FOUNDATIONAL METADATA FOR IMAGE BASED COGNITION

Gary L. Viviani

Insitu (A Boeing Company)
Bingen, WA 98605, USA
e-mail: (gary.viviani@insitu.com)

Edward J. Delp

Video and Image Processing Laboratory (VIPER)
School of Electrical and Computer Engineering
Purdue University, West Lafayette, Indiana, USA

ABSTRACT

The need to create useful information from full motion video gathered by drones is a significant motivation for devising methods to approximate human cognitive behaviors. Additionally, the regulatory needs associated with drone systems has spawned the requirement to be able to confirm, or audit, the activities of such devices. A conditional approach, as compared with a generalized video processing environment, is presented that associates practical and realistic constraints to simplify the problem of finding useful information from video acquired by a drone into something that is tractable and consistent with real-world requirements. A primary contribution of this paper is to introduce the concept of continuous cognition from a theoretical perspective, followed by a practical application derived from an operational system.

Index Terms— Metadata, Statistical Pattern Recognition, Multi-dimensional probability density functions, orthonormal series approximations

1. INTRODUCTION

This paper provides a means for enhanced creation and utilization of full motion video by examining the need of encoding the video with critical information (metadata) [1, 2]. It is not realistic to expect to be able to derive a complete understanding of a particular scene of interest without additional metadata (or side information) that provides necessary insight into how the video was acquired [3]. Examples of metadata include location, altitude, speed and heading. The nature of this insight is explored in this paper. The amount of additional effort to encode the metadata with the video is trivial compared to the added value, which comes in essentially two forms:

- The metadata confirms continuous custody (or not) of the object of interest of the video; *i.e. on a frame to frame basis if target information is invariant, the metadata confirms that the object of interest remains in the field of view*.
- The metadata provides insight into the context of the video itself, which is often lost after the initial creation; *i.e. on a frame to frame basis, metadata would typically provide location and other information that provides useful context that is often lost when video is archived for later use [4]*.

While these two capabilities may appear to be obvious, we are not aware of any applications that actually accomplish them. Importantly, creation of related metadata must be accomplished in real-time at the point of video creation. In practice, this additional complexity allows for an important requirement that is routinely neglected [5].

To better perceive this concept, think in terms of conventional control system concepts that have foundations in controllability and observability [6]. To see this relationship, recognize that if there is an automatic or manually operated camera in the drone, then both the pose and location of the platform is almost as important as the image itself [3]. Moreover, the most important element is the subject of the image from the perspective of the imaging device. Generally, imaging devices do not provide random scenes and orientations. Hence, it is important to recognize the imaging process as a system that is subject to the governing scientific principles of mathematical control. If in fact purposeful images are being constructed, then this necessitates controllability and observability relationships, which can be exploited for cognitive and analysis purposes. Roughly, the greater the ability to understand the cognitive properties of a video, the greater the ability to derive a relationship involving controllability and observability of the object of interest and the associated pixels. Such capabilities improve the basic data gathering, as well as the overall operation of the system, as will be seen.

Using this approach we are able to create a video based representation that is both controllable and observable in a classical sense. This capability is the essence of what is is known as continuous cognition. Formalizing the space in which a system is both controllable and observable has the desired attribute of extending the usefulness of such a system, especially in systems which are subject to continuous operations [7]. Alternatively, this can be thought of as a means for continuously conveying information versus just pixel data. One tangible manifestation of the benefits of such an approach are exemplified in Figure 4, where full scale hands-off flight operations (pilotless) were achieved and the system performed self-observability and controllability with no human involvement for an extended period of time through various environmental conditions. Such a capability is a precursor to full-scale autonomy combined with a completely trustworthy device capable of providing information containing objects of interest and associated meta data with no human interaction.

2. COGNITION OVERVIEW

The fact that a human is continually "aware" of their surroundings is a diverse subject [8, 9, 10]. We have no interest in broadly discussing the more general aspects of the human visual system. However, by focusing on a subset of human cognitive tasks, important and pertinent results are achievable. We claim that human based continuous cognition encapsulates what is essentially the real-time ability to combine sensory inputs and decision making in order to maintain a satisfactory and safe existence. For practical reasons, we will apply the concept of "continuous cognition" to a generalizable limited subset of human capabilities. In this paper, continuous cognition indicates that a robot (or human) is able to continuously confirm

critical object of interest parameters such as identity and location. This is broadly applicable, yet limited as compared with all human situations of interest, in that continuous cognition (robotic) is both feasible and achievable, as will be demonstrated. We take advantage of the fact that when electronic images are under consideration there is typically a purpose behind the motivation to "gather and collect" light or other radiation and assemble it in the form of arrays of pixels. We will call this information gathering robot an InfoBot. This can be thought of as a machine or an algorithm as shown in Figure 1. The term "target" will be applied to the primary purpose or subject matter for which the radiation (light) is being gathered (see [11] page 17 for example).

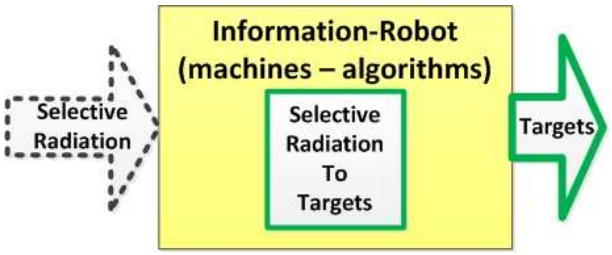


Fig. 1. InfoBot - A Machine (algorithm) that converts Radiation to Objects of Interest (Targets)

2.1. Illustration for an InfoBot

The combination of a manned aircraft and its pilot operating with no dependencies on any navigational aids - so called visual flight rules - can be thought of as an illustration of the concept of an InfoBot. The pilot relies on his ability to gather radiation with his eyes and convert recognizable objects of interest into a means for self-navigating from point to point. This is a particular type of InfoBot yet it encompasses the key elements of a system that is both controllable and observable. Because of essentially unencumbered freedom of motion and the ability to point their eyes at whatever object is of interest, the system is both controllable and observable and the stream of "video" that enter the pilot's brain is very purposeful. For more generalized problems of interest that would be associated with an InfoBot, the ability to both:

- **auto-recognize** - maintain an audit trail confirming identity of an object of interest (can be thought of as memory of recognizable temporal and spatial observations for associated objects of interest),
- **auto-navigate** - maintain a meaningful object of interest in the field of view

is foundational in nature. Hence, it is reasonable to expect that useful video information must also be able to provide such capabilities in order to at least be minimally effective at providing purposeful information (at least for a large subset of generalized targets of interest).

For real problems of interest, the associated augmented metadata of interest can be described as follows:

$$PIK = [\mathbf{T_{ID}}, \mathbf{P}, L, W, \gamma, \alpha, DATE, \mathbf{T_R}]$$

where

PIK - Precision Information Kernel

$\mathbf{T_{ID}}$ - is the metric which determines the identity of the object of interest

$\mathbf{P}$ - is the location of the center of the field of view, which corresponds to the object of interest (note: location of the platform can equally well be expressed in terms of the location of the object of interest)

L, W - is the size of the object of interest

γ - the associated dimensions of the array of pixels for the indicated range

α - Angle of the pose vector relative to the airborne vehicle

$DATE$ - is a time stamp

$\mathbf{T_R}$ - is the separation (range) between the object of interest and the vehicle

In the context of what we described above, the usefulness of the PIK metadata is apparent. It becomes the useful "commentary" for a frame by frame description of key cognitive considerations in the video which are equivalent to generalized auto-recognition and auto-navigation.

So the key problems to achieve such capabilities reduce to determining:

- $\mathbf{T_{ID}}$
- γ

To achieve desired system level functionality all elements of the PIK are required. The next section will describe means for determining a metric for $\mathbf{T_{ID}}$. Other elements of the PIK, which are less challenging, will be omitted.

3. DETERMINING PIK PARAMETERS

Previously, we postulated that by determining the significant statistics for a sub-image associated with a target of interest, there would be a reasonable representation for creating a relative signature [12], [13], [14]. Such a stochastic process is a fundamental characteristic for information flow associated with specific $\mathbf{T_{ID}}$'s and as such their ought to exist a suitably chosen discriminator.

The MGS (Multivariate Gram-Charlier Series) discussed in[15] is a suitable means in its multi-dimensional form for representing a probability density function (p.d.f). (statistically) based relative signature for the $\mathbf{T_{ID}}$'s of interest [12],

$$f(\vec{X}) = \sum_{s_1=0}^{\infty} \cdots \sum_{s_n=0}^{\infty} E[\prod_{i=1}^{n} H_{s_i}(x_i)] \prod_{p=1}^{n} \frac{H_{s_p}(x_p) G(x_p)}{s_p!} \quad (1)$$

where

$$G(x) = \frac{1}{\sqrt{2\pi}} exp^{(-\frac{x^2}{2})}.$$

The Hermite polynomials are defined implicitly by

$$exp(tx - \frac{t^2}{2}) = \sum_{i=0}^{\infty} t^i \frac{H_i(x)}{i!}.$$

We initially concern ourselves with a vector $\vec{X} \, \epsilon \, \mathbf{R^2}$ such that we have the following expression for a 2-D p.d.f. that is likely to be associated with planar properties of interest,

$$f(x_1, x_2) = \sum_{s_1=0}^{\infty} \sum_{s_2=0}^{\infty} \frac{E[H_{s_1}(x_1) H_{s_2}(x_2)]}{s_1! s_2!} \times H_{s_1}(x_1) H_{s_2}(x_2) G(x_1) G(x_2).$$

The first 6 Hermite polynomials are expressed as:

$$H_0(x) = 1$$

$$H_1(x) = x$$

$$H_2(x) = x^2 - 1$$

$$H_3(x) = x^3 - 3x$$

$$H_4(x) = x^4 - 6x^2 + 3$$

$$H_5(x) = x^5 - 10x^3 + 15x$$

and E[ ] represents the probability space of the Expected Value Operator.

4. PIK ALGORITHM

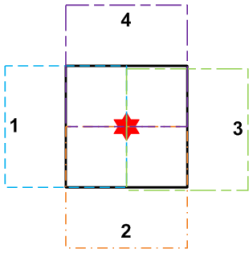


Fig. 2. Butterfly Pattern: Illustrates 5 different arrays of pixels; The first one is not numbered and has the "star" (representing a $\mathbf{T_{ID}}$) at its center, an array of the same dimensions shifted left is numbered 1, shifted down is numbered 2, shifted right is numbered 3 and shifted up is numbered 4, for a total of 5 different arrays related to the single $\mathbf{T_{ID}}$

In this section the generalized approach will be described, followed by some numerical results. We assume that the inputs to the signature generating algorithm (derived from (1)) are represented by Figure 2. The central array of pixels corresponds to a known target of interest at time t_{k-1}. Labeling the full set of pixels at time t_{k-1} as $\mathbf{\Gamma}_{k-1}$ and the associated set of pixels associated with the target of interest as $\mathbf{T}_{k-1}$ we choose the cardinality of $\mathbf{\Gamma}$ to assure that $\mathbf{T}_{k-1} \subseteq \mathbf{\Gamma}_{k-1}$. We also assume that the physical dimensions of the pixels corresponding to $\{\mathbf{\Gamma}_i \; \forall \; 0 \leq i \leq n\}$ where n is the number of time epochs of interest, is constant. For the illustration involving a flying vehicle, this is straightforward to accomplish through adjustment of range to target and or focal point zoom. It is important to note that this constraint is important since there is no assumed a priori target model. Assuming that the frame rate is sufficiently high, then at time t_k one of four sets of arrays indicated as $\{1, 2, 3, 4\}$ in Figure 2 will contain the indicated set $\mathbf{T}_k$.

Since we assume that rotationally invariant statistics for a particular $\mathbf{T_{ID}}$ will not be Gaussian, then (1) becomes a convenient means for representing the associated p.d.f. It provides a means to represent a convergent orthonormal approximation to the p.d.f. that is precisely associated with the array in question. Most importantly, it is a means to positively indentify an object of interest from one frame to the next. By assuring $\forall \; k$ that $\mathbf{T}_k \subseteq \mathbf{\Gamma}_k$ and by applying (1) to successive $\mathbf{\Gamma}_k$ we have a means for measuring successive p.d.f's and comparing them to ones known to represent the desired $\mathbf{T}_{ID}$. Again, in order for the statistics to remain at least approximately stationary, the cardinality and associated physical dimensions of the pixels must remain approximately constant.

For each time epoch, j, according to Figure 2, both $\mathbf{T}_j$ and $\mathbf{\Gamma}_j$ can be further generalized as $\mathbf{T}_j^p$ and $\mathbf{\Gamma}_j^p$ where $\{0 \leq p \leq 4\}$. Hence, assuming $\mathbf{\Gamma}_{k-1}^0$ contains the object of interest, we wish to find which of the subsequent $\mathbf{\Gamma}_k^p \; \forall \; \{0 \leq p \leq 4\}$ that contains the object of interest.

To accomplish this, we introduce the concept of a "signature" where

$$\vec{S}_{original}(\mathbf{\Gamma}_0^0) =$$
$$[(L^1\{f(\vec{X}_0\} + L^2\{f(\vec{X}_0)\}), E[\vec{X}_0], (E[\vec{X}_0^2] - (E[\vec{X}_0])^2)$$

corresponds to the $\mathbf{T}_{ID}$ for the initial object of interest. By also creating the comparison operation defined as

$$\vec{S}_{compare}(\mathbf{\Gamma}_k^p) =$$
$$[(L^1\{f(\vec{X}_k)\} + L^2\{f(\vec{X}_k)\}), E[\vec{X}_k], (E[\vec{X}_k^2] - (E[\vec{X}_k])^2)]$$
$$\forall \; \{0 \leq p \leq 4\}$$

we wish to find the subsequent, p, to

$$minimize\{\vec{S}_{original} - \vec{S}_{compare}\} \forall \{p\} \tag{2}$$

Hence, the PIK Algorithm is one of continually re-finding the statistically significant signature of interest, derived from (1), by determining $\mathbf{T}_k$ which solves (2) to find $\mathbf{T}_{ID}$, in real-time (frame-by-frame).

4.1. Numerical Illustrations

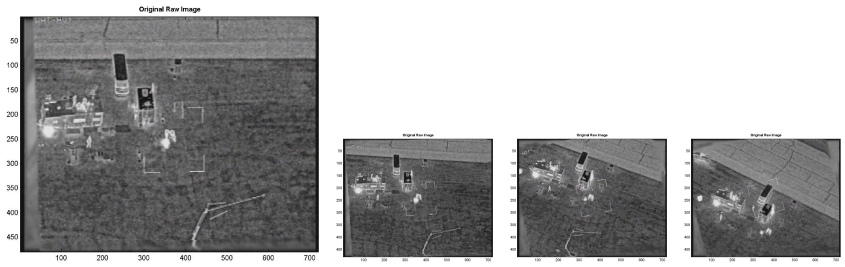
Figure 3 indicates entire scenes for frames 0, 1, 2 and 3. The original $\mathbf{\Gamma}_0^0$ is found in Figure 3(a), which gives rise to $\vec{S}_{original}(\mathbf{\Gamma}_0^0)$. Developing $\vec{S}_{original}$ is accomplished by using (1) along with the intermediate results associated with this numerical calculation as shown in Table 1 . For successive frames results of using (2) are shown in Table 2 where the minimum corresponds to the $\mathbf{T_{ID}}$.

MGS terms	C0	C1	C2	C3	C4
	88.000	87.987	68.981	32.850	17.425
MGS terms	C5	C6	C7	C8	C9
	68.981	93.273	57.844	24.878	32.850
MGS terms	C10	C11	C12	C13	C14
	57.844	53.311	24.601	17.425	24.8778
MGS terms	C15	C16			
	24.601	14.593			
Statistics	Mean	Variance			
	115.855	1.217e+003			

Table 1. $\vec{S}_{original}(\mathbf{\Gamma}_0^0)$ for Figure 3(a); MGS coefficients for truncated $f(\vec{X})$ with maximum s = 4 (17 terms in series) and some associated statistics as determined by a two dimensional form of (1).

5. RESULTS

By combining a means for real-time frame by frame solution to (2) (PIK Algorithm) we are able to both maintain the target of interest in



(a) (b) (c) (d)

Fig. 3. Sample Frames: (a) - Frame 1, (b) - Frame 2, (c) - Frame 3, (d) - Frame 4

$\vec{S}_{compare}(\mathbf{\Gamma}_1^p) - \vec{S}_{original}(\mathbf{\Gamma}_0^0)$	zero shift **266.125**	left shift 1214.9813	right shift 1382.1202	down shift 1486.6939	up shift 899.3213
$\vec{S}_{compare}(\mathbf{\Gamma}_2^p) - \vec{S}_{original}(\mathbf{\Gamma}_0^0)$	zero shift **152.601**	left shift 1130.3552	right shift 1354.6514	down shift 11880.1486	up shift 488.3406
$\vec{S}_{compare}(\mathbf{\Gamma}_3^p) - \vec{S}_{original}(\mathbf{\Gamma}_0^0)$	zero shift 1000.2937	left shift 366.974	right shift 867.3253	down shift **211.6414**	up shift 1381.7229

Table 2. Comparison Norms Results (best fit for each row is bold face type) with descriptive phrases for each "p"

the field of view, as well as determine how to maintain constant pixel size (adjust vehicle position and or camera zoom) in order to provide both a continuous stream of information rich video, as well as assure the system remains both controllable and observable without any human involvement. Figure 4 confirms these results which can be thought of as robotic continuous cognition. The associated precision exceeds what can be achieved by a human pilot, with superior cognitive abilities. The robot can obviously not compete with a human in generalized cognitive abilities, however, for specific repetitive ones it can exceed combined human based perception and control functions. Moreover, these capabilities were confirmed over an extended period of time, which suggests that even if a human could achieve such a level of competence, it would be very difficult to maintain such a level of performance over an equivalent period of time.

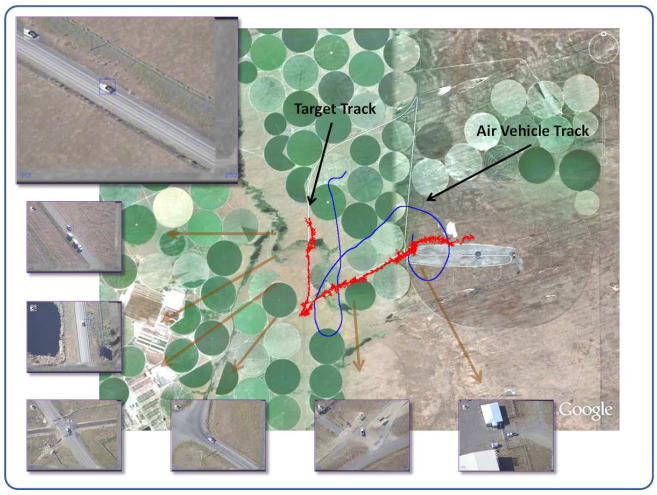


Fig. 4. Example of Continuous Cognition via auto-navigation and auto-recognition based on (1)

6. CONCLUSION

By combining a PIK Algorithm with full motion video, we have presented a means for increased cognitive abilities. As has been presented, such an approach is consistent with a well-formulated control problem that simultaneously satisfies conditions of observability and controllability under a wide variety of circumstances. Confirmation of the principles is characterized by a "replay" function that is able to retain all significant information in manner that is consistent with (or else exceeds) human cognitive abilities to assure delivery of trustworthy and verifiable information.

7. REFERENCES

[1] B. Bennett, C. Dee, and C. Meyer, "Emerging methodologies in encoding airborne sensor video and metadata," *Proceedings of the IEEE Military Communications Conference*, pp. 1–6, October 2009, Boston, MA.

[2] H. Zhou, L. Jia, and Y. Qin, "Metadata specification of railway video information and its application in video monitoring system for qinghai-tibet railway," *Proceedings of the International Symposium on Computer Network and Multimedia Technology*, pp. 1–4, January 2009, Wuhan, China.

[3] H. E. Neely, R. S. Belvin, and M. J. Daily, "Modeling threat behaviors in surveillance video metadata for detection using an analogical reasoner," *Proceedings of the IEEE Aerospace Conference*, pp. 1–9, March 2010, Big Sky, MT.

[4] S. McCloskey and P. Davalos, "Activity detection in the wild using video metadata," *Proceedings of the International Conference on Pattern Recognition*, pp. 3140–3143, November 2012, Tsukuba, Japan.

[5] I. Joo, T. Hwang, and K. Choi, "Generation of video metadata supporting video-gis integration," *Proceedings of the IEEE International Conference on Image Processing*, vol. 3, pp. 1695–1698, October 2004, Singapore.

[6] C. Chen, *Linear System Theory and Design*, ser. Oxford series in electrical and computer engineering. Oxford University Press, 1984.

[7] G. Viviani, N. Parisi, W. Callahan, and L. Zimmermann, "Achieving next generation performance of ion implanters with the varian control system (VCS)," *Proceedings of the 2000 International Conference on Ion Implantation Technology*, pp. 419–422, September 2000, Alpbach, Austria.

[8] S. Pinker, "Visual cognition: An introduction," *Cognition*, vol. 18, no. 13, pp. 1–63, December 1984.

[9] S. Pinker, *How the Mind Works*. Norton and Company, 1997.

[10] J. B. Carroll, *Human Cognitive Abilities*. Cambridge University Press, 1993.

[11] S. S. Skiena, *The Algorithm Design Manual*, 2nd ed. Springer-Verlag, 2008.

[12] G. L. Viviani, "On orthonormal representations for targets of interest," *Internal Insitu Report*, December 2010.

[13] T. Mundhenk, K. Ni, K. Kim, and Y. Owechko, "Detection of unknown targets from aerial camera and extraction of simple object fingerprints for the purpose of target reacquisition," *Proceedings of the SPIE Intelligent Robots and Computer Vision XXIX: Algorithms and Techniques*, vol. 8301, no. 1, pp. 83 010H–1–83 010H–14, January 2012, Burlingame, CA.

[14] K. Ni, T. Mundhenk, K. Kim, and Y. Owechko, "Manifold-based fingerprinting for target identifications," *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–6, June 2012, [Providence, RI.

[15] G. L. Viviani and G. T. Heydt, "Stochastic optimal energy dispatch," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-100, no. 7, pp. 3221–3228, July 1981.