Multi-modal Human-computer Interaction

Attila Fazekas

Attila.Fazekas@inf.unideb.hu

SSIP 2008, 9 July 2008

Hungary and Debrecen

9 July,2008
SSIP'09

Debrecen – Big Church

University of Debrecen

Coming Soon

➡ Summer School on Image Processing 2009

Coming Soon

- Summer School on Image Processing 2009
- 

Coming Soon

- Summer School on Image Processing 2009
- 

- http:\\www.inf.unideb.hu\~ssip

Road Map

- ➡ Multi-modal interactions and systems (main categories, examples, benefits)
- ➡ Turk-2 – Multi-modal chess player
- ➡ Face detection, facial gestures recognition
- ➡ Experimental results
- ➡ Examples

Defining Multi-Modal Interaction

- There are two views on multi-modal interaction:

Defining Multi-Modal Interaction

- There are two views on multi-modal interaction:
  - The first focuses on the human side: perception and control. There the word modality refers to human input and output channels.

Defining Multi-Modal Interaction

- There are two views on multi-modal interaction:
  - The first focuses on the human side: perception and control. There the word modality refers to human input and output channels.
  - The second view focuses on synergistic using two or more computer input or output modalities.

Human-Centered View

➡ The focus is on multi-modal perception and control, that is, human input and output channels.

Human-Centered View

- The focus is on multi-modal perception and control, that is, human input and output channels.

- Perception means the process of transforming sensory information to higher-level representation.

The Modalities

➡ We can divide the modalities in seven groups

The Modalities

- We can divide the modalities in seven groups
  - Internal chemical (blood oxygen, etc.)

The Modalities

- We can divide the modalities in seven groups
  - Internal chemical (blood oxygen, etc.)
  - External chemical (taste, etc.)

The Modalities

- We can divide the modalities in seven groups
  - Internal chemical (blood oxygen, etc.)
  - External chemical (taste, etc.)
  - Somatic senses (touch, etc.)

The Modalities

- We can divide the modalities in seven groups
  - Internal chemical (blood oxygen, etc.)
  - External chemical (taste, etc.)
  - Somatic senses (touch, etc.)
  - Muscle sense (stretch, etc.)

The Modalities

- We can divide the modalities in seven groups
  - Internal chemical (blood oxygen, etc.)
  - External chemical (taste, etc.)
  - Somatic senses (touch, etc.)
  - Muscle sense (stretch, etc.)
  - Sense of balance

The Modalities

- We can divide the modalities in seven groups
  - Internal chemical (blood oxygen, etc.)
  - External chemical (taste, etc.)
  - Somatic senses (touch, etc.)
  - Muscle sense (stretch, etc.)
  - Sense of balance
  - Hearing

The Modalities

- We can divide the modalities in seven groups
  - Internal chemical (blood oxygen, etc.)
  - External chemical (taste, etc.)
  - Somatic senses (touch, etc.)
  - Muscle sense (stretch, etc.)
  - Sense of balance
  - Hearing
  - Vision

System-Centered View

➡ In computer science multi-modal user interfaces have been defined in many ways.

System-Centered View

- In computer science multi-modal user interfaces have been defined in many ways.
- Chatty's explanation of multi-modal interaction is the one that most computer scientist use. With the term multi-modal user interface they mean a system that accepts many different inputs that are combined in a meaningful way.

Definition of the Multimodality

➡ "Multi-modality is the capacity of the system to communicate with a user along different types of communication channels."

Definition of the Multimodality

- "Multi-modality is the capacity of the system to communicate with a user along different types of communication channels."

- Both multimedia and multi-modal systems use multiple communication channels.

Definition of the Multimodality

- "Multi-modality is the capacity of the system to communicate with a user along different types of communication channels."
- Both multimedia and multi-modal systems use multiple communication channels. But a multi-modal system strives for meaning.

Two Types of Multi-modal Systems

➡ The goal is to use the computer as a tool.

Two Types of Multi-modal Systems

- The goal is to use the computer as a tool.
- The computer as a dialogue partner.

History

➡ Bolt's Put-That-There system

History

➡ Bolt's Put-That-There system. In this system the user could move objects on screen by pointing and speaking.

History

- ➡ Bolt's Put-That-There system. In this system the user could move objects on screen by pointing and speaking.
- ➡ CUBRICON is a system that uses mouse pointing and speech.

History

- Bolt's Put-That-There system. In this system the user could move objects on screen by pointing and speaking.
- CUBRICON is a system that uses mouse pointing and speech.
- Oviatt presented a multi-modal system for dynamic interactive maps.

Benefits

➡ Efficiency follows from using each modality for the task that it is best suited for.

Benefits

- Efficiency follows from using each modality for the task that it is best suited for.
- Redundancy increases the likelihood that communication proceeds smoothly because there are many simultaneous references to the same issue.

Benefits

- Efficiency follows from using each modality for the task that it is best suited for.
- Redundancy increases the likelihood that communication proceeds smoothly because there are many simultaneous references to the same issue.
- Perceptability increas when the tasks are facilitated in spatial context.

Benefits

➡ Naturalness follows from the free choice of modalities and may result in a human-computer communication that is close to human-human communication.

Benefits

➡ Naturalness follows from the free choice of modalities and may result in a human-computer communication that is close to human-human communication.

➡ Accuracy increases when another modality can indicate an object more accurately than the main modality.

Benefits

- Synergy occurs when one channel of communication can help refine imprecision, modify the meaning, or resolve ambihuities in another channel.

Applications

➡ Mobile telecommunication

Applications

- ➡ Mobile telecommunication
- ➡ Hands-free devices to computers

Applications

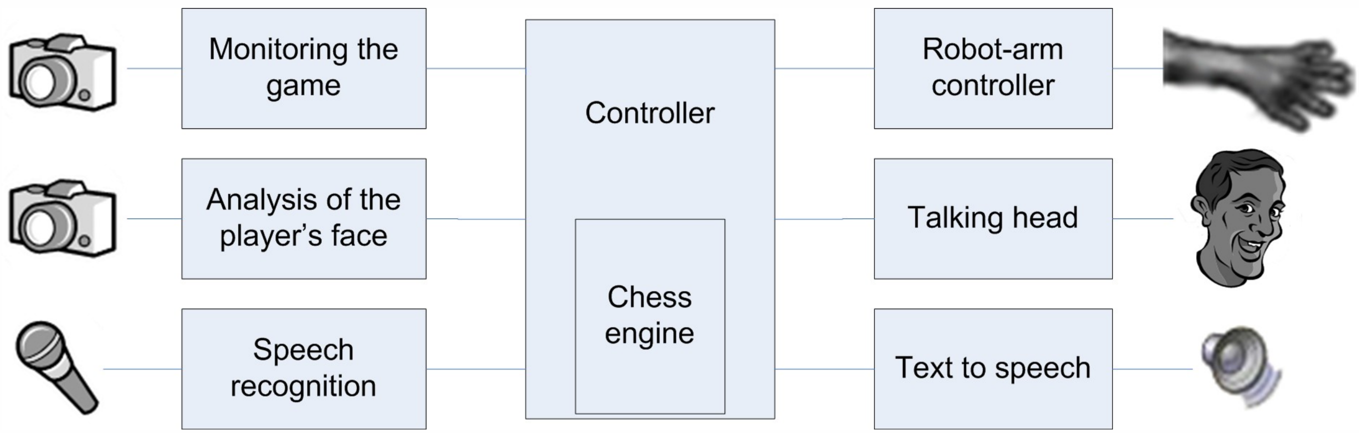
- ➡ Mobile telecommunication
- ➡ Hands-free devices to computers
- ➡ Using in a car

Applications

- Mobile telecommunication
- Hands-free devices to computers
- Using in a car
- Interactive information panel

Turk-2

System Components



Introduction

➡ Faces are our interfaces in our emotional and social live.

Introduction

- Faces are our interfaces in our emotional and social live.
- Automatic analysis of facial gestures is rapidly becoming an area of interest in multi-modal human-computer interaction.

Introduction

- Faces are our interfaces in our emotional and social live.
- Automatic analysis of facial gestures is rapidly becoming an area of interest in multi-modal human-computer interaction.
- Basic goal of this area of research is a human-like description of shown facial expression.

Introduction

- The solution of this problem can be based on the idea of some face detection approaches.

Related Research Topics

➡ Face detection (one face/image)

Related Research Topics

- Face detection (one face/image)
- Face localization (more faces/image)

Related Research Topics

- ➡ Face detection (one face/image)
- ➡ Face localization (more faces/image)
- ➡ Facial feature detection (eyes, mouth, etc.)

Related Research Topics

- Face detection (one face/image)
- Face localization (more faces/image)
- Facial feature detection (eyes, mouth, etc.)
- Facial expression recognition

Related Research Topics

- Face detection (one face/image)
- Face localization (more faces/image)
- Facial feature detection (eyes, mouth, etc.)
- Facial expression recognition
- Face recognition, face identification

Related Research Topics

- Face detection (one face/image)
- Face localization (more faces/image)
- Facial feature detection (eyes, mouth, etc.)
- Facial expression recognition
- Face recognition, face identification
- Face tracking

Problems of the Face Detection

➡ Pose: The images of a face vary due to the relative camera-face pose.

Problems of the Face Detection

- **Pose:** The images of a face vary due to the relative camera-face pose.
- **Presence or absence of structural components** (beards, mustaches, glasses etc.).

Problems of the Face Detection

- Pose: The images of a face vary due to the relative camera-face pose.

- Presence or absence of structural components (beards, mustaches, glasses etc.).

- Facial expression: The appearance of faces are directly affected by the facial expression.

Problems of the Face Detection

➡ Occlusion: Faces may be partially occluded by other objects.

Problems of the Face Detection

- **Occlusion:** Faces may be partially occluded by other objects.
- **Image orientation:** Face images vary for different rotations about the optical axis of the camera.

Problems of the Face Detection

- Occlusion: Faces may be partially occluded by other objects.
- Image orientation: Face images vary for different rotations about the optical axis of the camera.
- Imaging conditions (lighting, background, camera characteristics).

Face Detection in a Singe Image

➡ Knowledge-based methods (G. Yang and T.S. Huang, 1994).

Face Detection in a Singe Image

- Knowledge-based methods (G. Yang and T.S. Huang, 1994).
- Feature invariant approaches (T. K. Leung, M. C. Burl, and P. Perona, 1995), (K. C. Yow and R. Cipolla, 1996).

Face Detection in a Singe Image

➡ Template matching methods (A. Lanitis, C. J. Taylor, and T. F. Cootes, 1995).

Face Detection in a Singe Image

- ➡ Template matching methods (A. Lanitis, C. J. Taylor, and T. F. Cootes, 1995).
- ➡ Appearance-based methods (E. Osuna, R. Freund, and F. Girosi, 1997), (A. Fazekas, C. Kotropoulos, I. Pitas, 2002).

Face Detection

➡ Scanning of the picture by a running window in a multiresolution pyramid.

Face Detection

- Scanning of the picture by a running window in a multiresolution pyramid.
- Normalize of the window.

Face Detection

- ➡ Scanning of the picture by a running window in a multiresolution pyramid.
- ➡ Normalize of the window.
- ➡ Hide some parts of the face.

Face Detection

- ➡ Scanning of the picture by a running window in a multiresolution pyramid.
- ➡ Normalize of the window.
- ➡ Hide some parts of the face.
- ➡ Normalize of the local variance of the brightness on the picture.

- Equalization of the histogram.

Face Detection

- ➡ Equalization of the histogram.
- ➡ Localization of the face (decision).

➡ Let us consider a set of the facial pictures.

Face Gesture Recognition

➡ Let us consider a set of the facial pictures.

➡ Let us set up a finite system of some features related the pictures.

Face Gesture Recognition

- Let us consider a set of the facial pictures.
- Let us set up a finite system of some features related the pictures.
- It is known any pictures is related to only one class:

Face Gesture Recognition

- Let us consider a set of the facial pictures.
- Let us set up a finite system of some features related the pictures.
- It is known any pictures is related to only one class: face with the given gesture,

Face Gesture Recognition

- Let us consider a set of the facial pictures.
- Let us set up a finite system of some features related the pictures.
- It is known any pictures is related to only one class: face with the given gesture, face without the given gesture.

Face Gesture Recognition

➡ The problem to find a method to determine the class of the examined picture.

Face Gesture Recognition

- The problem to find a method to determine the class of the examined picture.
- One possible way to solve this problem: Support Vector Machine.

Experimental Results

➡ For all experiments the package SVMLight developed by T. Joachims was used. For complete test, several routines have been added to the original toolbox.

Experimental Results

- For all experiments the package SVMLight developed by T. Joachims was used. For complete test, several routines have been added to the original toolbox.

- The database recorded by our institute was used.

Experimental Results

➡ Training set of 40 images (20 faces with the given gesture, 20 faces without the given gesture.).

MMHCI
Turk 2
Face detection
Face gesture
Examples
SVM

Experimental Results

- ➡ Training set of 40 images (20 faces with the given gesture, 20 faces without the given gesture.).
- ➡ All images are recorded in 256 grey levels.

Experimental Results

- ➡ Training set of 40 images (20 faces with the given gesture, 20 faces without the given gesture.).
- ➡ All images are recorded in 256 grey levels.
- ➡ They are of dimension 640×480.

Experimental Results

➡ The procedure for collecting face patterns is as follows.

Experimental Results

- The procedure for collecting face patterns is as follows.

- A rectangle part of dimension 256×320 pixels has been manually determined that includes the actual face.

Experimental Results

➡ This area has been subsampled four times. At each subsampling, non-overlapping regions of 2×2 pixels are replaced by their average.

Experimental Results

- ➡ This area has been subsampled four times. At each subsampling, non-overlapping regions of 2×2 pixels are replaced by their average.
- ➡ The training patterns of dimension 16×20 are built.

Experimental Results

➡ The class label $+1$ has been appended to each pattern.

Experimental Results

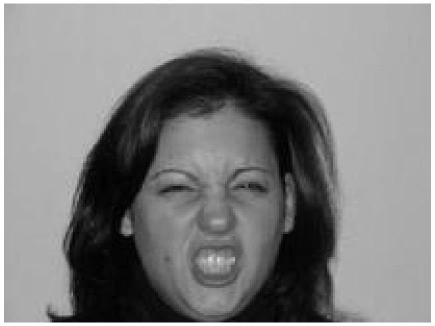
- The class label $+1$ has been appended to each pattern.
- Similarly, 20 non-face patterns have been collected from images in the same way, and labeled -1.

Facial Gesture Database

Surprising face

Smiling face

Sad face



Angry face

Classification Error

Angry	Happy	Sad	Serial	Suprised
22.4%	10.3%	11.8%	9.4%	18.9%

Examples

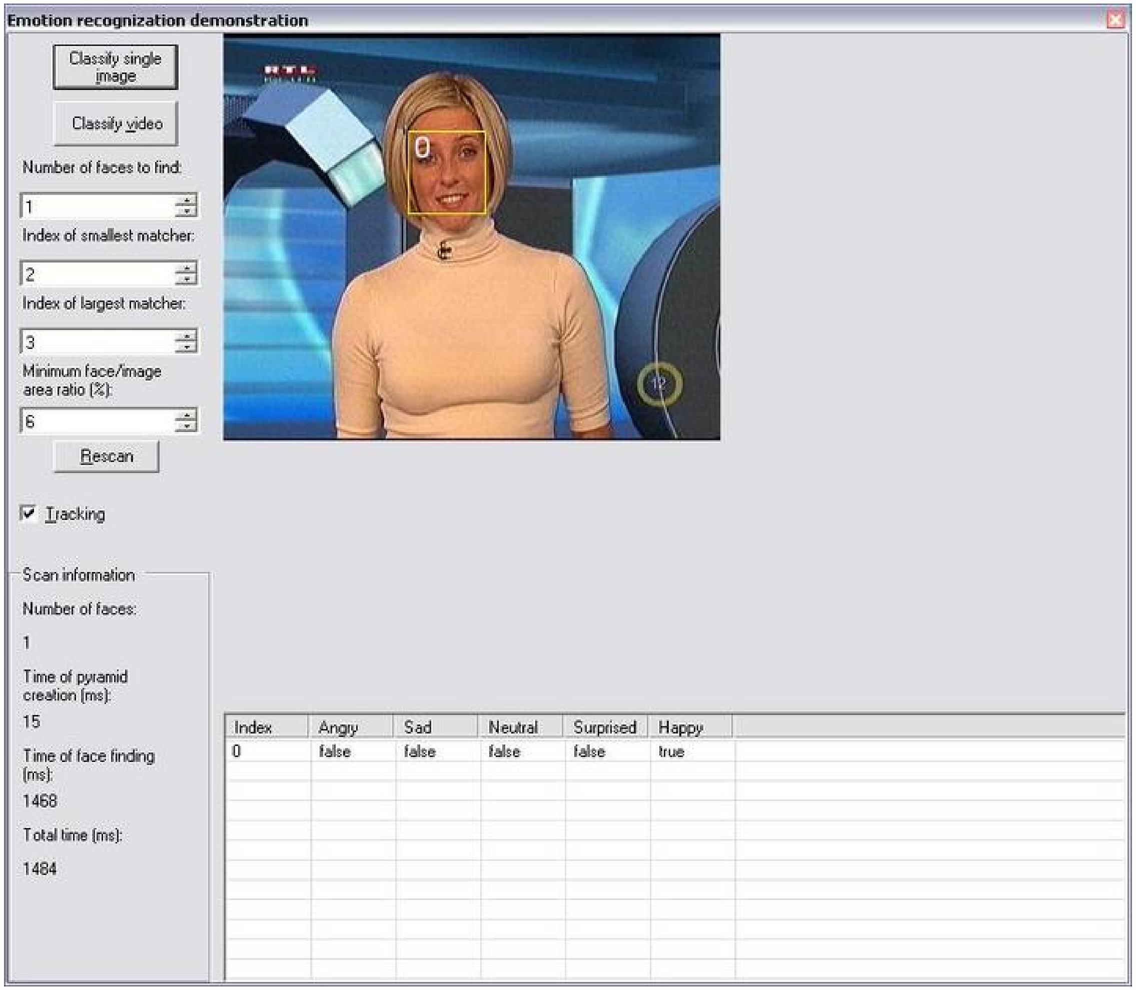
Examples


Emotion recognization demonstration
Classify single image
Classify video
Number of faces to find:
1
Index of smallest matcher:
2
Index of largest matcher:
3
Minimum face/image area ratio (%):
6
Rescan
Tracking
Scan information
Number of faces:
1
Time of pyramid creation (ms):
15
Time of face finding (ms):
1406
Total time (ms):
1421
IRAN'S NUCLEAR PROGRAM
CNN
MEXICO 243.95
U.S. FEDERAL RESERVE DISCUSSES INTEREST RATE HIKES
Index Angry Sad Neutral Surprised Happy
0 false true false false false

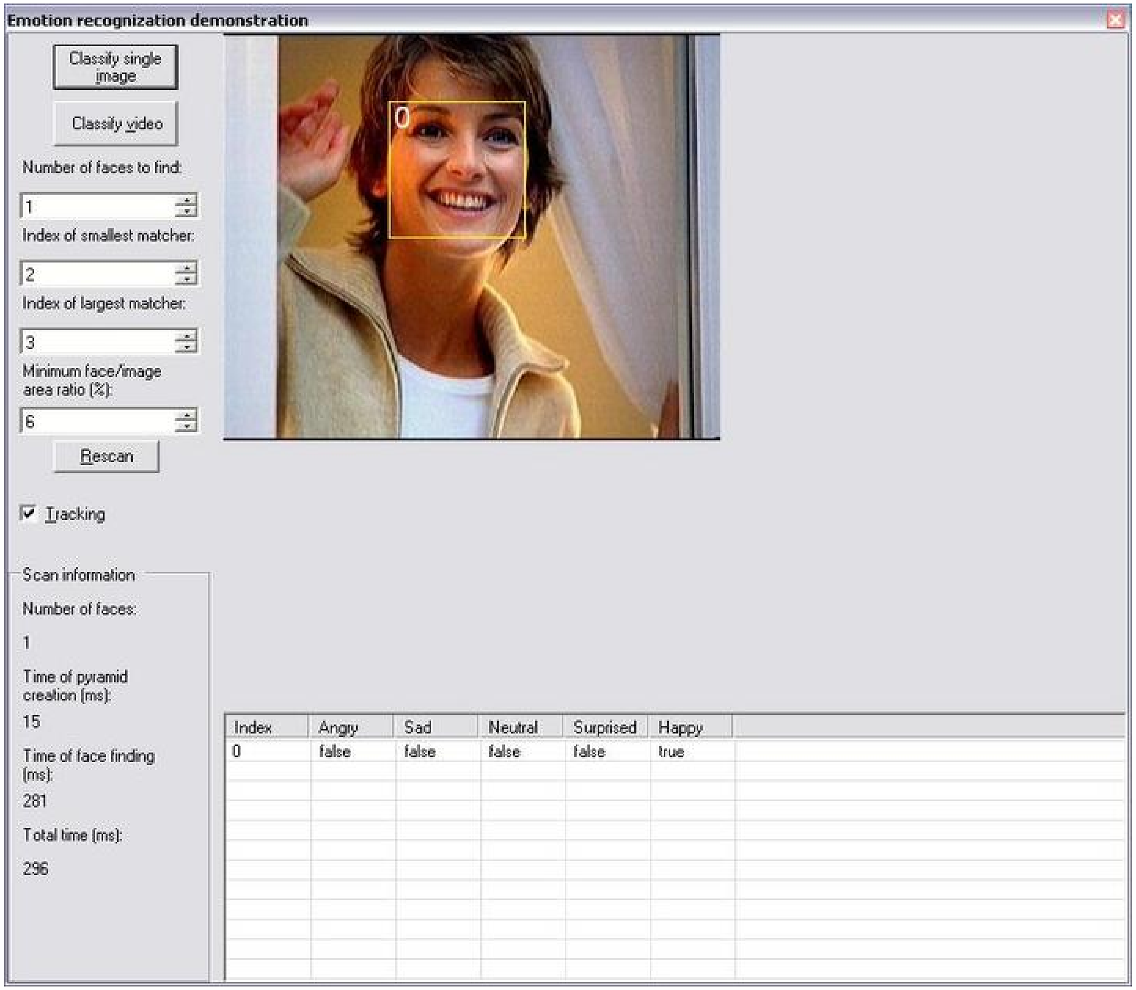
Examples

Emotion recognization demonstration
Classify single image
Classify video
Number of faces to find:
1
Index of smallest matcher:
2
Index of largest matcher:
3
Minimum face/image area ratio (%):
6
Rescan
Tracking
Scan information
Number of faces:
1
Time of pyramid creation (ms):
0
Time of face finding (ms):
281
Total time (ms):
281
0
BUY NORWEGIAN EXPLOSIVES COMPANY DYNO NOBEL FOR $1.5 BN.
Index
Angry
Sad
Neutral
Surprised
Happy
0
false
false
true
false
false

Examples


Emotion recognization demonstration
Classify single image
Classify video
Number of faces to find:
1
Index of smallest matcher:
2
Index of largest matcher:
3
Minimum face/image area ratio (%):
6
Rescan
Tracking
Scan information
Number of faces:
1
Time of pyramid creation (ms):
15
Time of face finding (ms):
1468
Total time (ms):
1484
Index Angry Sad Neutral Surprised Happy
0 false false false false true

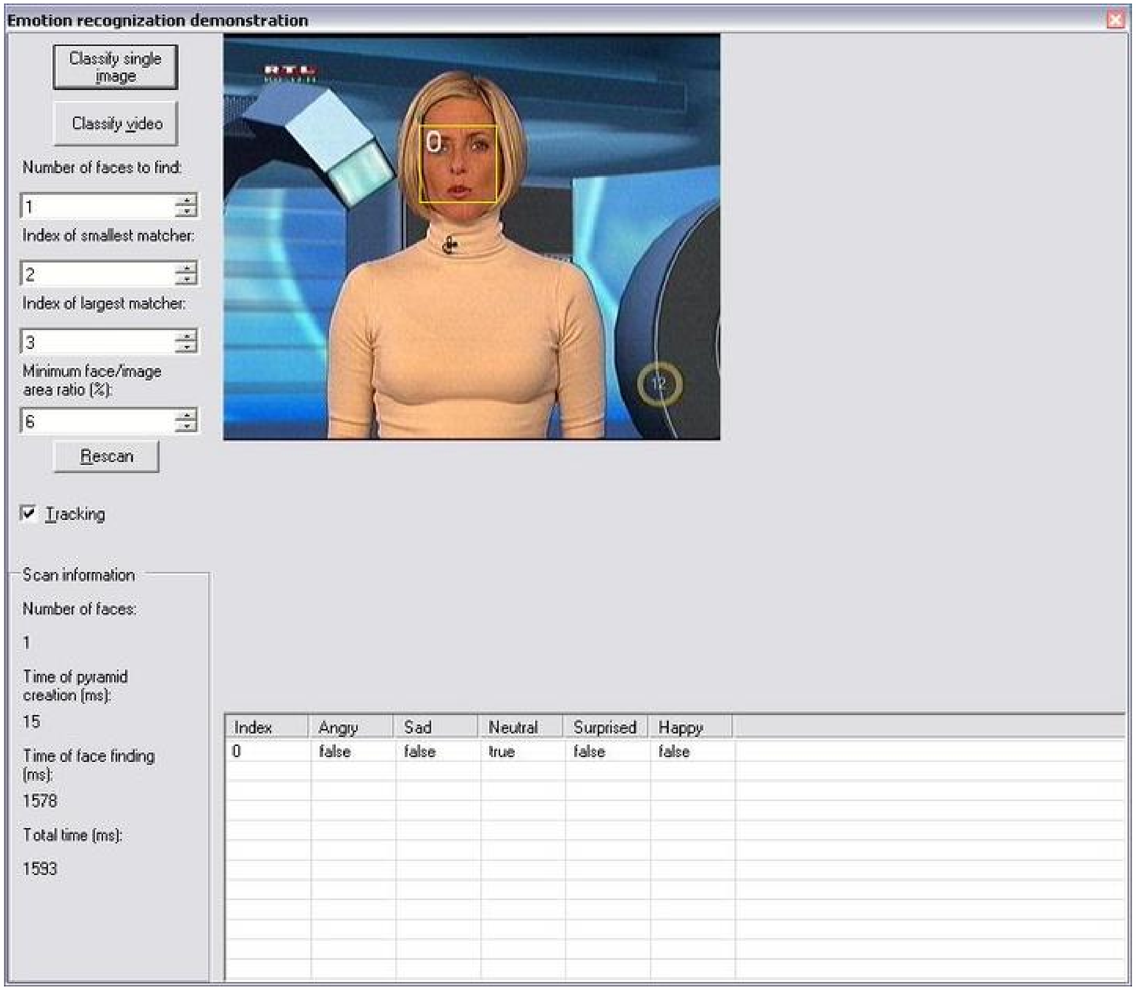
Examples


Emotion recognization demonstration
Classify single image
Classify video
Number of faces to find:
1
Index of smallest matcher:
2
Index of largest matcher:
3
Minimum face/image area ratio (%):
6
Rescan
Tracking
Scan information
Number of faces:
1
Time of pyramid creation (ms):
15
Time of face finding (ms):
281
Total time (ms):
296
0
Index Angry Sad Neutral Surprised Happy
0 false false false false true

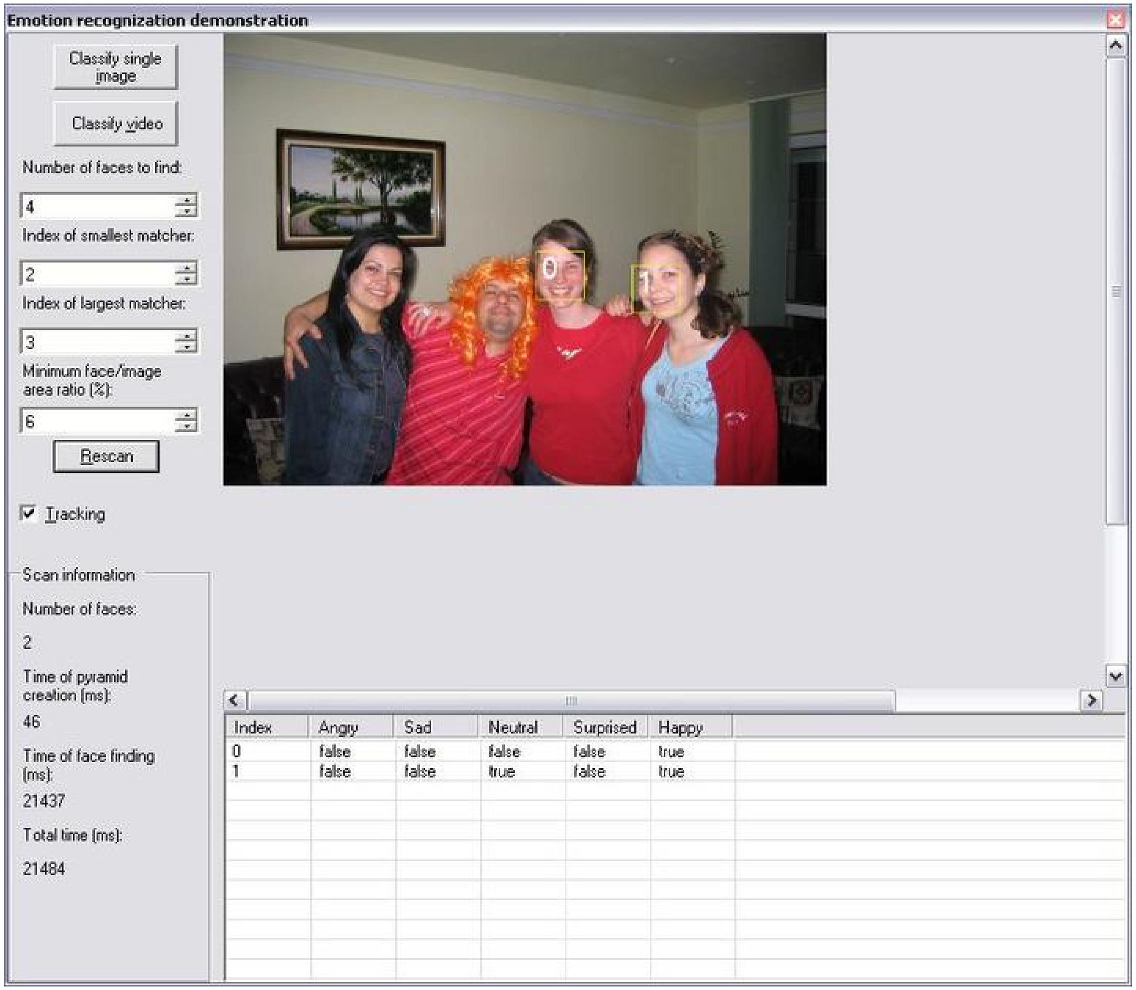
Examples


Emotion recognization demonstration
Classify single image
Classify video
Number of faces to find:
1
Index of smallest matcher:
2
Index of largest matcher:
3
Minimum face/image area ratio (%):
6
Rescan
Tracking
Scan information
Number of faces:
1
Time of pyramid creation (ms):
0
Time of face finding (ms):
671
Total time (ms):
671
CNN
MEXICO 243.95
ROWING TO LEAVE OUTPUT QUOTAS UNCHANGED. U.S. FEDERAL
Index Angry Sad Neutral Surprised Happy
0 false false false false false

Examples


Emotion recognization demonstration
Classify single image
Classify video
Number of faces to find:
1
Index of smallest matcher:
2
Index of largest matcher:
3
Minimum face/image area ratio (%):
6
Rescan
Tracking
Scan information
Number of faces:
1
Time of pyramid creation (ms):
15
Time of face finding (ms):
1578
Total time (ms):
1593
Index
Angry
Sad
Neutral
Surprised
Happy
0
false
false
true
false
false

Examples



MMHCI
Turk 2
Face detection
Face gesture
Examples
SVM

Support Vector Machine

➡ Statistical learning from examples aims at selecting from a given set of functions $\{f_\alpha(\mathbf{x}) \mid \alpha \in \Lambda\}$, the one which predicts best the correct response.

Support Vector Machine

➡ This selection is based on the observation of l pairs that build the training set:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \;\; \mathbf{x}_i \in \mathbb{R}^m, y_i \in \{+1, -1\}$$

which contains input vectors $\mathbf{x}_i$ and the associated ground "truth" given by an external supervisor.

Support Vector Machine

- This selection is based on the observation of l pairs that build the training set:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \;\; \mathbf{x}_i \in \mathbb{R}^m, y_i \in \{+1, -1\}$$

which contains input vectors $\mathbf{x}_i$ and the associated ground "truth" given by an external supervisor.

- Let the response of the learning machine $f_\alpha(\mathbf{x})$ belongs to a set of indicator functions.

Support Vector Machine

- If we define the loss-function:

$$L(y, f_\alpha(\mathbf{x})) = \begin{cases} 0, & \text{if } y = f_\alpha(\mathbf{x}), \\ 1, & \text{if } y \neq f_\alpha(\mathbf{x}). \end{cases}$$

The expected value of the loss is given by:

$$R(\alpha) = \int L(y, f_\alpha(\mathbf{x}))p(\mathbf{x}, y)d\mathbf{x}dy,$$

where $p(\mathbf{x}, y)$ is the joint probability density function of random variables $\mathbf{x}$ and y.

Support Vector Machine

➡ We would like to find the function $f_{\alpha_0}(\mathbf{x})$ which minimizes the risk function $R(\alpha)$.

Support Vector Machine

- ➡ We would like to find the function $f_{\alpha_0}(\mathbf{x})$ which minimizes the risk function $R(\alpha)$.
- ➡ The basic idea of SVM to construct the optimal separating hyperplane.

Support Vector Machine

➡ Suppose that the training data can be separated by a hyperplane, $f_\alpha(\mathbf{x}) = \alpha^T\mathbf{x} + b = 0$, such that:

$$y_i(\alpha^T\mathbf{x}_i + b) \geq 1, \;\; i = 1, 2, \ldots, l$$

where α is the normal to the hyperplane.

Support Vector Machine

➡ Suppose that the training data can be separated by a hyperplane, $f_\alpha(\mathbf{x}) = \alpha^T\mathbf{x} + b = 0$, such that:

$$y_i(\alpha^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, l$$

where α is the normal to the hyperplane.

➡ For the linearly separable case, SVM simply seeks for the separating hyperplane with the largest margin.

Support Vector Machine

➡ For linearly nonseparable data, by mapping the input vectors, which are the elements of the training set, into a high-dimensional feature space through so-called kernel function.

Support Vector Machine

- For linearly nonseparable data, by mapping the input vectors, which are the elements of the training set, into a high-dimensional feature space through so-called kernel function.

- We construct the optimal separating hyperplane in the feature space to get a binary decision.

Thank you for your attention!