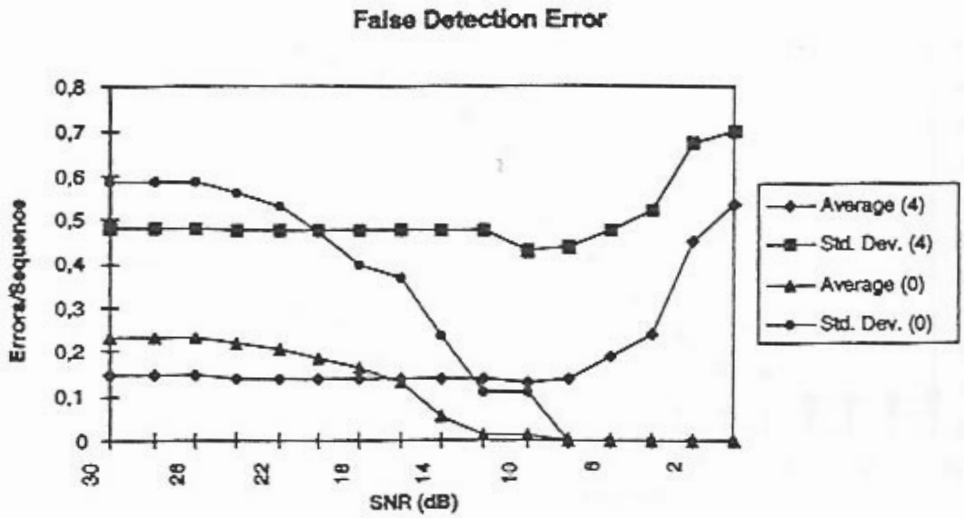

Figure 9: False detection error.

[3] J.M.Corridoni and A.Del Bimbo *Multi-perspective Indexing and Access of Movie Information Content*, submitted for publication.

[4] K.M.Uz, M.Vetterli and D.J.LeGall *Interpolative Multiresolution Coding of Advanced Television with Compatible Subchannels*, IEEE Trans. on Circuit and Systems for Video Technology, CSVT-1(1), 86-99 (Mar. 1991).

[5] C.E.Jacobs, A.Finkelstein and D.H.Salesin *Fast Multiresolution Image Querying*, Proc. SIGRAPH.

[6] W.Niblack, R.Barber, W.Equitz, M.Flickner, E.Glasman, D.Pektovic, P.Yanker, C.Faloutsos and G.Taubin *The QBIC project: Querying images by content by using color, texture and shape* Storage and Retrieval for Video Image Databases, 173-187, SPIE (1993).

[7] Y.Gong, H.Zhang, H.C.Chuan and M.Sakauchi *An image database system with content capturing and fast image indexing abilities*, Proceedings of the International Conference on Multimedia Computing and Systems, 121-130, IEEE (1994).

[8] T.Kato *Database architecture for content-based image retrieval*, Proceedings of the SPIE - The International Society for Optical Engineering, 1662, 112-123, SPIE (1992).

[9] S.W.Smoliar and H.J.Zhang *Content-based video indexing and retrieval*, IEEE Multimedia Magazine, 1, 62-72 (1994).

[10] D.M.Gavrila and L.S.Davis *Fast Correlation Matching in Large (Edge) Image Database*, DACA76-92-C-009 (1994).

[11] G.Ravichandran and D.Casasent *Advanced In-Plane Rotation-Invariant Correlation Filters*, IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-16(4), 415-420 (Apr. 1994).

[12] R.Chellappa, C.L.Wilson and S.Sirohey *Human and Machine Recognition of Faces: A Survey*, Proceedings of the IEEE, 83(5), (May 1995).

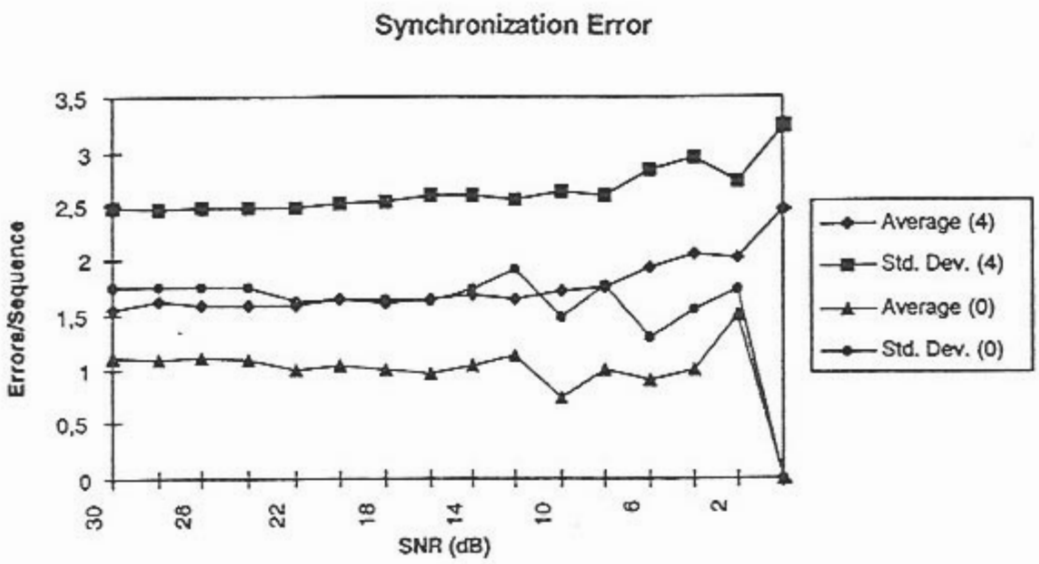

Figure 10: Synchronization error.