entropy

MDPI

Article

An Analysis of Information Dynamic Behavior Using Autoregressive Models

Amanda Oliveira [1,*], Adrião D. Dória Neto [2] and Allan Martins [3]

1 Center of Exact and Natural Sciences, Federal Rural University of the Semi-Arid Region, Mossoro 59625-900, RN, Brazil
2 Department of Automation and Computer Engineering, Federal University of Rio Grande do Norte, Natal 59078-970, RN, Brazil; adriao@dca.ufrn.br
3 Department of Electrical Engineering, Federal University of Rio Grande do Norte, Natal 59078-970, RN, Brazil; allan@dee.ufrn.br
* Correspondence: amandagondim@ufersa.edu.br or amandagondim@gmail.com; Tel.: +55-84-3317-8398

Received: 20 September 2017; Accepted: 10 November 2017; Published: 18 November 2017

Abstract: Information Theory is a branch of mathematics, more specifically probability theory, that studies information quantification. Recently, several researches have been successful with the use of Information Theoretic Learning (ITL) as a new technique of unsupervised learning. In these works, information measures are used as criterion of optimality in learning. In this article, we will analyze a still unexplored aspect of these information measures, their dynamic behavior. Autoregressive models (linear and non-linear) will be used to represent the dynamics in information measures. As a source of dynamic information, videos with different characteristics like fading, monotonous sequences, etc., will be used.

Keywords: information theory; dynamics process; information potential

1. Introduction

Recent research [1,2] has been studying Information Theoretic Learning (ITL) fundamentals to extend useful concepts of linear and second order statistical techniques to higher order and nonlinear statistical learning. ITL research began in the 1990s and its basic principle is the use of information theory measures estimated directly from the data to replace the conventional statistical descriptors of mean, variance and covariance. For this, it is necessary to measure quantities such as entropy and mutual information from samples, without knowledge of the probability distributions that generated such samples. Thus, the two major tools proposed in the ITL area are the information potential and the correntropy. These quantities indirectly quantify the entropy and cross-entropy of a data source using directly samples without a previous knowledge of their distribution.

The possibility of estimating statistical quantities, such as entropy from the samples, facilitates the study of environments where there are stochastic processes that evolve over time. This fact is very interesting when we remember that there is a strong interaction between the theory that studies dynamic processes and the one that studies probabilities and stochastic models, since the great majority of real systems are dynamic and affected by noise. Thus, stochastic dynamic processes have been gaining increasing interest last decades due to the diversity of problems whose modeling includes some probabilistic aspect.

Using the techniques present in ITL, it is possible to estimate quantities as entropy at a time slot and study its evolution. Besides entropy, it is possible to define other information estimates, such as mutual information between blocks of data at the same time slot. Given this, we can think of defining a time series that involves these measures of information, which consequently brings up some questions:

Are there dynamics in these estimates? Is there any relationship between the estimates and their past values? Is there any temporal structure?

Main researches in ITL include regression, neural network training, classification and clustering [3–10], parameter estimation, system identification and signal processing [11–17] areas.

In all these studies, focus has always been to find system dynamics using minimum error entropy (MEE) as optimization criterion in learning, replacing the traditional mean square error. Works like [18,19] show the robustness of the MEE criterion and the convergence of the fixed-point MEE algorithm, respectively.

In addition, the objective of this work is to investigate the information dynamic measured in the system considering time change perspective and estimate its behavior. It is worth noting that no similar methods were found to this information dynamics modeling in other works, being this, therefore, the contribution of this work. The organization of this article is as follows. In Section 2 some concepts of information theory are presented. In Section 3 we have a description of our testing methodology. In Section 4 the results of the simulations are presented and, finally, the conclusion is given in Section 5.

2. Information Theoretic Learning (ITL)

Information theory is a branch of mathematics that studies information quantification. This theory had its pillars established by Claude Shannon in 1948 that formalized concepts with applications in the theory of communication and statistics and currently have been used in many other areas in which data are the factor of analysis [20].

Briefly, information theory defines measures that aims to quantify the information contained in a given random variable (RV) or even the amount of information that a random variable has on another.

According to [20], learning by information theory (ITL) is a field of study that uses information theory descriptors estimated from the data to replace conventional statistical descriptors such as variance and covariance. Recently, ITL techniques have been used with great success in unsupervised learning problems. The great success of ITL use is mainly due to the fact that it is a technique based on high order statistics, since its descriptors are estimated directly from data samples.

Entropy is the best-known descriptor of information theory. According to Shannon, the entropy of a set X is the sum of the uncertainties of all messages weighted by the probability of each one of them. According to [21], the Shannon entropy for a continuous random variable X is given by Equation (1):

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) \, \mathrm{d}x \tag{1}$$

where $p(x)$ is the probability density function (pdf) of RV X in question.

An important feature of the measure is that the combination of uncertainties is weighted by its probabilities, which makes up the essence of the concept of entropy. In this sense, unlikely events have much information but are discounted by their rare occurrence in the calculation of the product [22].

In 1950, Alfred Renyi [23] introduced a parametric family of entropies as a mathematical generalization of the Shannon entropy known as Renyi entropy of order α, presented in Equation (2):

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} p^\alpha(x) \, \mathrm{d}x \tag{2}$$

The logarithm argument in Equation (2) is known as Information Potential (IP), which is an information measure also widely used today [20]. Thus, the IP of the RV X is computed by:

$$V_\alpha(X) = \int_{-\infty}^{\infty} p^\alpha(x) \, \mathrm{d}x \tag{3}$$

The quadratic Renyi's entropy $(H_2(X))$ shows particular interest due to the possibility of estimating the probability density function of X directly from the samples through the use of a nonparametric method such as the Parzen windows. Equation (4) defines $H_2(X)$:

$$H_2(X) = -\log(\int_{-\infty}^{\infty} p^2(x)\,\mathrm{d}x) \tag{4}$$

The estimate of $pdf(X)$, according to the method of Parzen [24], using an arbitrary kernel function $K_\sigma(.)$ is given by:

$$\hat{p}_X(x) = \frac{1}{N}\sum_{i=1}^{N} K_\sigma\,(x - x_i) \tag{5}$$

where σ is the *kernel size* and $\{x_1, x_2, x_3, ..., x_N\}$ are the N samples of the random variable X.

By performing some algebraic manipulations from Equations (4) and (5) and assuming a Gaussian kernel $G_\sigma(.)$ with standard deviation σ, we obtain a new equation for the estimation of the quadratic entropy of Renyi and consequently for the information potential, presented in Equations (6) and (7):

$$\hat{H}_2(X) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i)\right) \tag{6}$$

$$\hat{V}_2(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i) \tag{7}$$

where

$$G_\sigma(\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} exp - \frac{||\mu||^2}{2\sigma^2}$$

and d is the dimension of the RV X.

According to [9], Equation (6) represents one of the main results of learning by information theory, since it shows that IP can be estimated directly from the samples with an exact calculation of the integral over the random variable for Gaussian kernels.

3. Information Dynamics

The term stochastic processes is used to describe the evolution in time of a random phenomenon according to probabilistic laws [25]. In a way, they represent an extension of the concept of random variable for dynamic processes.

According to [26], a stochastic process is a family of random variables that evolve over time. That is, a stochastic process X is a set $\{X_t : t \in T\}$, where each X_t is a random variable and T is a fully ordered set representing the time. Figure 1 presents a general scheme of a stochastic process $X(t)$, where each $X(t_i)$ represents a random variable and each ξ_j represents the result of a random experiment.

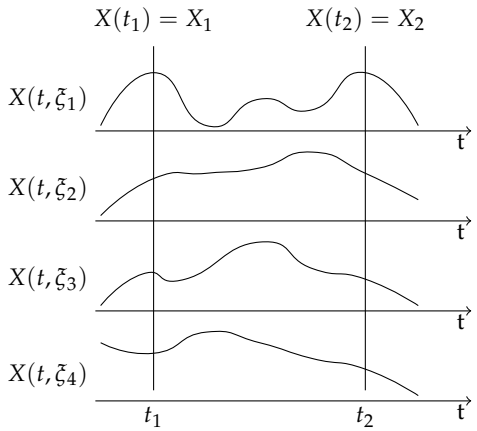

Figure 1. Stochastic Process General Schema.

With the evolution of time, the statistical properties of the process, such as mean, variance and probability density function (pdf) can undergo changes. As with conventional statistical measures,

process information measures, such as entropy and information potential, also change with time. This is because these descriptors measure the information contained in random variables and these variables are evolving over time.

The aim of this work is to investigate the way these measures of information change over time, that is, to discover the dynamics that rule the behavior of information. Dynamic behavior of information is understood as the variation of time space of information, quantified by measures of information theory.

A fairly common example of stochastic environments are videos. Consider the following scenario: A video with n frames, where there are w pixels with color information for each video frame present in the image that compose it. These pixels can be considered as samples of a given X random variable that evolves over time. Thus, this scenario gives us a data source of the X random variable at each time slot, which allows us to estimate different measures of information for each slot, and IP among those measures. Figure 2 presents a schematic for this scenario, where video frames represent the random variables in a stochastic process.

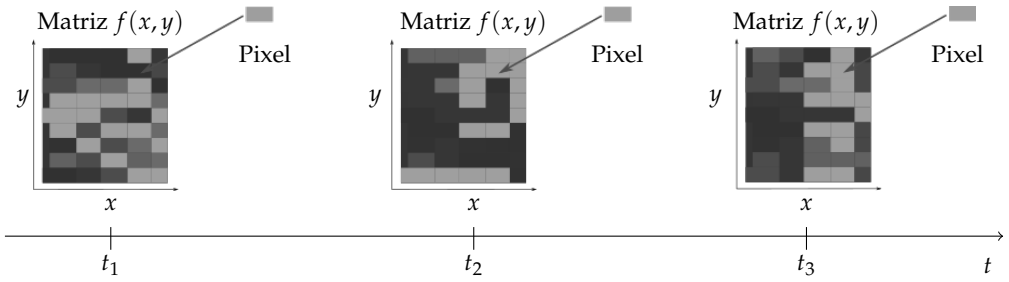

Figure 2. Video as an Example of a Stochastic Process.

In this context, the information contained in each video frame changes with time and one must investigate if there is any dynamics that rules this change.

Let $\mathbf{h}[k]$ be a vector with information measures of the random variable X of a stochastic process. The existence of dynamics in the information behavior of this RV is conditioned to find a function f such as

$$\mathbf{h}[k+1] = f(\mathbf{h}[k], u[k]) \quad k = 0, 1, 2, ...$$

where k is the time and u is called input. In this case, the vector $\mathbf{h}[k]$ is a state vector constructed from information in random variable X.

The methodology used in the tests was based on the analysis of information dynamics using videos as examples of dynamic environments. As seen in Section 2, there are several descriptors in information theory that measure the information of a given random variable and therefore could be used to compose the vector $\mathbf{h}[k]$ in this work. For convenience, we will consider that $\mathbf{h}[k]$ is one-dimensional and we will choose the information potential ($V_2(.)$) as our measure of information hereafter. For simplicity, we will call $V_2(.)$, $V(.)$.

The initial procedure of the simulations consisted in obtaining a sequence with the IPs of each frame of the video. The calculation of the IP was performed per Equation (7). From there, it became necessary to use some technique capable of finding the existing dynamics in the data. There are many techniques for this purpose, depending on the system characteristics. In this initial modeling, we chose to make the identification and prediction of the systems through a NARX-NN (Nonlinear Autoregressive with Exogenous Inputs Neural Network). The NARX-NN model is a recurrent dynamic network that is commonly used in time-series modeling. The analogy between the ARIMA model and the NARX-NN model is clear as both models use previous samples of the output signal and previous values of an independent input signal to compute its output $y[k]$ [27]. Equation (8) presents the definition for the NARX-NN model:

$$y[k] = f(y[k-1], y[k-2], \ldots, y[k-N], u[k], u[k-1], u[k-2], \ldots, u[k-Q]) \tag{8}$$

In this work, the signal $y[k]$ will be represented by the information potential of the frame at time k. In addition, we do not use exogenous inputs in the NARX modeling, i.e., the next value of the output signal $y[k]$ is regressed on previous values of the output signal. In other words, for $N = 4$, we have:

$$\hat{V}[k] = f(\hat{V}[k-1], \hat{V}[k-2], \hat{V}[k-3], \hat{V}[k-4]) \tag{9}$$

Figure 3 presents a schematic of the NARX-NN architecture using IP as the input signal and without using exogenous input.

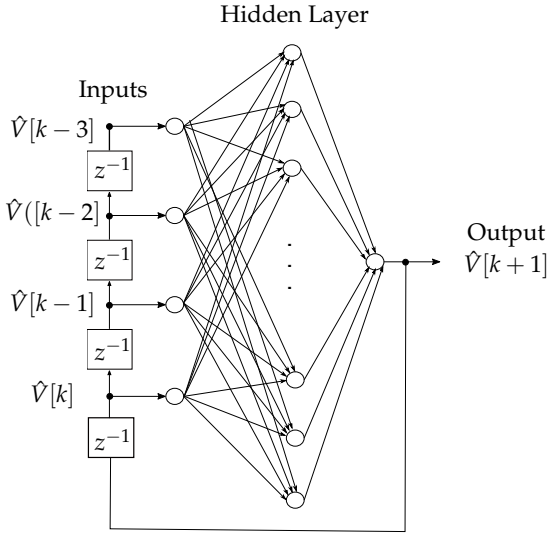

Figure 3. Schematic of the Nonlinear Autoregressive with Exogenous Inputs Neural Network (NARX-NN) architecture.

For the simulation of the NARX network the sigmoid (φ) was used as a function of the neurons activation. Considering that the network has q neurons in the input layer and p neurons in the hidden layer, the output $\hat{V}[k+1]$ is expressed in Equation (10):

$$\hat{V}[k+1] = \sum_{j=1}^{p} \varphi(\sum_{i=1}^{q} w_{ij}\, \hat{V}[k+1-i] + b_j) \tag{10}$$

where b_j is bias, w_{ij} is the weight between neurons i and j and

$$\varphi(v) = \frac{1}{1 + exp(-v)}$$

The Levenberg-Marquardt algorithm was used for network training and the mean square error (MSE) as performance function in order to find the vector of weights for the neural network that presents the minimum mean square error. Let $\boldsymbol{w}$ be the vector with the weights of the neural network, the MSE of $\boldsymbol{w}$ is defined as follows:

$$MSE(\boldsymbol{w}) = \underset{\boldsymbol{w}}{\operatorname{argmin}}\ \mathbf{E}\{(\hat{V}[k] - V[k])^2\}$$

where the symbol **E** denotes the operation of expected value or mathematical hope.

Analyzes of the error autocorrelation function (e) were carried out, in order to evaluate the model suitability to the systems tested. The network error (e) is the difference between the output estimated by the NARX network ($\hat{\mathbf{V}}$) and its real value ($\mathbf{V}$). According to [28], the autocorrelation function

measures the correlation between $e[k]$ and $e[k+l]$, where $l = 0, ..., L$ and $e[k]$ is a stochastic process. The definition of autocorrelation for lag l is given by

$$r_l = \frac{c_l}{c_0}$$

where c_0 is the sample variance of the time series, $\bar{e}$ is the sample mean of the $e[k]$ and

$$c_l = \frac{1}{K-1} \sum_{k=1}^{K-l} (e[k] - \bar{e})(e[k+l] - \bar{e})$$

The output of the NARX-NN, after training, will present a prediction of the next IP value in function of its past values, proving the presence of information dynamics in the video over time. The results obtained from the experiments are presented in Section 4.

4. Results and Discussion

This section presents the results obtained in experiments with three videos with different characteristics. The first video presents fairly similar frames over time, without large color variations; The second video is a very colorful animation with frames with quite random behavior and the third video is a commercial with effects of fadein and fadeout. The IP behavior was analyzed in each of these videos, highlighting the peculiarities of each one of them. In addition, we also performed an analysis of the existence of information dynamics of these videos over time with the help of an autoregressive neural network.

4.1. Experiment with Video 1—Constant IP

The first analysed video is an animation where a fourier machine is simulated. The video features 300 frames at an update rate of 25 frames/s. At Figure 4, the image of one of the frames of this mentioned video.

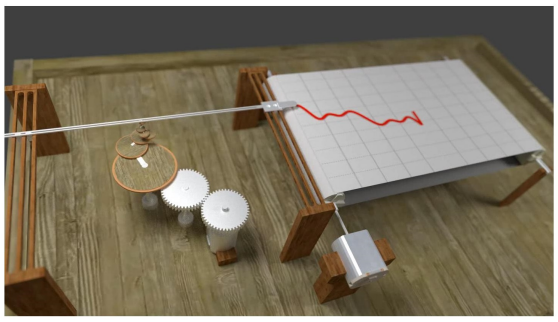

Figure 4. Video Frame 1—Fourier Machine.

As the video only simulates plotting the red curve over time, its frames are quite similar, without major color changes as time passes. Figure 5 shows the behavior of IP for video 1.

Note that the IP values are practically constant over time. This behavior was already expected because, as previously mentioned, the information potential of each time slot is calculated as a function of RGB values found in each pixel of the frame, which do not change with time. Figure 6 shows a sequence of video frames with their respective RGB clouds.

Note in Figure 6 that all clouds are quite similar in shape. This is due to the fact that the images along the video are practically the same, differing only in the format of the red curve. This repetitive color pattern is what makes the IP values constant in Figure 5.

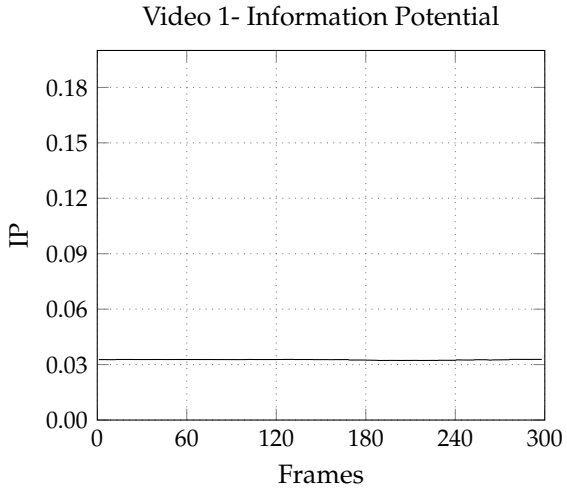

Figure 5. Information Potential (IP) behavior throughout video 1.

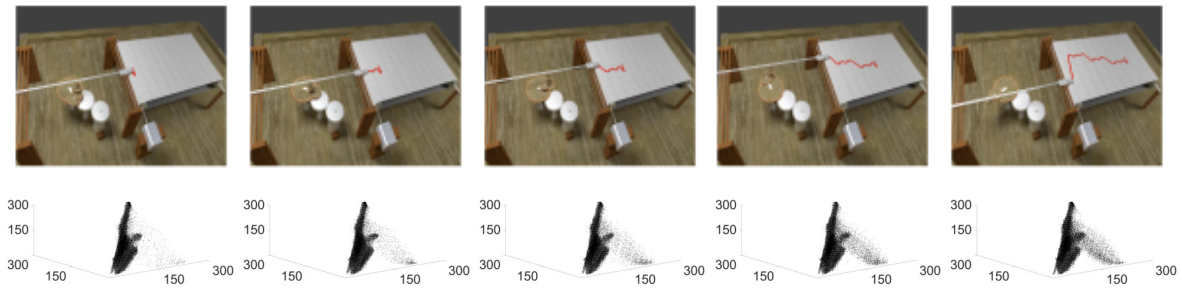

Figure 6. Sequence of frames and RGB graphics of video 1.

4.2. Experiment with Video 2—Dramatic Changes in IP

The second video [29] analyzed is a very colorful animation with 3000 frames and an update rate of 25 frames/s. Figure 7 shows the image of one of the frames of the mentioned video.

Figure 7. Frame from video 2. Source: El Espantapájaros—Estudios de Animacion ICAIC.

As previously mentioned, the information potential of each time slot is calculated as a function of RGB values found in each pixel of the frame. Figure 8 shows the behavior of the IP for video 2. Note that IP values have a random behavior over time and that in addition, remains in values below 0.03 in the vast majority of frames.

An important fact to notice is the presence of peaks in IP values at some points in the video. This is justified by a dramatic change in color pattern presented throughout the video. In Figure 8,

two points were highlighted because they present a substantial change in the information potential. Point 1 is located in frame 735, where a small sequence of almost black frames occurs, while point 2 is located in frame 1250 and also presents a sequence of video frames darker than the average.

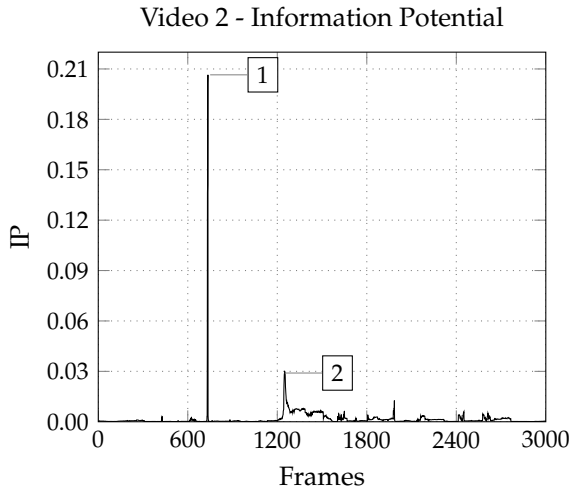

Figure 8. IP behavior throughout video 2.

The sequences of frames with their respective RGB graphs values, referring to points 1 and 2, can be observed in Figures 9 and 10.

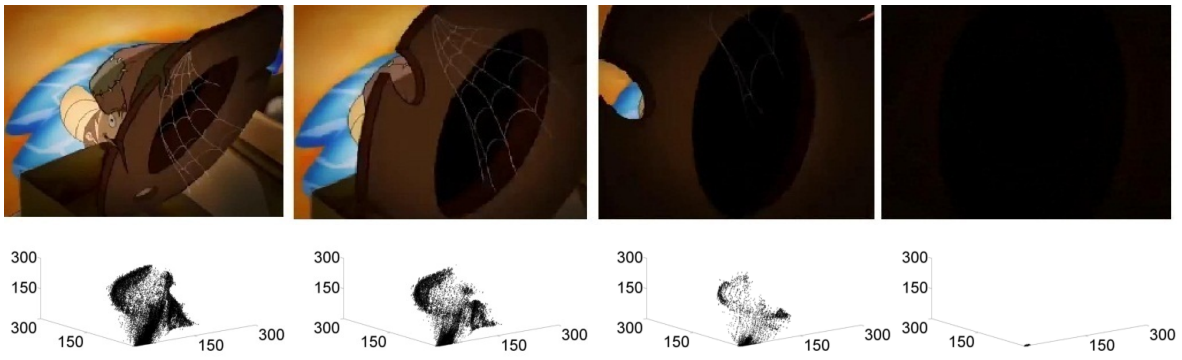

Figure 9. Sequence of frames and RGB graphics at point 1.

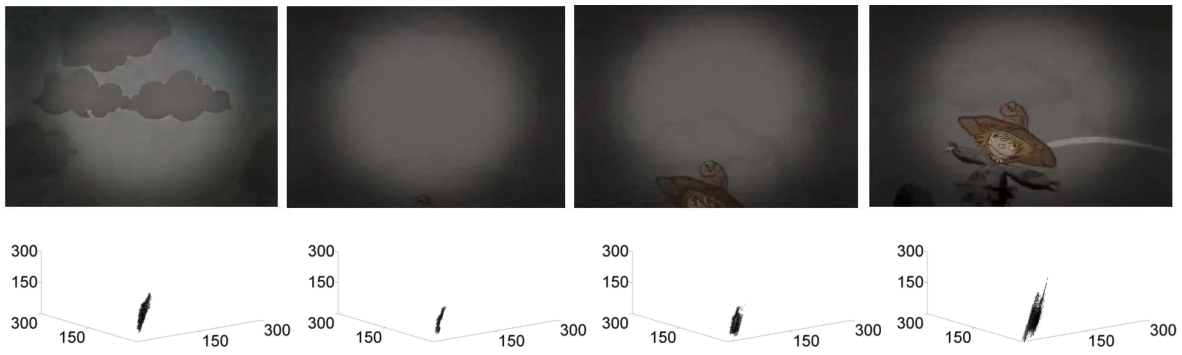

Figure 10. Sequence of frames and RGB graphics in point 2.

Notice in Figure 9 that the RGB cloud referring to the first frame is quite sparse, reflecting the color of it. However, as the video presents darker frames, the RGB cloud tends to decrease, getting its

points very close to the origin of the axes. It is at this moment that the peak represented by point 1 happens, since the IP of a random variable is inversely proportional to the data variance.

Observe in Figure 10 that all clouds are compact with very similar format. This is justified by the greyish color pattern that prevails throughout the sequence shown in Figure 10. These dark tones in the frames make the IP value higher than the average because the information is being calculated per RGB values of each pixel.

The video IP's sequence shown in Figure 8 was used in NARX training network and the test results are presented in Figure 11. From the tests, we can perceive that the neural network was able to carry out the data prediction in a very acceptable way, presenting a sum of the mean squared errors of 9.61×10^{-5}.

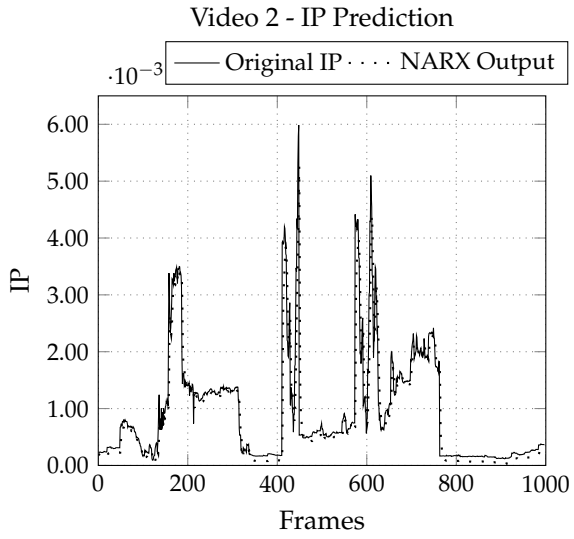

Figure 11. NARX tests result—Video 2.

From the error autocorrelation function, shown in Figure 12, we can see that the network was able to identify an over time dynamics present in the system information. The presence of dynamics confirmation in data is shown as a very interesting result, since it allows us to use it as information in compression or analysis in video processes.

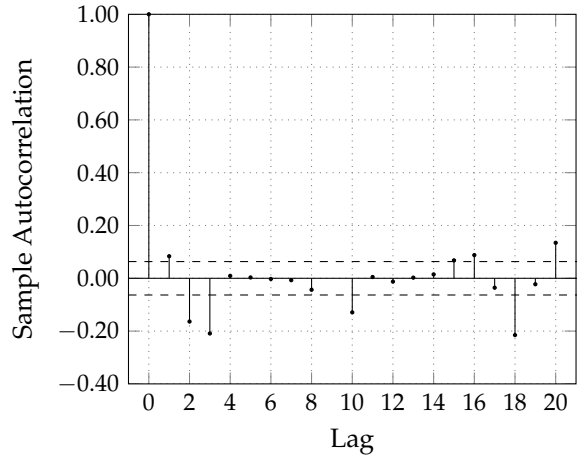

Figure 12. Video 2—Autocorrelation Function.

4.3. *Experiment with Video 3—Fadeout and Fadein Effects*

The third video used in tests is a commercial video with 420 frames and a 9 frames/s rate. Figure 13 shows the image of one of the frames of video 3.

Figure 13. Video 3 Frame—Source: Funny Commercials Part 13—Canal Chuva na Nuca.

Each frame IP behavior over time can be observed in Figure 14. For video 3, we can observe a greater variation in IP values over time when compared to the video previously tested. Note that IP values that are much higher than the average at the beginning and at the end of the video are presented, where there are fadeout and fadein effects, respectively, presenting almost totally black frames in both cases.

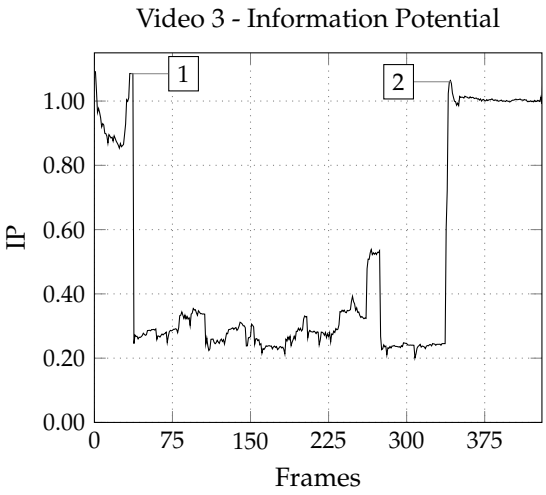

Figure 14. IP behavior throughout video 3.

In Figures 15 and 16 we can see frame images with fadeout and fadein effects in video 3, as well as their respective RGB values graphs, represented by points 1 and 2 in Figure 14.

From Figure 15 we can observe RGB cloud pixels concentrating near the axes origin as the frame becomes darker due to Fadeout effect, causing the peak in IP value represented by point 1 in Figure 14. In Figure 16, we can see the opposite effect, where the initially concentrated RGB cloud gradually expands as the fadein effect occurs.

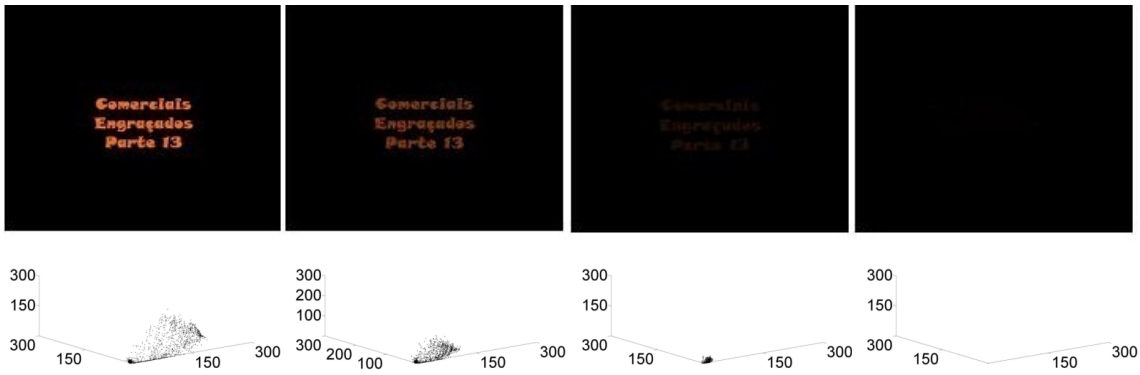

Figure 15. Frames Sequence and RGB Graphics Sequence—Point 1: Fadeout effect.

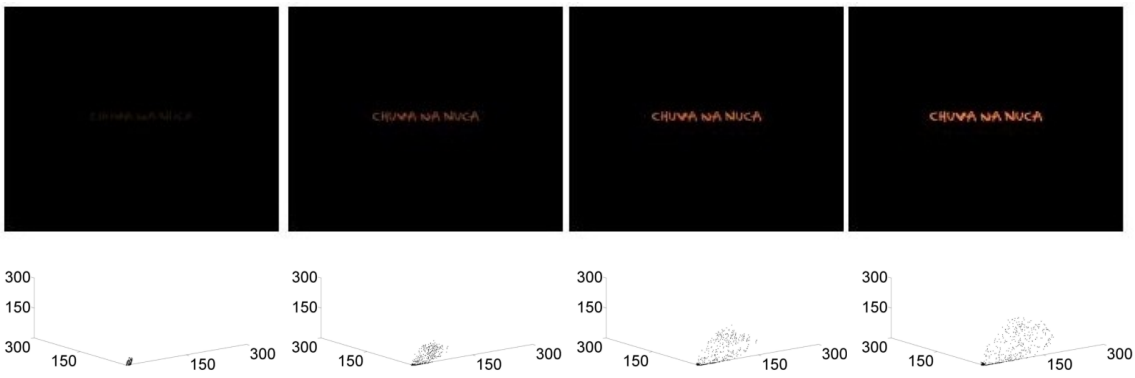

Figure 16. Frames Sequence and RGB Graphics Sequence—Point 2: Fadein effect.

The results of tests with the NARX network are presented in Figure 17. From the tests, we can see that the neural network proved to be efficient in predicting data, presenting a sum of mean squared errors of 3.41×10^{-1}.

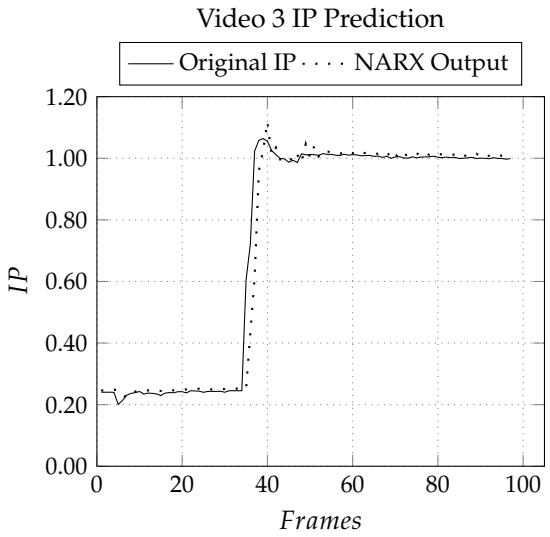

Figure 17. NARX Tests Result—Video 3.

Figure 18 shows the autocorrelation function of the residue between the calculated IP and the estimated one through the network. Analogous to the second video, this graph leads us to again conclude that there is also a dynamic that is present at third video which the NARX network was

able to identify it. Despite the good result, we can notice through the initial lags that the network configuration was not able to extract all the dynamics present in data, suggesting an adequacy in its parameters to obtain better results.

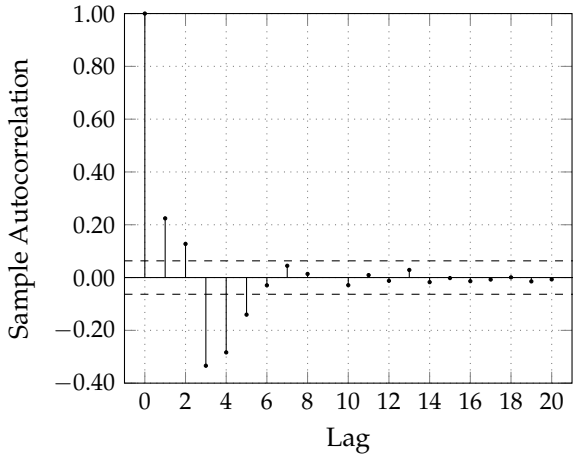

Figure 18. Autocorrelation Function—Video 3.

5. Conclusions

In this article we performed an analysis on dynamic information behavior over time. The experiments were carried out using videos as examples of real dynamic processes. Three case studies were carried out, and in all of them, a statistical magnitude of information theory, known as information potential, was estimated for every time slot and the behavior of this measure over time was analysed, taking into account the different characteristics of each video. In addition, analyses were performed through an autoregressive neural network with the purpose of generating a dynamic behavior of the information present in the systems. This result was very interesting because it showed that the NARX network was able to identify and measure the information dynamics present in data in an appropriate way. Moreover, it presents itself in a promising way, since it opens a range of possibilities for applications in the area of information theory within the context of dynamic systems. Among the several possible areas for applications, we can highlight image monitoring systems, where alerts are triggered by significant changes in information content. We can also suggest the use of information as an additional variable in dynamic systems modeling by means of traditional techniques.

Author Contributions: Amanda Oliveira conducted the experiments, presented the illustrative example and wrote the paper. Alllan Martins e Adriao Duarte Doria Neto polished the language and were in charge of technical checking. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, W.; Pokharel, P.; Principe, J. Correntropy: A Localized Similarity Measure. In Proceedings of the IEEE International Joint Conference on Neural Networks, Vancouver, BC, Canada, 16–21 July 2006; pp. 4919–4924.
2. Santamaria, I.; Pokharel, P.; Principe, J. Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization. *IEEE Trans. Signal Process.* **2006**, *54*, doi:10.1109/TSP.2006.872524.
3. Miranda, V.; Santos, A.; Pereira, J. State Estimation Based On Correntropy: A Proof of Concept. *IEEE Trans. Power Syst.* **2009**, 24, 1888–1889.
4. Hild, K.; Erdogmus, D.; Torkkola, K.; Principe, J. Feature Extraction Using Information-Theoretic Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1385–1392.
5. Araújo, D.; Dória, A.; Martins, A. Information-Theoretic Clustering: A Representative and Evolutionary Approach. *Expert Syst. Appl.* **2013**, *40*, 4190–4205.

6. Shimoji, S.; Lee, S. Data Clustering with Entropical Scheduling. In Proceedings of the 1994 IEEE International Conference on IEEE World Congress on Computational Intelligence, Orlando, FL, USA, 28 June–2 July 1994; pp. 2423–2428.
7. Rao, S.; Martins, A.; Príncipe, J.C. Mean Shift: An Information Theoretic Perspective. *Pattern Recognit. Lett.* **2009**, *30*, 222–230.
8. Gokcay, E.; Principe, J. Information Theoretic Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 158–171.
9. Principe, J.C.; Xu, D.; Zhao, Q.; Fisher, J.W., III. Learning from Examples with Information Theoretic Criteria. *J. VLSI Signal Process.* **2000**, *26*, 61–77.
10. Steeg, G.V.; Galstyan, A.; Sha, F.; Simon, D. Demystifying Information-Theoretic Clustering. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
11. Wu, Z.; Peng, S.; Chen, B.; Zhao, H.; Principe, J. Proportionate Minimum Error Entropy Algorithm for Sparse System Identification. *Entropy* **2015**, *17*, doi:10.3390/e17095995.
12. Erdogmus, D.; Principe, J. An Error-Entropy Minimization Algorithm for supervised training of Nonlinear Adaptive Systems. *IEEE Trans. Signal Process.* **2002**, *50*, doi:10.1109/TSP.2002.1011217.
13. Ren, M.; Zhang, J.; Fang, F.; Hou, G.; Xu, J. Improved Minimum Entropy Filtering for Continuous Nonlinear Non-Gaussian Systems using a Generalized Density Evolution Equation. *Entropy* **2013**, *15*, 2510–2523.
14. Liu, W.; Pokharel, P.; Principe, J. Correntropy: Properties and Applications in Non-Gaussian Signal Processing. *IEEE Trans. Signal Process.* **2007**, *55*, doi:10.1109/TSP.2007.896065.
15. Chen, B.; Xing, L.; Liang, J.; Zheng, N.; Principe, J. Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion. *IEEE Signal Process. Lett.* **2014**, doi:10.1109/LSP.2014.2319308.
16. Izanloo, R.; Fakoorian, S.; Yazdi, H.; Simon, D. Kalman Filtering Based on the Maximum Correntropy Criterion in the Presence of Non-Gaussian Noise. In Proceedings of the Annual Conference on Information Science and Systems (CISS), Princeton, NJ, USA, 15–18 March 2016.
17. Vaidya, U.; Sinha, S. Information Based Measure for Influence Characterization in Dynamical Systems with Applications. In Proceedings of the American Control Conference (ACC), Boston, MA, USA, 6–8 July 2016.
18. Chen, B.; Xing, L.; Xu, B.; Zhao, H.; Principe, J. Insights Into the Robustness of Minimum Error Entropy Estimation. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, doi:10.1109/TNNLS.2016.2636160.
19. Zhang, Y.; Chen, B.; Liu, X.; Yuan, Z.; Principe, J. Convergence of a Fixed-Point Minimum Error Entropy Algorithm. *Entropy* **2015**, *17*, 5549–5560.
20. Principe, J.C. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*; Springer: Berlin/Heidelberg, Germany, 2010; p. 462.
21. Chen, B.; Wang, J.; Zhao, H.; Principe, J. Insights into Entropy as a Measure of Multivariate Variability. *Entropy* **2016**, *18*, 196.
22. Martins, A.; Duarte, A.; Dantas, J.; Principe, J.C. A New Clustering Separation Measure Based on Negentropy. *J. Control Autom. Electr. Syst.* **2015**, *26*, 28–45.
23. Rényi, A. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1960.
24. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.
25. Haykin, S. *Adaptative Filter Theory*; Pearson: London, UK, 2013.
26. Campos, M.; Rêgo, L.; Mendonça, A. *Métodos Probabilísticos e Estatísticos com Aplicações em Engenharias e Ciências*; Grupo Gen-LTC: Rio de Janeiro, Brazil, 2016.
27. Fleifel, R.; Soliman, S.; Hamouda, W.; Badawi, A. LTE Primary User Modeling Using a Hybrid ARIMA/NARX Neural Network Model in CR. In Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, USA, 19–22 March 2017.
28. Box, G.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis: Forecasting and Control*; Prentice Hall: Upper Saddle River, NJ, USA, 1994.
29. El Espantapájaros. Available online: https://www.youtube.com/watch?v=sGI7mm9fKrM (accessed on 15 April 2017).