Optical Engineering

SPIEDigitalLibrary.org/oe

Re-establish the time-order across sensors of different modalities

Ming Kai Hsu
Ting N. Lee
Harold Szu

Re-establish the time-order across sensors of different modalities

Ming Kai Hsu
Ting N. Lee
George Washington University
Department Electrical and Computer Engineering
Washington, DC 22002
E-mail: mkhsu@gwmail.gwu.edu

Harold Szu
U.S. Army Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia 22060

Abstract. Modern cameras can cut passengers' faces into boxes in 0.04 s per frame in parallel without time stamps. Unfortunately, that creates random storage without the tracking capability, and one can no longer meet the 5 W's challenge—"who speaks what, where and when." We develop a time-order reconstruction methodology which sorts the boxes as follows. i. A morphological image preprocessing to overcome the facial changes is based on the peripheral invariance of a human visual system when focusing on a maximum overlapping central region. ii. Replacing the Wiener matched filter desired output with an averaged but blurred long exposure, one can select the best matched sharp short exposures called the anchor faces β's. iii. The time-order neighborhood chaining is done by an iterative self-affirmation logic that demands a mutually agreed-upon minimum distance: whether or not the two nearest neighbors of β, namely face A and face C, also consider β to be their two nearest neighbors. The reconstruction procedure mathematically amounts to a product of two triple correlation functions sharing an intermediate state. We have thus demonstrated the time-order helps us associate a video submanifold with the acoustic manifold that solves the 5 W's challenge. [DOI: 10.1117/1.3562322]

Subject terms: long exposure; time-order; associate a video submanifold and acoustic manifold.

Paper 100670PRR received Aug. 23, 2010; revised manuscript received Feb. 13, 2011; accepted for publication Feb. 14, 2011; published online Apr. 14, 2011.

1 Introduction

The global village enjoys the availability of data rich environment, but sometime suffers from an explosive growth of useless data known as the curse of dimensionality in the digital age. Thus, the scientific community has developed various pre- and post-processing techniques to deal with the curse of dimensionality. For example, a smart organization principle called compressive sensing.[1,2] Specifically, when one does not need the frequency information of those sinusoidal components, one shall not sense it with the Fourier transform (FT), but with a physical group wave packet called the wavelet transform (WT). Furthermore, the wavelet transform in compressive sensing can be viewed upon as preprocessing technique replacing the FT with the adaptive WT,[3] which allows several super-mother wavelet kernels that each match the time evolving signal content, and therefore create a natural spatiotemporal preprocessing. For post-processing, the conventional data compressions or removing the redundancy in data such as JPEG, MPEG, face detection, etc., and dimensionality reduction (DR) algorithms, i.e., ISOMAP,[4] local linear embedding,[5] Laplacian eigenmap[6,7] and diffusion map.[8,9] They have been developed to discover the nonlinear manifold underlying complex datasets that traditional dimensionality reduction methods, e.g., principle component analysis[10] and multidimensional scaling[11] could not do. In addition, more DR algorithms[12–15] have been developed by encoding objects as matrices or tensors or arbitrary order. A generalized framework for dimensionality reduction has been presented in Ref. 16.

The generic DR for video datasets is defined as follows. Since every lexicographical order of pixels of the i'th video frame, $L \times M$ pixels, forms a long vector $\vec{X}_t \in O(L \times M)$, subscript $t = 1, \ldots, N$, has its own intensity values, the video itself becomes an exceedingly high dimensional set of light spots $O(L \times M \times N$ time-frames). All these light spots are contributing to the Haken's cooperative and competitive self-organization phenomena[17,18] yielding the collective or master degrees of freedom, e.g., facial pose angles, which drive all the underlying slaver degrees of freedom.

Szu[19] has derived asymmetric weights $W_{i,j} \neq W_{j,i}$ which have the positive diagonal $W_{i,i} \geq 0$ and a Kirchhoff-like graph theory energy function K as the weighted least mean square (LMS) $\min \cdot K = \min \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} W_{i,j} |\vec{X}_i - \vec{X}_j|^2 \geq 0$ where an in-bound risk $R_j = \sum_i W_{i,j} \geq 0$ is different from the out-bound popularity $P_i = \sum_j W_{i,j} \geq 0$, $\sum_i |\vec{X}_i|^2 P_i \neq \sum_j |\vec{X}_j|^2 R_j$ and the off-diagonal terms are $-2 \sum_i \sum_j W_{i,j} \vec{X}_i \cdot \vec{X}_j$.[19] Thus, the Kirchhoff energy $K \equiv (\vec{X}_i, K_{i,j}, \vec{X}_j) \geq 0$ becomes the inner product of the Kirchhoff matrix operator $K_{i,j} \equiv [(R_j + P_i/2)\delta_{i,j} - W_{i,j}]$, $\vec{X}_i$ and $\vec{X}_j$. In the traditional video pose nonlinear DR, the Gaussian distance among pixels of video images becomes symmetric, and then the Kirchhoff matrix operator $K_{i,j}$ is reduced to the graph-Laplacian matrix: $L_{ij} \equiv [d_i \delta_{ij} - W_{ij}]$ where $d_i \equiv (R_j + P_i/2)$. The ground state zero eigenvalue is the facial membership functions indicating the same person[7] (a degenerated ground state will indicate more than one person). The eigenvectors of the next eigenvalue become the master order parameters whose inner products with each facial image are the directional cosine

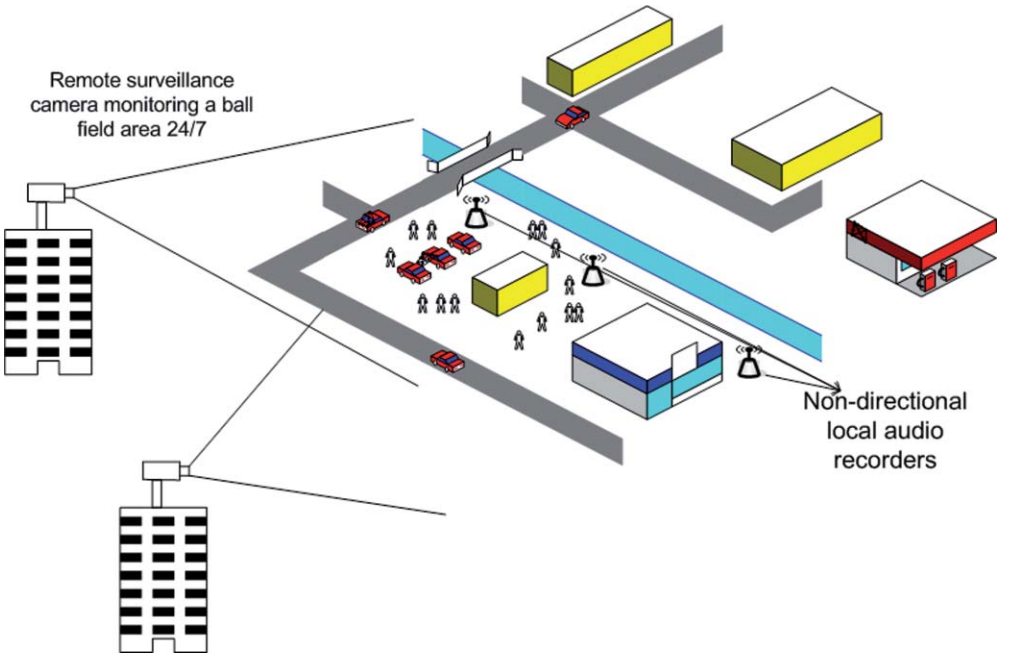


Fig. 1 In most persistent surveillance systems, multiple remote cameras are used for monitoring a ball field-sized area with local nondirectional audio recorders 24/7. In such a setup, hundreds of people are under surveillance at a time. The challenge for security personnel is to track people of interest within the crowds in accumulated video database over time. This challenging problem can be solved with the strategy we suggested in the beginning of this paper: removing redundancy but keeping the faces of passengers. The trade-off is that the collections of random faces from several cameras stored in the CPU for further processing are in random order.

angles representing the turning pose ordering angles. Such a pose-sorting similarity measure refers to the heading angles such as left-to- right and up-to-down.

To deal with the curse of dimensionality in the persistent surveillance, different strategies mentioned above can be applied, but in the end, a potential loss of association among different sensory tracks, i.e., video and audio, may occur. The observer cannot answer "who speaks what, when and where." In this paper, a video organization principle is given and demonstrated in a controlled persistent surveillance environment, which can be generalized for other applications such as i. smart grid infrastructure, ii. biomedical wellness web, and iii. hyperspectral data for precision farming, etc.

In persistence surveillance, video cameras are remotely set up for security reasons while low-cost audio recorders are hidden locally within the persistent surveillance field shown in Fig. 1. To reduce the video storage size requirement, one can extract facial poses from each video frame and store those facial boxes instead of the video, which reduces the storage size by 6 orders of magnitude (see Table 1). The

Table 1 Persistent surveillance is commonly used from military to civil applications. The storage requirement of persistent surveillance is huge and content organization is complex due to the huge amount of video data. It is a challenge for security personnel to sort and search for specific intruders among enormous video data. Since human faces are the most typical surveillance targets, saving faces of passengers/intruders without redundant background information saves up to 6 orders of magnitude when considering storage volume.

Size of Video Frames (pixels), 30Hz	640×480, 8 bits gray vlue	1024×768, 8 bits gray value
Storage requirement of video (Week)	5.58×10^{12} Bytes	1.43×10^{13} Bytes
Size of facial boxes (pixels), 30Hz.	60×60	90×90
Total time of Intruders revealed in video (α)	10%	10%
Average number of intruders recorded in video (β)	5	5
Storage requirement of saving facial boxes (Week)	3.27×10^{10} Bytes	7.3×10^{10} Bytes
Frontal facial boxes recorded in different resolutions (γ)	5	5
Total Number of intruders recorded in video (week)	1000	1000
Storage requirements of saving frontal facial boxes (week)	1.8×10^{7} Bytes	4.05×10^{7} Bytes
Size of recorded audio (week), 8 kHz Mp3 8kbit/s	6×10^{8} Bytes	6×10^{8} Bytes

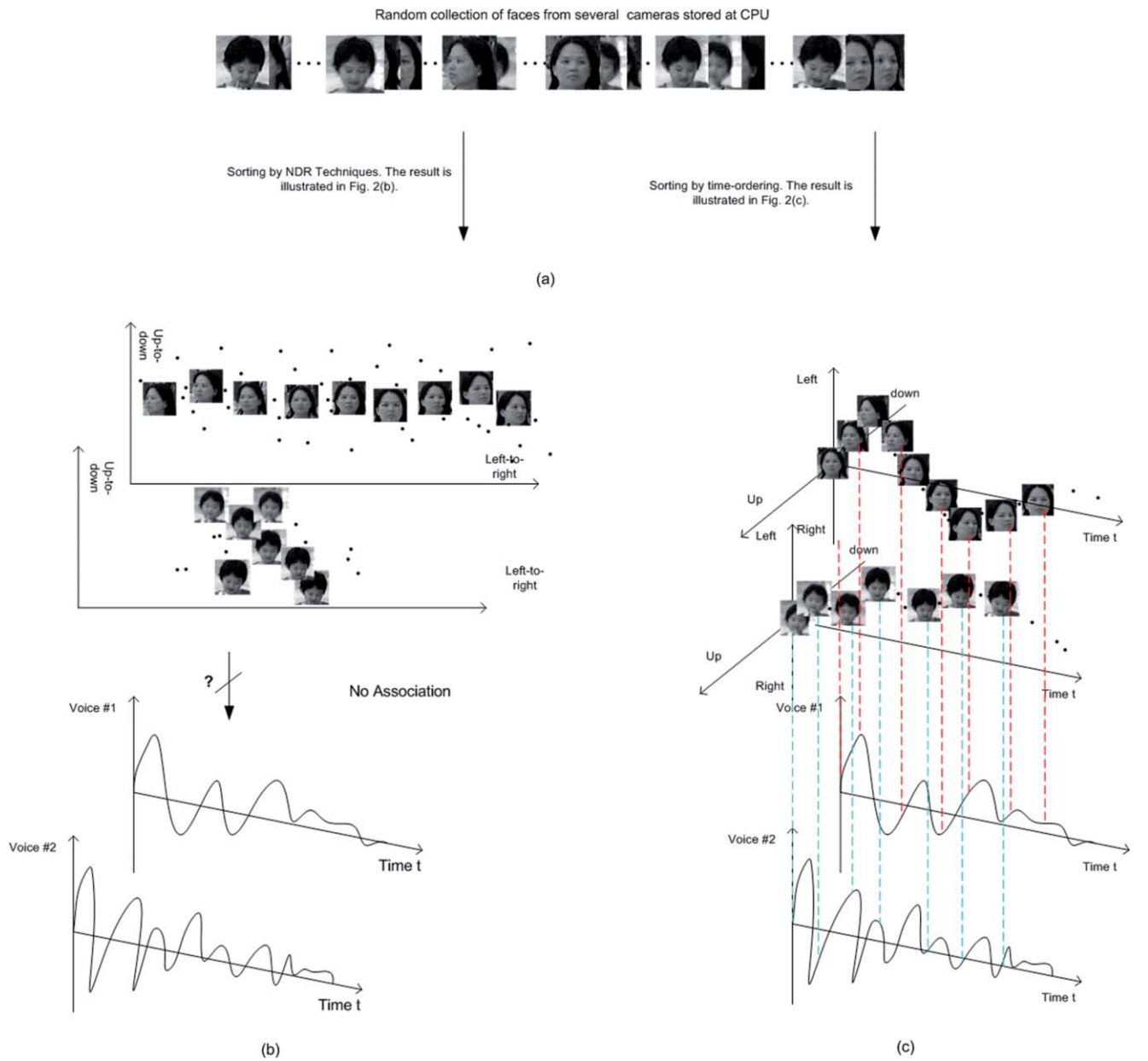


Fig. 2 This collection of facial poses of different passengers is stored in a randomized order in the computer for further processing. Current NDR techniques (sorting facial poses by similarity, left-to-left, right-to-right, etc.), yield results that do not reconstruct the time-ordering and cannot solve cross track association of video and audio. On the contrary, if we can precisely reconstruct the time-ordering of facial poses, the result will provide a method for cross track association between passengers and time-ordered audio recordings.

face detections, by facial color hue, and extractions in video are implemented by a face detection (FD)[20] system on chip. Adding time stamps on facial boxes is not recommended because it may be insufficient to reconstruct time-order and takes extra time. For an example, when there is a crowd in a single camera, faces of a crowd may be blocked by each other, obstacles, or detected without enough size, these faces are ignored by face detection techniques. Moreover, when passengers move past each other, the face of unwanted people momentarily fills the location of the face of the person of interest. In the consequences, facial boxes cut from each video frame are labeled by time stamps but the order of facial boxes from each frame are still in random order. Due to the reasons mentioned above, the collection of facial poses in video over 10 s is stored in a randomized order for further process. Such a loss of temporal correlation requires a re-establishment of the association across different sensor modalities that may help to ascertain "who said what, when, and where."

For an example, a passenger was walking in an airport transit looking left, right, and left again for the exit sign and speaking to her son about "where is the exit?" Both of their facial boxes were collected in Fig. 2(a). In Fig. 2(b), facial poses were sorted from left to right and top to bottom by graph embedded dimensionality reduction algorithms such as Laplacian eigenmap, etc., that is good to pick out mug-shots of each person, but they cannot reconstruct the time-ordering of walking passengers which is necessary to associate with an acoustic manifold finding out what may have been

spoken. On the opposite, in Fig. 2(c), video and audio were correlated by the reconstructed time-ordering of passengers in a video. Furthermore, the walking speed of a passenger could be determined from a set of time-ordered facial boxes alone, assuming only few facial boxes were ignored by the FD system on a chip. In the latest studies,[21,22] walking speed is correlated with the personality.

A time-order reconstruction requires the careful differentiation between the continuous notion of time versus the discrete sampling of time flowing known as the time-ordering in this paper. Thus, we begin with a concept of a clock, Newton's concept of time, which can mark 1 s time unit with a "tick and tock." Then, the unit of time must be followed by another unit tick and tock to indicate the 2 s time-order. This fact of directional flow may be described by a product of triple correlation functions, $W_{t_i,t_j,t_k} W_{t_k,t_l,t_m}$, where the common node t_k represents a discrete sampling point of time.

In Sec. 2, we approximate the triple correlation products with a product of pair correlation functions for a computationally efficient sorting algorithm, which we call the zipper chain (ZC) algorithm. It works like a zipper, two rows of multiple "teeth," mounted across each other alternatively. The underlying topology may be visualized as the time flow along a geodesic line, where two intertwined least mean square (LMS) minimums intersect. The ZC algorithm describes the product of paired correlations verified by a zip-chain event—an anchor face, say frontal β, has two nearest neighbors, A and C, while A and C have another two nearest neighbors, respectively. The time-order is determined if the anchor face β happens to be also the nearest neighbor of A and C. To choose an anchor face, the traditionally supervised desired output of the LMS Weiner matched filter becomes unsupervised, and is self-determined by the facial box dataset itself in two passes: the so-called local instances of good seeing determined by a self-referencing matched filter (SRMF) by Szu and Blodgett[23] in 1982. The first pass takes a uniform average of all faces, producing a blurred, but statistically correct, face at a specific pose with respect to the setup between multiple camera angles and the walking corridor [emulating the so-called astronomical imaging long exposure (LE) of a double star in order to catch an instance of good seeing to take a turbulent-free snapshot]. For our digital video post-processing, the second pass chooses a sharp short exposure (SE) to be the anchor face (which could be the frontal face for a narrow corridor and head-on setup of videos).

Another phenomenon of facial boxes extracted by face detection techniques is that facial boxes extracted from video are in different sizes and coordinates. A preprocessing of viewing transformation and registration for each facial box is necessary before time-ordering reconstruction. This preprocessing is introduced to overcome the change of facial boxes by a maximum overlapping central region by a morphological image processing focusing the binarized center of each face box to achieve the scale invariant peripheral vision[24] in Sec. 3. The ZC algorithm for time-ordered reconstruction of facial poses for association across different sensory is in Sec. 3. In Sec. 4, the impact of leak-through background information in maximum overlapped face boxes for the neighborhood determination is discussed. In addition, when the product of triple correlation is replaced by a pair-wise correlation, we have derived a fast approximation of embedding in line for facial pose similarity sorting in a short duration. The simulation of ZC is compared step-by-step against simpler similarity sorting results without the iterative re-check necessarily for the time-order sorting in Sec. 5. In Sec. 6, several possible applications of the algorithmic framework developed in this paper are briefly introduced before the conclusion in Sec. 6.

2 Theory of Time-Order Reconstruction of Pictures

Understanding time through the simple observation of a clock is the key. One tick-tock of a clock is a record of time, but it may take more than one set of tick-tock's to tell the time-order. This time-order is used to track the flow of time by the changes. That is to say tick-tock and tick-tock. On the other hand, a static face can be regarded as a collection of pixels ($L\times M$) with light intensities representing each pixel. This collection of pixels specifies a high dimensional vector space $\vec{x}_j \in O(L\times M)$ with respect to a set of $L\times M$ Cartesian coordinates, one per pixel. Given any specific face, $\vec{x}_j$, there are two nearest neighbor frames of facial poses, $\vec{x}_i$, $\vec{x}_k$ except the first and last frames. In other words, we may need to introduce more than one pair correlation $W_{i,j}(\vec{x}_i, \vec{x}_j)$ of the i'th tick and the j'th tock time subscript; e.g., $W_{i,j}(\vec{x}_i, \vec{x}_j)W_{l,m}(\vec{x}_l, \vec{x}_m)$ where the l'th tick must follow j'th tock as the necessary condition with two more faces $(\vec{x}_l, \vec{x}_m)$. To make explicit this restrictive condition, time-order reconstruction is an iterative sorting of finding the maximum chain product of triple image correlation $W_{i,j,k}W_{k,l,m}$ over the time-ordered facial image space. The triple correlation is defined as

$$W_{i,j} = R(\vec{x}_i, \vec{x}_j) = \iint \vec{x}_i \vec{x}_j f_{\vec{x}_i \vec{x}_j}(\vec{x}_i, \vec{x}_j) d\vec{x}_i d\vec{x}_j$$

$$\equiv \sum_{x=1}^{L} \sum_{y=1}^{M} I_{x_i}(x, y) I_{x_j}(x, y) \tag{1}$$

$$W_{i,j,k} \equiv \sum_{x=1}^{L} \sum_{y=1}^{M} I_{x_i}(x, y) I_{x_j}(x, y) I_{x_k}(x, y) \tag{2}$$

where $f_{\vec{x}_i \vec{x}_j}(\vec{x}_i, \vec{x}_j)$ is a joint probability function (j-p.d.f.) of a random process $x(t)$ and each facial box is the function of $x(t)$ at a specific time $t = 1, \ldots,$n. I_{x_i}, I_{x_j} and I_{x_k} are facial boxes.

2.1 *Theory of Time-Order Reconstruction*

One of our working assumptions about time-order is that video cameras are fixed and continuously image moving people, causing changes in facial sizes, norm distances, etc. We observed that the tick-tock of a clock follows another tick-tock. Similarly, the time-order in face images takes a minimum of three time points between frames. Two similar faces becoming smaller, or vice versa bigger, do not imply a correct time-ordering of whether they are walking closer or further away, unless a third similar face becomes even smaller or vice versa even bigger. In other words, in a set of N facial image vectors where $\vec{x}_i \in O(L\times M)$ $i = 1, \ldots, N$, a tick-tock $(\vec{x}_{t_i}, \vec{x}_{t_j}, \vec{x}_{t_k})$ is determined by triple face image vectors $(\vec{x}_i, \vec{x}_j, \vec{x}_k)$ and $W_{i,j,k}(\vec{x}_i, \vec{x}_j, \vec{x}_k) \neq 0$ in the directional order $(t_i \geq t_j \geq t_k)$. The time-order in tandem is determined by a tick-tock, $(\vec{x}_{t_i}, \vec{x}_{t_j}, \vec{x}_{t_k})$, with another tick-tock, $(\vec{x}_{t_k}, \vec{x}_{t_l}, \vec{x}_{t_m})$, that satisfies $W_{i,j,k}(\vec{x}_i, \vec{x}_j, \vec{x}_k)W_{k,l,m}(\vec{x}_k, \vec{x}_l, \vec{x}_m) \neq 0$. This may lead to a higher order graph theory that is not addressed herein

but in a George Washington University PhD thesis which requires higher order correlations to automate the cross association among different sensor modality tracks, achieved by time-ordered reconstruction for each sensor.

2.2 *Self-Reference Matched Filter*

For simplicity, the square (Euclidean) norm distance, also called the LMS distance, in Eq. (3) represents a measurement of similarity (or for neighborhood determination) between two images,

$$Min\ e = Min||\vec{x}_i - \vec{x}_j||^2, \tag{3}$$

where $\vec{x}_i$ and $\vec{x}_i$ are images and represented by a single vector O($L\times M$) of image size $L\times M$ consisting of all gray pixel values of the whole image. Although in general the norm distance may not necessarily be the accurate metric to determine the similarity (e.g., other pertinent features: color of clothes, background of the person). In our applications, Eq. (3) seems to be adequate for image sorting of facial surveillance videos provided that we have a preprocessing called binarized centroid registration of each short exposure in a facial box, inspired by a graceful degradation of scaling in the human visual system (HVS) (see Appendix).

If we explicitly combine the maximum likelihood cost function together with the graph-Laplacian diffusion distance, the similarity metrics becomes identical to our self-reference matched filter (SRMF) least mean square energy as follows:

$$e = Min||\vec{x}_D - \vec{x}_{t_i}||^2/2\sigma = -\log[G_{ij}(\vec{x}_i, \vec{x}_j)]$$
$$= -\log\left[\exp\left(-\frac{||\vec{x}_i - \vec{x}_j||^2}{2\sigma}\right)\right]. \tag{4}$$

Coifman et al. have adopted the graph-Laplacian spectral theory $G_{ij}(\vec{x}_i, \vec{x}_j) \equiv \exp(-||\vec{x}_i - \vec{x}_j||^2/2\sigma)$ to sort facial poses as an example of nonlinear dimensionality reduction.[8,9] Without creating any confusion, we have associated, for two sets of readerships familiar with video processing and with nonlinear dimensionality reduction (NDR) graph theory, a snapshot time index $\vec{x}_{t_i}$ freely with the vortex node $\vec{x}_i$ of a face image. The minimization process, Min in Eq. (4), is due to the opposite algebraic signs. The first term $\vec{x}_{t_i}$ represents an individual face image and $\vec{x}_D$ is the first pass average through all faces as the statistically centroid correct but blurred LE defined as:

$$\vec{x}_D \cong \frac{1}{N}\sum_{t=1}^{N}\vec{x}_t. \tag{5}$$

The instantaneous output $\vec{x}_{t_i}$ is the SE facial images. The filter can select a SE with minimum square norm distance error to long exposure.[23,25] An example is illustrated in Fig. 3.

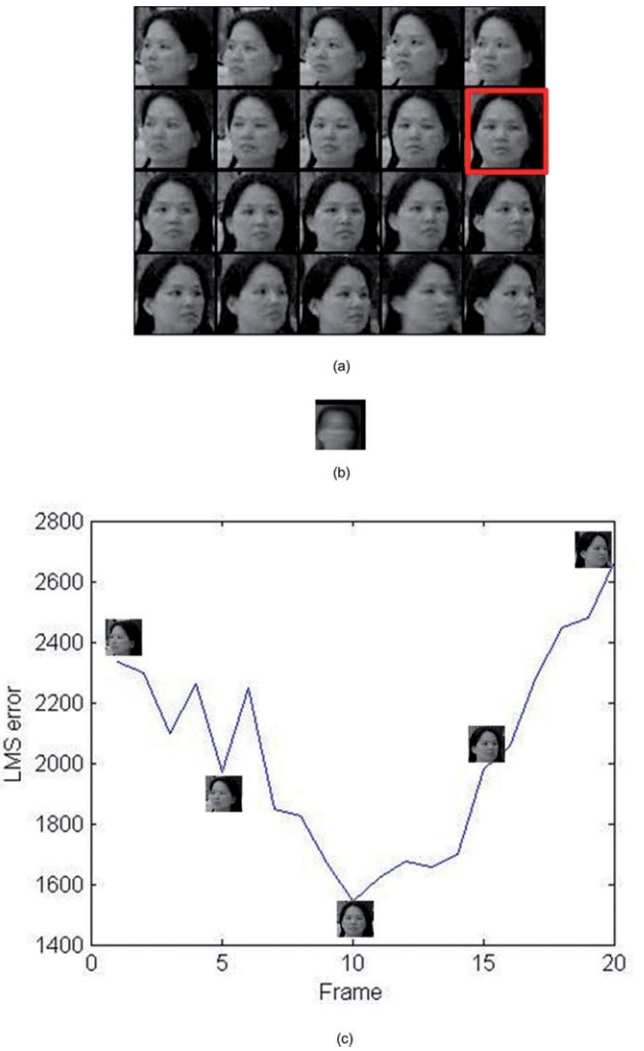


Fig. 3 SRMF illustration: In this single video setup, the passenger is limited to walking through a narrow corridor toward the camera. (a) SE of a video from 70th to 89th frame with different sizes, frames from 70 to 79 have 58×58 pixels and from 80 to 89 have 59×59 pixels (8 bits dynamic range). (b) For frames from 70 to 79, the pixels in the 59th row and column are set to 0. From Eq. (5), LE is blurred but is the statistically correct centroid of the frontal pose, 59×59 pixels (8 bits DR). (c) Square norm distance error between SE and LE. The minimum norm distance error of SE happens on the 79th frame and the bit error per pixel is 0.05 due to size and background difference.

3 Implementations

In remote video surveillance, facial data acquired from video is generally a lower resolution than photographic biometric resolution, which typically must have at least 12 pixels between the eyes defined.[26] Our time-order reconstruction algorithm must follow this constraint as well. A static face is regarded as a collection of pixels with light intensity on each pixel. This collection of pixels also specifies the Cartesian coordinates of a point with respect to a set of axes. Therefore, a point can represent a face in an abstract image space.[4,5,27] For facial images of walking passengers, they can be visualized as lying on a three-dimensional manifold that can be parameterized by two pose angles and changes of resolution, assuming light intensity is constant in a short duration, perhaps lasting over 10 s.

3.1 *Solving Variable Faces by Focusing at the Binarized Centroid for a Graceful Degradation of Facial Scaling*

Our goal is to sort the time-order of changing faces among backgrounds. We are motivated from HVSs to derive a morphologically efficient preprocessing technique. It allows us to track maximum overlapping common foreground faces among changing perspectives and size distances. The outputs of a face detection system on chip, in Fig. 9(a), are

boxes in different sizes. Some boxes have a lot of background information and some boxes do not. Before comparison between facial boxes, image preprocessing is needed to deal with the different sizes of boxes and any unwanted background information. Mathematically speaking, our approach is equivalent to a video version of gray-scaled self-reference Wiener matched filter. Physiologically speaking, we have 150 million night vision cones distributed nonuniformly in polar-exponentially grids at the fovea, which is densely packed in the middle of the fovea and exponentially dropping out in the peripheral. The location of each rod is denoted by a vector r_i. There are about 100 rods that share an integrator called ganglion firing through a single neurofiber at the fovea output denoted as a vector u_j to the cortex 17. Altogether there are millions of neurofibers that are closely packed together into a uniform output bundle toward the lateral geniculate nucleus.[24] For example, the input i'th pixel of facial boundary is $r_i = \exp(u_j)$, which is equivalent to the output $u_j = \log(r_i)$ achieving a massive parallel flow-through logarithmic scaling transform without explicit computation. In other words, when a face changes its size with a scale factor s at $r_i' \equiv r_i s = \exp(u_i')$ implying $u_i' = \log(r_i') = \log(r_i s) = \log(r_i) + \log(s) \cong u_j$. For example, a size of half $s = 2$ change gives $\log_e(2) = \ln(2) = 0.69$ which is a small fraction over many pixels. This fan-in polar exponential grid (PEG) architecture[24] facilitates the logarithmic computation of millions of rods' coordinates in a parallel flow through the PEG. This real-time "algo-tecture" is one of the wonders of the human visual system. This is the reason why a hunter running after a prey in the moon light can gracefully tolerate the rapid changes of the prey, that allows the hunter to integrate continuously for a better signal to noise ratio (SNR) over several hundred photons falling upon the bundle of 100 rods via another fan-in architecture connecting hundreds of rods to a single ganglion toward lateral geniculate nucleus (LGN) and cortex 17. Incidentally, a rod detecting a single photon at SNR $= 40$ due to the wavelength at night is about 1 μm, equivalent to 1 eV, while the body temperature at 37°C is equivalent to 1/40 eV. The spatial spreading of a single micron photon can cover a 100 rod bundle at 100 nm each.

To achieve a morphologic registration of the facial invariant centroid, we will approximate the aforementioned color hue for face detection by the gray scale intensity thresholding. For example, we choose the Heaviside step function of the gray scale g by an arbitrary θ threshold: $u = Heaviside|\frac{g}{\theta} - 1| = 1$ if $g \geq \theta$; zero if otherwise. Instead of the logarithmic transform of a million coordinates for the scale invariance, a real-time morphologic preprocessing is a simple binarized centroid determination at the middle of unity counts, as shown in Fig. 4.

Binarized centroid registrations: For N face boxes (SE) in different sizes, a complete segmentation of a face box R is a finite set of regions $R_1, \ldots, R_s$,

$$R = \cup_{i=1}^{s} R_i, \quad R_i \cap R_j = \phi, \quad i \neq j. \tag{6}$$

The morphologic segmentation likewise resulted from the threshold operation that transforms an input gray-scale image f to an output binary image g as follows:

$$\begin{aligned} g(i,j) &= 1 \quad for \quad f(i,j) \geq T, \\ &= 0 \quad for \quad f(i,j) < T, \end{aligned} \tag{7}$$

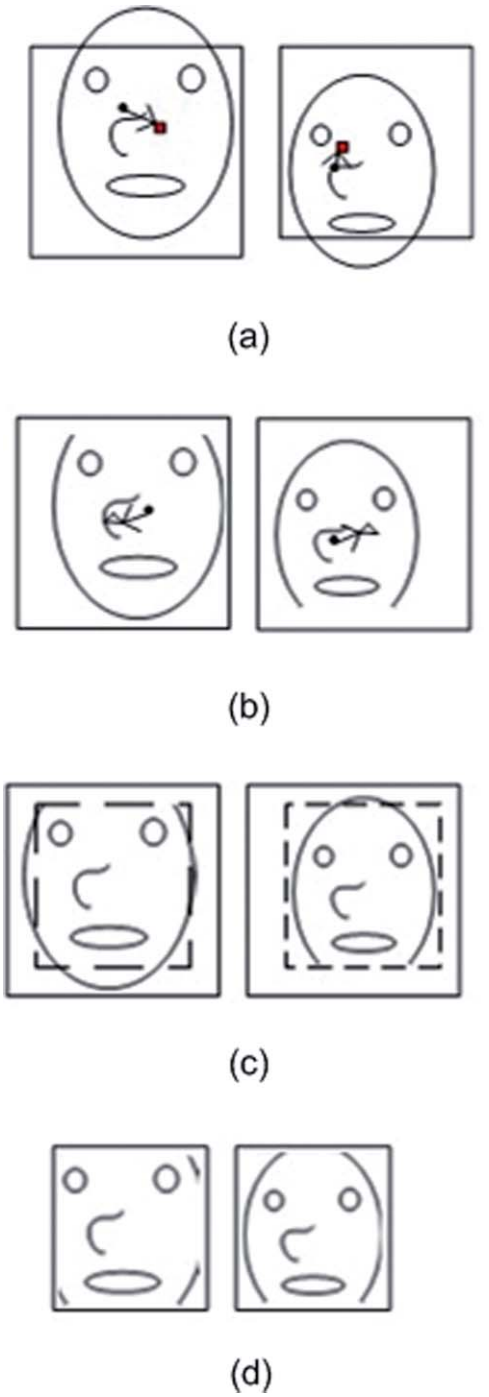


Fig. 4 Technique to find a common central facial portion. Facial center allows us to cut an identical central portion of each face. (a) Translation from facial center to box center: two facial boxes are in different sizes. The dot is the binarized centroid of the face and the square is the geometry centroid of the box. The first translation is to move the face from binarized centroid of the face to the geometry centroid of its box. After that, a transformation is applied to every face box from its original coordinates to the laboratory coordinates. (b) Registration of every face box to the centroid of the long exposure: the triangle is the binarized centroid of the long exposure in Eq. (5). (c) Choose common portions of faces: after the two translations, major facial features such as the eyes, nose, and mouth of every face box are relatively close under the same coordinates. It is crucial so that norm distance or Gaussian diffusion distance can be adequate for a similarity measurement. (d) Clean cut foregrounds: every face box is cut by a rectangle template to get rid of the unwanted background information.

where T is the threshold, $g(i, j) = 1$ for the face, and $g(i, j) = 0$ for the background.

$$\overrightarrow{bc}_i = \frac{1}{n}\sum_{j,k}^{n} g(j,k), \quad where \quad g(j,k) = 1, \tag{8}$$

$$(\overrightarrow{bc}_i - \overrightarrow{gc}_i) = 0, \quad for \quad i = 1, \ldots, N, \tag{9}$$

$$\overrightarrow{bc}_i^{*} - \overrightarrow{bc}_{LE} = 0, \quad for \quad i = 1, \ldots, N, \tag{10}$$

where $\overrightarrow{bc}_i$ is the binarized centroid of the face in i'th face box (SE), $\overrightarrow{gc}_i$ is the geometric centroid of the i'th face box, $*$ in $\overrightarrow{bc}_i^{*}$ indicates the translation from the i'th coordinate to the laboratory coordinates system that covers every face box, and $\overrightarrow{bc}_{LE}$ is the binarized centroid of the long exposure, Eq. (5), determined by Eq. (8). The first pass centroid registration is demonstrated in Eqs. (9) and (10) is the second pass centroid registration. The determination of T's gray scale value is critical in this algorithm. For optimization of the threshold, different strategies have been developed for different scenarios.[28] The robustness of the binarized centroid of facial boxes is discussed in the Appendix.

3.2 *Binarized Centroid Registration Algorithm*

Step 1: Find the binarized centroid of every face and translate the binarized centroid of face to the geometry centroid of each face box.

Step 2: Then, translate every face box to the laboratory coordinate

Step 3: Registration of every face box to the centroid of the long exposure from Eq. (5).

Step 4: Remove unwanted background by rectangle template.

In two-pass binarized centroid registrations, we used the properties of the benchmark for face detection techniques; the eye(s), nose, and the mouth have to be inside the facial boxes if they are in video. Otherwise, the detection result is incorrect. When we developed and tested our algorithms, we assumed that every cut-out facial box was correct. In other words, every facial box should have eye(s), a nose, and a mouth in it. It also implied that no matter what size the facial boxes were, they were partially overlapped in content, area of eye(s), nose, and mouth even when they were in different sizes and resolutions. With the condition mentioned above, in most facial poses, eyes, nose, and mouth are always around the face centroid in the facial boxes except poses heading left-most and right-most.

3.3 *ZC Algorithm*

The theory of time-order from Sec. 2 is implemented in an iterative research fashion herein. The video sequence captured by the camera satisfies the following inequality, Eq. (11), in an abstract manifold:

$$\|O_t - O_{t\pm1}\| < \|O_t - O_{t\pm s}\|, \; t > s > 1 \; and \; t, s \in \text{Integer}, \tag{11}$$

where O_t is a node representing a facial box and the subscript labels t and s are the time indices of the time-ordering of nodes on the manifold. The $\|\cdot\|$ represents the geodesic distance between nodes of the original time-ordering of nodes on the manifold. If Eq. (11) holds true, it implicitly indicates that a trajectory of time-ordering (of a passenger) cannot intersect itself and other trajectories of time-orderings (of other passengers) on the manifold. It also indicates that finding the time-order is equivalent to find out the geodesic order which has the lowest total neighborhood geodesic distances. We plot the geodesic distance (simulated by LMS distance) as the contour map over the N^2 dimensionality surface in Fig. 5.

In implementation, we adopt LMS distance between facial boxes to approach the geodesic distance on the manifold. Explicitly, this approach may be skewed sometimes. If a video sequence cannot satisfy Eq. (11) all the time, some frames may be in an incorrect time-order. The impact of incorrect alignment, partial frames in sequence, and of time-order is discussed in Sec. 7. The meta algorithm is stated as follows:

3.4 *Steps of the ZC Algorithm*

Step 1: Apply two-pass binarized centroid registrations introduced in Sec. 3.1 to approximate the area of the eye(s), nose, and mouth in every face box. The approximation result has to meet Eq. (11) to insure the result of the ZC algorithm is correct.

Step 2: Determine the anchor pose: The long exposure is calculated by Eq. (5). The anchor pose is the nearest neighbor of the long exposure, selected from the input facial boxes.

Step 3: Construct a two nearest neighborhood matrix between each node and an adjacency graph by the two nearest neighborhood matrix.

Step 4: Determine the weights of connections in the adjacency graph. The weight of the nearest neighbor connection is 3 and the weight of the next nearest neighbor connection is 1. Every node is supposed to connect two neighbors except the head and tail nodes that we do not know. When a node has more than two relations with other nodes, the connections are selected by the top two highest weights of relations.

4 Discussion

In this section, two phenomenon are discussed. The first is the impact of leak-through background information in face boxes for the neighborhood determination. The second is what the consequence is if the step of iterative two neighborhoods comparisons in ZC algorithm is replaced by a direct neighborhood determination by LMS distance.

4.1 *Impact of Background Information of Face Boxes for Neighborhood Determination*

Many spectral clustering methods[4–9,27,29] are developed for sorting facial poses at a fixed distance with clear backgrounds. As mentioned in Sec. 2, norm distance or LMS distance has been used for the neighborhoods (or similarity) determination of facial poses in adjacency matrix.[4–9,16,27,29] However, in video surveillance, when passengers move, the facial resolutions of passengers change over time with respect to a fixed video camera. Moreover, the backgrounds of face boxes are not clear. In this case, the pixel-by-pixel LMS distance could not determine the correct neighborhoods due to the leak-through background information.

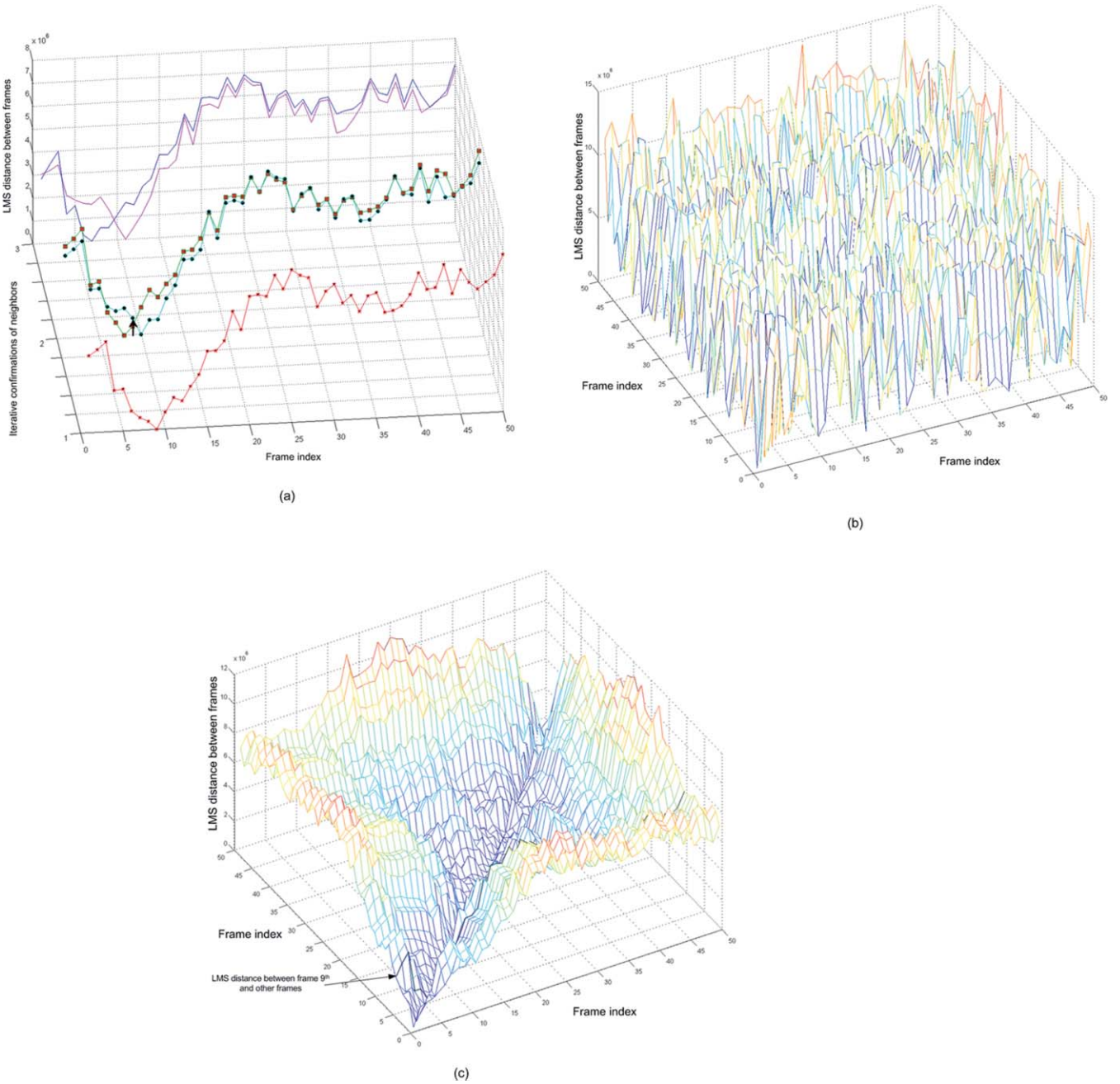


Fig. 5 Iterative confirmation logic for time-ordering reconstruction and contour map of N nodes over the N^2 dimensionality surface. (a) Iterative confirmation logic of s-curve poses ZC algorithm: In this example, facial boxes are sorted in a manner of time-order. The first confirmation is to select the two nearest neighbors in LMS distance of the 9th frame, that are the 8th and 10th frames. The second confirmation is to determine whether the 9th frame is also one of the two nearest neighbors of the 8th and 10th frames in LMS distance. The third confirmation check is to determine another neighbor of the 8th and 10th frames, that is the 7th and 11th frames, respectively, and so on. The poses in the 9th and 37th frames are similar. However, with the morphological preprocessing and the iterative confirmation logic given in this paper, we can find the correct neighbors of the 9th frame without confusion. The axes are frame index, iterative confirmation of neighbors, and LMS distance between frames. (b) Contour map of N random nodes over the N^2 dimensionality surface. N is 50 in this example. The axes are frame index, frame index, and LMS distance between frames. (c) Contour map of N time-ordered nodes over the N^2 dimensionality surface. N is 50 in this example. The axes are frame index, frame index, and LMS distance between frames. The correct time-ordering of facial boxes is found when the summation of all neighborhood pair-wise distances between nodes happens to be at the absolute minimum on the manifold.

The under-sampled real world surveillance images of faces shown in 14 facial poses in a 4 s window are in Fig. 6. The LMS distance may not be sufficient to determine the similarity between faces when the changes of background are drastic for each face box shown in Fig. 6(c). However, with binarized centroid registration preprocessing applied before sorting, the result became much more robust to be able to tolerate a factor of 2 in sizes over about 10 s walking time.

4.2 *Consequence of Replacing the Product of Two Triplet Correlation Checks in ZC Algorithm*

As mentioned in Sec. 2, we assumed that the time-ordering is a product of triple image correlation $W_{i,j,k}W_{k,l,m}$ over the time-ordered facial image space. In the implementation of time-ordering reconstruction, a product of triple image correlation is equivalent to iterative two neighborhoods comparisons in ZC algorithm presented in Sec. 3. What is the consequence if we replace the product of triple image correlation by a pair-wise correlation? In other words, what is the sorting result if we replace iterative two neighborhoods comparisons in ZC algorithm by a neighborhood determination by LMS distance? The answer is that the result is a fast approximation of embedding in line for facial poses similarity sorting in a short duration, several seconds. This fast approach of similarity sorting algorithm is stated as follows:

4.3 *Steps of the Similarity Sorting Algorithm*

Step 1: Apply two-pass binarized centroid registrations introduced in Sec. 3.1 to approximate the area of the eye(s), nose, and mouth in every face box.

Step 2: Determine the anchor pose: The long exposure is calculated by Eq. (5). The anchor pose is the nearest neighbor of the long exposure, selected from the input facial boxes.

Step 3: Neighborhood determination by LMS (or norm) distance. The connectivity of the anchor pose and other poses is determined by the LMS (or norm) distance. Other choices of distance, such as face, feature, distance, etc., are considerable if the LMS distance is not adequate to present the similarity. The result is a fast approximation of embedding in line for similarity sorting in a short duration, several seconds.

5 Simulations

The ZC algorithm was coded in MATLAB on a PC with an Intel i-7 920 processor and 12 GB memory. The time of loading facial images is excluded in the execution time. The first simulation is a similarity sorting test of 60 facial poses of 1 person in different order. The face boxes are in different sizes and resolutions over time. In this simulation, it shows the robustness of our similarity sorting algorithm. No matter what orders the facial poses are, the sorting result is the same, as illustrated in Fig. 7. It is a robust and real-time sorting approximation of moving objects in a short time window, based on the two-pass centroid registrations and SRMF proposed in this paper.

For a simulation of time-ordered reconstruction, detailed steps are illustrated in Fig. 8. There are seven facial boxes in a random order. By the zipper chain algorithm introduced in Sec. 3, the norm distance matrix is built up by the input facial boxes. After that, the two nearest neighbor matrix are constructed in Fig. 8(d). The adjacent graph is determined by the two nearest neighbor matrix and the result is in Fig. 8(f).

6 Applications

In this section, we briefly consider several possible video surveillance applications for the algorithmic framework developed in this paper.

6.1 *Video Organization Principle*

Electro-optical/infrared videos with the help of sleep-wake acoustic/seismic trigger for the persistent surveillance can last about 7 days and 24 hours per day. The video database captured by several surveillance cameras can easily exceed peta-bytes worth of data, and the inspectors have to spend a lot of time sorting out useful information from such an enormous database. An efficient strategy for reducing video database storage and an efficient sorting algorithm of the content of video database has been the theme of this paper.

In general, surveillance looks for the most wanted passengers at traveler transit locations. The facial boxes of all passengers have been extracted in parallel per frame into the video storage from megabyte per frame to kilobyte of facial boxes in Fig. 9(a).[30] In addition, an automatic selection of frontal views for each person can save another 3 orders of magnitude, e.g., walking through an inspection corridor already shown in 10 min transit time per person in Table 1.

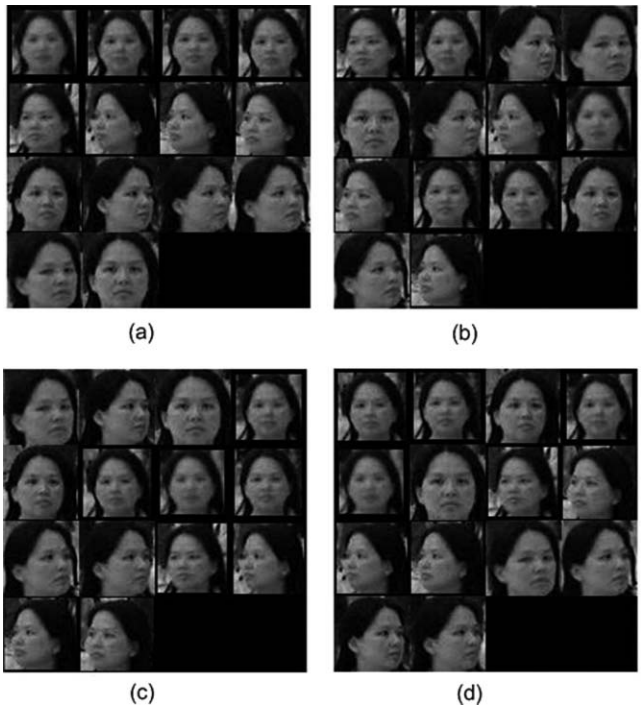


Fig. 6 Impact of background information for neighborhood determination (a) Fourteen selected time-ordered facial boxes of a passenger in different sizes, resolutions, backgrounds, and poses. The time window between the first frame and the last frame is 4 s. In (a), the change of image resolutions and backgrounds are significant. (b) Random order: The fourteen facial boxes are in a random order. (c) The sorting result by LMS distance when the difference of resolutions and background information between each image is significant. The result can be improved by two pass binarized centroid registrations. (d) Before sorting, two pass centroid registrations were applied and background information was removed as mentioned in Sec. 3.1. This result is in a much better shape. It shows for neighborhood determination that the LMS distance is less sensitive to the resolution difference than leak-through background information between each frame.

6.2 *Aided Target Recognition for Video Surveillance with Mug Shots*

In an effort to strengthen National Security, spotting people of interest in airport surveillance is critical. Currently, airport surveillance relies on security personnel to tediously match a wanted person in a crowd with a mug-shot. An automated video surveillance and recognition system is desirable to

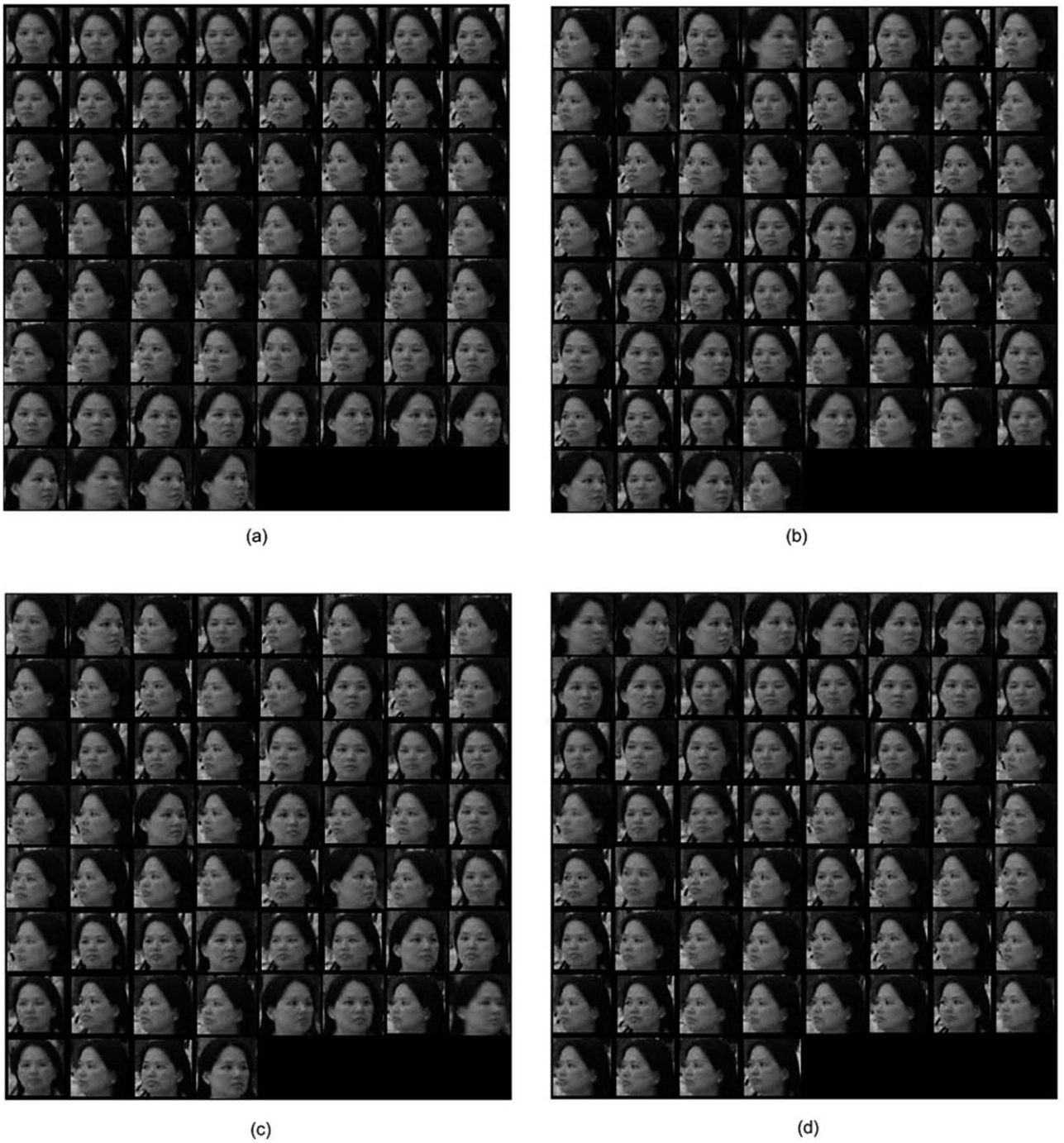


Fig. 7 Sixty different facial poses in two random orders in (b) and (c). The sorting result of (b) and (c) by the similarity sorting algorithm proposed in this paper is in (d). The execution time is 0.65 s including two pass centroid registration preprocessing in average. The sorting results from different order of facial poses are the same.

boost efficiency. However, passengers' facial biometric data captured by video can be in different resolutions, poses, and even be disguised from mug-shots.[14,26,31,32] It is how and why, until then, there is not any automated video surveillance system adopted for airport surveillance. A man in the loop for surveillance is still necessary. Thus, a database-cueing aided-target recognition (AiTR) video surveillance system is recommended to aid security personnel for better recognition performance.[33]

The fundamental concept is to maximize the overlapping pixels on target between faces and mug shots at matched poses and inspectors can be prompted for a closer look at a specific passenger. To achieve this goal, the AiTR system has to detect and sort passengers poses in real-time through surveillance cameras in a corridor illustrated in Fig. 9(b). The robust and real-time SRMF algorithm proposed in this paper is well suited for such an AiTR surveillance system in Fig. 9(c).

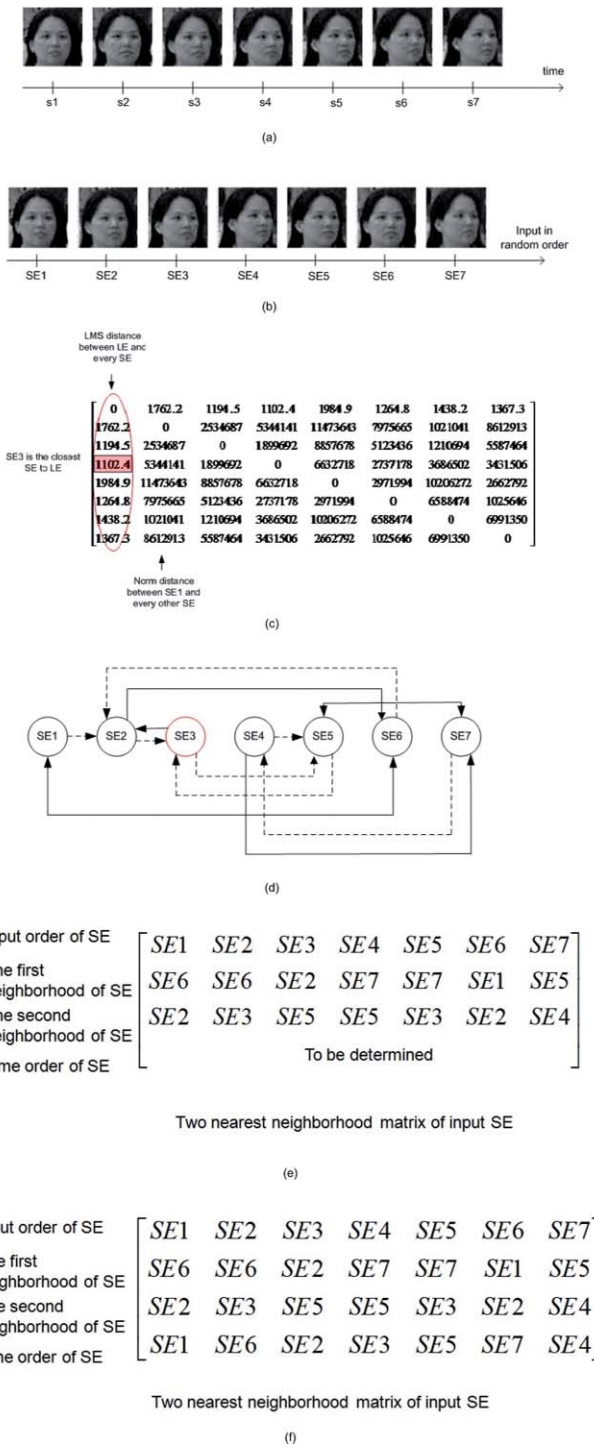


Fig. 8 (a) Seven facial boxes in a time-ordering. (b) The same seven facial boxes in (a) but in a randomized order. (c) The norm distance matrix of each facial box. The first column is the root of norm distance between the long term average and each facial box. It determined the 3rd facial box is the pivot point in this simulation. (d) The two nearest neighborhood matrix of each facial box. (e) The adjacent graph of seven facial boxes constructed by the two nearest neighborhood matrix in (d) and (f) The final result is the same as (a).

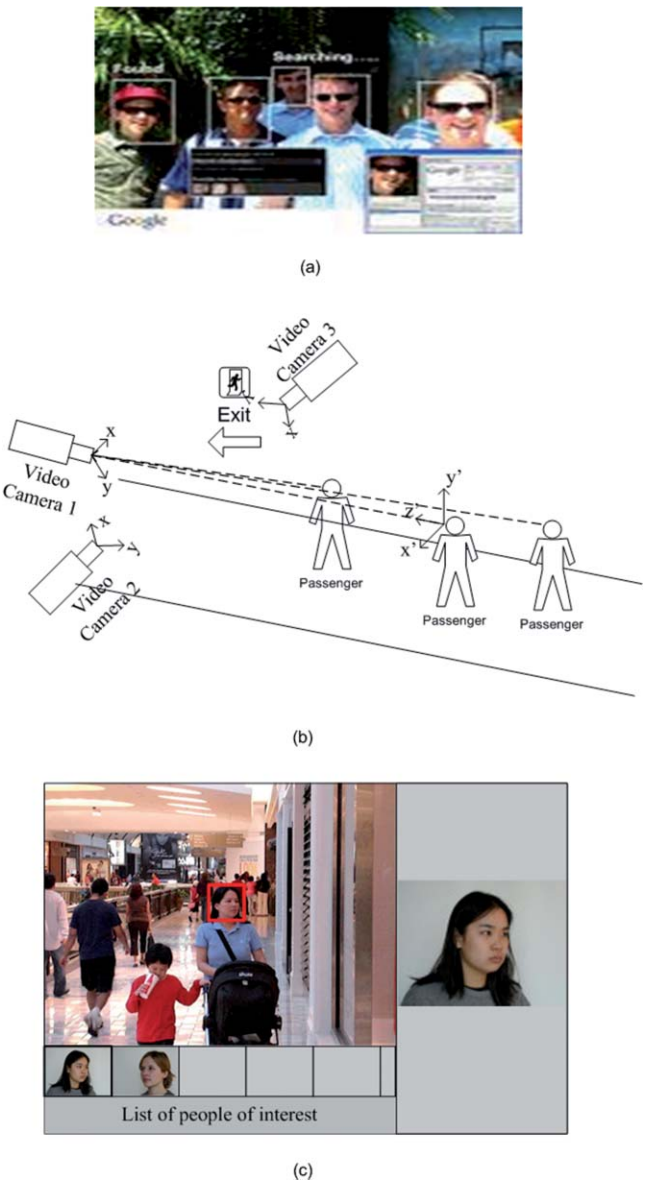
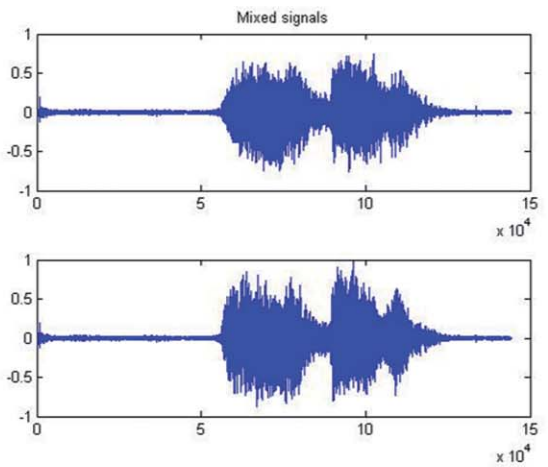


Fig. 9 (a) FD is well developed and commercialized in most modern cameras by different brands. This example of FD is from Google. (b) An example of multiple cameras adopted for surveillance in a long corridor for monitoring passengers. (c) An aided target recognition system for inspectors to take a closer second look at the specific passenger who has the same (or similar) pose of mug shots in the system. The two mug-shots are selected from MIT face databases.[33]

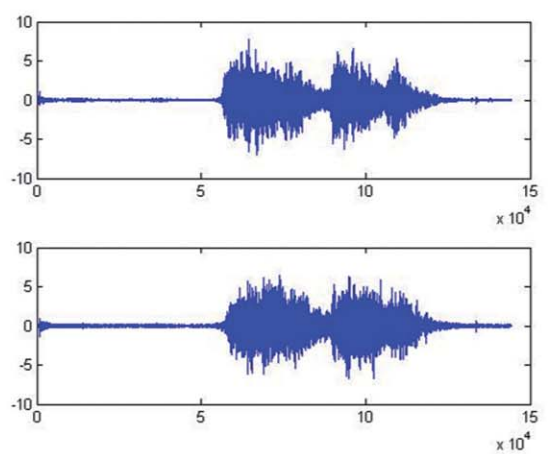
6.3 *Association Across Sensory Modalities*

The cross track association problem of archival video and audio becomes challenging after a different compression algorithm using current technologies. Without the basic spatiotemporal-order reconstruction, a close-up lip read and gesture motion cannot easily associate without ambiguity a remotely imaged passenger and with *in situ* audio recording by local microphones. The facial boxes of a crowd can be detected efficiently in random order by the face detection system on the chip mentioned in this paper. The next step is to sort out those facial boxes in a time-ordering so that we may be able to read lips, gestures, and other body languages in the right order. For example, in Fig. 10(b), "where is exit?" and "I am tired." are recorded by dual microphones with airport music and background noise. In Fig. 10(c), two speeches are separated, e.g., by fast independent component analysis (ICA) algorithm based on different positive Kurtosis values.

(a)



(b)



(c)

Fig. 10 (a) Two passengers are walking in a corridor. (b) Mixed signals of audio from two passengers, recorded by dual microphones. (c) Signals separated by fast ICA.

For a passenger who is looking for an exit, the passenger has a higher chance to look around, turning to face right and left, instead of looking down at the floor. The time-ordering sorting algorithm proposed in this paper is useful for cross track association of video and audio.

7 Conclusion

In this paper, we have presented an efficient time-order reconstruction algorithm of sorting facial boxes in order to associate remote video and local audio for answering the surveillance challenge—"who speaks what, where and when." We show that it is crucial to have a morphological image preprocessing which begins at a binarized centroid registration, in order to find a maximum overlapping common region from all facial boxes in different sizes and coordinates. Then a ZC algorithm is developed by considering the products of pair correlations functions in tandem verification. Consequently, each face box is linked together in a time-order, as if the zipper's teeth are mounted overlapping one another. This ZC algorithm reconstructs the time-order linking the video manifolds of walking facial boxes and audio manifolds of *in situ* voice recorders. Independent of the cameras setup configuration, the video sampling rate is an important parameter. Our experience shows to be sufficient at 20 to 30 Hz for an average walking pace of 3 to 6 miles per hour. However, if a passenger runs faster than the walking limit, the camera sampling rate must be increased to reduce the image blur. We combine the anchor face that is automatically discovered by means of self-reference matched filter with the time-order ZC sorting capability, and we hope to find the time-order of each person to allow us to associate each with his or her own voices recording. In general, the average speaking speed is 3 to 4 words per second. In the movie industry, the technician can sync the voice recording by a lip reading technique when a close up facial detail is available. However, the remote video may not have the detail. In this situation, our goal is to identify the corresponding speakers according to the recorded audio.

The time-order reconstruction of facial poses belongs to a nondeterminant polynomial-time (NP) complete problem and the proof is in Appendix B. Given N image poses, the ordering index runs from 1 to N. Then, there are $N!$ possibilities to assign the ordering index to each pose. We plot the LMS distance as the contour map over the N^2 dimensionality surface in Sec. 3.3. The time-order reconstruction is equivalent to find the geodesic order that has the lowest total neighborhood distances on an abstract manifold. ZC algorithm is a heuristic solution for this NP complete problem. Equation (11) is the constraint of the solution that each time-ordering trajectory on an abstract manifold is prohibited to intersect with itself or other time-ordering trajectories.

Appendix A: Real World Challenges of Cross-Track Association: Time-Ordered Reconstruction

Cross sensory track association is an unsolved Empire 2010 sensor challenge problem proposed by the Office of Naval Research. The real world field test is much more difficult from a laboratory test. In the real world, there are many more uncontrollable factors that may affect test results. For example, on a rainy day, an outdoor video may be unclear, or passengers in the video may have heavy shadows on a sunny day. In addition, passengers may be obscured by each other or obstacles such as trees, pillars, cars, etc. To overcome those uncontrollable environmental factors, we take advantage of the strategy adopted in this paper: cutting facial boxes and removing redundant background information can save up to

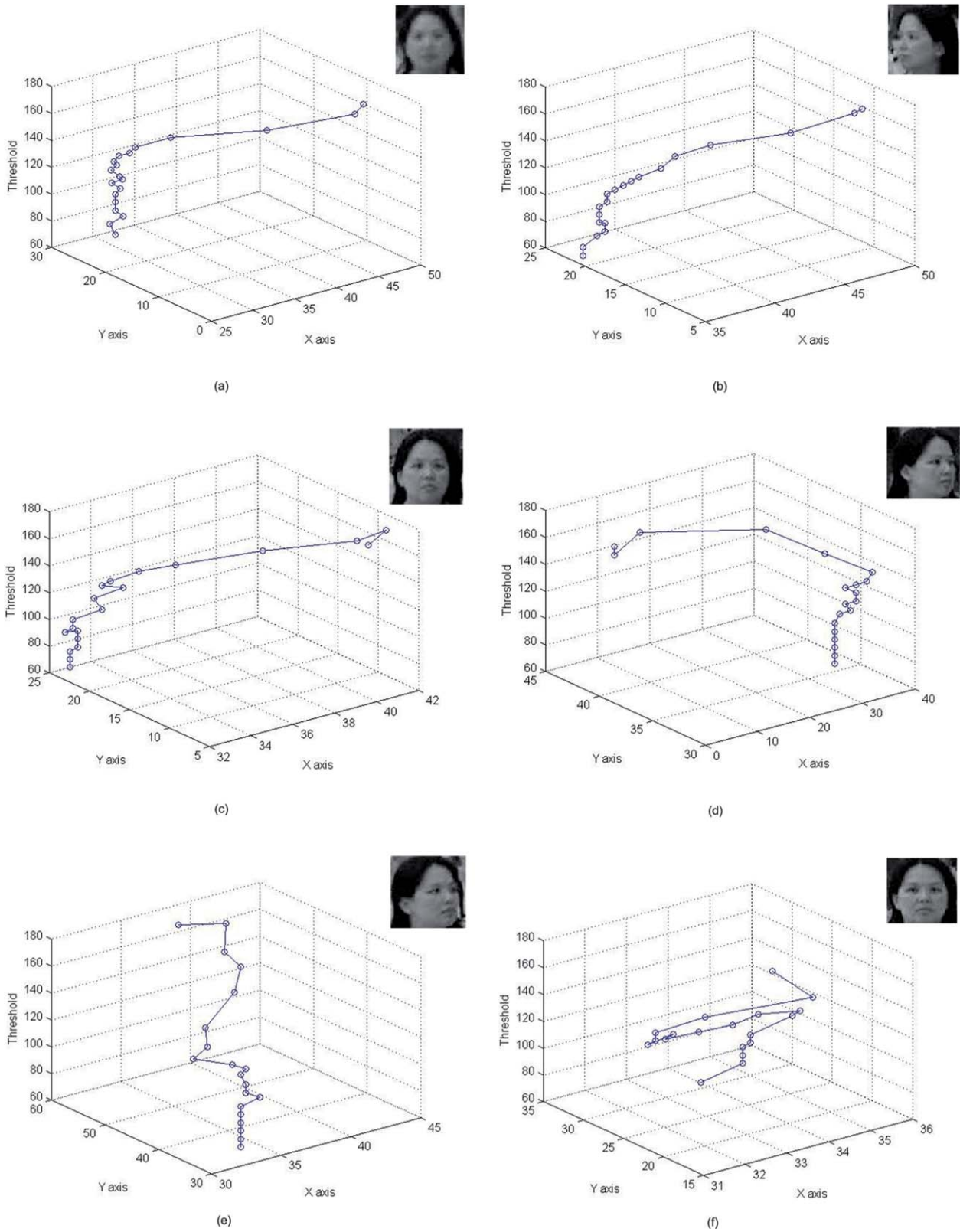


Fig. 11 The tested poses are in different sizes and resolutions. The *X* axis and *Y* axis are the coordinates of binarized centroid calculated by different thresholds. The threshold is varied from 60 to 174 (gray value) under 256 dynamic ranges. In the test, the binarized centroid of most poses has a small variation with different thresholds until the threshold is greater than the 130 to 140 range. In other words, the binarized centroid of every pose is fairly close between the threshold range of 60 to 140, over 256 dynamic ranges.

6 orders of magnitude in data storage (see Table 1), and by increasing cameras, coverage density is increased. The problem of facial continuity in occluding and unpredictable environmental factors may be mitigated by an increase in the sensor density.

It is known that our HVS gracefully degrades the scale of faces by focusing on the centroid of the faces, say nose and/or eyes. This is accomplished by taking a log transform along the radial direction of the facial center so that the peripheral boundary sizes are less sensitive to our eye. Thus, if we wish to ignore a varied facial size, we take a binary truncation of a gray-scale image for efficiency reasons (equivalent to zero, when the log of unity at the threshold, and to one, when the gray value is above the threshold). If the centroid of a face is so chosen within 1 or 2 pixels, the peripheries of the face become less sensitive as demonstrated in Fig. 11, where binarized centroid sensitivity tests of different faces in different poses and sizes are given with different thresholds. In the results, the binarized centroids of facial boxes are gracefully degraded with threshold changes between 60 and 140, over 256 dynamic ranges. In our analysis, it is because the area of the face is over at least 50% of each face box so that the binarized centroid is insensitive to threshold changes. In general, most correct face boxes from FD systems on chip also follow the condition mentioned above with very few exceptions.

In our ZC algorithm, the product of triple correlations with common nodes in the middle, i.e., $W_{i,j,k}W_{k,l,m}$ is equivalent to the products of pair correlations in a double verification, which saved computation costs.

Time difference, time-ordering, and arrow of time are different. Time-ordering is the change of the time difference, while the arrow of time is the flow of the time difference. In this paper, we have presented a theory and a heuristic implementation of time-ordered image reconstruction. Understanding the arrow of time may be achieved by calculating the Boltzmann entropy of time-ordered images; the direction of flow as the increase of entropy.

We have re-established the time-ordering in an airport persistent surveillance cases of those random collection of facial poses at arbitrary sizes, after they have been cut automatically by a video system on chip using the facial color hue. Given multiple cameras in the transit corridors, we have demonstrated that we can determine their cross-sensor modality correlation by the re-establishment of time-order of each sensor. However, for Navy large area Marine harbor and Department of Homeland Security open boarder surveillances there remains the challenge to be formulated and tackled.

Appendix B: Geodesic Neighborhood Re-Ordering is a NP Complete Problem

Lemma:

Geodesic spatiotemporal neighborhood re-ordering of N projected poses of a moving person is a NP complete problem.

There are combinatorial possibilities in the re-ordering of N poses: the first place has N choices, and the second place has $N-1$, etc. The total number is explosive as N increases:

$$\text{total number of ordering} = N(N-1)(N-2)\ldots 1/2 = N!/2 \tag{12}$$

*Theorem***:** The geodesic neighborhood ordering of video projections of a moving person in space and time must be at the absolute minimum of the total difference distances between all successive order.

Proof: By induction

Given $N = 3$ poses. Assume the correct spatiotemporal neighboring order is A, B, C, then

$$|\mathrm{A}-\mathrm{B}|^2 + |\mathrm{B}-\mathrm{C}|^2 \le |\mathrm{A}-\mathrm{C}|^2 + |\mathrm{C}-\mathrm{B}|^2 \qquad \text{Q.E.D.} \tag{13}$$

If the inequality is satisfied, then the correct neighborhood order is A, B, C.

Assume N to be true, it is trivial to prove $N+1$ is true.

Q.E.D.

Acknowledgments

We thank Mr. Louis Larsen and Mr. Jeffrey Jenkins at NVESD and Professor Harrington, Professor Carroll, and Professor Doroslovacki at ECE Department in GWU for useful discussions.

References

1. G. G. Walter, "A sampling theorem for wavelet subspaces," *IEEE Trans. Inf. Theory* **38**(2), 881–884 (1992).
2. D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006).
3. H. H. Szu and B. A. Telfer, "Mathematics of adaptive wavelet transforms: relating continuous with discrete transforms," *Opt. Eng.* **33**(7), 2111–2124 (1994).
4. J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science* **290**, 2319–2323 (2000).
5. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science* **290** 2323–2326 (2000).
6. M. Belkin and P. Niyogi,"Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.* **15**(6), 1373–1396 (2003).
7. U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.* **17**(4) (2007).
8. R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, **21** 5–30 (2006).
9. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps," *Proc. Natl. Acad. Sci. U.S.A.* **102**(21),74267431 (2005).
10. K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philos. Mag.* **2**(6), 559–572 (1901).
11. I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed., Springer-Verlag, New York (2005).
12. X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," *Proc. 10th IEEE Int. Conf. on Computer Vision*, pp. 1208–1213 (2005).
13. S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang, "Discriminant analysis with tensor representation," *Proc. Inter. Conf. Computer Vision and Pattern Recognition*, **1**, 526–532 (2005).
14. D. Zhong J. Han, X. Zhang, and Y. Liu, "Neighborhood discriminant embedding in face recognition," *Opt. Eng.* **49**(7), 77203 (2010).
15. F. R. K. Chung, *Spectral Graph Theory*, AMS Publications, Rhode Island (1997).
16. S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 40–51 (2007).
17. H. Haken, *Cooperative Phenomena*, Springer-Verlag, New York (1973).
18. H. Haken, *Information and Self-organization*, Springer-Verlag, Berlin (1988).
19. H. Szu, "Asymmetric GT of Social Networks," *Proc. SPIE* **7703**, 770304 (2010).
20. C. Zhang and Z. Zhang, "A survey of recent advances in Face Detection," Technical Report, Microsoft Research, June 2010, http://www.research.microsoft.com.
21. M. I. Tolea, et al, "Sex-Specific correlates of walking speed in a wide age-ranged population,"*J. Gerontol. B Psychol. Sci. Soc. Sci.* **65B**(2), 174–184 (2010).
22. T. Osaragi, "Modeling of pedestrian behavior and its applications to spatial evaluation," *Proc. of AAMAS*, **2** (2004).

23. H. Szu and J. A. Blodgett, "Self-reference spatio-temporal image-restoration technique," *Opt. Soc. Am.* **72**, 1666–1669 (1982).
24. H. H. Szu and R. A. Messner, "Adaptive invariant novelty filters," *Proc IEEE* **74**, 518–520 (1986).
25. H. Szu, J. A. Blodgett, and L. Sica, "Local Instances of Good Seeing," *Opt. Commun.* **35**, 317–322 (1980).
26. D. Gorodnichy, "Video-based framework for face recognition in video," in *Proc. 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, Victoria, British Columbia, Canada, pp. 330–338 (2005).
27. H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science* **290**, 2268–2269 (2000).
28. M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*, Pacific Grove, CA: Brooks/Cole publishing company, 2nd edition (2003).
29. I. Fodor, "A survey of dimension reduction techniques," US DOE Office of Scientific and Technical Information (2002).
30. http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html
31. R. Willing, "Airport anti-terror systems flub tests face-recognition technology fails to flag'suspects'," in *USA TODAY* (September 4, 2003). Available at http://www.usatoday.com/usatonline/20030902/5460651s.htm.
32. K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based faces recognition using probabilistic appearance manifolds," in *Proc 2003 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2003)*, Madison, Wisconsin, pp. 313–320, Vol. 1 (2003).
33. M. K. Hsu, T. N. Lee, and H. Szu, "Video surveillance of passengers with mug shots," *Proc. SPIE* **7703** 770305 (2010).

Ming Kai Hsu received his MS degree in computer engineering from Tatung University, Taiwan, in 1998 and his MS degree in electrical engineering from George Washington University (GWU), Washington DC, in 2003. He is a PhD candidate in electrical engineering at GWU (final stage of dissertation defense). Besides academia research, he has rich industrial and information technology working experience. In 2001, he worked in the IT industry as a research engineer in Taiwan. From 2003 to 2005 he was a technical assistant and information specialist in SEAS computing facility at GWU. From 2005 to 2007, he was a research assistant in GWU. He has been involved in Internet web delivery system, FPGA firmware, system on chips, and smart sensor on chip processing. He has worked in neural networks, unsupervised learning, chaos, fuzzy, security codec, and moving platform video imaging. His current focus and interests are multiple band video images processing for biomedical wellness and surveillance applications.

Ting N. Lee is a professor in the Department of Electrical and Computer Engineering at George Washington University (GWU). His main research interests are in systems theory and network synthesis. Recent projects have centered on optimal distributed network theory and neural network theory, digital, and analog.

Harold Szu received his PhD in theoretical physics in 1971 from the Rockefeller University, New York. He worked at the Naval Research Lab (1977 to 1990) and was the leader of Naval Surface Warfare Center at White Oak (1990 to 1996) followed by his relocation to Dahlgren, Virginia (1997 to 2008). He is now a senior scientist at the Army Night Vision and Electronic Sensors Directorate, located in Ft. Belvoir, Virgina. He is a founder, former president, and a current governor of the International Neural Network Society (INNS), receiving the INNS Dennis Gabor Award in 1997 and the Eduardo R. Caianiello Award in 1999 from the Italian Academy. Recently, SPIE awarded him with the Nanoengineering Award and the Biomedical Wellness Engineering Award. He is a fellow of American Institute of Medicine & BioEngineering 2004, a fellow of IEEE (1997), a foreign academician, Russian Academy of Nonlinear Sciences, 1999, a fellow of the Optical Society of America (1996), a fellow of International Optical Engineering, a Fellow of SPIE (1995), and a fellow of INNS (2010).