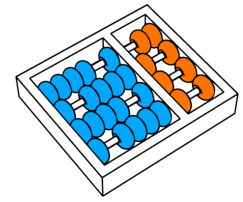


Image and Video Analysis: Applications, Challenges, Research Opportunities

Hélio Pedrini

Institute of Computing

University of Campinas (UNICAMP)

http://www.ic.unicamp.br/~helio

March 17th 2017

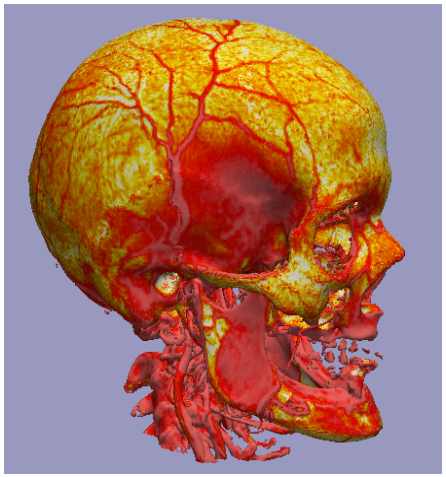
Summary

1. Introduction
   - Motivation
   - Objectives
   - Challenges
   - Applications
2. Research Problems
3. Trends

Introduction

Motivation

- Images and videos convey rich information.
- The essence of image and video data can be extracted from visual patterns.



Introduction

Objectives

- Automatic extraction, analysis and understanding of useful information from a single image or sequence of images.
- Search for automation of tasks that the human visual system can do.
- Development of theoretical and algorithmic basis to achieve automatic visual understanding.

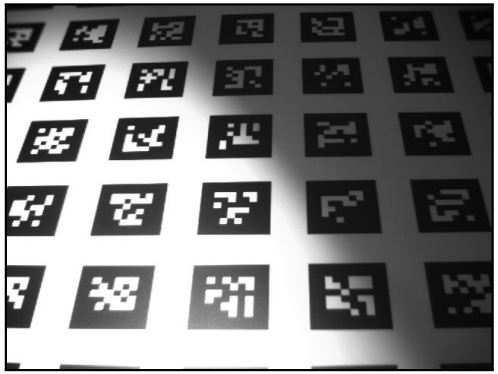
- Input space is typically high dimensional.
- Diversity of sensors (vision, audio, light, temperature, direction, acceleration).
- Efficient mapping from images/videos to flexible and scalable models.
- Comprehension of how the brain processes visual stimuli in humans to solve vision-related tasks.
- Generalization from few examples.
- Finite computational resources: memory and processing.
- Interdisciplinary/multidisciplinary field: psychophysics, cognitive neuroscience, neuroanatomy, computer vision, machine learning, signal processing, computer science, information theory, geometry, statistics, visualization, among others.

- Continuous advances in digital technology, associated with the development of new algorithms, have allowed an increasing number of applications.
- Examples:
  - medicine
  - microscopy
  - biology
  - agriculture
  - remote sensing
  - industrial automation
  - astronomy
  - military
  - security and surveillance
  - forensic science
  - art

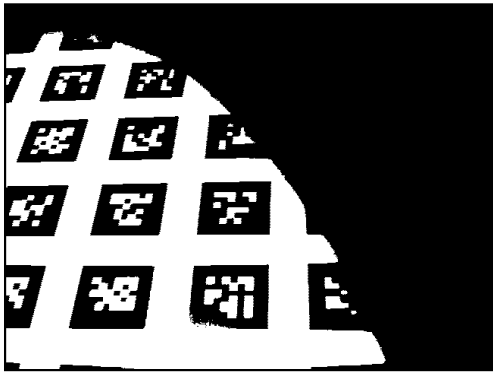
Research Problems

- Image thresholding improved by optimization methods.
- Image approximation based on triangular meshes.
- 3D object reconstruction from contour data.
- Adaptive human skin segmentation.
- Identification of computer-generated images $\times$ digital photographs.
- Text identification in images and videos.
- Video caption detection.
- Abnormal event detection.
- Recognition of human actions / activities.
- Emotion classification through facial expression analysis.
- Video summarization.
- Video stabilization.

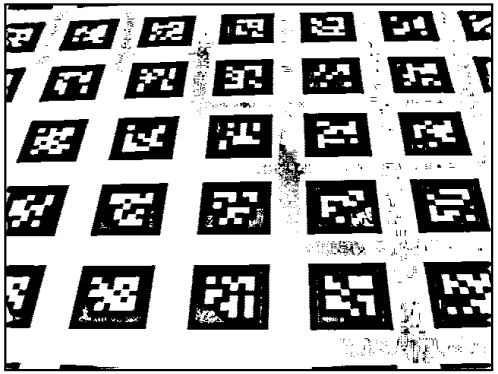
Image Thresholding Improved by Optimization Methods



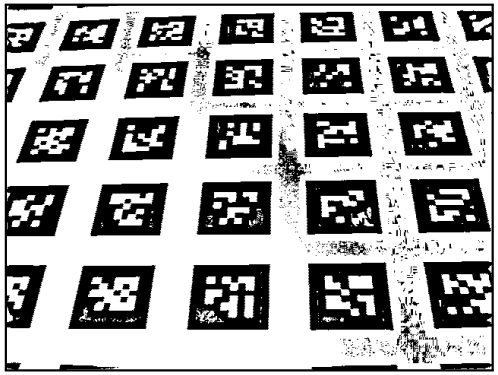
Fiducial (original)



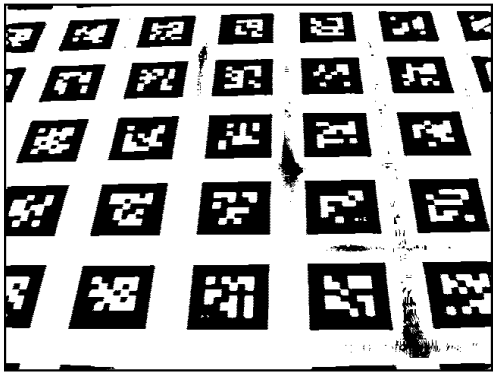
Otsu
(SSIM: 0.246552)



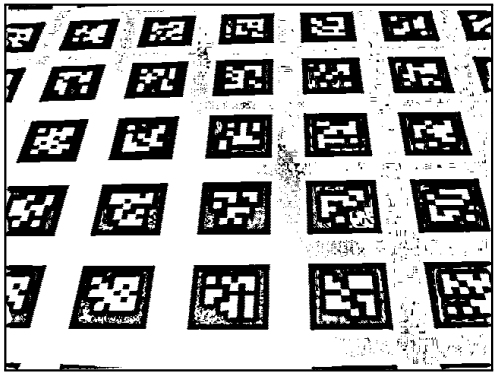
Genetic Algorithm
(SSIM: 0.289617)



Particle Swarm Optimization
(SSIM: 0.289777)



Simulated Annealing
(SSIM: 0.287376)

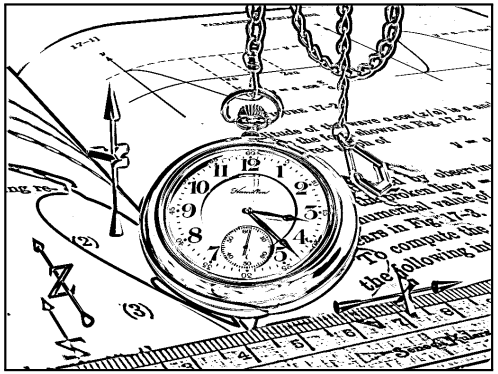


Pattern Search
(SSIM: 0.290501)

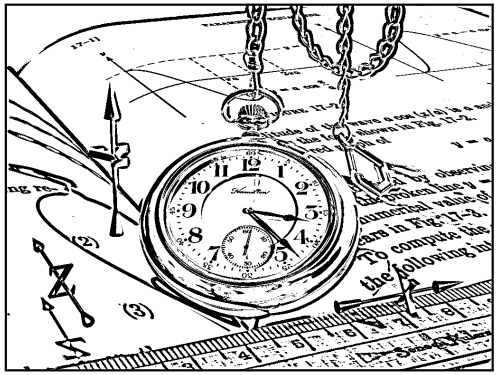
Image Thresholding Improved by Optimization Methods

Watch (original)

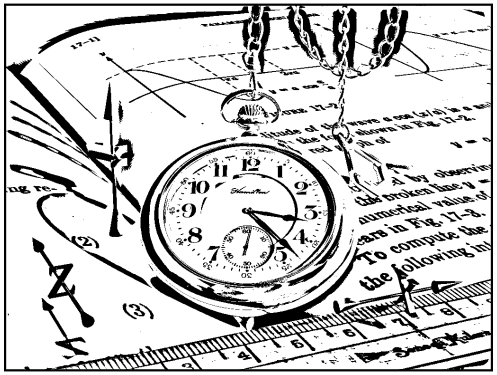
Otsu (SSIM: 0.094174)



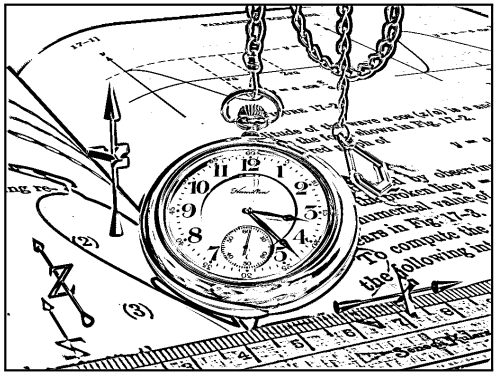
Genetic Algorithm
(SSIM: 0.125003)



Particle Swarm Optimization
(SSIM: 0.125195)



Simulated Annealing
(SSIM: 0.122471)

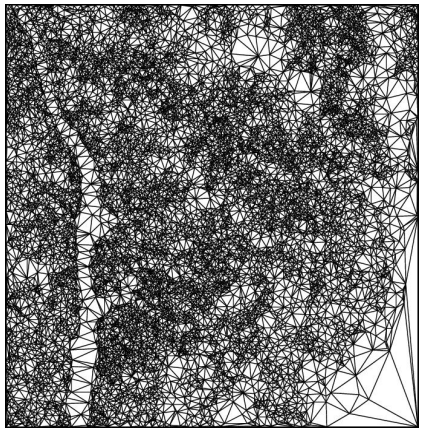


Pattern Search
(SSIM: 0.125426)

Image Approximation Based on Triangular Meshes

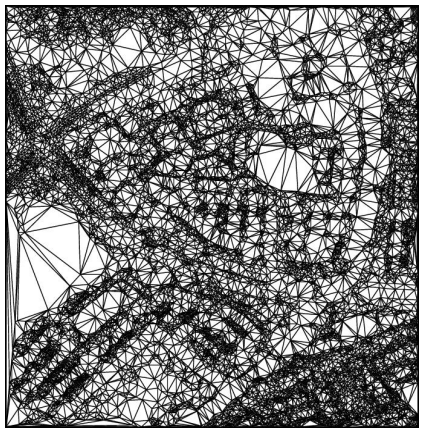
Refinement and Simplification of Meshes

image



triangular mesh

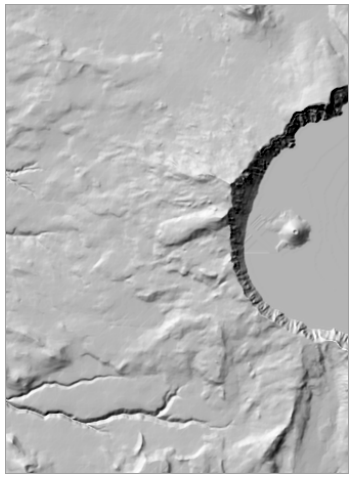
image



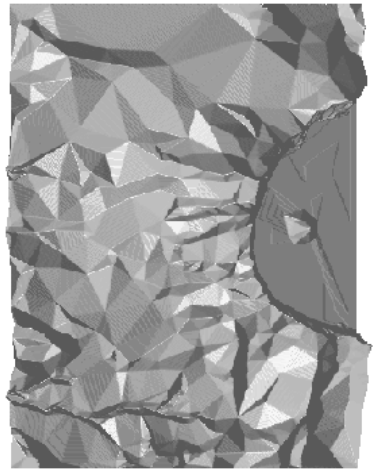
triangular mesh

Image Approximation Based on Triangular Meshes

Refinement and Simplification of Meshes

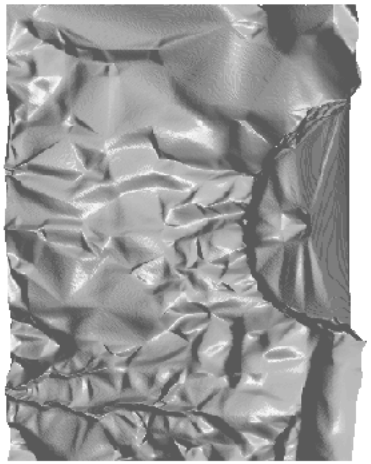


original



linear

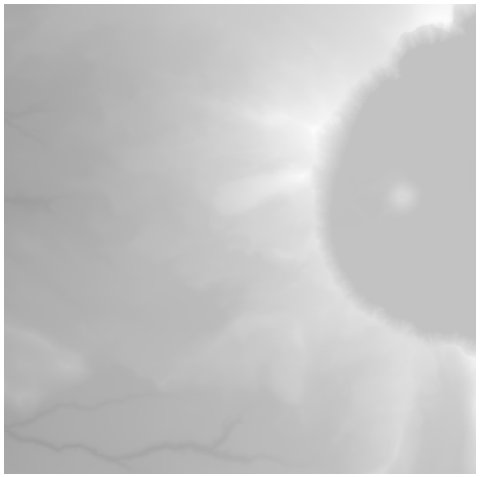
quadratic



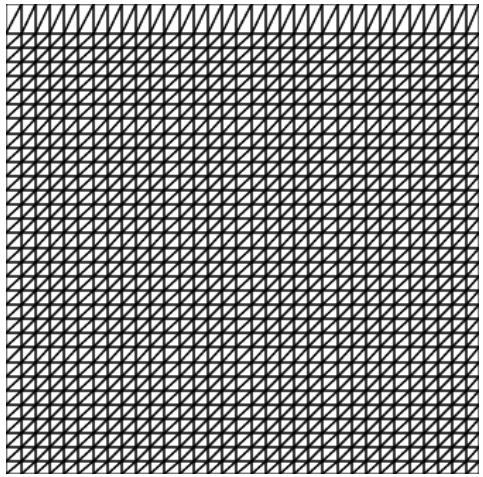
cubic

Image Approximation Based on Triangular Meshes

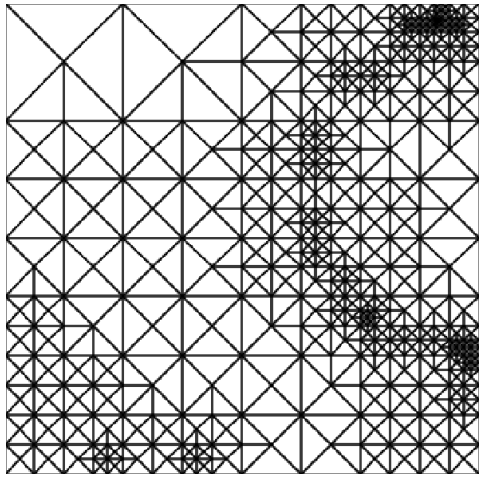
Refinement and Simplification of Meshes



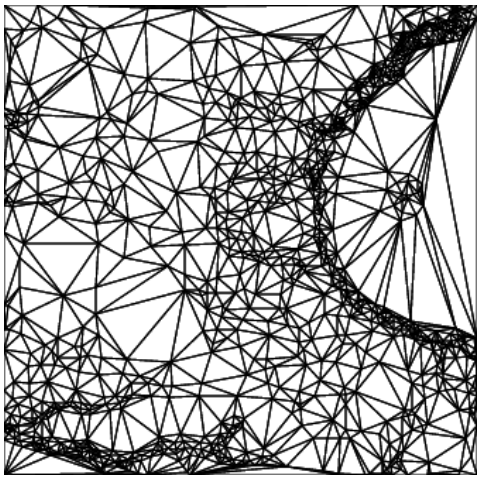
original



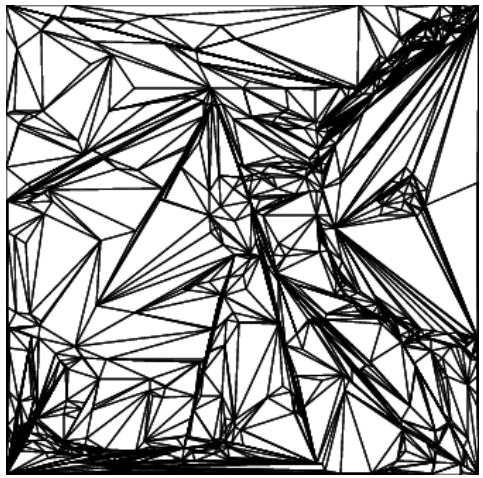
regular



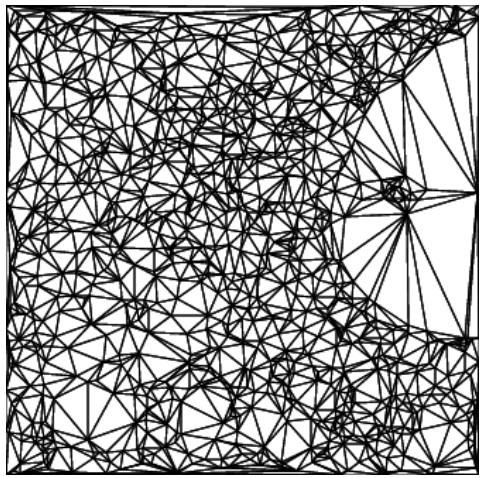
quadtree



Delaunay irregular



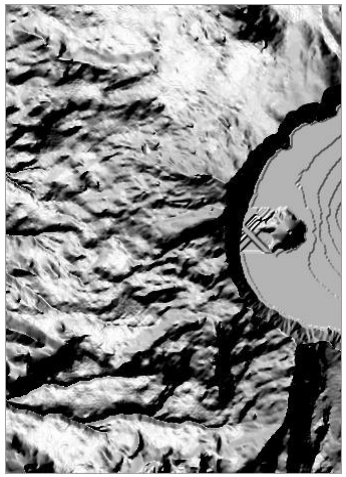
data dependent irregular



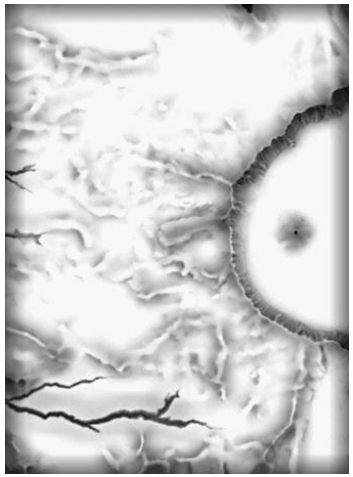
irregular based on features

Image Approximation Based on Triangular Meshes

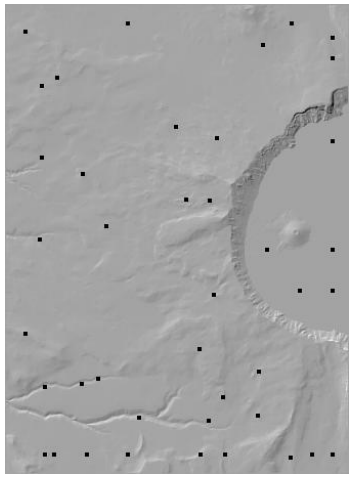
Refinement and Simplification of Meshes



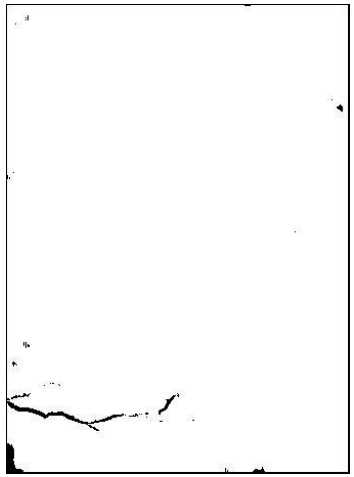
original



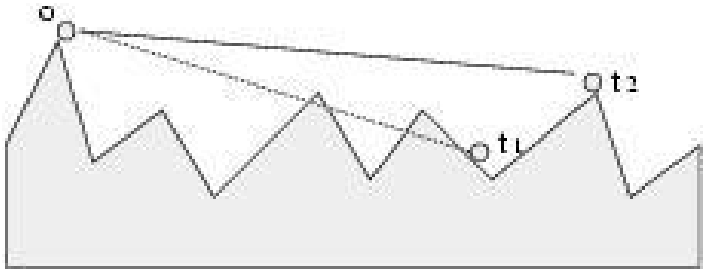
visibility



41 observers

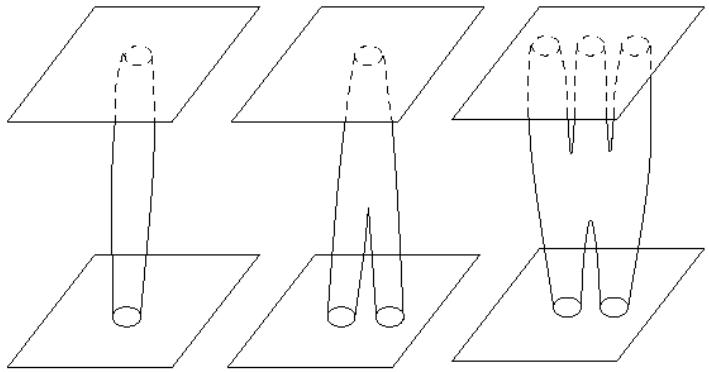


regions without coverage

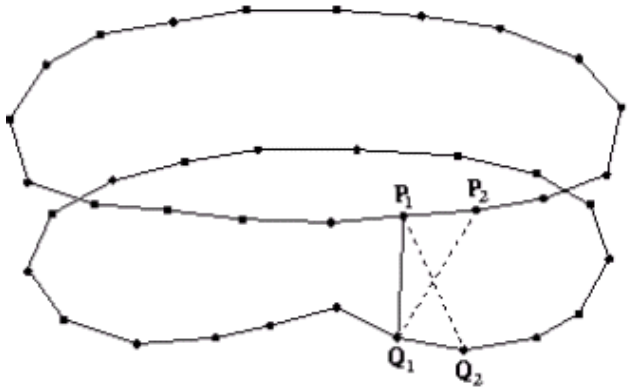


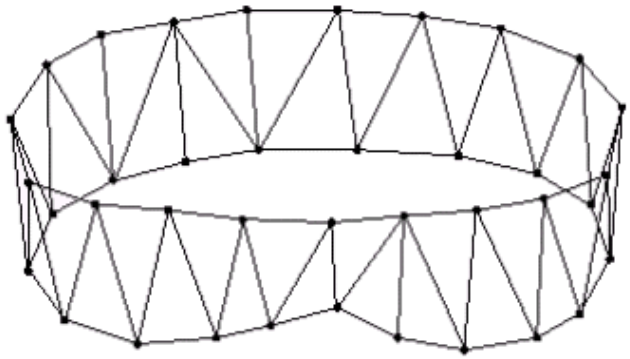
visibility test

3D Object Reconstruction from Contour Data

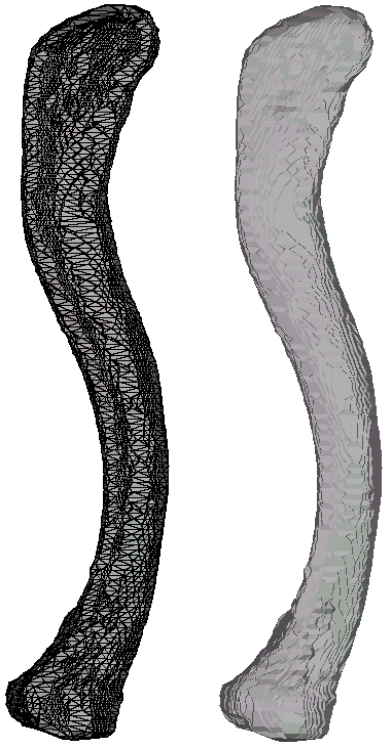


correspondence between contours





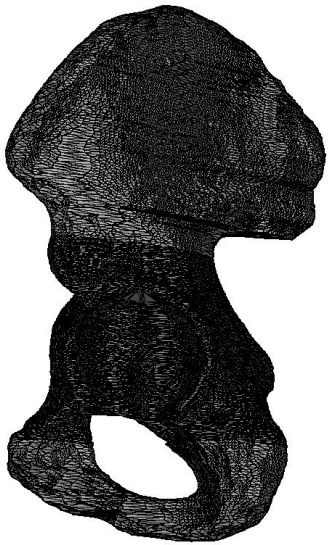
3D Object Reconstruction from Contour Data



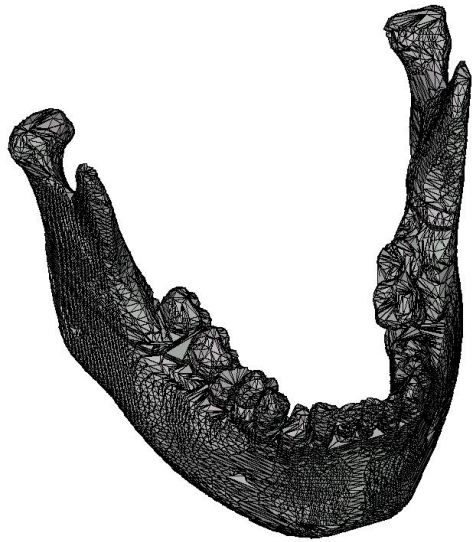
clavicle

humerus

3D Object Reconstruction from Contour Data

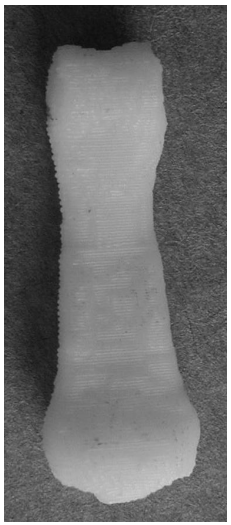


iliac

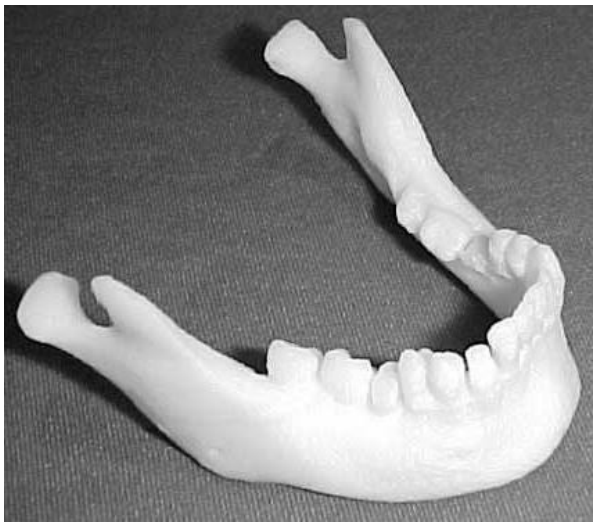


mandible

scapula



phalanx



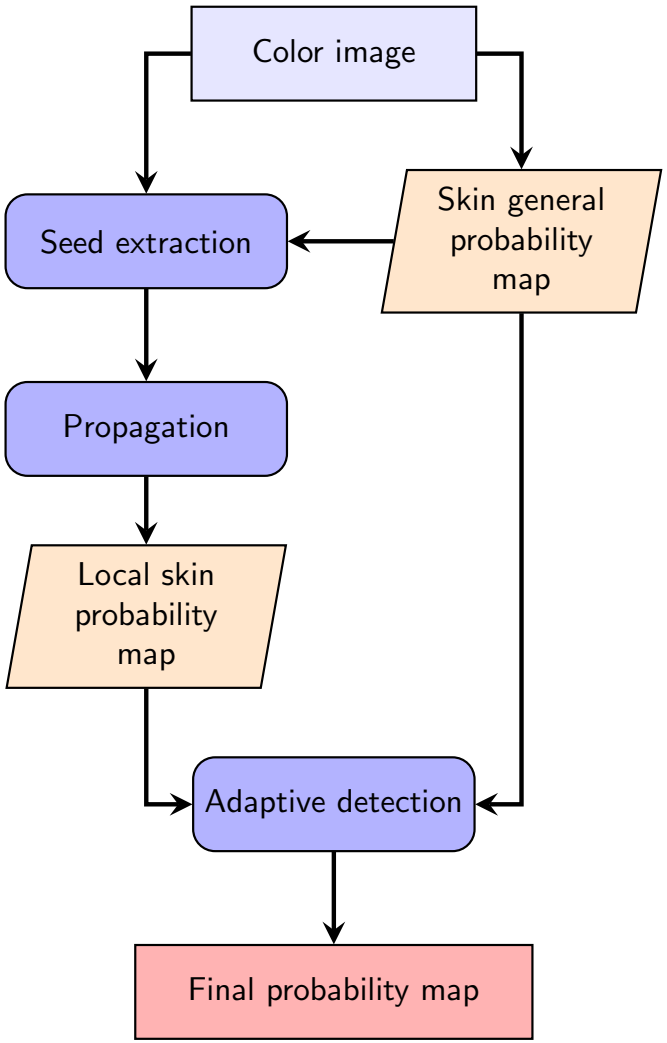
mandible

Adaptive Human Skin Segmentation

- Human skin detection in digital images is a crucial part in several applications:
  - face detection
  - gesture analysis
  - content-based image retrieval
  - nudity detection (consequent adult content filtering)
- Skin detection can be seen as a classification problem, whose purpose is to determine which image pixels belong to the skin or non-skin classes.
- Challenges:
  - variation in scene illumination interferes with the appearance of the skin
  - different cameras produce distinct colors
  - ethnic diversity promotes various skin tones

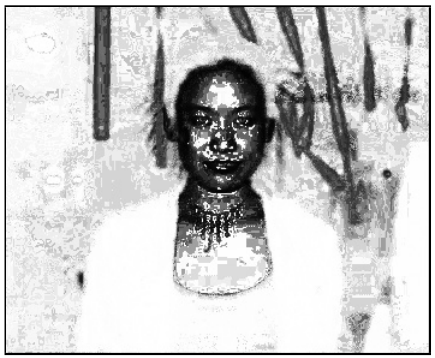
Adaptive Human Skin Segmentation

- Method for skin segmentation that combines spatial analysis and adaptive models for better skin probability estimation.



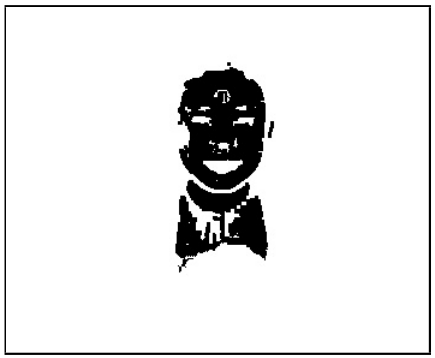
Adaptive Human Skin Segmentation

original

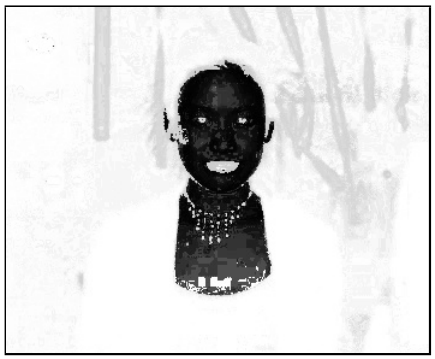


general probability map

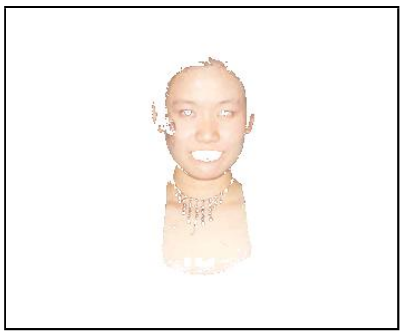
seeds



blobs after propagation



final probability map



segmentation result

Identification of Computer-Generated Images × Digital Photographs

- Methods for differentiating computer-generated images (CG) from real photographs (PG)
- Collection of approximately 4850 PGs and 4850 CGs with large diversity of image content and quality.

computer-generated image (CG)

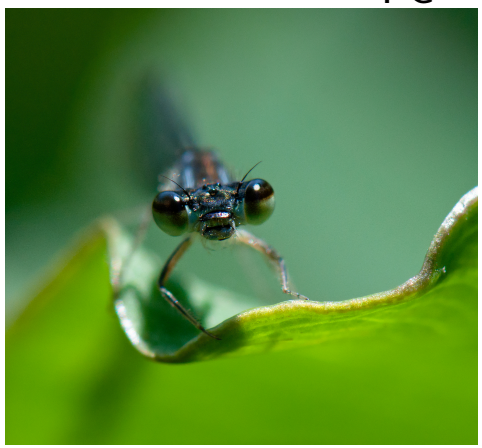
photograph (PG)

Identification of Computer-Generated Images × Digital Photographs

- The website *Fake or Foto* requests every visitor to identify, among 12 images, which ones are PGs or CGs.

CG

PG



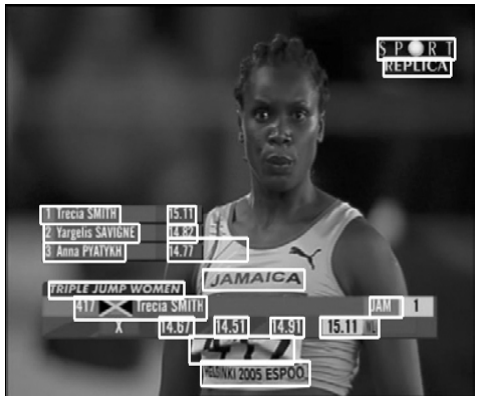
CG

CG

PG

Text Identification in Images and Videos

Text Identification in Images and Videos



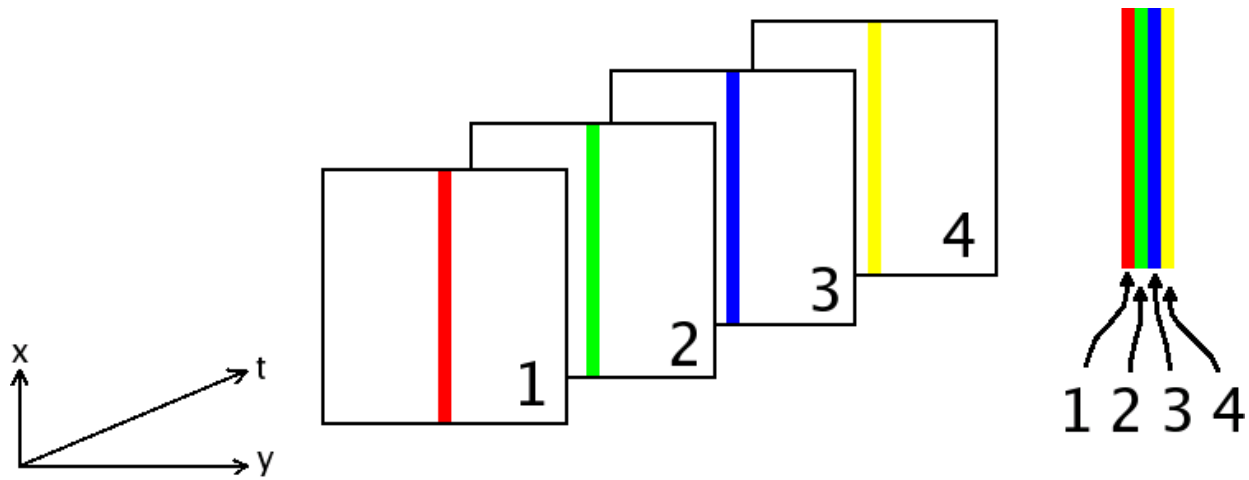
Video Caption Detection

Visual Rhythm or Video Slice

- The visual rhythm is defined as the union of all video slices:

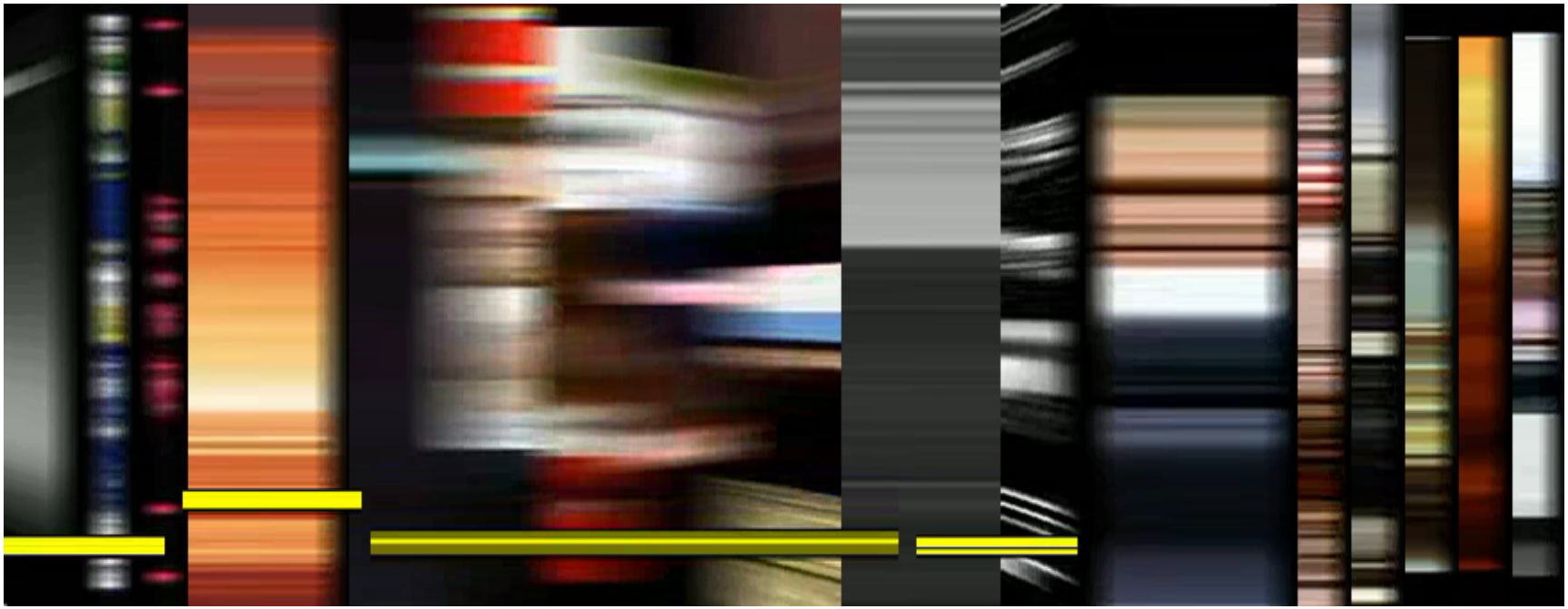
$$\boxed{rhythm(i,t) = \cup\, slice(i, F_t)}$$

where t is the time unit and F_t is the video frame for such instant.



Video Caption Detection

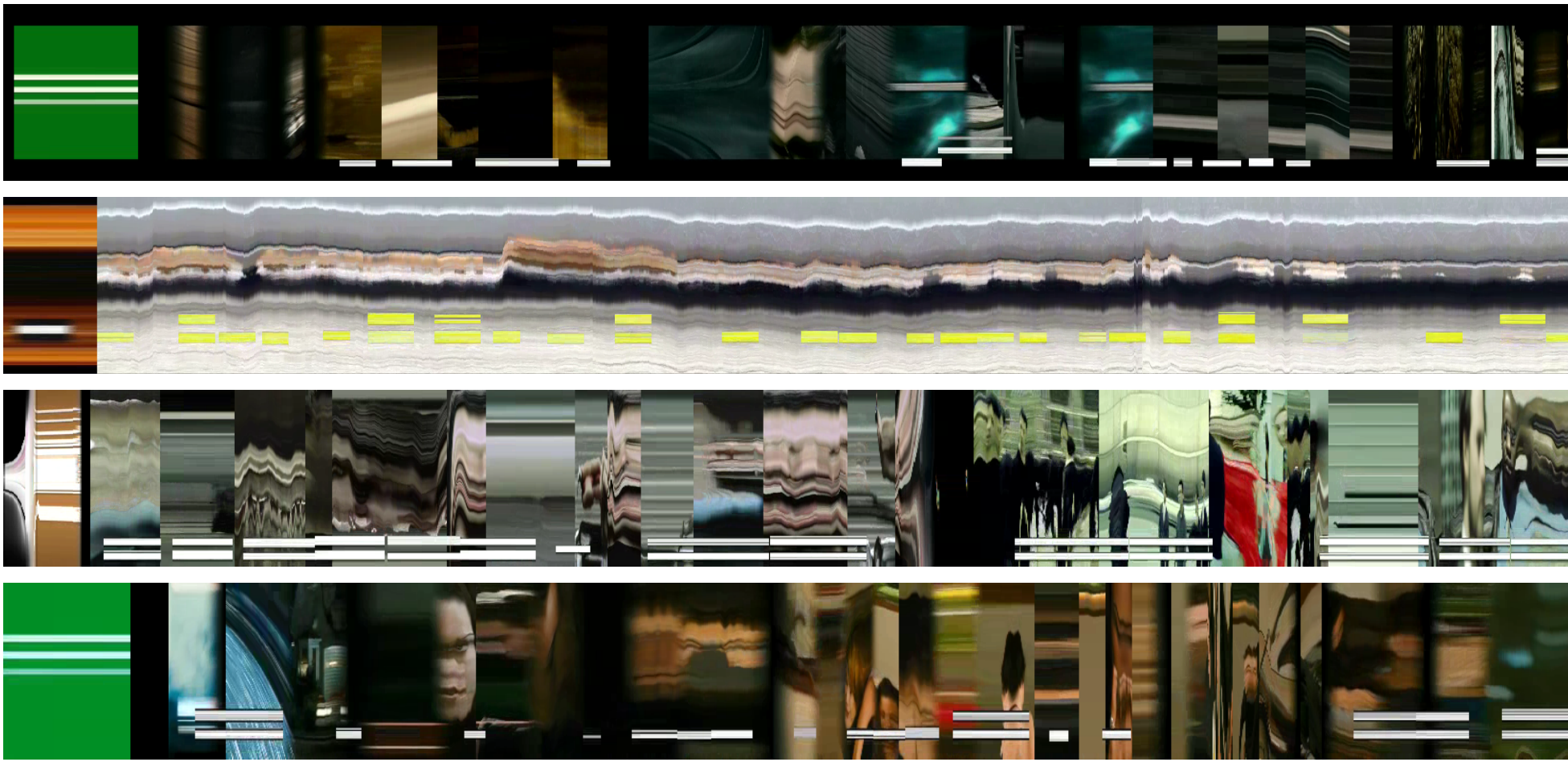
Visual Rhythm or Video Slice



Visual rhythm constructed through a vertical line at the center of the frame, applied to a video with 872 frames, each one with dimension of 448×336 pixels.

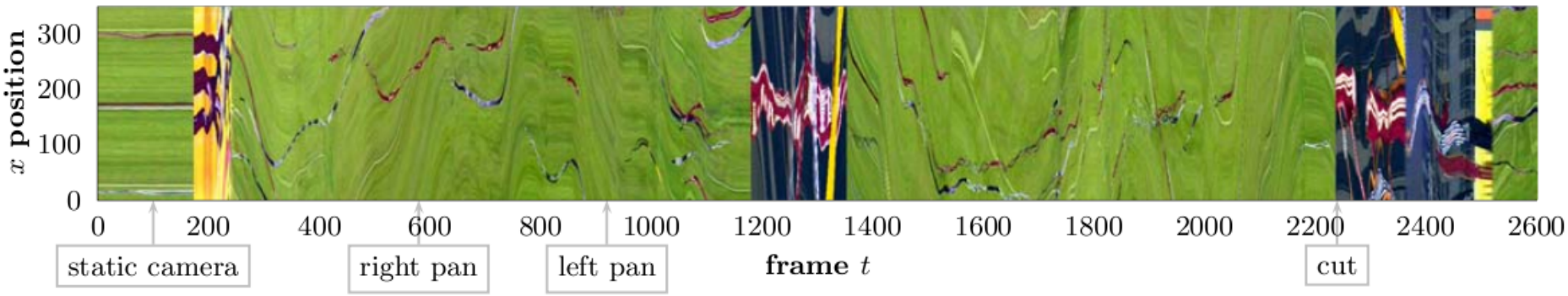
Video Caption Detection

Visual Rhythm or Video Slice



Video Caption Detection

Visual Rhythm or Video Slice



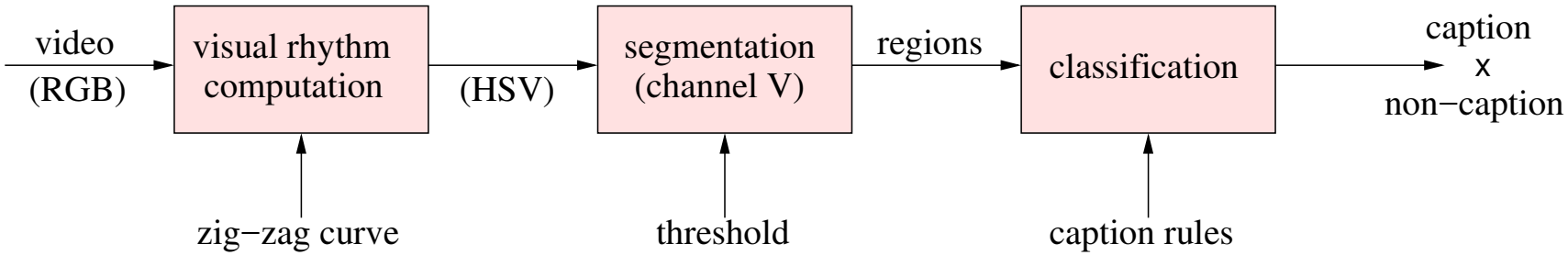
Video Caption Detection

Visual Rhythm or Video Slice

- Assumptions for subtitled texts:
  - The lowest line of the caption should be at least 1/12 of the screen height just above the bottom of the screen.
  - The minimum duration of a single-word subtitle should be 3/2 seconds.
  - For the spectator to clearly perceive a caption transition, at least 1/4 of a second needs to be inserted between two consecutive subtitle.

Video Caption Detection

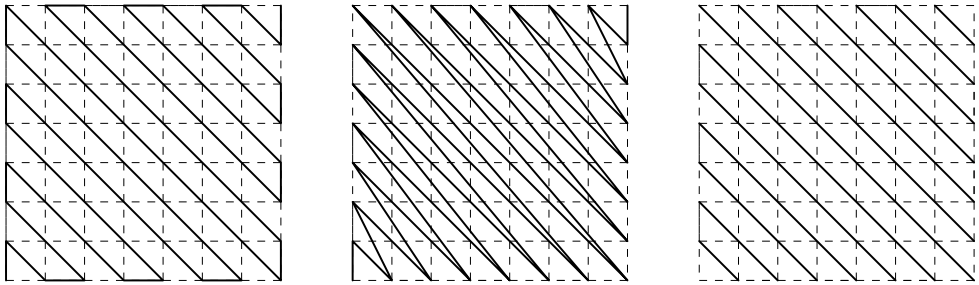
Visual Rhythm or Video Slice



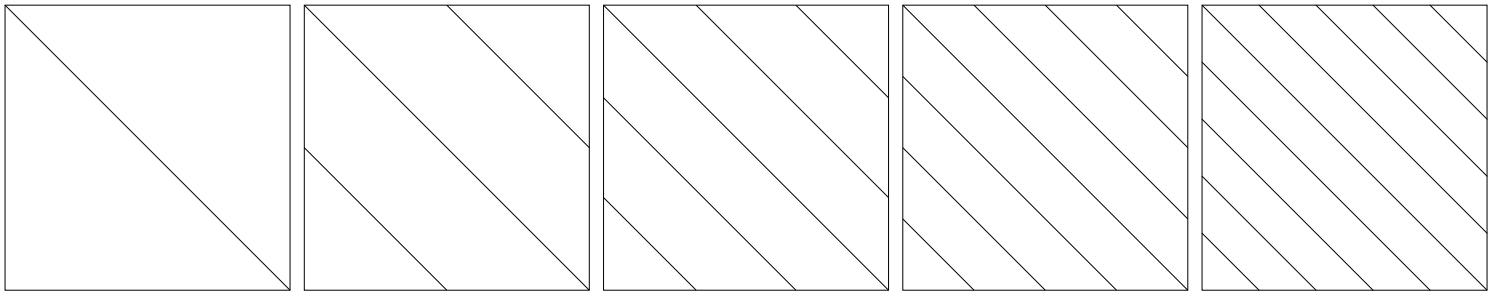
Pipeline for video caption detection.

Video Caption Detection

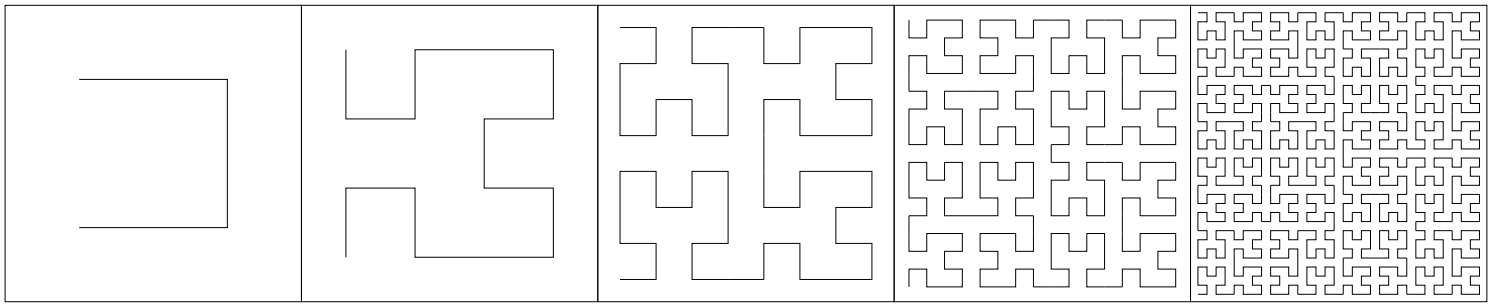
Visual Rhythm or Video Slice



different zig-zag curves



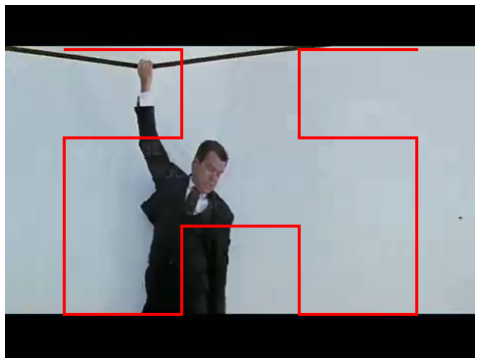
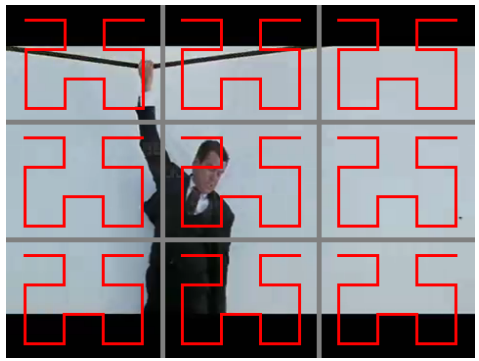
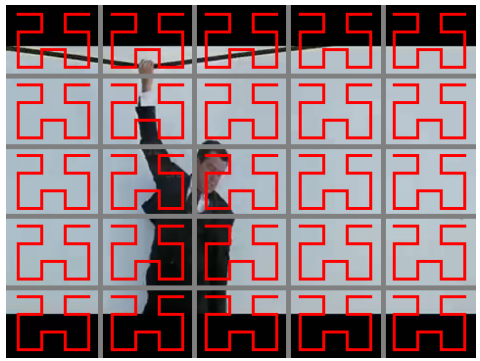
zig-zag curves for scales 1, 3, 5, 7 and 9, respectively



Hilbert curve for five approximations

Video Caption Detection

Visual Rhythm or Video Slice

Abnormal Event Detection

- United Kingdom has 20 percent of the world's closed-circuit television (CCTV) cameras and only 1 percent of the world's population.
- According to Metropolitan Police, approximately one crime is solved for every 1,000 CCTV cameras per year.

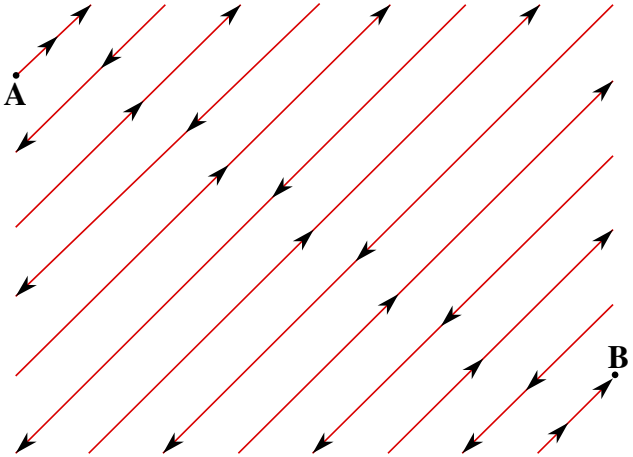
Source: Drew Swainston – `http://www.securitynewsdesk.com`

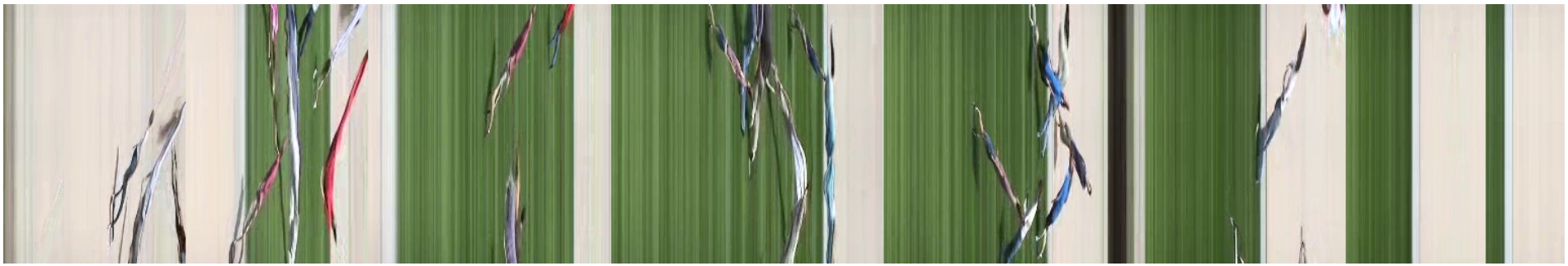
Abnormal Event Detection

http://securitycameraking.com

http://www.eastlondonadvertiser.co.uk

Abnormal Event Detection



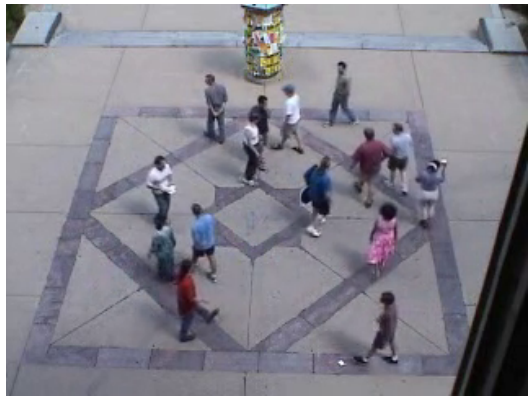


Visual rhythm constructed through the zig-zag path.

Abnormal Event Detection

Source: `https://www.youtube.com/watch?v=os-habmUBuU`

Abnormal Event Detection

UMN Crowd Activity Data Set



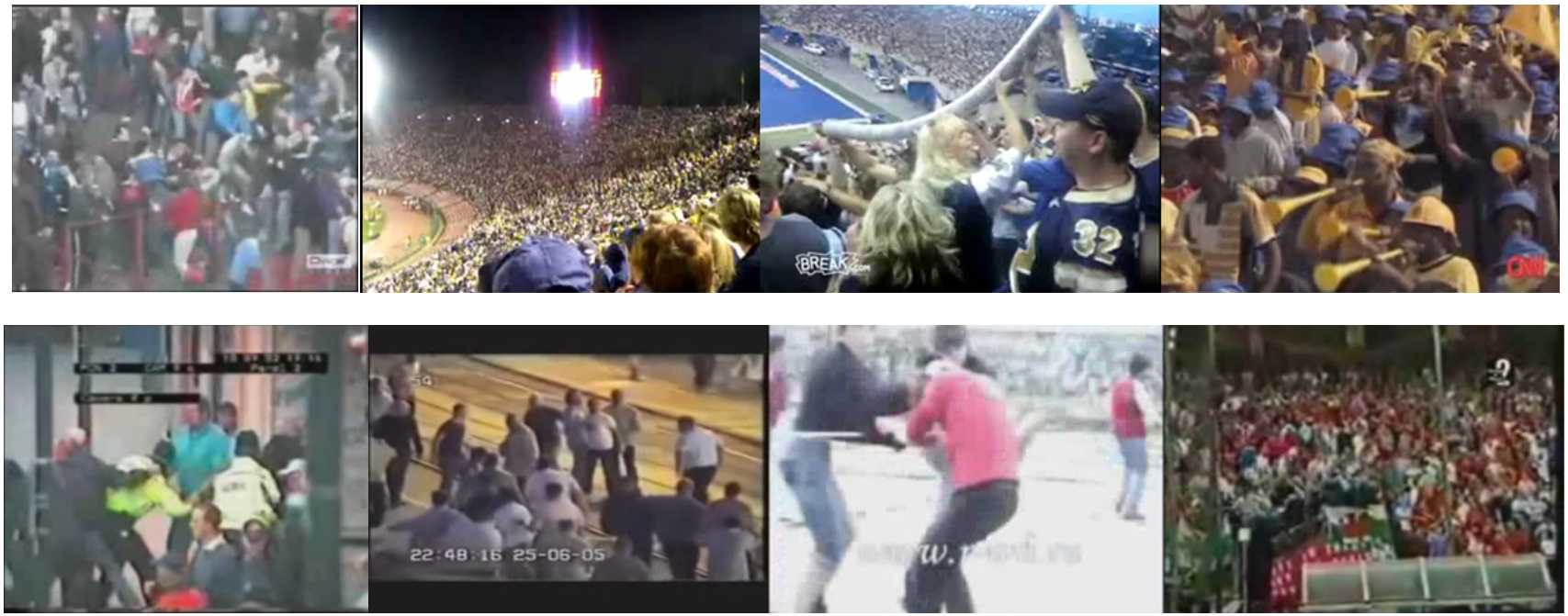
UCSD-Pedestrian Data Set (Ped1 and Ped2)

Abnormal Event Detection

Subway Data Set

Recognition of Human Actions / Activities

- Violent Flows data set: contains 246 crowded scenes categorized into two groups, each one with 123 violent and 123 non-violent scenes.



Samples of frames extracted from the Violent Flows data set.

Recognition of Human Actions / Activities

- Hockey Fights data set: contains 1000 clips categorized into two groups, 500 related to fight scenes and the other half to common hockey scenes. The video sequences were distributed into 5 folds, each one with 100 clips with fight scenes and 100 clips with no fight scenes.

Samples of frames extracted from the Hockey Fight data set.

Recognition of Human Actions / Activities

Samples of frames extracted from the Weizmann data set.

Recognition of Human Actions / Activities

Samples of frames extracted from the KTH data set.

Recognition of Human Actions / Activities

Samples of frames extracted from the MuHAVi data set.

Recognition of Human Actions / Activities

LIRIS

IXMAS

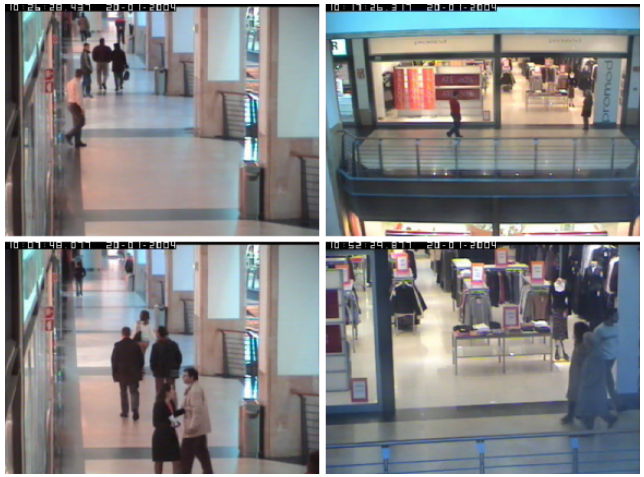
MSR Action

UCF-Sports

Samples of frames extracted from the LIRIS, IXMAS, MSR Action, and UCF-Sports data sets.

Recognition of Human Actions / Activities

VIRAT

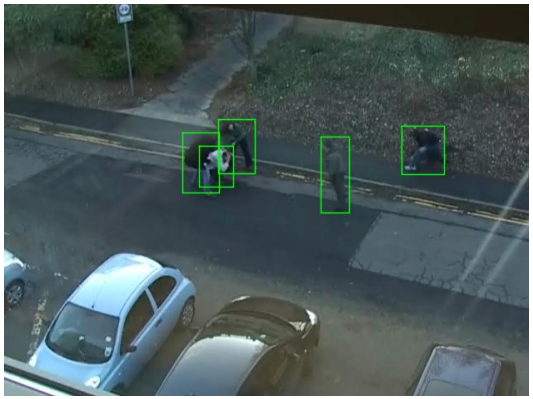


CAVIAR

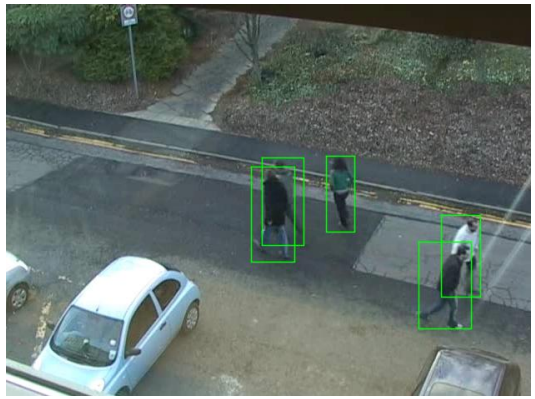
Daily Living

Samples of frames extracted from the VIRAT, CAVIAR, and Daily Living data sets.

Recognition of Human Actions / Activities



BEHAVE (fight)



BEHAVE (people walking)

TV Human Interaction

Base UT Interaction

Samples of frames extracted from BEHAVE, UT Interaction, and TV Human Interaction data sets.

Recognition of Human Actions / Activities

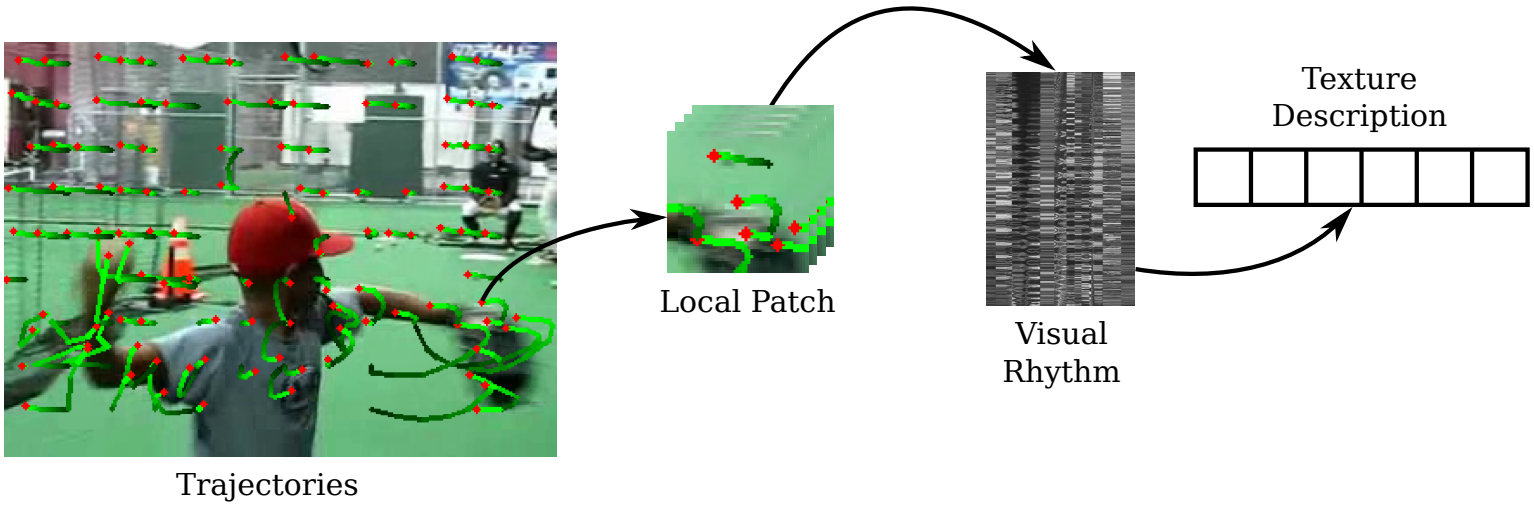

DogCentric Activity Dataset

Recognition of Human Actions / Activities

- original gray scale images: the original domain conveys appearance information.
- intensity gradients (x and y): gradients are often used to represent shape and have shown to provide discriminative information for action recognition.
- optical flow (x and y): motion information has been shown to be complementary to shape, also contributing with discriminative power.
- motion boundaries (x and y): defined as the gradient of optical flow images, they convey information about both shape and motion.

Recognition of Human Actions / Activities

- Construction of seven visual rhythm images for each video segment.
- A visual rhythm texture descriptor is the concatenation of texture features for every visual rhythm image of a given video segment.



Construction of the visual rhythm descriptors.

Emotion Classification Through Facial Expression Analysis

- Facial expressions are an important demonstration of human's humor and emotions.
- Emotions can be demonstrated through facial expressions, body movement, voice intonation, heart rhythm, among other forms.
- Facial expressions, which enable communication of emotions and intentions in a more easy, simple and natural way.
- Applications:
  - image retrieval
  - human-computer interaction
  - action recognition

Emotion Classification Through Facial Expression Analysis

JAFFE Data Set

- The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The photos were taken at the Psychology Department in Kyushu University.
- Basic expressions: happiness, sadness, surprise, anger, disgust, fear.

Samples from JAFFE data set. (a) neutral facial expression; (b) facial expression for anger; (c) facial expression for happiness.

Emotion Classification Through Facial Expression Analysis

AR Face Data Set

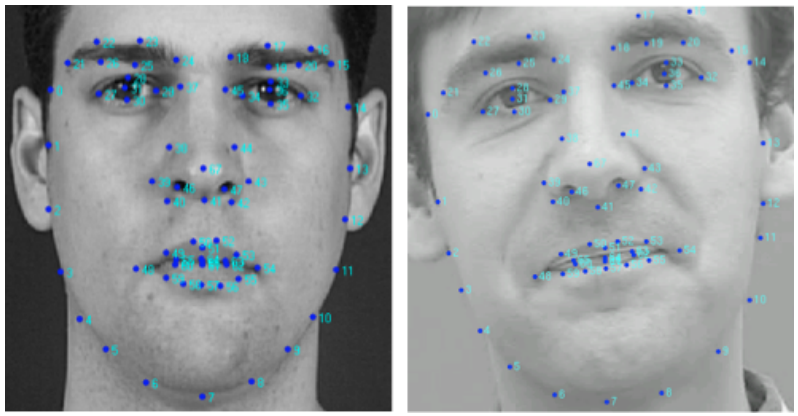
- The AR Face database, from the Ohio State University, contains 4,000 color images corresponding to 126 people's faces (70 men and 56 women).
- Images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf).

Images with occlusions from AR Face data set. (a) facial expression for happiness + glasses; (b) face with sun glasses; (c) face with glasses + scarf.

Emotion Classification Through Facial Expression Analysis

BioID Face Data Set

- The BioID Face data set, from HumanScan AG, Switzerland, consists of 1521 gray level images with a resolution of 384×286 pixels. Each one shows the frontal view of a face of one out of 23 different test persons. For comparison reasons, the set also contains manually set eye positions.
- The data set contains additional information of points that highlights facial characteristics of 1521 images. Such images were labeled by doctorate students from ISBE (Image Science and Biomedical Engineering) at the University of Manchester, United Kingdom.



Images from the BioID Face data set.

Emotion Classification Through Facial Expression Analysis

Cohn-Kanade AU Coded Facial Expression Data Set

- Subjects in the released portion of the Cohn-Kanade AU-Coded Facial Expression Database are 100 university students. They ranged in age from 18 to 30 years. Sixty-five percent were female, 15 percent were African-American, and three percent were Asian or Hispanic. Subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units and combinations of action units. Image sequences from neutral to target display were digitized into 640 $\times$ 480 or 490 pixel arrays with 8-bit precision for grayscale values.

Images from Cohn-Kanade data set. Facial expressions for surprise, sadness and happiness.

Emotion Classification Through Facial Expression Analysis

- Related recognition problems:
  - Age.
  - Gender.
  - Ethnicity.
- Extensions:
  - 3D data sets.
  - Video sequences.

Video Summarization

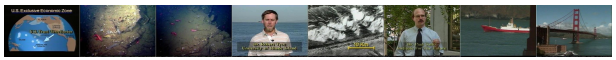
- Demand for developing techniques that are capable of manipulating digital videos in an automatic, efficient and accurate way, concerning the issues of searching, browsing, retrieval and analysis of contents.
- Video summarization: consists of deriving a short version from a given video, preserving as much relevant information as possible, such that the users can grasp the message transmitted by the original video.
- Challenging task since there is a variety of video genres, such as sports, movies, news programs, documentaries, and home movies in general. Even to humans, it is hard to reach a consensus to know how good a summary is, since what is relevant to ones may not be to others.

Video Summarization

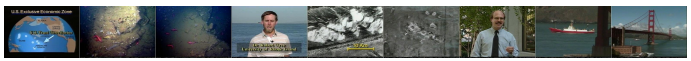
- A collection of 50 videos of several genres from Open Video Project (OVP).
- All videos are in MPEG format, with 30 frames per second and 352×240 pixels for each frame. Together, they have approximately a total duration of 75 minutes (where each video varies from 1 to 4 minutes of duration) and 150,000 frames (352×240 pixels).

Video Summarization

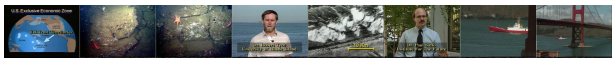
Open Video Project

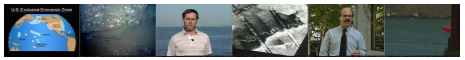
USER 1 

USER 2 

USER 3 

USER 4 

USER 5 

DT (0.706) 

STIMO (0.621) 

OVP (0.631) 

VSQUAL (0.714) 

VISON (0.801) 

VSUMM (0.708) 

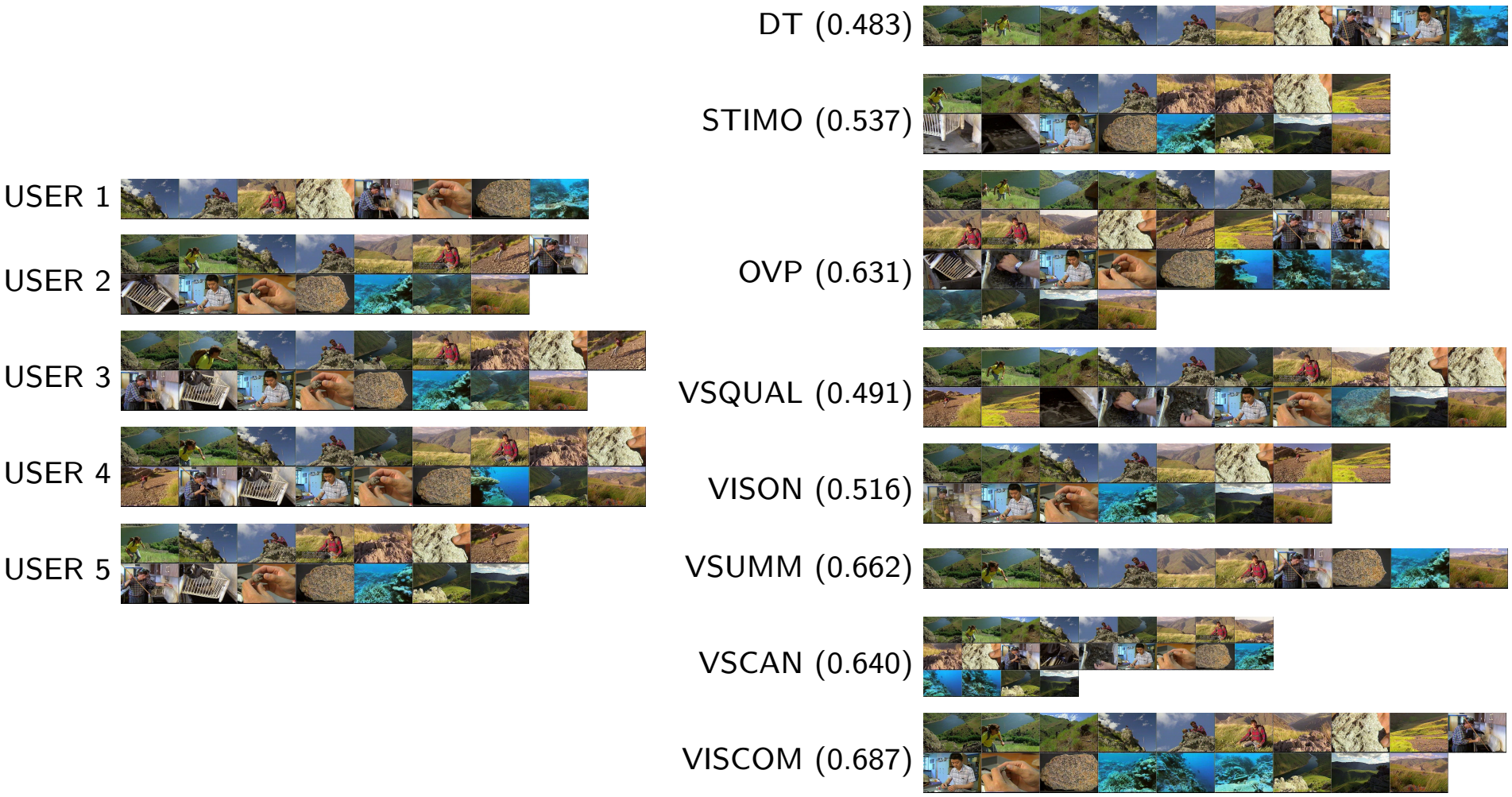
VSCAN (0.729) 

VISCOM (0.819) 

User summaries and automatic summaries of each method from the video *America's New Frontier, Segment 10*, along with the respective F-measures.

Video Summarization

Open Video Project

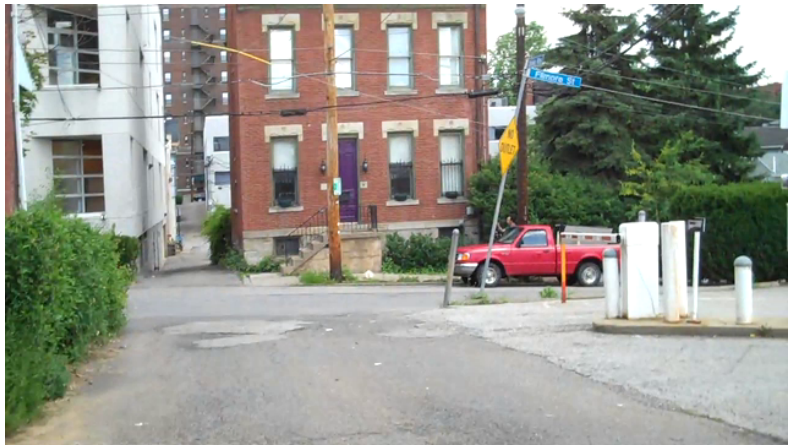


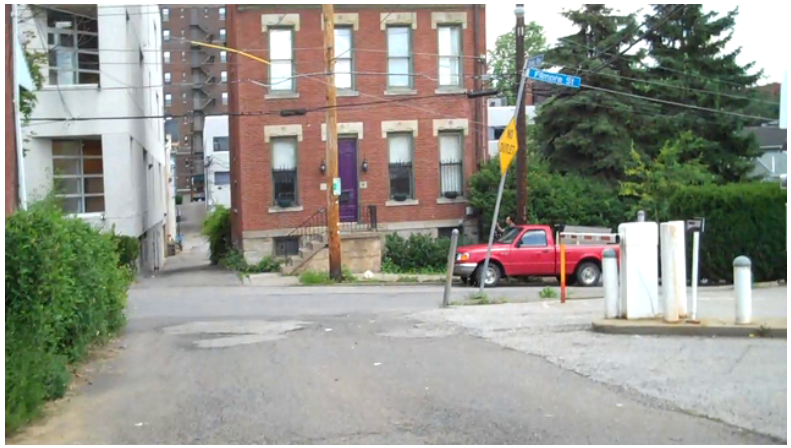
User summaries and automatic summaries of each method from the video *Exotic Terrane, Segment 04*, along with the respective F-measures.

Video Stabilization

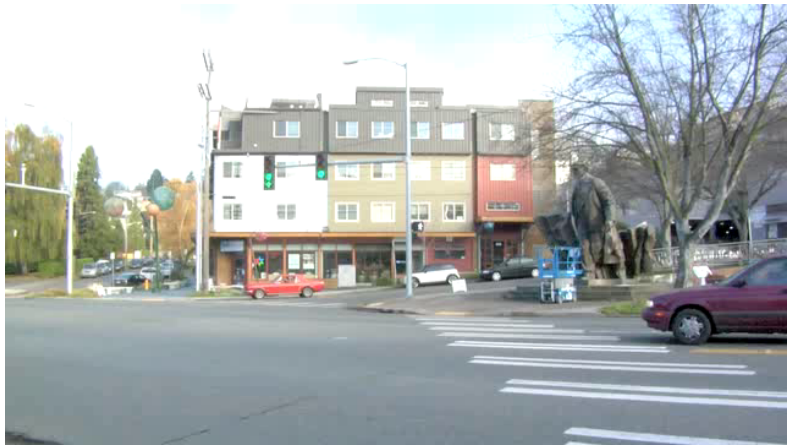
- It aims to correct problems caused in the acquisition process, particularly when the cameras are non-static and handled by amateurs.

Video Stabilization

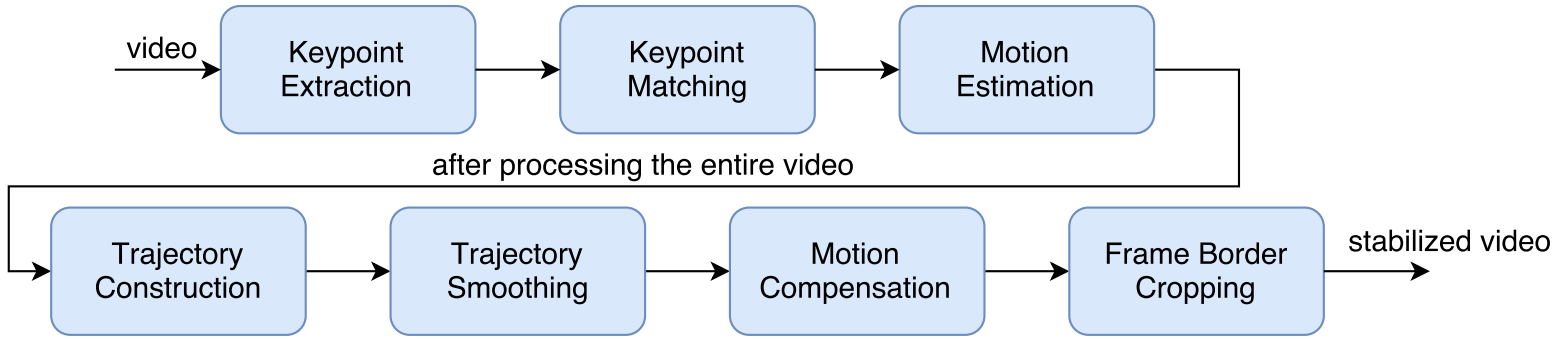




Video Stabilization

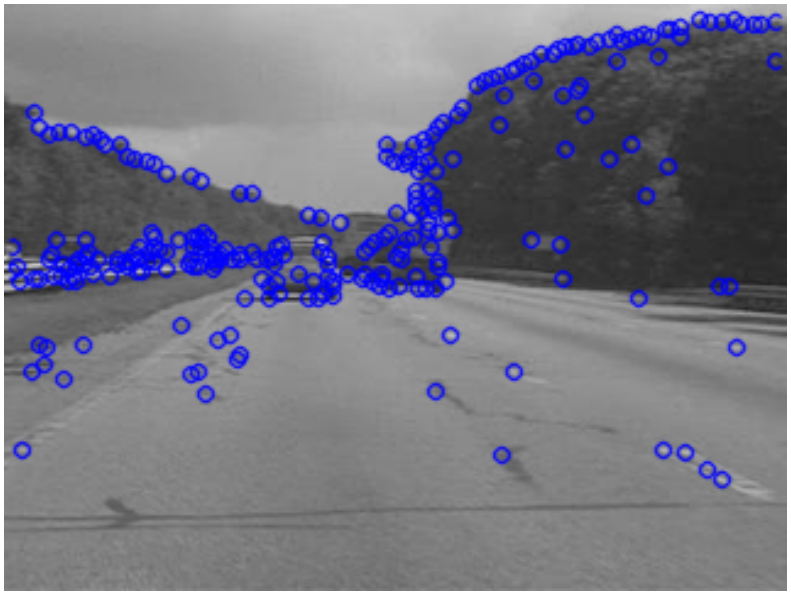


Video Stabilization



Method for digital video stabilization.

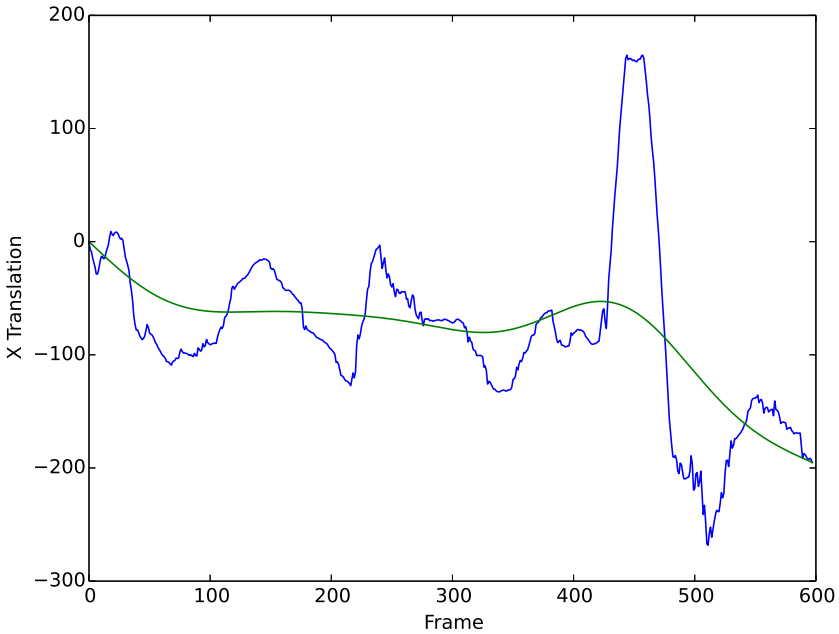
Video Stabilization



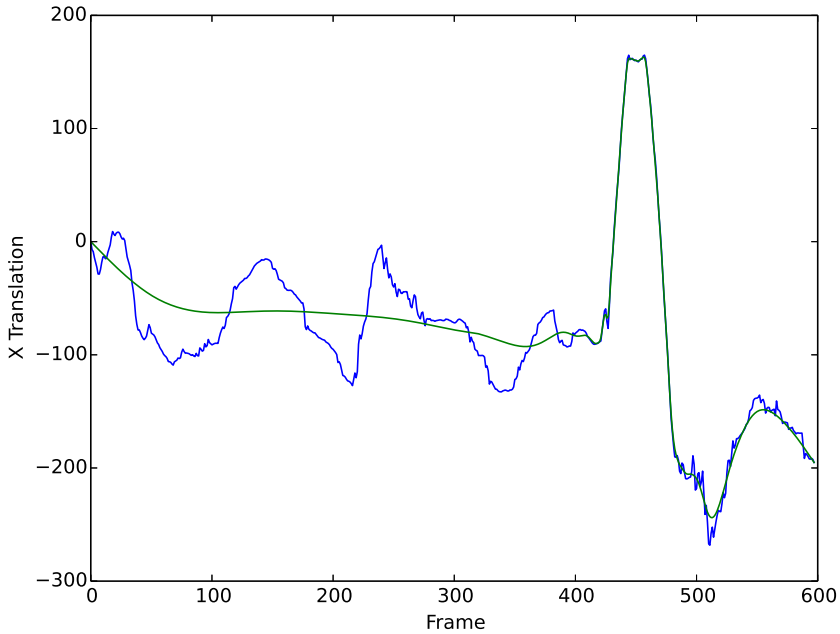
keypoint detection

keypoint matching

Video Stabilization



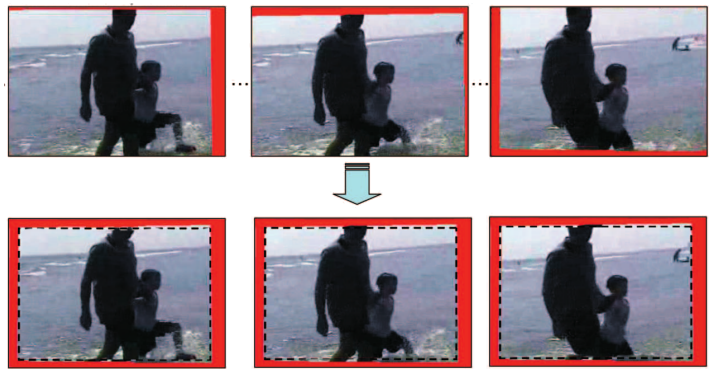
Gaussian filter with σ=40



adaptive Gaussian filter

Video Stabilization

transformed frame



frame cropping

- Interframe Transformation Fidelity (ITF):

$$\text{ITF} = \frac{1}{N-1}\sum_{k=1}^{N-1}\text{PSNR}(k)$$

where N is the number of video frames and Peak Signal to Noise Ratio (PSNR) is expressed as:

$$\text{PSNR}(f_i, f_{i+1}) = 10\log_{10}\frac{WHL_{max}^2}{\sum_{x=1}^{W}\sum_{y=1}^{H}[f_i(x,y) - f_{i+1}(x,y)]^2}$$

where f_i and f_{i+1} are two consecutive frames, W and H are the width and height of each frame, respectively, and L_{max} is the maximum intensity value present in the frame.

Trends

- Spatio-temporal features.
- Multiscale analysis.
- Contextual information.
- Applications for mobile devices (smartphones, tablets).
- Deep learning:
  - layer design.
  - activation functions.
  - loss function.
  - regularization.
  - fast computation.
- Faster and more accurate acquisition hardware:
  - vision processing units (specialized for video encoding/decoding).
  - graphics processing units (GPU).
  - mobile devices.
  - wearable sensors (eyeglasses, gyroscopes, accelerometers, heart rate monitors, blood pressure monitors).
  - ambient sensors (cameras, switches, light sensors).