The Demise of the Frame: A Media Archaeology of Motion Prediction

Ricardo Cedeño Montaña

Humboldt-Universität zu Berlin
Department of Cultural History and Theory
Berlin, Germany
cemontar@hu-berlin.de

Abstract

Prediction theory emerged during the WWII in order to improve anti-aircraft artillery and resulted in algorithms devised to statistically predict airplanes and missile paths. Although today prediction is the backbone of the video compression, the historical and technical connection between this mathematical theory and contemporary imaging technologies has not been sufficiently determined. Using a media archaeological approach this paper discusses how the implementation during the 1990s of prediction algorithms to video compression has generated an entirely new type of moving images. I argue that the consequence of turning each displayed picture into a rigid grid and its construction into the statistical prediction of the pixel's values is dramatic because it renders the temporal coincidence of all pixels within the frame unnecessary. On the surface there is no change. Yet, using prediction a video codec such as H.264/AVC has turned the frame into an address where chunks of pixels coming from different moments in time are put together. At the coding level, prediction has banished the frame. The elimination of that basic unit of all moving images, not only miniaturized video but it also has had ontological consequences for the image that are not yet fully understood.

Keywords

Motion Prediction, MPEG, Digital Video, Video Compression, Statistics, Ballistics.

Surface and Subface

As an algorithmic medium, digital video has two sides. On the surface and immediately before us, we perceive colours, shapes, and changes on the screen. A raster image. Underneath, on the *subface* (Nake, 2005, p. 50), occurs a complex series of operations. It is on the subface of digital video that one encounters some of the operations that reduced digital video so it could be packaged into a mobile phone. Such operations include de-constructing, quantizing, encoding, grouping, locating, predicting, compensating, transposing, decoding, and reconstructing. Since the last decade we have witnessed the sudden flood of online video, making video something of the everyday and in need of a new set of conceptual tools to see its ontological transformation (Treske, 2015, p. 18). I argue that on the subface lies one possible answer to that question. Observing the subface of digital video leads into the history of the algorithms that are used today every time a digital video file is recorded or reproduced.

Subface is a concept that aligns with the epistemic aim of media archaeology to reveal the nature of media by observing and dealing with their inner micro-operations instead of their macro discursive effects (Ernst, 2012, p. 16). It is in that contact with the subface that media history is able to render the concrete time-axis operations and mathematical principles of media graspable.

This paper focuses on one time-axis operation of digital video that greatly affects the size of the file: motion prediction. One development in mathematics during the 1940s is key in making possible the implementation of prediction algorithms as means to transmit video over telephone net-works since the end of 1980s. Among other techniques, digital video formats implement descendants of algorithms devised not for image processing but for statistically predicting airplanes and missiles paths during the WWII.

Prediction

Ballistics

During the WWII, the superiority of the *Luftwaffe*, combined with the Allies slow anti -aircraft fire control on the ground, made the calculations for ballistics a problem of tremendous military importance. The solution was straight-forward: the gunner, the firing table, and on-the-spot calculations had to be replaced with an apparatus that would follow an airplane, compute its distance, determine the length of time before a shell could reach it, and figure out where it would be at the end of that time.

It was during WWII that the U.S. NDRC engaged Nor-

bert Wiener, a Harvard PhD in mathematics at the age of nineteen, to tackle "the most difficult mathematical problem in fire control: prediction" (Mindell, 2002, p. 278). Wiener's task was to formulate a method to trace the path of an airplane and estimate its future location at a given time. Wiener was able to contribute to the solution of this problem in part because of his previous work on the equations to solve situations in which two regions are separated by a given barrier and the data gathered on one region influences the behaviour in the other, but not vice versa. Wiener had observed that in such problems the 'present' acts as a buffer between the influencing 'past' and the indeterminate 'future' (Masani, 1990, p. 134). This continual elaboration of the new stems from Henri Bergson's concept of temporal duration and proved important during Wiener's work on the anti-aircraft fire control (Ernst, 2012, p. 275). His basic idea was "to use electrical networks to determine, several seconds in advance, where an attacking plane would be and use that knowledge to direct artillery fire" (Galison, 1994, p. 234).

After the war, Wiener generalised the principles governing his prediction system as mathematical formulations beyond the purpose of taking down airplanes. Text, speech, and pictures flowing through wires would soon be generalized as continuous signals, and later as discrete chains of numbers, that can be tracked and predicted. They'd become the statistical input data that constitutes a series in time, such as prices in the stock market, sounds in a mobile phone, and pixels in an online streaming video.

Statistics

A predictive operation has two pillars: time series and communication engineering. The first is a discrete or continuous sequence of "quantitative data assigned to specific moments in time and studied with respect to the statistics of their distribution in time" (Wiener, 1950, p. 1). The second component, communication engineering, treats a message as data and detached from its medium or physical carrier.

Speaking of communication engineering, Wiener considered that to be useful an apparatus should be designed to operate on a general set of messages and not exclusively on a single message (Wiener, 1950, p. 4). Thus, he designed and implemented an algorithm not to take down one particular airplane but any airplane in the set of airplanes. Furthermore, he later generalised it to statistically predict the output of any message sampled on a time series, be it words, sounds, images or any signal. When motion and then vision became statistically analysed in time series a new type of imagery emerged, one that is the result of data processing techniques that continuously predict the next values on the numerical chain.

Based on the continuous storage and statistical accountability of past results, Wiener's algorithm promised to predict a possible future result. The predicted position of an airplane *gh* at time *t* thus became the infinite sum of all its past positions *f(t-τ)* and a derivative of the past errors affecting it *dWh*:

$$g_h(t) = \int_0^{\infty} f(t-\tau)\, dW_h(\tau)$$

Similarly, the MPEG video decoder produces a *predicted picture* based solely on the statistical data gathered from a set of previous pictures. Future pictures are thus estimated by comparing the direction and distance of any motion between the present and past pictures. This decoder is able to make such calculation because each picture is a rigid raster grid whose points are addresses both in space and time as it was any airplane in the night sky over London in 1943.

Indexing

The substitution of videotapes with digital files marked a fundamental change in the reproduction of moving images. In order to go to one point in time to another, the former has to be wound or rewound, whereas the latter has indexes and addresses to jump directly to any particular frame. The new control thus gained over the reproduction of moving images, that is the ability to go from index to index and frame to frame without shuttling over the storage surface, abolishes the continuous and linear reading of film and videotape.

Indexing time and simultaneously targeting any of the frames echoe Friedrich Kittler's argument that the Roman codex was more revolutionary in the genealogy of writing than the invention of the printing press because it granted non-linear access to specific content by fragmenting the scroll into a series of discrete leaves of papyrus fastened together (Kittler, 1993, p. 178). Centuries later, each *pagina* got a number, an address,

that allows the leaves to be compiled in the proper order and any detail in the text to be precisely referenced and accessed in a non-linear fashion. Similarly, a video player on a computer offers a timeline with a cursor to freely move the playback head to any point in time. This operation does not take the time-axis as a continuous line for strict sequential access from a beginning to an end, rather it treats it as a discrete series of binary addresses for instant non-linear retrieval.

Compression

MPEG: Interframe Prediction

MPEG is a data compression format whose main implementation has been in tapeless video. One of its core features is motion prediction, which is an application of predictive coding to estimate and interpolate the changes between successive frames.

Predictive coding has been part of video codecs since their very beginning. The pioneering H.120 recommendation for *Codecs for Videoconferencing over Telephone Net-works* issued by the International Telecommunication Union (ITU) in 1988 already included the motion-compensation between two adjacent frames for a still or slowly moving area (ITU, 1988, p. 18). In a video sequence, objects tend to move in predictable patterns, changing little from frame to frame. Consequently, their motion trajectories can be traced over time and their future positions can be predicted frame-by-frame (Haskell & Puri, 2011, p. 7).

Predictive coding not only changed the way video is transmitted and processed but also affected the relation between the frame and the time-axis and redefined the inner structure of the frame itself.

Early versions of MPEG and H.26X converted the time-axis from the site for the mere succession of frames to the site for tracking and estimating motion in order to achieve higher compression rates. Frame references and predicted frames emerged from decoupling the synchronization between the encoded and the displayed image. In tapeless digital video, the order of the frame encoding and the order of the frame display do not correspond to each other. While frames are still fastened to a time-axis, the prediction of the location of pixels forces a H.26X/MPEG encoder to sequence and transmit the image stream by placing reference frames before predicted frames, i.e. future before present frames. This is a sequence that the decoder reorders to display a coherent video sequence.

The vast majority of video conferences, camera phone, and action camera videos are composed of one single shot with very little changes from frame to frame. In order to compress the amount of bits, a motion prediction algorithm takes advantage of such little change to track the positions of groups of pixels, called macroblocks, belonging to two adjacent frames, estimate their differences, and transmit that difference. When a cut occurs and the scene is replaced entirely, the algorithm establishes a new reference index and the process starts anew.

The H.262/MPEG-2 (1994) standard, used in DVDs, specifies three types of frames, Intra-coded (I), Predictive-coded (P), and Bidirectionally predictive-coded (B), each of which uses a different encoding method. I-frames are coded using only its own data. These are frames in the traditional sense and are used as reference for the motion prediction. P -frames are coded using motion compensated prediction from a past reference frame. And B-frames are coded using motion compensated prediction from a past and/or future reference frame (ITU, 1995, p. 13). A sequence of decoded frames might be: I_BB_P_ BBB_P_ (ITU, 1995, p. v), where there are 7 predicted pictures from I to I.

Using a time series of two frames, the H.262/MPEG-2 encoder first searches for matches between successive frames, then, if found, it estimates the changes in position. If the differences are small, no prediction is transmitted. If there is sufficient change, then the encoder calculates the difference between both positions and transmits it as a motion vector. Such vectors describe the movement of macroblocks within the frame. As the data in the difference is smaller than in the macroblock, that difference can be represented with fewer bits. Motion vectors produce frames (P and B) that contain data about changes to be made with regard to other frames, but do not contain any data about light intensities and their locations. These frames have little to do with physically separated photographs or electronically scanned lines, since they are places for the granular analysis and assembly of motion.

A video encoder generates a series of motion vectors that the decoder reconstructs as predictive pictures that are not pictures at all but rather “a set of instructions to convert the previous picture into the current picture. If the previous picture is lost, decoding is impossible”(Watkinson, 2004, p. 256). The lack of picture references has been exploited by visual artists

like Takeshi Murata in works such as *Untitled (Pink dot)*, 2007, that result in the organic flow of video artifacts and glitches produced by the lack of reference frames (I) in a MPEG video file. Thus, the desire to hide the operations carried out on the subface of a homogeneous surface is countered by hacking the motion-prediction function. And the dirty guts of digital video lay there wide open.

Motion prediction made the frame, the old container of light intensities, into a specialised place for the machinic synthesis of movements, in which every block is tracked, computed, predicted, and coded. The detection and assemblage of vectors turned the moving image into an ever changing image, where the blocks that constitute it move within the frame forming new images until their changes are so dramatic that blocks with new luma and chroma data replace them and the morphing process begins anew. This produces smooth transitions between the frames of a moving image that, as Adrian Mackenzie observes, never flickers (Mackenzie, 2008, p. 53).

Sending only the motion data of macroblocks instead of their picture data allows any H.26X codec to produce pictures that can only exist as the sum of transformations that a past or future picture undergoes. Motion prediction redefined the frame and its relation to the time- axis, thus altering the material storage of moving images. In order to encode and decode each frame, predictive coding requires the non-linear access to the time-axis granted by the indexes and addresses of tapeless video. As a consequence, the magnetic tape, with its strict sequential writing and track-by-track manner of reading, was rendered an unsuitable storage medium for this new type of moving image.

Until H.262/MPEG-2 the frame was still the cohesive geometrical structure for one instant of time in digital video. More recent codecs like H.264/AVC (2003) and HEVC/H.265 (2013) changed this by creating patchwork-like frames made up of fragments of present data and fragments of data located elsewhere on the time-axis. A video frame, as shot with a camera phone or played via online streaming, is the site for the assemblage of macroblocks each of which originates at a different point in time.

Slices

One approach formed at the end of the 1990s, frame partitioning with multi-frame prediction, has, at least on the subface, rendered the frame obsolete as the cohesive unit for storing moving images.

Today, from the large HDTV screens in our drawing rooms to the small mobile phone displays in our pockets, H.264/AVC is a ubiquitous presence. Since around 2005, H.264/AVC has become one of the most frequently used codecs for recording, compressing, and distributing HD video files. It is behind DVB transmissions, HDV disk storage, and low and high resolution video streaming over the Internet. When a camera phone records a video it depends on a H.264/AVC encoder, implemented in C, to perform the cascade of operations that determine the outcome. Permanently connected to a network via the phone, the encoder produces a moving image that is immediately transmissible. Today, it is very common to get video footage of breaking news such as natural disasters or terrorist alerts coming first from a camera phone rather than from a TV camera.

Sean Cubitt pinpoints the rise of H.264/AVC as the dominant online video codec behind the decision of YouTube to drop a variation of the H.263 codec owned by Adobe in favour of H.264 (Cubitt, 2014, p. 246). Higher compression factors have allowed this codec to store HD videos in memory cards inside mobile phones, pocket digital camcorders, and action cameras. One of the reasons for the higher compression rate of H.264 is that, in contrast to H.262, it makes references to a time series composed of multiple frames in order to estimate the changes in the current video frame (Wiegand & Girod, 2001, p. 4).

Instead of using frames as geometrical containers for time, H.264/AVC partitions each video picture into slices of different sizes and types of coding. And although it assembles them into a frame for display, H.264/AVC's basic temporal container is the slice. Figure 1 illustrates the video bitstream of this codec. On the left side from top to bottom, it shows the hierarchical data structure from the smallest processing unit, the macroblock, up to the largest unit, the video sequence. On the upper right side, it shows macroblocks of 4 × 4 pixels for motion prediction on the time-axis. This small macroblock allows this codec to track subtle changes in areas with more detail. H.264 has no pictures in coding terms. Instead, it has an "imaginary Picture structure that is composed of one or more Slices," where 'imaginary' means "that there is no Picture layer in the bitstream structure, but a picture is generated through the Slice decoding process" (Lee & Kalva, 2008, p. 79). Pictures in the moving image sequence are only formed at the very moment of decoding, a formation that is signalled

in the bitstream by an indication of the types of slices that might assemble it.

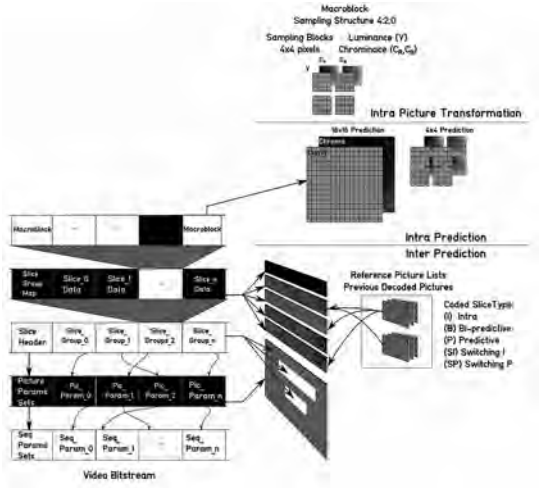

Figure 1. H.264/AVC (MPEG-4 part 10) Video Bitstream. Illustration based on data by (ITU, 2003) and (Lee & Kalva, 2008, p. 80)

H.264/AVC defines several types of slices according to their coding. There are slices without motion prediction (I), slices containing macroblocks with motion prediction based on multiple past frames (P and B), and Switching slices (SI and SP) that facilitate switching between high- and low-bit -rate streams on the decoder side by creating and storing an artificial picture. In previous codecs, "such switching would be impossible because of temporal coding" (Watkinson, 2004, p. 331). Using Switching slices, a H.264/AVC decoder adds decoded "data to the last decoded picture of the old bitstream and this converts the picture into what it would have been if the decoder had been decoding the new bitstream since the beginning" of the stored sequence (Watkinson, 2004, p. 331). A bunch of slices with their own temporal references and positions in the time-axis are enough to create a picture in the memory of the decoder. These 'imaginary' pictures are only 'seen' by the decoder in order to adapt the bitstream to this or that screen or to this or that resolution. Philip K. Dick's androids might not dream of electric sheep but a digital video decoder might.

During the preparation phase for establishing this standard, time series greater than two references emerged in order to use long-term statistical data in digital video. This was the last step in order to banish the frame from the subface of digital video. A long-term storage process for multiple frames was added to the codec to predict the motion of each 16 × 16 macroblock. Thus, from this codec on, "motion vectors are determined by *multi-frame motion estimation* which is conducted via block matching on each frame [in the] memory" (Wiegand & Girod, 2001, p. 38). Just as in Wiener's problem about shooting down airplanes, H.264/AVC statistically estimates the present position of a macroblock on the screen based on a finite time series of its past positions.

Slicing the frame into regions with multiple references in the time-axis (time series > 2) had two effects on the moving images shot with any camera phone today. On the surface, this time-interwoven digital frame confirms Paul Virilio's conclusion about speed that, "the delineation between past, present, and future, between here and there, is now meaningless except as a visual illusion" (Virilio, 1994, p. 31). On the subface, the absence of a synchronous frame enables the creation of pictures only for the machine. On the subface, the moving image has been restructured as a paradoxical object in which the number of time-axes is the same as the number of blocks it contains, with each axis starting at a different time and being of a different duration. On the subface, there are no frames.

Conclusion

In this paper, I have been suggesting that the ubiquitous linear predictive coding can be considered the key algorithm in the digital moving image. Predictive coding was used as the basis for early compression codecs for video-conferences, such as H.120 at the end of the 1980s. And although today it is hidden behind the smooth flow of HD video, prediction remains at the heart of digital video in codecs such as the H.264/AVC. Predictive coding, however, did not emerge out of any research into imaging techniques. As Wolfgang Schäffner has pointed out, the seemingly smooth relationship between images and computers is indeed more rocky than it appears on the surface, because the computer is not a technology for images but for mathematical operations with symbols (Schäffner, 2008, p. 127). It is in the subface, not on the surface, where every individual pixel is constructed out of several computations with endless strings of bits. Digital video is only possible due to algorithms that sample the incoming video signal and each pixel is located on the screen by algorithms that give them precise addresses.

Acknowledgements

This paper was possible thanks to the support of the Deutscher Akademischer Austauschdienst (DAAD) and the Interdisziplinäres Labor Bild Wissen Gestaltung at the Humboldt-Unversität zu Berlin.

References

Cubitt, S. (2014). *The Practice of Light: A Genealogy of Visual Technologies from Prints to Pixels*. Cambridge, MA: MIT Press.

Ernst, W. (2012). *Chronopoetik: Zeitweisen und Zeitgaben technischer Medien*. Berlin, Deutschland: Kulturverlag Kadmos.

Galison, P. (1994). The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision. *Critical In-quiry*, *21*(1), 228–266. https://doi.org/10.2307/1343893

Haskell, B., & Puri, A. (2011). MPEG Video Compression Basics. In L. Chiariglione (Ed.), *The MPEG Representation of Digital Media*. New York, NY: Springer.

ITU. (1988). *H.120: Codecs for Videoconferencing Using Primary Digital Group Transmission.* (Technical Standard Recommendation) (p. 62). Geneva, CH: ITU.

ITU. (1995). *ITU-T H.262: Information Technology - Generic Coding of Moving Pictures and Associated Audio Information: Video* (Technical Standard Recommendation) (p. 211). Geneva, CH: ITU.

ITU. (2003). *ITU-T H.264: Advance Video Coding for Generic Audiovisual Services.* (Technical Standard Recommendation) (p. 282). Geneva, CH: ITU.

Kittler, F. (1993).Geschichte der Kommunikationsmedien. In J. Huber & A. M. Müller (Eds.), *Raum und Verfahren: Interventionen* (pp. 169–188). Basel ; Frankfurt am Main: Basel ; Frankfurt am Main : Stroemfeld/Roter Stern.

Lee, J.-B., & Kalva, H. (2008). *The VC-1 and H.264 Video Compression Standards for Broadband Video Ser-vices*. New York, NY: Springer Science & Business Media.

Mackenzie, A. (2008). Codecs. In M. Fuller (Ed.), *Software studies: a lexicon* (pp. 48–55). Cambridge, Mass.: MIT Press.

Masani, P. R. (1990). *Norbert Wiener, 1894-1964*. Basel, CH: Birkhäuser.

Mindell, D. A. (2002). *Between Human and Machine: Feedback, Control, and Computing Before Cybernetics*. Baltimore, Maryland: Johns Hopkins University Press. Retrieved from https://books.google.de/books?id=sExvSbe9MSsC

Nake, F. (2005). Das doppelte Bild. In M. Pratschke (Ed.), *Digitale Form* (Vol. 3.2, pp. 40–50). Berlin, Ger-many: Berlin : Akad.-Verl.

Schäffner, W. (2008). La Revolución Telefónica de la Imagen Digital. In J. La Ferla (Ed.), *Artes y Medios Audiovisuales: Un Estado de Situación II. Las Prácticas Mediáticas Pre Digitales y Post Analógicas* (pp. 127–34). Buenos Aires, AR: Nueva Librería.

Treske, A. (2015). *Video Theory: Online Video Aesthetics or the Afterlife of Video*. Bielefeld, Germany: transcript Verlag.

Virilio, P. (1994). *The Vision Machine*. Indiana University Press.

Watkinson, J. (2004). *The MPEG Handbook: MPEG-1, MPEG-2, MPEG-4*. Elsevier/Focal Press.

Wiegand, T., & Girod, B. (2001). *Multi-Frame Motion-Compensated Prediction for Video Transmission*. Norwell, MA: Kluwer Academic Publishers.

Wiener, N. (1950). *Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications.* Cambridge, MA: Technology Press of the Massachusetts Institute of Technology. Retrieved from http://catalog.hathitrust.org/Record/010056247

Author Biography

Ricardo Cedeño Montaña (Bogotá, Colombia, 1976). His background spans the fields of media history, media art, and design. His current research interest revolves around technical media and the history of knowledge with a particular focus on imaging techniques. He holds a PhD in Cultural History and Theory from the Humboldt-Universität zu Berlin, Germany (2016), an MSc in Digital Media from the Hochschule Bremerhaven, Germany (2009), a degree in Multimedia Creation from the Universidad de los Andes, Colombia (2003), and a degree in Industrial Designer from the Universidad Nacional de Colombia (1999).