Image and Video Modeling and Understanding

Fang Liu
The MIT Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
fliu@media.mit.edu

Abstract

Seven digital library related projects conducted in the Vision and Modeling Group of the MIT Media Laboratory are reviewed. These projects address a large variety of issues that are essential to building sophisticated, efficient, and user friendly tools for image and video library applications. The problems on which these projects focus include feature extraction, feature combination, similarity comparison, image and video understanding, and learning from man-machine interaction.

1 Introduction

This talk reviews seven digital library related projects completed or in progress in the Vision and Modeling Group of the MIT Media Laboratory. These projects address a large variety of issues that are essential to building sophisticated, efficient, and user friendly tools for image and video library applications.

In the existing image and video database applications, color and texture are the most commonly used features. The first work presented is a texture model that provides perceptually sensible image features. This model has been applied to both spatial and temporal texture modeling for image retrieval and video analysis.

Using low-level image features to achieve high-level image understanding is a challenging problem. The second project reviewed is a scene classification system, which successfully uses both color and texture information to classify indoor and outdoor consumer photographs.

Face recognition is a classic digital library application. The third project presented is a face detection and recognition system that uses visual learning. This system placed first in the 1996 FERET contest.

The ability to track and interpret human action is very important for video analysis and understanding. Three projects in this area are presented: Pfinder, American Sign Language recognition, and discourse video analysis. Pfinder provides a means to extract the gesture information from a video stream. The gesture information is then used to interact with artificial life or to interpret American Sign Language. The discourse video analysis algorithm uses both gesture and audio information to detect semantic patterns in monologue videos. This system can pick out stand-up comedians' punch lines!

Most of the existing retrieval systems require users to select similarity measures alone with a query. Relevance feedback is a more natural form of man-machine interaction. The last project presented is the FourEyes system. This image browser learns continuously from user feedback and incorporates a variety of models for representing the content of image and video.

Below is an extended abstract of the texture modeling work. Short descriptions of the other six projects are also provided. Related papers can be found at "http://www.media.mit.edu/vismod/", under "Publications".

2 Modeling Spatial and Temporal Textures (Liu and Picard)

2.1 Overview

Image texture features have been widely used in digital library applications. For image retrieval, a computer system is expected to return to its user database images that resemble the visual properties of the prototypes. To build such a system, it is important that the computational features used for pattern comparison are faithful to those used by humans in comparing patterns.

A random field decomposition theory named after statistician H. Wold allows the decomposition of a homogeneous texture pattern to be decomposed into three orthogonal components: harmonic, evanescent,

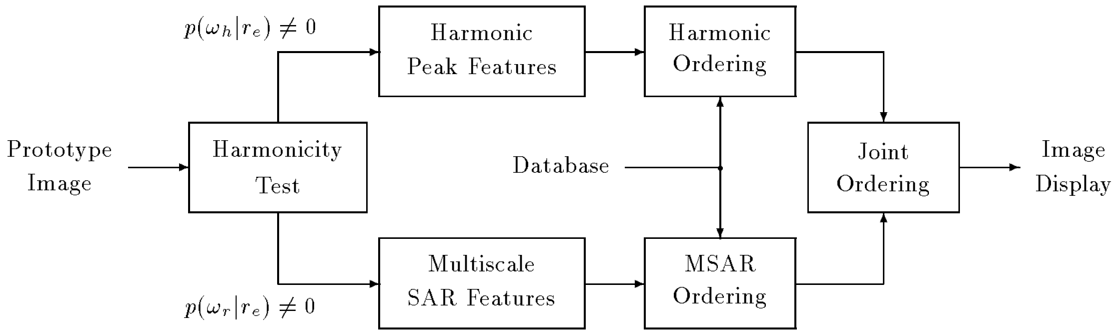


Figure 1: Flow-chart of the image retrieval system based on Wold texture modeling.

and indeterministic. The perceptual properties of the components can be described respectively as “periodicity”, “directionality”, and “randomness”, agreeing closely with that of the top dimensions of human texture perception [13]. Hence, perceptually salient features can be constructed based on the Wold theory.

A textured image database retrieval system has been developed. The core of the system is a Wold-based shift, rotation, and scale invariant texture model. When compared to two other texture models, the Wold model appears to offer perceptually more satisfying retrieval results.

To investigate the perceptual properties of the Wold texture models, a psychophysical study was conducted. A highly significant correlation was found between the human and computer texture ranking data, suggesting that the component energy resulting from the 2-D Wold decomposition of an image is a good computational measure for the most salient dimension of human texture perception, the dimension of repetitiveness vs. randomness.

Applying the principle of Wold-based texture modeling to the temporal dimension, an algorithm is developed to simultaneously detect, segment, and characterize spatiotemporal periodicity. This technique is robust to noise and computationally efficient, providing a useful tool for video analysis and understanding.

In the following subsections, the Wold-based texture modeling work is briefly described. Details can be found in [5] and [6].

2.2 Textured Image Database Retrieval

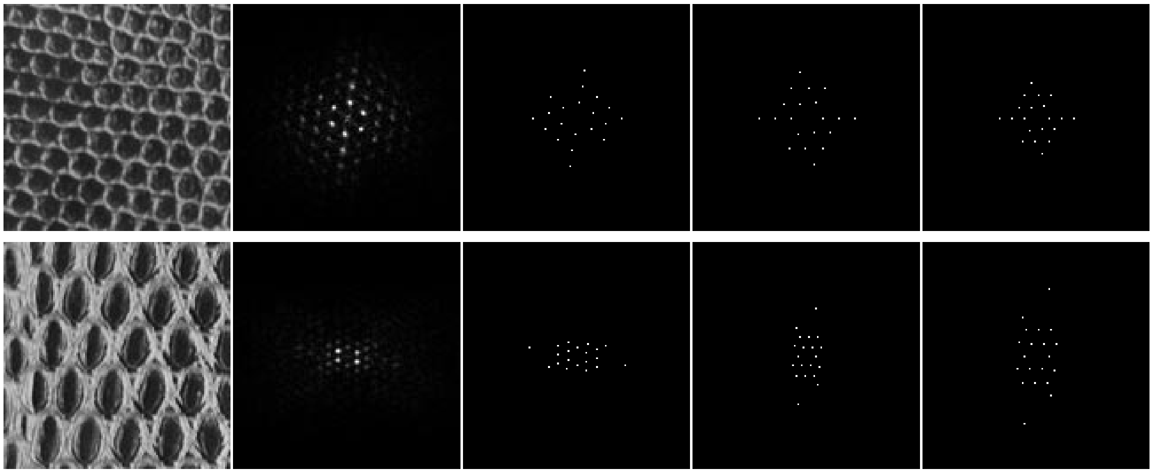
The textured image database retrieval system is based on Wold texture modeling. Given a texture pattern, its repetitive structure is represented by its spectral harmonic peaks, and its randomness modeled by a multiresolution simultaneous autoregressive (MRSAR) fitting. Shown in Figure 1, the retrieval system consists of four stages. Given a prototype image, its level of repetitiveness is first examined by a harmonicity test. The Wold features are then extracted to characterize the periodic and the random components of the image separately. Based on each type of features, the entire database is ordered according to the image similarity to the prototype. (The Wold features of the database images are pre-computed.) Finally, the database orderings are combined for final query return.

The retrieval system was evaluated on the Brodatz Texture Database. This database contains 1008 eight-bit gray scale images cropped from the Brodatz album [3]. Each page of the album contributes nine images.

2.2.1 Harmonicity Test

The harmonicity of a textured image (*i.e.,* the amount of repetitive structure in the image) is determined by examining the energy distribution of the image autocovariance function. The autocovariance energy of a highly structured texture has periodic concentration throughout the 2-D displacement plane, while that of a random-looking texture concentrates in the small-displacement region. The ratio between the autocovariance energy in the small-displacement region and the total energy (total sum of the absolute value of the autocovariance function) can be used as a measure of image harmonicity.

The autocovariance energy ratio r_e was computed for each image in the Brodatz database. The histogram of the ratios has a bi-modal structure. Gaussian assumptions were made to model the energy ratio data using an expectation and maximization (EM) procedure [6]. Denote the resulting classes as ω_h (harmonic) and ω_r (random). The EM procedure gives the posterior probabilities $P(\omega_h|r_e)$ and $P(\omega_r|r_e)$, which can be used as the confidence measure of characterizing the image as highly structured and relatively unstructured, respectively.



Original Fourier Mag. Peak Feature Rotated Rot. & Scaled

Figure 2: Harmonic peak feature rotation, and scale invariance. Top row: Reptile skin. Bottom row: Lizard Skin. Although the original patterns are in different scale and have relative rotation, their harmonic peak features allow rotation and scale invariant similarity comparison.

2.2.2 Feature extraction

In the second stage, the spectral harmonic peak features and the MRSAR features of the prototype image are estimated.

The harmonic peak feature set consists of the frequencies and the magnitudes of ten largest spectral harmonic peaks. When extracting the spectral peaks, the harmonic relationship among the peak frequencies is explicitly examined.

The harmonic peak features inherit from the Fourier spectral magnitude the property of spatial shift-invariance. To provide the ability of comparing images with respect to relative rotation and scaling, the peak feature set is rotated to align the lowest fundamental frequency to a chosen orientation (horizontal in this work) and scaled such that the distance between the lowest fundamental and the zero frequency is some chosen value (10 in this work). An example is shown in Figure 2.

The relatively unstructured texture components are characterized by using the MRSAR method introduced by Mao and Jain [8]. A second-order symmetric MRSAR model is fit to the image at three scales, resulting a 15-parameter feature vector and its covariance matrix.

2.2.3 Image Similarity Comparison

Two orderings of the entire database are generated in this stage. For each ordering, image similarities are measured by either the harmonic peak matching or the MRSAR feature Mahalanobis distances, and the database is sorted by the descending order of the image similarity to the prototype.

2.2.4 Joint Ordering

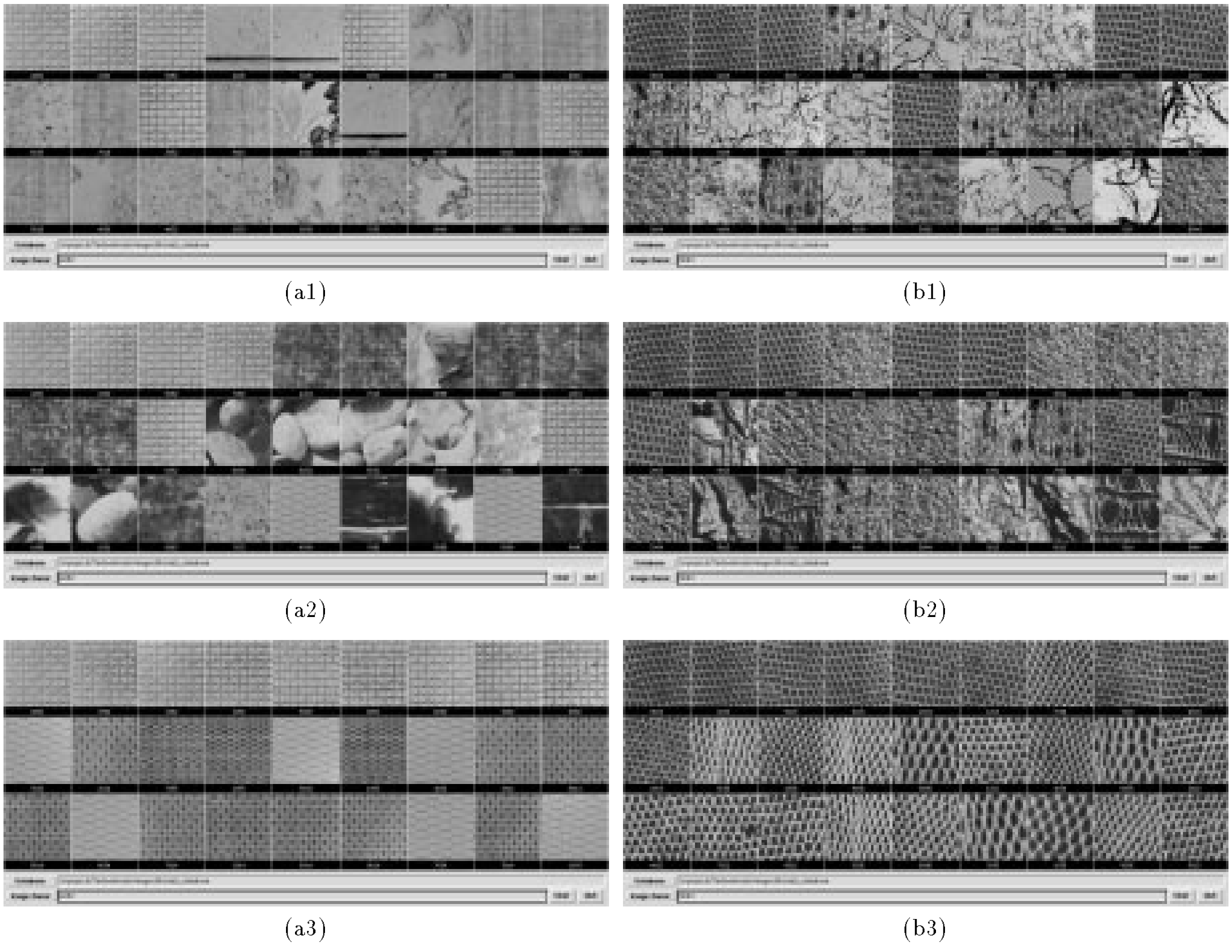
In the final stage, the two database orderings are combined using the confidence measures generated by the harmonicity test. Denote the rank of a database image in the harmonic ordering as O_h and the one in the MRSAR ordering as O_r. The joint rank of the image is computed as

$$O_{joint} = O_h P(\omega_h|r_e) + O_r P(\omega_r|r_e).$$

The final similarity ordering of the database is formed by sorting images in the ascending order of their joint rank values.

2.2.5 Image Retrieval Examples

Figure 3 shows two sets of image retrieval results of using the shift-invariant principle component analysis (SPCA) [12], the MRSAR, and the Wold-based methods over the Brodatz Database. In each picture, the upper left image is the prototype, and the retrieved images are shown by descending similarity to the prototype in raster-scan order. The Wold method demonstrates superior qualitative and quantitative performance by offering both "intra-class" accuracy and perceptually more satisfying "inter-class" similarity. The benchmarking results of five texture models, where the Wold model placed the first, can be found in [6].



(a1) (b1)

(a2) (b2)

(a3) (b3)

Figure 3: Image retrieval examples. (a1)-(a3): straw cloth pattern. (b1)-(b3): reptile skin pattern. Three methods are compared: SPCA ((a1),(b1)), MRSAR ((a2),(b2)), and Wold ((a3),(b3)). In each picture, the images are raster-scan ordered by their similarities to the image in upper left.

2.3 Natural Scene Representation

A K-means-based MRSAR feature clustering algorithm was introduced in [6] to segment natural scene images for homogeneous regions. The Wold features of these regions can be used for subsequent content identification and similarity comparison.

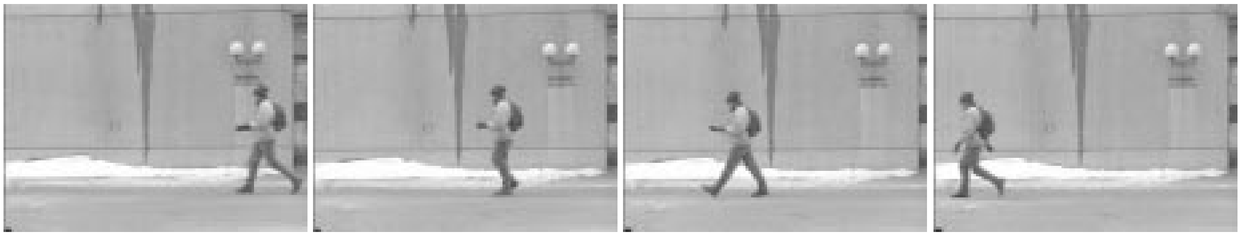
2.4 Perceptual Properties of Wold-based models

In the image retrieval experiment, the Wold texture model appears to offer perceptually more satisfying results. To further investigate the perceptual properties of Wold-based modeling, a psychophysical study was conducted [5].

Rao and Lohse identified the most important dimension of human texture perception as repetitiveness vs. randomness [13]. In the current study, humans and a computer program order a set of texture samples along this dimension. The correlation between the averaged human ordering and the computer ordering are used to gauge how well the computational model captures the perceptual properties of the images with respect to the perceptual axis.

2.4.1 Human Experiment

In the experiment, 32 human subjects (equal number of men and women) ordered a set of 20 Brodatz texture samples between two sets of adjectives: repetitive, non-random, directional, regular, locally oriented, and uniform vs. non-repetitive, random, non-directional, irregular, non-oriented, and non-uniform. (In Rao and Lohse's results, these adjectives label the two ends of the top perceptual axis.) The human ranking scores were then averaged and the samples re-ordered based on their average ranks to produce the final human ordering.



Frame 20 Frame 40 Frame 60 Frame 80

Figure 4: Example frames of the Walker sequence, with frame size 320 × 240.

2.4.2 Computer Experiment

A computer program ordered the same set of images using the Wold computational model. The orthogonal Wold components of an image have distinctive visual properties. It is conceivable that these components can be used to represent the perceptual properties of a texture pattern. Based on models of human early vision system [1][2], the total energy of the Wold components was used as the physical quantity to measure the perceptual strength of the components.

For each test sample, the computer program first performs a spectral Wold decomposition [5] to obtain the orthogonal image components. Then the signal energy of the components are computed. The ratio between the deterministic energy (including both harmonic and evanescent components) and the total energy is used for image ordering.

2.4.3 Data Analysis and Conclusions

Both Spearman and Kendall rank correlation coefficients were used to assess the correlation between the final human ranking and the computer ranking. To ensure that the ranking based on the averaged ranks is the best estimate of the "true" human ranking, the concordance of the human data was also evaluated.

The Spearman correlation coefficient for the human and computer rankings is $r_s = 0.9504$ with statistic $t = 12.96$ and significance $p < .001$, while the Kendall coefficient is $\tau = 0.7474$ with statistic $z = 4.61$ and significance $p < .001$. The Kendall concordance coefficient for the 32 sets of human ranking data is $W = 0.7874$ with statistic $\chi_r^2 = 478.72$ and significance $p < .001$. Therefore, the human and the computer rankings are significantly correlated.

The following conclusions can be drawn from the experimental results:

1. The highly significant correlation between the human and the computer texture ranking data suggests that the component energy resulting from the 2-D Wold decomposition of an image is a good computational measure for the most salient dimension of human texture perception, the dimension of repetitiveness vs. randomness.
2. The highly significant concordance of the human rankings indicates the following:
   (a) There exists a common interpretation to the semantic labels (the adjectives) associated to the perceptual dimension.
   (b) These labels indeed correspond to certain underlying criteria, upon which the human subjects agree, for texture similarity measurement.

2.5 Temporal Texture Modeling for Video Analysis

2.5.1 Overview

In this work, the principle of Wold texture modeling is applied to spatiotemporal dimensions to detect, segment, and characterize periodic phenomenon in image sequences.

Figure 4 shows four frames of a video sequence Walker, where a person walking across the image plane. Regarding the sequence as a data cube of three-dimensions: X (horizontal), Y (vertical), and T (temporal), the XT and YT slices of the cube can reveal the temporal behavior usually hidden from the viewer. Figure 5 shows the head and ankle level XT slices of the Walker sequence. In (a), the head leaves a non-periodic straight track while the walking ankles in (b) make a crisscross periodic pattern. As it is, the periodicity in (b) is difficult to characterize.

The algorithm presented here has two stages: object tracking by frame alignment, which transforms data into a form in which periodicity can be easily detected and measured; (2) simultaneous detection and segmentation of spatiotemporal periodicity. The latter stage generates a periodicity template that not only

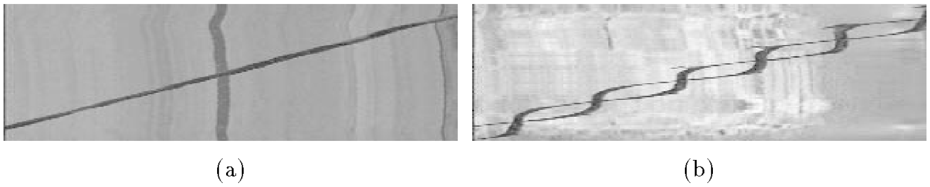


Figure 5: Head and ankle level XT slices of Walker sequence. (a) Head level. (b) Ankle level. As it is, the periodicity in (b) is difficult to characterize.

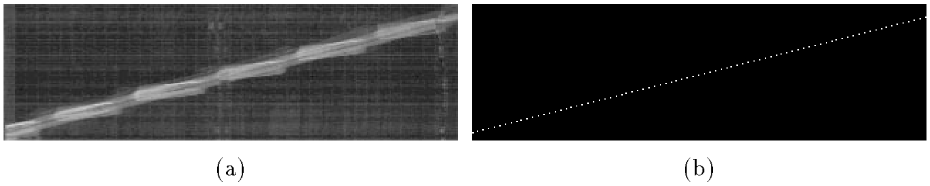


Figure 6: (a) Averaged XT image of the Walker sequence after background removal. (b) Line found in (a) by using a Hough transform method.

indicates the presence and the location of a periodic event, but also contains the fundamental frequencies and an accurate quantitative measure of how periodic the event is. Decoupling object tracking and periodicity detection conceptually modularizes the analysis process and allows the use of other tracking algorithms. In the following, the Walker sequence will be used to illustrate the key technical points. More detailed explanations of the algorithm can be found in [7].

2.5.2 Frame Alignment

In this work, a procedure is developed for the alignment of image sequences that involves little ego-motion and contains objects moving approximately frontoparallel to the camera along a straight line and at a constant speed. After frame alignment with respect to a moving object, the object should as a whole be moving in place.

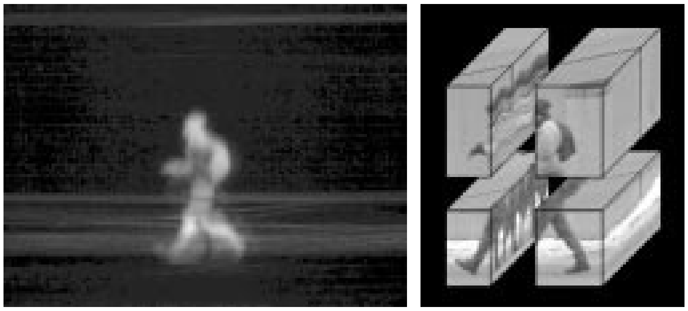
The trajectories of moving objects in the 3-D data cube is first detected. Applying 1-D median filtering along the temporal dimension, the output has mostly the still background of the sequence. The *difference sequence* between the original and the background contains mainly the moving objects. Since the object trajectories in consideration are approximately linear, the projections of the trajectories onto the XT and the YT planes (averaged XT and YT images of the difference sequence) are straight lines. These lines can be detected via a Hough transform to give the X or the Y positions of the moving objects in each frame. These position values are the *alignment indices*. The averaged XT image of the Walker difference sequence and the line found by the Hough transform method are shown in Figure 6. Each horizontal line of the pictures represents a frame, and the diagonal white line marks the object X location in each frame. Note that multiple object trajectories can be detected simultaneously using this procedure.

Using the alignment indices, image frames in a sequence are repositioned to center a moving object to any specified position in the XY plane. The aligned sequence can be cropped to save computation in subsequent processing. The location and size of the cropping window can be estimated from the average XY image of the *aligned* difference sequence. Figure 7 shows such XY image of the Walker sequence and the aligned and cropped original sequence with splits near the center of the frames to show the inside of the data cube.

2.5.3 Generating Periodicity Templates

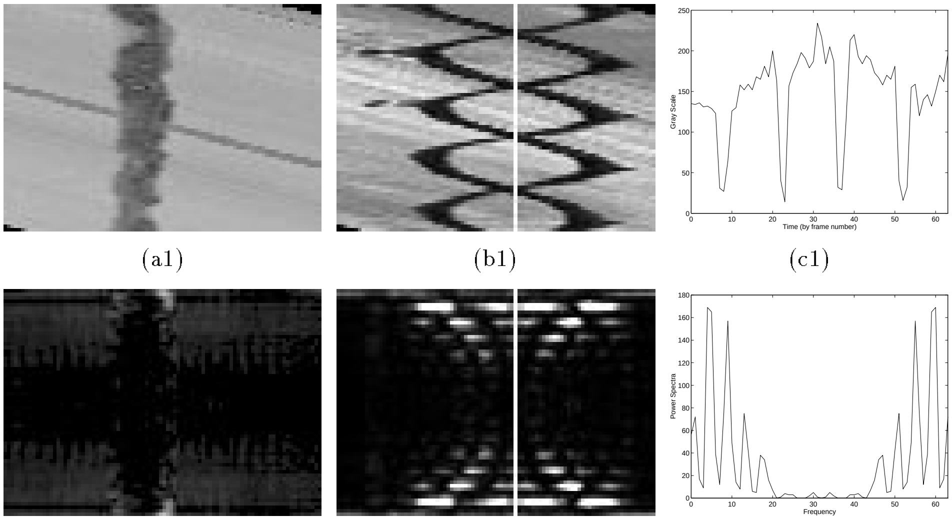
Now consider an aligned and cropped data cube. Frame pixels with the same X and Y locations form straight lines in the cube. Call these lines the *temporal lines*. Since the object of interest moves in place, its cyclic motion is reflected as re-occurring signals on some of the temporal lines. After computing the power spectrum of a temporal signal via a 1-D Fourier transform, the spectral harmonic peaks are detected and used to compute the *temporal harmonic energy* of the signal. A periodicity template is generated by using the extracted fundamental frequencies and the ratios between the harmonic energy and the total energy (the *temporal harmonic energy ratio*) along each of the temporal lines (*i.e.,* for every frame pixel locations).

Figure 8 (a1) and (b1) show the head and the ankle level XT slices of 64 frames (Frame 17 to 80) of the data cube in Figure 7 (b). Each column in the images is a temporal line. These images are the aligned and cropped version of the two XT slices in Figure 5. Columns in Figure 8 (a2) and (b2) are the 1-D power spectra of the corresponding columns in (a1) and (b1), normalized among all temporal lines in the data cube.



(a) (b)

Figure 7: (a) Averaged XY image of aligned Walker difference sequence. The area of interest is clearly shown. (b) Aligned and cropped Walker sequence with splits near the center of the frames to show the inside of the data cube.



(a1) (b1) (c1)

(a2) (b2) (c2)

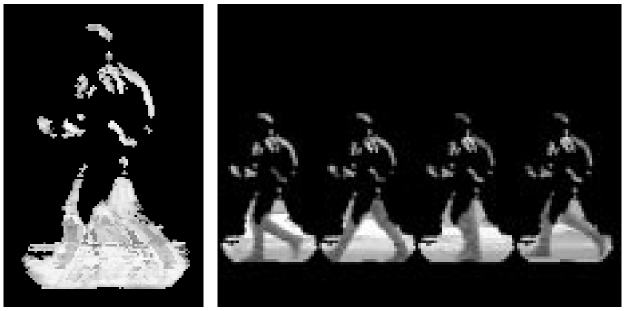
Figure 8: Signals and their power spectra along temporal lines (columns in images). (a1) and (b1): head and ankle level XT slices of aligned and cropped Walker sequence. (a2) and (b2): each column is the 1-D power spectra of the corresponding column in (a1) and (b1). (c1) and (c2): details along the white vertical lines in (b1) and (b2). Periodicity in (b1) is reflected by the spectral harmonic peaks in (b2).

Figure 8 (c1) and (c2) show details along the white vertical lines in (b1) and (b2). While the head level slice in (a1) shows no harmonicity, the periodicity of the moving ankles in (b1) is reflected by the spectral harmonic peaks in (c2).

The temporal harmonic energy ratio values of the periodicity template for the Walker sequence are shown in Figure 9 (a). The larger the energy ratio value, the more periodic energy at the location. As expected, the brightest region is the wedge shape created by the walking legs. The head, the shoulder, and the outline of the backpack are detected because the walker bounces. The hands appear at the front of the body since in most parts of the sequence the walker was fixing his gloves and moving his hands in a rather periodic manner. Note that the moving background and parts of the walker do not appear in the template since there is no periodicity present in those areas. Using the alignment indices generated at the first stage, the periodicity template can be used to mask the original sequence for the regions of periodicity in each frame. Figure 9 (b) shows the four frames in Figure 4 after they are masked and then stacked together.

The algorithm discussed above is not limited to periodicity caused by human activities. Shown in Figure 10, Wheels is a 64 frame sequence of a car passing by a building. Near the top of the building, two spinning wheels are connected by a figure 8 belt. One side of the belt is patterned and appears periodic. Every region with periodicity should be captured: the hub caps, the wheels, and one side of the belt. The algorithm accomplishes just that.

More examples can be found in [7]. Those examples demonstrate that the algorithm is well suited for detecting multiple periodicities and is robust in the presence of noise.



(a) (b)

Figure 9: (a) Temporal harmonic energy ratio values of the aligned Walker sequence. High value indicates more periodic energy at the location. (b) Using the alignment indices, the four frames in Figure 4 are masked by the template shown in (a) and then stacked together.

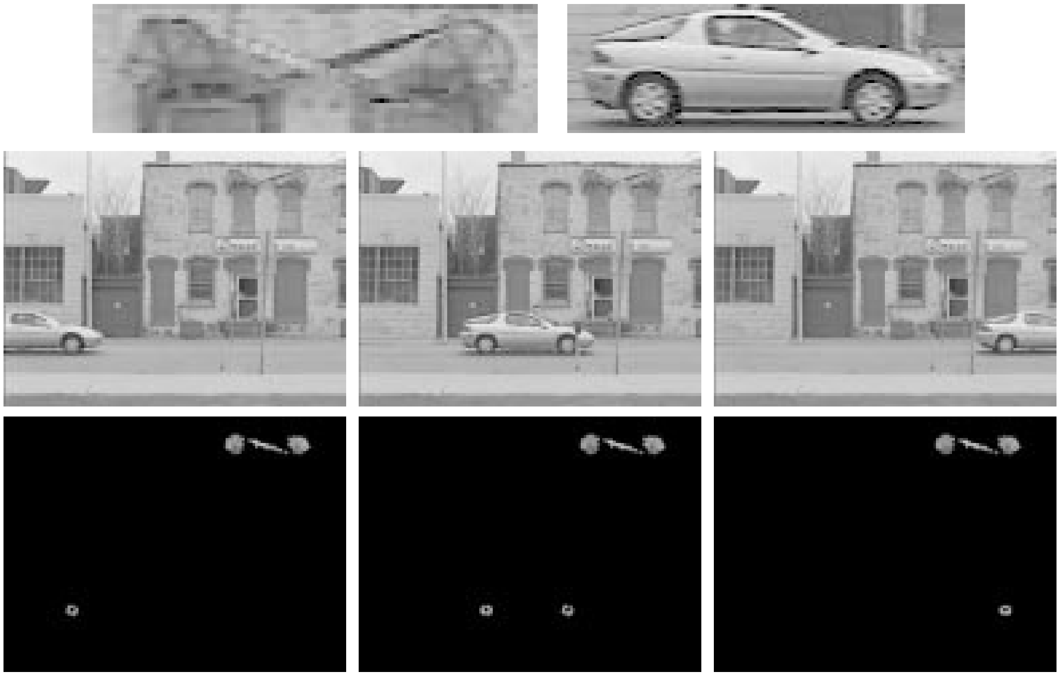


Figure 10: Wheel sequence. Top row: details of spinning wheels and car. Two bottom rows: the algorithm captures all regions with periodicity — the hub caps, the wheels, and one side of the belt.

2.6 summary

A texture model based on the 2-D Wold decomposition of homogeneous random fields is applied to image database retrieval. The Wold-based model characterizes a texture by its periodicity, directionality, and randomness, approximating what are indicated to be the three most important dimensions of human texture perception. Compared to two well-known texture models, the Wold model appears to offer a perceptually more satisfying measure of pattern similarity.

The results of a psychophysical study suggests that the component energy resulting from the 2-D Wold decomposition of an image is a good computational measure for the most salient dimension of human texture perception, the dimension of repetitiveness vs. randomness. The highly significant concordance of the human data also verifies that the top perceptual dimension found by Rao and Lohse indeed corresponds to certain underlying criteria, upon which the human subjects agree, for texture similarity measurement.

The Wold-based modeling is also applied to temporal textures. An algorithm is developed for finding periodicity in space and time. This method allows simultaneous detection, segmentation, and characterization of periodic motion in data. The resulting periodicity templates carry information on the location, frequency, and relative energy of periodic motion in a video sequence. This algorithm can also be considered as a periodicity filter, providing a model of low-level periodicity perception.

3 Scene Classification (Szummer and Picard)

Classifying images into high-level semantic classes is a very difficult task for a computer. This work [15] shows how one particular scene classification problem – classifying indoor and outdoor scenes of consumer photographs – can be approached.

Color histograms and multi-resolution autoregressive texture model coefficients are used as low-level image features. The images are also tessellated to incorporate coarse spatial position information. The classifier first classifies all the subimages. Then the full-size images are classified based on their sub-image classifications. For a collection of 518 photographs (306 outdoor and 212 indoor), this algorithm classifies 92.5% of the pictures correctly.

4 Face Detection and Recognition (Moghaddam and Pentland)

The face detection and recognition algorithm uses features resulted from the image eigenspace decomposition. The feature space consists of the eigenspace dimensions that correspond to the largest eigenvalues. For face and facial feature (eyes, nose, and mouth) detection, an unsupervised learning technique is developed [10]. This learning technique uses either a multivariate Gaussian or a mixture-of-Gaussian model to characterize the feature space. The location of a face in an image is found by using the maximum-likelihood ratio test over multi-scale. After a face is located, it is normalized to a fixed size and the facial features are detected. Using the location of the facial features, the face image is warped to align to the shape of a canonical model. Then the facial region is extracted, normalized for contrast, and projected onto a set of eigenfaces to obtain a feature vector, which is subsequently used for similarity comparison to other faces.

This face detection and recognition system placed first in the 1996 FERET (Face Recognition Technology) contest [11], which uses over 3000 images taken of people at different times and with different facial expressions.

5 Pfinder (Wren, Azarbayejani, Darrell, and Pentland)

Using a static camera, Pfinder (Person Finder) [16] is a real-time system that can find and track a person and the person's head, hands, and body while the person moves around a room.

The system uses a maximum a posteriori probability based approach. A person is modeled as a connected set of blobs (two for hands, two for feet, and one each for head, shirt, and pants). The feature set for each blob includes a pixel-level support map and the Gaussian distribution of the blob spatial location and color. The background scene is modeled as a textured surface, where every pixel is associated with a Gaussian-distributed color model. Newtonian dynamic models are used to predict the blob's position and velocity. Contour analysis is used to help initializing the blob models. For steady state tracking, the likelihood of each pixel being a member of each of the blobs and the scene is computed at each frame. The likelihood values are then used to update the support maps. Finally, all the statistical and dynamic models are updated.

The Pfinder is computationally efficient – it runs on a standard SGI Indy computer in real time. This technique has been used in applications such as gesture recognition and interactive entertainment.

6 Real-time Recognition of American Sign Language (Starner and Pentland)

American Sign Language (ASL) consists of a complex set of hand gestures. This recognition system uses Pfinder for hand tracking and hidden Markov modeling (HMM) for recognition [14].

Using one color camera the hand tracking process produces a coarse description of hand shape, orientation, and trajectory. The hand tracking data are then sent to a four-state HMM for sentence-level ASL recognition. This system interprets in real-time a forty-word subset of ASL with 99% accuracy.

7 Analysis of Discourse Video (Casey and Wachman)

This project explores ways of combining features extracted from both audio and video data for video understanding [4]. Syllabic inter-onset intervals are used for temporal segmentation. Other features include the position and velocity of the hands (use Pfinder) and the value and change in pitch. Unsupervised analysis of video is conducted via clustering in the feature space. Using this technique, a Joke Detector is built to pick out stand-up comedians Jay Leno and David Letterman's punch lines!

8 FourEyes (Minka and Picard)

Most of the existing retrieval systems require the selection of similarity measures alone with a query. This is usually a difficult task for a user. Using relevance feedback eliminates the task, and provides a more natural form of man-machine interaction.

FourEyes [9] is an extensible and self-improving interactive learning system that assists users in digital library image and video segmentation, retrieval, and annotation. The system makes tentative groupings of the data using user relevance feedback and features provided by a variety of computational models. Users no longer have to choose features or set feature control knobs. Instead, they provide positive and negative examples that allow the system to choose similarity measures automatically. FourEyes is capable of continuous learning and learns at multiple scales: on a small scale from each interaction and on a larger scale across multiple interactions.

References

[1] J. R. Bergen and E. H. Adelson. Visual texture segmentation based on energy measures. *J. Opt. Soc. of Amer. A*, 3(13), 1986.

[2] J. R. Bergen and E. H. Adelson. Early vision and texture perception. *Nature*, 333:363–364, 1988.

[3] P. Brodatz. *Textures: A Photographic Album for Artists and Designers.* Dover, New York, 1966.

[4] M.A. Casey and J.S. Wachman. Unsupervised cross-modal analysis of professional monologue discourse. In *Proc. Workshop on the Integration of Gesture in Language and Speech*, 1996.

[5] F. Liu. *Modeling Spatial and Temporal Textures.* PhD thesis, Media Arts and Sciences, MIT, Cambridge, Sept. 1997.

[6] F. Liu and R. W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE T. Pat. Analy. and Machine Intel.*, 18(7):722–733, July 1996.

[7] F. Liu and R. W. Picard. Finding periodicity in space and time. In *Proc. Int. Conf. on Computer Vision*, Bombay, India, January 1998. To appear.

[8] J. Mao and A. K. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Patt. Rec.*, 25(2):173–188, 1992.

[9] T. Minka. An image database browser that learns from user interaction. Master's thesis, Dept. of EECS, MIT, Cambridge, MA, 1996.

[10] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. In S.K. Nayar and T. Poggio, editors, *Early Visual Learning*, pages 99–130. Oxford Univ. Press, 1996.

[11] P.J. Phillips *et al.*. The FERET September 1996 database and evaluation procedure. In *Proc. First Intl. Conf. on Audio and Video-based Biometric Person Authentication*, Crans-Montana, Switzerland, March 12-14, 1997.

[12] R. W. Picard and T. Kabir. Finding similar patterns in large image databases. In *Proc. Int. Conf. on Acous., Speech, and Signal Proc.*, pages V–161–V–164, Minneapolis, MN, 1993.

[13] A. R. Rao and G. L. Lohse. Towards a texture naming system: identifying relevant dimensions of texture. *Vision Research*, 36(11):1649–1669, 1996.

[14] T. Starner and A. Pentland. Real-time American Sign Language recognition from video using hidden Markov models. Perceptual Computing Section Technical Report No. 375, MIT Media Lab, Cambridge, MA, 1996.

[15] M. Szummer and R.W. Picard. Indoor/outdoor image classification. Perceptual Computing Section, MIT Media Lab, April 1997. Unpublished article.

[16] C. Wren *et al.*. Pfinder: real-time tracking of the human body. *IEEE T. Pat. Analy. and Machine Intel.*, 19(7):780–785, July 1997.