IDIAP COMMUNICATION

Annotation of face detection: description of XML format and files

Sébastien Marcel [a] Yann Rodriguez [a]
Maël Guillemot [a] Andrei Popescu-Belis [b]

IDIAP–Com 06-06

July 2006

[a] IDIAP Research Institute, rue du Simplon 4, 1920 Martigny, Switzerland
[b] ISSCO - Université de Genève, 40, bd. du Pont-d'Arve, 1211 Genève 4, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

1 Face detection

The goal of face detection is to determine whether or not there are any faces in the image and, if present, their location. Face detection is the crucial first step of any application that involves face processing systems including face recognition, face tracking, pose estimation or expression recognition.

In most applications, the face region is represented by a bounding box, such as a rectangle or a square. In this document, we assume that every image is processed by a face detector, which decides, according to specific settings, the location of faces. When a face is detected (or if there is no significant change in the image with respect to the previous one) the location and size of the bounding box (top-left coordinate x, top-left coordinate y, width w and height h) and a confidence on the detection are returned.

2 Access to files

The face detection annotation files are named using the following format:
meetingID.cameraID.facepos.xml (e.g. `ES2002a.Closeup1.facepos.xml`). It is likely that these files belong to a folder named 'faceDetection', and possibly a sub-folder bearing the name of the detection algorithm used to create the annotation.

Currently, the 'facepos' files are available from the following URL:
http://mmm.idiap.ch/protected/idiap/facedetection/. The files corresponding to the scenario meetings from the three AMI/IM2 institutions are stored in separate folders.

Video files are accessible from the following URLs:

- M4 meeting videos: http://mmm.idiap.ch/publicMeetings.html,
- AMI [IS, ES, TS] meeting videos: http://corpus.amiproject.org.

For more information about video signals, please contact `mmmAdmin[at]idiap.ch`.

3 XML file format – version 1.0

The XML annotation format for face detection information is composed of a header with metadata information about the algorithm and the video signal, and a body which contains the face detection information proper, with one `frame` element for each detected face in each video frame. A sample file is provided below.

3.1 Header

The header of each XML file provides the annotation type, a reference to the name of the processed video file as well as detailed information about the video source and the face detection method. Information like the size of every image in the video, the total number of images in the video, the frame rate (typically 25 frames per second), the name of the face detection algorithm (here MCT-MS-MF-MV-1) and the value of the parameters are encoded in the header. The structure is defined in the DTD.

3.2 Body

The body provides on every line information about the position of ONE and only ONE face detected in one frame (image), as a `frame` element. Hence, if no face is detected in one image, then no `frame` element will be output for that image; conversely, if more than one face is detected, there will be more than one `frame` element for the corresponding image. The `frame` elements are ordered chronologically and appear on separate lines (though this is irrelevant to the XML format).

Images are identified by an absolute number (n='000000') starting at zero for the first image of the video. For every image, the inserted VITC timecode is provided (when available). This timecode (t='00:00:10:03') allows synchronization with other video or audio streams.

Using the following sample line,

```
<frame n='000003' t='00:00:10:06' x='442' y='237' w='47' h='47'
                                                  score='0.642835'/>
```

the description of the attributes is the following:

- `n` is the absolute image number starting at zero.
- `t` is the VITC timecode. t='00:00:11:18' means 0 hours, 0 minutes, 11 seconds and 18 frames (ranging from 00 to 24 in a second).
- `x` is the top-left X coordinate of the bounding box of a face.
- `y` is the top-left Y coordinate of the bounding box of a face.
- `w` is the width of the bounding box of a face.
- `h` is the height of the bounding box of a face.
- `score` is the "confidence" of the detected face.
- Some images are missing (i.e. they do not have a corresponding `frame` element) because faces are not detected or are not present.
- `w` equals `h` on every line because in the present case the bounding box for faces was set to be a square.
- `w`, `h` and `score` can be identical across a number of consecutive images because no significant changes have been detected and therefore the previous detections are maintained.

3.3 Sample file

This is an excerpt from a valid XML file.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE video SYSTEM "facepos.dtd">
<!-- author: Sebastien Marcel (http://www.idiap.ch/~marcel) -->
<video annotation-type="face detection"
       mode="automatic"
       id="Scripted-Meeting-TRN-05.Cam1">
   <source
      filename="/data/Scripted-Meeting-TRN-05/Cam1/
                            Cam1_T000010.120_T000439.320.avi"
      width="720"
      height="576"
      nframes="6983"
      fps="25" />
   <method name="MCT-MS-MF-MV-1">
      <parameter name="min size" value="38"/>
      <parameter name="max size" value="74"/>
      <parameter name="motion"   value="800"/>
      <parameter name="modulo detection" value="25"/>
   </method>
   <frame n='000000' t='00:00:10:03' x='442' y='237' w='47' h='47'
                                                  score='0.642835'/>
   <frame n='000001' t='00:00:10:04' x='442' y='237' w='47' h='47'
                                                  score='0.642835'/>
...
   <frame n='006974' t='00:00:10:11' x='190' y='254' w='57' h='57'
                                                  score='0.799428'/>
   <frame n='006981' t='00:00:10:18' x='478' y='240' w='57' h='57'
                                                  score='0.697516'/>
</video>
```

The DTD follows the structure below (note that since the names of the parameters of the method cannot necessarily be known in advance, they are encoded as generic `parameter` elements):

```
<!ELEMENT video (source, method?, frame+)>
<!ATTLIST video annotation-type CDATA #FIXED "face detection"
                id CDATA #REQUIRED
                mode CDATA #IMPLIED>
<!ELEMENT source EMPTY>
<!ATTLIST source filename CDATA #REQUIRED
                 width CDATA #IMPLIED
                 height CDATA #IMPLIED
```

```
                 nframes CDATA #IMPLIED
                 fps CDATA #IMPLIED>
<!ELEMENT frame EMPTY>
<!ATTLIST frame n CDATA #REQUIRED
                t CDATA #IMPLIED
                x CDATA #REQUIRED
                y CDATA #REQUIRED
                w CDATA #REQUIRED
                h CDATA #REQUIRED
                score CDATA #IMPLIED>
<!ELEMENT method (parameter*)>
<!ATTLIST method name CDATA #REQUIRED>
<!ELEMENT parameter EMPTY>
<!ATTLIST parameter name CDATA #REQUIRED
                    value CDATA #IMPLIED>
```