Emotion challenge fact sheet

iCV

February 2017

1 Team details

- Team name: CMTech
- Team leader name: Federico Sukno
- Team leader address, phone number and email: Roc Boronat 138, Barcelona, Spain - federico.sukno@upf.edu
- Rest of the team members: Dmytro Derkach, Adria Ruiz
- Team website URL (if any): http://fsukno.atspace.eu/Research.htm
- Affiliation: Universitat Pompeu Fabra, Department of Information and Communication Technologies

2 Contribution details

- Title of the contribution: Head pose estimation based on 3-D facial landmarks localization and regression
- Final score:At the learning phase, the sum of avrage angle errors on the evaluation data was 21.49 degrees.
- General method description: The method consists of 4 steps: 1) The head region (including a variable part of the shoulders) is separated from the background by simple clustering of the 3D points; 2) A state of the art 3D landmarking algorithm (SRILF) is applied to the head region in search for 12 facial landmarks; 3) The detected landmarks are used to produce both a geometry-based and a regression-based estimates of the head pose. Upon agreement of these two estimates, the average of both methods is returned. Otherwise, we proceed to the next step; 4) A bag-of-words approach using a modified 3D Shape Contexts descriptor is used to estimate the head pose angles from a set of descriptors randomly sampled on the head region.

- References: Sukno, F.M., Waddington, J.L. and Whelan, P.F. (2015). *3-D facial landmark localization with asymmetry patterns and shape regression from incomplete local features.* IEEE transactions on cybernetics, 45(9), 1717-1730.

- Representative image / diagram of the method: Figure 1

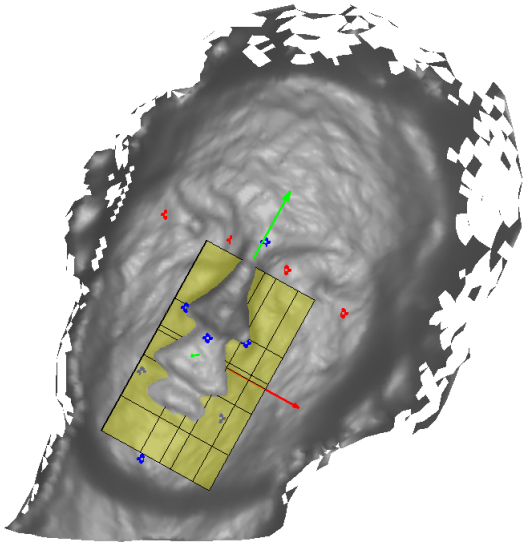

Figure 1: The Mesh built from the depth map with automatically detected landmarks. Blue points are used for fitting a plane and estimation of pitch and yaw angles (green arrows), and red point are used for roll angle estimation (red arrow).

- Describe data preprocessing techniques applied (if any): The depth map input is firstly converted to a 3D point cloud so that all subsequent processing steps are performed based on a full-3D representation instead of using the 2.5D representation of the depth map.

3 Head Pose estimations

3.1 Features / Data representation

The 3D landmark detection software uses Asymmetry Pattern Shape Contexts (APSC) to describe the local geometry. For the regression steps we have used 3D Shape Contexts (3DSC) with the view-point fixed to coincide with the camera viewpoint (hence independent from the normals) and a fixed initial position for the azimuth bins. These modifications make the resulting descriptor dependant on the orientation of the surface (and hence are informative about the head

pose). In the standard formulation of 3DSC (as well as other descriptors such as spin images or APSC) the orientation of the descriptors depends on the surface normals, aiming to provide rotational invariance, which is clearly not our goal.

There were two stages of the system where descriptors were used: 1) to describe the local geometry around the detected landmarks and perform regression to estimate the head angles: in this case, only one descriptor per landmark point is computed; 2) to describe the global geometry by means of a dictionary of local descriptors, with no correspondences between surfaces (bag-of-words approach): in this case the surface was sampled randomly with an umbrella operator to limit the number of descriptors per surface.

3.2 Dimensionality reduction

At the landmark-regression stage we performed dimensionality reduction using Principal Component Analysis on the concatenated descriptors to regularize the estimated regressor.

In the bag-of-words approach we performed quantization of the descriptors randomly sampled over the whole head surface ending up with a histogram of 500 bins that was set as input to the regression step. This representation reduced the dimensionality of the input between 1,000 and 2,000 times.

3.3 Learning strategy

Learning strategy applied (if any)
The geometry based approach does not need any learning strategy, as it is based on fitting planes a lines to the detected landmarks. Both regressors, however, need training. We used linear regression for the landmark-based regressor and ridge linear regression for the bag-of-words approach. For the latter, we also needed to construct the dictionary, which we did by applying k-means clustering on a relatively small subset of the training data (approx 800 surfaces). The regressors, on the other hand, were trained using all training data except the surfaces used to build the dictionary) and parameters were adjusted in cross validation setting.

3.4 Other techniques

Other technique/strategy used not included in previous items (if any)
We used Shape Regression with Incomplete Local Features (SRILF) for landmark detection on the generated point clouds (containing head and, possible, part of the shoulders). The algorithm generates sets of candidate locations from feature detectors and performs combinatorial search constrained by a flexible shape model. A key assumption of our approach is that for some landmarks there might not be an accurate candidate in the input set. This is tackled by detecting partial subsets of landmarks and inferring those that are missing, so

that the probability of the flexible model is maximized. The ability of the model to work with incomplete information makes it possible to limit the number of candidates that need to be retained (which reduces computational load) but it also makes it tolerant to incomplete surfaces (e.g. holes, self-occlusions). This is very helpful when working with depth data captured from a single camera, as it is the case of the Kinect 2 images used in this challenge.

The results reported here have been obtained using the free implementation of the SRILF algorithm (SRILF 3D Face Landmarker), available at: http://fsukno.atspace.eu/Data.htm#SRILF_3dFL

4 Global Method Description

- Total method complexity: all stages: Approximately 10 seconds per image (on average), running on an Intel i7-4770 processor at 3.4 GHz with 16 Gb or RAM.
- Which pre-trained or external methods have been used (for any stage, if any):
  The results reported here have been obtained using the free implementation of the SRILF algorithm (SRILF 3D Face Landmarker), available at: http://fsukno.atspace.eu/Data.htm#SRILF_3dFL
- Which additional data has been used in addition to the provided training and validation data (at any stage, if any): None

5 Other details

- Language and implementation details (including platform, memory, parallelization requirements):
  MATLAB R2016a, running on Windows 10 platform (should be compatible at least back to Windows 7, but did not test).
- Detailed list of prerequisites for compilation: Install SRILF 3D Face Landmarking software, available at: http://fsukno.atspace.eu/Data.htm#SRILF_3dFL
- Human effort required for implementation, training and validation? About 4 weeks.
- Training/testing expended time? Approximately 10 seconds per image (on average), running on an Intel i7-4770 processor at 3.4 GHz with 16 Gb or RAM.
- General comments and impressions of the challenge? There were some issues with the ground truth of training data at the initial stages of the challenge, which reduced the actual time of the learning phase. The majority of these issues seem solved but we have still identified a few wrong angles in the (revised) ground truth.