Compressive Video Classification for Decision Systems with Limited Resources

George Tzagkarakis*, Pavlos Charalampidis†‡, Grigorios Tsagkatakis†, Jean-Luc Starck*, and Panagiotis Tsakalides†‡
*Commissariat à l'Énergie Atomique (CEA), Centre de Saclay, F-91191 Gif-Sur-Yvette cedex, France
†Institute of Computer Science (ICS) - Foundation for Research & Technology - Hellas (FORTH), Crete, Greece
‡Department of Computer Science, University of Crete, Greece
e-mail: {georgios.tzagkarakis, jstarck}@cea.fr, {pcharala, greg, tsakalid}@ics.forth.gr

***Abstract*—In this paper, we address the problem of video classification from a set of compressed features. In particular, the properties of linear random projections in the framework of compressive sensing are exploited to reduce the task of classifying a given video sequence into a problem of sparse reconstruction, based on feature vectors consisting of measurements lying in a low-dimensional compressed domain. This can be of great importance in decision systems with limited power, processing, and bandwidth resources, since the classification is performed without handling the original high-resolution video data, but working directly with the set of compressed measurements. The experimental evaluation verifies the efficiency of the proposed scheme and illustrates that the compressed measurements in conjunction with an appropriate decision rule result in an effective video classification scheme, which meets the constraints of systems with limited resources.**

I. Introduction

Modern high-resolution sensing devices, with signal processing and communication capabilities largely based on the seminal Shannon and Nyquist studies, have enabled the acquisition, storage, and transmission of ever increasing amounts of data. Apart from reconstructing the original signal, several tasks such as detection and classification are also of paramount importance in signal processing applications. Focusing on the classification task, the problem consists in finding the correct class of the sensed signal among a set of candidate classes.

An area which could benefit significantly by the introduction of efficient computational models is *video classification*. With the advent of digital TV and the availability of large digital video databases, it is desirable to classify and retrieve high-resolution video content automatically. Moreover, in a remote sensing application, the potentially limited power, storage, and bandwidth resources require the efficient representation of the video content in a precise and compact way for further decision making. A characteristic example in the later case is the design of unmanned aerial vehicles (UAVs) and terrestrial sensor networks, which have been increasingly used in surveillance and reconnaissance applications, where the captured video is exploited to classify a target of interest.

Techniques like *principal component analysis* (PCA), *independent component analysis* (ICA) and *linear discriminant analysis* (LDA) [1] have been employed widely in the framework of signal classification. At the core of these methods is the extraction of the salient information content in a low-dimensional space resulting in an improved classification performance. The framework of *compressive sensing* (CS) [2], acting simultaneously as a sensing and compression protocol, is based on non-adaptive linear incoherent projections for the representation and reconstruction of sparse signals. Furthermore, in contrast to reconstructing the original signal, the task of classification requires fewer compressed measurements, which have been shown to preserve the meaningful information of the acquired signal [3]. In this paper, our goal is to address the problem of video classification by working directly in the compressed measurements domain. Among several other applications, the use of sparse representations and/or compressed measurements have been also exploited successfully for signal/image classification [4], [5], [6] and face recognition [7], [8], revealing the advantages of this framework.

The effectiveness of a video classification system is determined by two factors, namely, i) the extracted feature vector (or signature), and ii) the selected classifier. Conventional approaches require full resolution video data for the generation of descriptors, such as, color histograms, optical flow vectors and shape features, while several techniques are employed for classification, such as, support vector machines (SVM), hidden Markov models (HMM), and Bayesian methods based on maximum a posteriori (MAP) estimation. However, the above procedures can be highly inefficient in the case of limited-resource sensing systems. In particular, the onboard processing of a high-resolution video for the generation of the associated features may be computationally and power demanding placing significant burden on the encoder's hardware, while on the other hand, a large bandwidth is required to transmit full-resolution data at a base station for further processing and classification.

In the present work, we address the above drawbacks by introducing a CS-based video classification approach working directly in the compressed domain. More specifically, we consider the futuristic scenario of a sensing system equipped with a *single-pixel camera* [9], having the ability to estimate the correct class without demanding the acquisition of the video data at full resolution. Instead, a suitable feature vector, associated with the captured video sequence, along with an appropriate decision rule, are expressed directly in terms of the compressed measurements.

The paper is organized as follows: in Section II, the standard CS-based signal model and classification framework are reviewed in brief. Section III describes in detail the proposed compressive video classification system, followed by an experimental evaluation of the classification accuracy in Section IV. Finally, conclusions and further extensions are outlined in Section V.

II. CS-based video acquisition model

Let $\mathbf{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_R\}$ be a video sequence consisting of R frames $\mathbf{x}_j$, $j = 1, \ldots, R$. For convenience, in the following we consider that each frame is expressed as a column vector, $\mathbf{x}_j \in \mathbb{R}^N$. Then, a vector of compressed measurements $\mathbf{g}_j$, $j = 1, \ldots, R$ is generated for each frame using a suitable measurement matrix $\mathbf{\Phi}$ (for simplicity we use the same matrix for each frame) as follows,

$$\mathbf{g}_j = \mathbf{\Phi}\mathbf{x}_j \ , \tag{1}$$

where $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is a random measurement matrix with $M < N$.

Common choices for $\mathbf{\Phi}$ are random matrices with independent and identically distributed (i.i.d.) Gaussian or Bernoulli entries, whose columns are normalized to unit ℓ_2-norm. In a decision system with

limited resources, some additional requirements should be posed on the choice of the desired matrix $\mathbf{\Phi}$, such as the use of minimal number of compressed measurements, and the fast and memory efficient computation along with a "hardware-friendly" implementation. A class of matrices satisfying these requirements, the so-called *structurally random matrices*, was introduced recently [10]. The block Walsh-Hadamard (BWHT) operator is a typical member of this family and is used in the subsequent evaluation.

The random projections of the frames onto the rows of $\mathbf{\Phi}$ result in a low-dimensional representation of the original video sequence,

$$\mathbf{V} \stackrel{\mathbf{\Phi}}{\longmapsto} \{\mathbf{g}_1, \ldots, \mathbf{g}_R\} \ . \tag{2}$$

Notice that with the use of a single-pixel camera the generation of CS measurements does not require the acquisition of frames at full resolution, thus reducing significantly the processing and storage expenses of the sensing device.

In the framework of compressive video classification (CVC), we consider that the given video sequence belongs to the class c, where $c \in \{1, \ldots, C\}$. Following a supervised learning approach, a set of training video samples is obtained for each class, $\mathcal{T}_c = \{\mathbf{V}_1^c, \ldots, \mathbf{V}_Q^c\}$. For simplicity, we consider that the number of training samples Q is equal for all the classes. Let also $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_C\}$ denote the overall set of training samples. The CVC problem is stated as follows: *Given a low-dimensional signature of the acquired video sequence, a training dictionary* $\mathbf{D}$*, and a measurement matrix* $\mathbf{\Phi}$*, estimate the correct class* $c \in \{1, \ldots, C\}$.

III. Proposed CVC system

A typical classification system consists of two main phases, namely, a *feature extraction phase*, where a more compact representation of the original information is generated in a low-dimensional space, with the goal of preserving a high discriminative power, and a *classification phase*, where the extracted feature vector of the given signal is compared with the corresponding features of the training samples by means of a suitable similarity criterion resulting in the estimated class. In the following sections, the main characteristics of the two phases are introduced in detail for the proposed CVC system, which is depicted in Figure 1.

A. Feature extraction

When working directly in the CS domain, a suitable signature for a given video sequence is given by (2). However, under the assumption of limited transmission bandwidth, even this representation of reduced dimensionality may be prohibitive for a large number of frames. A further reduction of the transmission cost can be achieved via the fusion of the set of measurement vectors in a single CS feature vector. More specifically, in the following we consider that a *feature vector (or signature)* is assigned to a given video sequence $\mathbf{V}$ as follows,

$$\mathbf{V} \longmapsto \mathbf{f}_{CS} = \frac{1}{R} \sum_{j=1}^{R} \mathbf{g}_j \ . \tag{3}$$

Using the above mapping for the database of training samples $\mathcal{T}$, the following *training dictionary* is formed,

$$\mathbf{D} = [\mathbf{f}_{CS,1}^1, \ldots, \mathbf{f}_{CS,Q}^1, \cdots, \mathbf{f}_{CS,1}^C, \ldots, \mathbf{f}_{CS,Q}^C] \ , \tag{4}$$

where $\mathbf{f}_{CS,l}^c \in \mathbb{R}^M$ denotes the feature vector for the l-th training video of the class c. Similarly, let $\mathbf{f}_{CS,query}$ denote the corresponding feature vector of the query video sequence.

B. Classification methods

Following the feature extraction step, the classification phase is performed by means of an appropriate decision rule. In the following, two categories of decision criteria are employed: the first exploits directly the feature vectors containing the average CS measurements, while the second is based on the solution of a convex optimization problem for the recovery of a sparse class-indicator vector.

Regarding the former category, the simplest decision rule for estimating the optimal class is given by the *nearest-neighbor* (NN) criterion defined by

$$c^* = \underset{c \in \{1, \ldots, C\}, \, \forall \, l}{\arg\min} \| \mathbf{f}_{CS,query} - \mathbf{f}_{CS,l}^c \|_2^2 \ . \tag{5}$$

The second classification scheme, which will be employed in the subsequent evaluation, is the widely used *support vector machine* (SVM), originally designed for binary classification. In our case, the multi-class SVM version is used. More specifically, let $\mathcal{D} = \{\mathbf{f}_{CS,l}^c, c\}_{l=1,\ldots,Q, \, c=1,\ldots,C}$ denote the labeled training data. A way to solve the problem of multi-class classification is to follow a one-against-one approach, where an SVM is constructed for every pair of classes by training it to discriminate them. The number of SVMs to be trained in this approach is equal to $C(C-1)/2$. Let also $(i, \, j)$ be a pair of classes and $d_{ij}(\mathbf{y})$ the associated discriminant function [11]. Then, given the query feature vector $\mathbf{f}_{CS,query}$, if $d_{ij}(\mathbf{f}_{CS,query}) > 0$ a vote is assigned to the i-th class, otherwise the vote is given to the j-th class. The process is repeated for each pair of classes and finally, the class with the *maximum number of votes* is assigned to the query $\mathbf{f}_{CS,query}$. When there are multiple classes with the same maximum number of votes, the class with the maximum value of the total magnitude of discriminant functions (TMDF) is assigned to the query, where the TMDF for the class i is given by

$$\mathrm{TMDF}_i(\mathbf{y}) = \sum_{j=1, \, j \neq i}^{C} |d_{ij}(\mathbf{y})| \ . \tag{6}$$

Regarding the later category of classification methods, an alternative way to estimate the class of the query video is obtained by reformulating the classification problem as a problem of recovering an appropriate sparse vector. More specifically, a class-indicator vector $\boldsymbol{\alpha}$ is introduced, where

$$\boldsymbol{\alpha} = [\alpha_1^1, \ldots, \alpha_Q^1, \ldots, \alpha_1^i, \ldots, \alpha_Q^i, \ldots, \alpha_1^C, \ldots, \alpha_Q^C] \in \mathbb{R}^{CQ} \ .$$

If the query video belongs to the i-th class, in the ideal case we expect that its feature vector $\mathbf{f}_{CS,query}$ will be similar to the corresponding training data of the i-th class, or equivalently to the corresponding columns of the training dictionary $\mathbf{D}$ (cf. (4)). Accordingly, the class-indicator vector has the following Q-sparse structure

$$\boldsymbol{\alpha} = [0, \ldots, 0, \alpha_1^i, \ldots, \alpha_Q^i, 0, \ldots, 0] \ , \tag{7}$$

with the non-zero components corresponding to the Q indices of the i-th class. Thus, the CVC problem is reduced to a problem of recovering the sparse support of $\boldsymbol{\alpha}$, which is expressed as the solution of a convex optimization problem as follows,

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^{CQ}}{\arg\min} \|\boldsymbol{\alpha}\|_1 \ , \ \text{s.t.} \ \|\mathbf{f}_{CS,query} - \mathbf{D}\boldsymbol{\alpha}\|_2 < \epsilon \ . \tag{8}$$

Numerous algorithms have been proposed for the solution of (8). Although the choice of the reconstruction algorithm affects the classification performance, an exhaustive comparison of several CS methods is beyond the scope of this study. Motivated by its simple

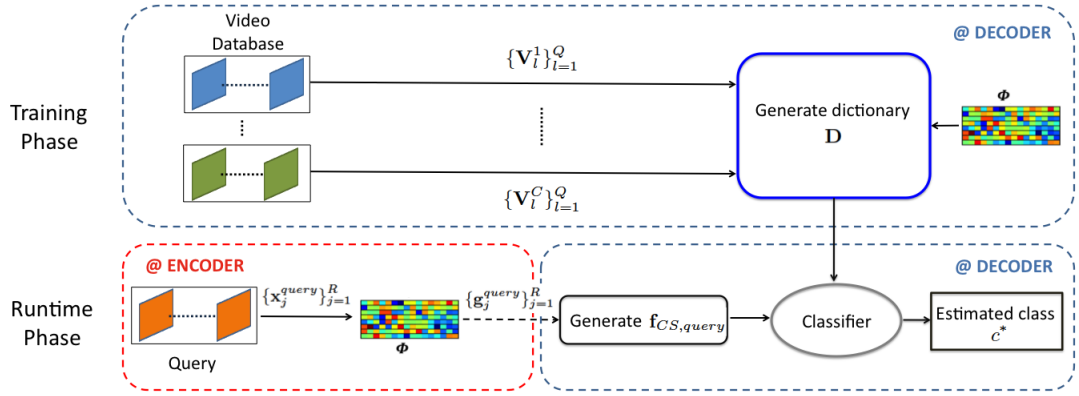


Figure 1. Feature extraction and classification during the training and runtime phases of the proposed CVC architecture.

and fast implementation, the orthogonal matching pursuit (OMP) algorithm[1] [12] is used in the subsequent evaluation.

Notice that in practice, especially in noisy conditions, the recovered class-indicator vector is not exactly Q-sparse. In this case, an additional step is applied to obtain the final class estimate by enforcing the Q-sparsity of $\boldsymbol{\alpha}^*$ as follows,

$$c^* = \arg\min_{c=1,\ldots,C} \|\mathbf{f}_{CS,query} - \mathbf{D}\delta_c(\boldsymbol{\alpha}^*)\|_2 \ , \tag{9}$$

where $\delta_c(\boldsymbol{\alpha})$ denotes the block-Kronecker operator, which sets to zero all the components of $\boldsymbol{\alpha}$ except for these corresponding to the Q indices of the i-th class, as shown in (7).

We emphasize once again that, in contrast to the standard approaches for video classification, the proposed one is based explicitly on the available information in a low-dimensional CS measurements domain, without requiring the acquisition of video frames at full resolution, neither for the training nor for the classification phase. As a result, the significantly reduced processing, storage and transmission costs satisfy the constraints of a decision system with limited resources, which is the main motivation for this study.

IV. Experimental evaluation

In the following, the performance of the proposed CVC method is evaluated and compared with the classification accuracy when using the raw intensity values of the original full-resolution frames. In particular, our database consists of 8 classes from the UCF50 dataset [2], namely, "Basketball", "Billiards", "Playing guitar", "Playing piano", "Rowing", "Rock climbing indoor", "Tennis swing" and "Skiing". This dataset includes videos categorized in classes corresponding to different actions and is particularly challenging due to large variations in camera motion, object appearance and pose, as well as the illumination conditions. Each class consists of 50 video sequences with 50 frames per sequence. A preprocessing step is applied on each frame, by converting into grayscale and downsampling at 128×128 pixels. For each class, we run 50 Monte-Carlo runs, where in each run a different separation of the 50 videos in K training and $50 - K$ testing samples is generated, with $K \in \{12, 24, 36\}$. The classification accuracy is expressed in terms of the average success rate, which is defined by

$$\textit{success rate} = \frac{\textit{number of correctly classified sequences}}{\textit{total number of query sequences}} \ . \tag{10}$$

In addition, a distinct block Walsh-Hadamard (BWHT) measurement matrix $\boldsymbol{\Phi}$ is used in each run, while the sampling ratio M/N varies in $[0.01,\ 0.20]$.

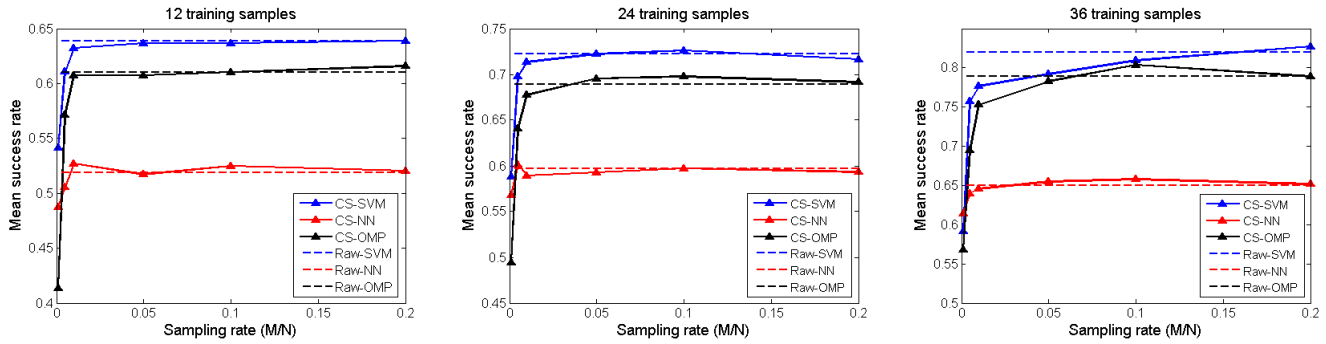
Figures 2(a), 2(b) and 2(c) show the overall success rate averaged over the 50 Monte-Carlo runs and the 8 classes, as a function of the sampling ratio M/N. The constant dashed lines correspond to the classification accuracy when the intensity of the full-resolution frames is used to generate the feature vectors, which is independent from the sampling ratio. Moreover, as mentioned in Section III, three classification methods are compared, namely, the NN, the multi-class SVM, and the sparse reconstruction using the OMP. As it was expected, the classification accuracy increases as the sampling ratio and the number of training samples increase.

Regarding the classification efficiency of the low-dimensional compressed features, when compared with the full-resolution ones, the proposed CVC approach achieves comparable or even superior classification accuracy even for small sampling ratios. In addition, we observe that the performance of the NN method is very close to the performance of the OMP, which is another indication that, according to the restricted isometry property [2] imposed in the theory of CS, the low-dimensional random projections via an appropriate measurement matrix $\boldsymbol{\Phi}$ preserve with high-probability the distances in the original signal space. Finally, in all cases the pairwise voting approach employed by the multi-class SVM appears to be more robust in misclassification errors, resulting in a superior performance when compared with the NN and the OMP, which rather resemble to an one-against-all comparison among the different classes.

Furthermore, Table I shows the confusion matrix between the 8 classes for $M/N = 20\%$ and 36 training samples per class using the multi-class SVM. We observe that in most cases the classification accuracy is relatively high. For example Rowing achieved 100% classification accuracy, while Basketball and Rock climbing where also correctly classified in most cases. Meanwhile in other classes such as Billiards and Tennis are more challenging. One reason for this type of confusion can be attributed to the lack of color information and the use of the luminance component only.

[1] Matlab code: http://www.di.ens.fr/willow/SPAMS/

[2] http://www.computervisiononline.com/dataset/ucf50-action-recognition-realistic-videos



(a) (b) (c)

Figure 2. Total mean success rate as a function of the sampling ratio, for 8 classes and three methods (NN, SVM, OMP) using CS and raw intensity features.

Table I
CONFUSION MATRIX FOR THE SVM METHOD WITH $M/N = 20\%$ AND 36 TRAINING SAMPLES.

	Basketball	Billiards	Guitar	Piano	Rock climbing	Rowing	Skiing	Tennis
Basketball	**90.00**	4.29	0	4.29	0	0	1.43	0
Billiards	11.43	**60.00**	0	11.43	0	1.43	12.86	2.86
Guitar	2.86	4.29	**84.29**	2.86	0	0	5.71	0
Piano	2.86	28.57	0	**60.00**	0	0	5.71	2.86
Rock climbing	1.43	0	0	1.43	**97.14**	0	0	0
Rowing	0	0	0	0	0	**100.00**	0	0
Skiing	5.71	12.86	5.71	2.86	0	0	**72.86**	0
Tennis	0	2.86	7.14	27.14	0	0	0	**62.86**

V. CONCLUSIONS AND EXTENSIONS

In the present work, a compressive video classification method is introduced. More specifically, the design of the proposed CVC system is primarily based on the assumption of limited resources, where the video data are captured directly in the CS domain using a single-pixel camera. A supervised learning approach is followed, where each column of the training dictionary is formed by simply averaging the CS measurement vectors over all the frames of a given training video sequence. Finally, the estimated class is obtained by means of typical classification methods, namely, the NN and the multi-class SVM. An alternative way is also tested, where the classification problem is reduced to a problem of reconstructing a sparse class-indicator vector as the solution of a convex optimization problem. The experimental results revealed that the classification performance is robust to the number of samples captures, where even at very low sampling rates at the order of 1%, significantly smaller than the rates required for solving the problem of sparse reconstruction. Moreover, it was shown that even the simple NN algorithm was quite efficient, achieving a classification accuracy very close to the accuracy of a CS-based sparse reconstruction approach using the OMP.

In the current implementation we do not exploit the sparsity of the original video data in an appropriate transform domain (DCT, DWT), which is at the core of the CS framework. As a direct extension, we expect that the sparsification of the original data before their embedding in the low-dimensional CS domain can enhance the discriminative power of the generated features, and consequently the classification accuracy. This can be done by employing standard linear dimensionality reduction methods, such as the PCA, adapted to the constraint that we work directly with the CS measurements without having access to the original information. Moreover, the generation of CS features using the color information can also increase the classification margin among the several classes. Finally, the property of common sparse support of the class-indicator vector for the video sequences belonging to the same class can be exploited in the framework of group sparse reconstruction, which has been shown recently to achieve a superior reconstruction performance in comparison to the single sparse reconstruction.

ACKNOWLEDGMENT

This work is supported by CS-ORION Marie Curie Industry - Academia Partnerships and Pathways (IAPP) project funded by the European Commission in FP7 (PIAP-GA-2009-251605).

REFERENCES

[1] C. Bishop, "Pattern recognition and machine learning," Springer, 2006.
[2] E. Candès, J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Info. Th.*, Vol. 52, No. 2, pp. 489–509, Feb. 2006.
[3] C. Hegde, M. Davenport, M. Wakin and R. Baraniuk, "Efficient machine learning using random projections," *NIPS Work. on Efficient Machine Learning*, Whistler, Canada, Dec. 2007.
[4] J. Haupt, R. Castro and R. Nowak, "Compressive sampling for signal classification," *in Proc. of 40th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2006.
[5] M. Davenport *et al.*, "The smashed filter for compressive classification and target recognition," *in Proc. IS&T/SPIE Symp. on Electronic Imaging: Computational Imaging*, San Jose, CA, Jan. 2007.
[6] D. Thanou and P. Frossard, "Compressed classification of observation sets with linear subspace embeddings," *in Proc. Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'11)*, Prague, May 2011.
[7] G. Tsagkatakis and A. Savakis, "Face recognition using sparse representations and manifold learning," *in Proc. Intl. Symp. on Visual Computing (ISVC)*, Nevada, NV, 2010.
[8] S. Cotter, "Sparse representation for accurate classification of corrupted and occluded facial expressions," *in Proc. Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'10)*, Dallas, TX, Mar. 2010.
[9] http://dsp.rice.edu/cscamera
[10] T. Do, T. Tran and L. Gan, "Fast compressive sampling with structurally random matrices," *in Proc. IEEE Int. Conf. on Ac., Speech and Sig.Proc. (ICASSP'08)*, Las Vegas, NV, 2008.
[11] R. Duda, P. Hart and D. Storck, "Pattern classification," J. Wiley & Sons, Inc. (2nd ed.), 2001.
[12] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. on Information Th.*, Vol. 53, No. 12, Dec. 2007.