Generating Facial Ground Truth with Synthetic Faces

Rossana Queiroz, Marcelo Cohen, Juliano L. Moreira, Adriana Braun, Júlio C. Jacques Júnior, Soraia Raupp Musse
Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS
Graduate Programme in Computer Science
Virtual Human Laboratory -www.inf.pucrs.br/∼vhlab
Porto Alegre, Brazil

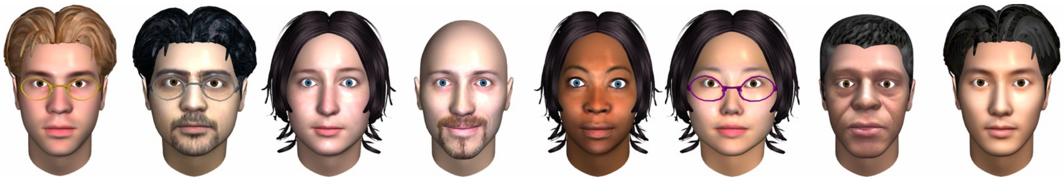


Figure 1. A sample of 3D faces generated by our prototype.

***Abstract*—This work describes a methodology for generation of facial ground truth with synthetic faces. Our focus is to provide a way to generate accurate data for the evaluation of Computer Vision algorithms, in terms of facial detection and its components. Such algorithms play a key role in face detection. We present a prototype in which we can generate facial animation videos using a 3D face models database, controlling face actions, illumination conditions and camera position. The facial animation platform allows us to generate animations with speech, facial expressions and eye motion, in order to approach realistic human face behavior. In addition, our model provides the ground truth of a set of facial feature points at each frame. As result, we are able to build a video database of synthetic human faces with ground truth, which can be used for training/evaluation of several algorithms for tracking and/or detection. We also present experiments using our generated videos to evaluate face, eye and mouth detection algorithms, comparing their performance with real video sequences.**

***Keywords*-Computer Vision; Ground Truth; Computer Animation.**

I. Introduction

Nowadays, the research on Computer Vision (CV) algorithms to detect, track and recognize faces and/or its attributes has been widely expressive due the possibility of their application in several fields, such as security, automatic photography and robotics [13]. Because of its non-rigidity and complex three-dimensional (3D) structure, the appearance of a face is affected by a number of factors including identity, face pose, illumination, facial expression, age, occlusion, and facial hair [5].

The development of robust CV algorithms needs databases that include varied and accurate data for their evaluation. Currently there are several databases for testing of different types of CV algorithms, but most of them do not include ground truth data, which define the actual position of each facial component. When a database is constituted of real images, the generation of ground truth usually requires the manual specification of face components, which is a time-consuming and error-prone task. When dealing with video, this is even more demanding: a few seconds means to have hundreds or thousands of images that should be manually segmented. Furthermore, all manual specification is also user-dependent: two persons analyzing the same scene may (and probably will) produce different ground truth data [7].

An approach that tries to fulfill the need for accurate and automatic ground truth generation is the extraction of the desired ground truth information from three dimensional (3D) Computer Graphics (CG) scenes. In the CG scenario, it is possible to obtain the exact position/shape of every single object in the scene, allowing a quantitative comparison of different algorithms [8] [14]. The use of synthetic images allows ground truth data to be readily available, but it also introduces the question of how realistic such images must be in order to be usable to evaluate CV algorithms. This question still is a topic of research. Thinking of simulation algorithms used to aid computer vision approaches, Andrade et al. [2] present a method for generating video evidence of dangerous situations in crowded scenes. The scenarios of interest are those with high safety risk such as blocked exit, collapse of a person in the crowd, and escape panic. Real visual evidence for these scenarios is rare or unsafe to reproduce in a controllable way. Thus, there is a need for simulation to allow training and validation of computer vision systems applied to crowd monitoring.

In this context, we present a methodology for generation of facial ground truth with easily animatable synthetic faces. We present a prototype in which we can generate facial animation videos with 3D MPEG-4 parameterized face models, controlling face actions, scene lights and camera position. The facial animation platform allows to generate animations with speech, facial expressions and eye motion, in order to approach realistic human behavior. In addition, our model provides ground truth of a set of facial feature points for each frame. Hence we are able to build a video database of synthetic human faces with ground truth, which can be used for the evaluation of several algorithms of feature detection and/or tracking. This paper also compares face and facial features detection results obtained from synthetic and real databases (including ground truth), in similar conditions. We expect that such obtained results should be coherent, with respect to the accuracy of the computer vision algorithm under evaluation, meaning that this approach is valuable and can further be used to provide images database.

The next section presents a review of the research related to 3D facial databases and the generation of ground truth using CG data. We made a review of the current synthetic face databases available for research, and found out that most of them do not include ground truth data. Thus we can say that our methodology, proposed in Section III is our main contribution, as the model allows the generation of a variety of image sequences (animations) with automatic ground truth. We also present in Section IV the results of testing some of our videos with known face, eye and mouth detection algorithms, in order to verify if our generated data can be used to evaluate CV algorithms, through the comparison of the performance of the algorithm with real videos. Finally, in Section V we discuss some considerations about the entire process, pointing for future work to improve the method.

II. Related Work

Recent research has produced facial databases for the purpose of CV evaluation. Gross [5] presents a review containing 27 public available databases for face and their components detection. The Face Recognition Homepage[1] also provides a large list of databases and a brief description about each one. However, few of these databases offer ground truth data. FERET[2] and BioID[3] are examples of real image databases that provide ground truth, which was manually included.

We investigated the current main synthetic face databases, in order to verify if (and how) they provide ground truth data for feature detection/tracking algorithms. There are several synthetic facial databases for different purposes of evaluation (such as biometric algorithms or face and facial expressions recognition) as listed in [1], [5]. Most of them do not explicitly include ground truth data, providing 3D models (such as the Extended M2VTS [4], GavaDB [5] and BU-3DFE [6] databases) as well some pose images. Concerning image sequence databases, the BU-4DFE (3D+time version) database offers sequential meshes and images of scanned individuals performing facial expressions. However, due to the expensive cost of storage, this database is small. A summary of the main features of these databases is shown in Table I. The last line shows the features of our proposed database (VHuF Database), which has two advantages over most of the others: it includes animation and ground truth, reinforcing the main contribution of this work.

Hu *et al.* [6] propose to build a large scale 3D face database with dense correspondence for variant face analysis research purposes. “Dense correspondence” means that the key facial points with semantic meanings are carefully labeled and aligned among different faces, which can be used for a broad range of face analysis tasks. The goal is to obtain a face database that provides ground truth for computational face related tasks such as face detection, tracking, recognition and animation. However, this database is not currently available.

The work of Woodward *et al.* [16] presents a methodology for facial ground truth generation for the purpose of evaluation of stereo algorithms. Ground truth data is the disparity maps generated from a 3D face surface by an accurate process of data acquisition. The paper focuses on the experimental evaluation of some stereo algorithms with the ground truth disparity map generated from one face.

Concerning tools for human ground truth generation, the OVVV system [14] provides a virtual environment to design and evaluate surveillance applications. The tool is capable of simulating multiple synchronized video streams from a variety of camera configurations in a virtual environment populated with virtual humans and vehicles.

Musse *et al.* [8] describe a method to generate synthetic data and ground truth of a 3D environment inhabited by virtual humans. The work focuses on the evaluation of human tracking algorithms, and allows the generation of a variety of data through the insertion of an arbitrary number of animated virtual humans, controlled by crowd simulation algorithms. Also, different illumination conditions can be created, generating situations that typically degrade the performance of tracking algorithms (such as shadows and noise).

In this context, our approach allows the construction of a facial database (the Virtual Human Faces Database – VHuF) that has two main advantages over most of the others: it includes the flexibility of computer animation presented through an interactive tool and the associated ground truth. Following the main ideas described in Musse *et al.* [8], we propose a methodology focused on the automatic generation of CG database and ground truth of facial components, described in details in the next section.

[1] http://www.face-rec.org/databases/

[2] http://www.itl.nist.gov/iad/humanid/feret/

[3] http://www.bioid.com/downloads/facedb/index.php

[4] http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

[5] http://gavab.escet.urjc.es/recursos_en.html

[6] http://www.cs.binghamton.edu/˜lijun/Research/3DFE

Table I
A REVIEW OF THE CURRENT DATABASES OF SYNTHETIC FACES.

Database	3D Model	Type	Texture image	Images	Animation	Ground Truth
MIT-CBCL Face Recognition Database	No	–	–	Yes	No	No
Extended M2VTS Database	Yes	Mesh	Yes	No	No	No
3D_RMA database	Yes	Cloud of Points	No	No	No	No
GavabDB	Yes	Mesh	No	Yes	No	No
FRAV3D	Yes	Mesh	Yes	Yes	No	No
Max Planck Institute for Biological Cybernetics Face Database	Yes	Range image	Yes	Yes	No	No
3D face database of York University	Yes	Range image	No	No	No	No
Notre Dame Biometric Dataset	Yes	Range image	Yes	Yes	No	No
BU-3DFE Database	Yes	Mesh	Yes	No	No	Yes
BU-4DFE (3D + time)	Yes	Sequential Meshes	Yes	Yes	Yes	No
VHuF Database	**Yes**	MPEG-4 parameterized mesh	**Yes**	**Yes**	**Yes**	**Yes**

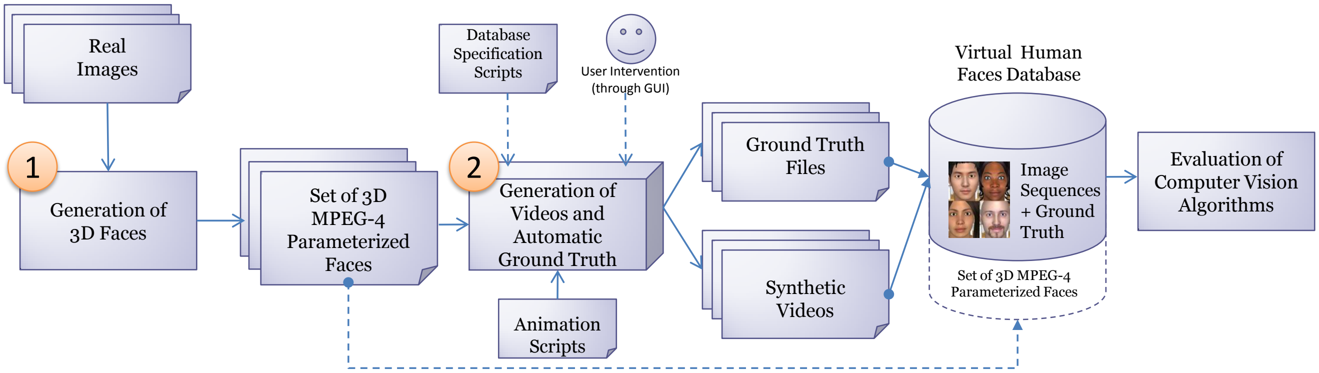

Real Images
1
Generation of 3D Faces
Set of 3D MPEG-4 Parameterized Faces
Database Specification Scripts
User Intervention (through GUI)
2
Generation of Videos and Automatic Ground Truth
Animation Scripts
Ground Truth Files
Synthetic Videos
Virtual Human Faces Database
Image Sequences + Ground Truth
Set of 3D MPEG-4 Parameterized Faces
Evaluation of Computer Vision Algorithms

Figure 2. Diagram illustrating our facial ground truth generation workflow.

III. METHODOLOGY

This section presents our model. Figure 2 illustrates our workflow.

First, we create 3D face meshes (1) according to the process described in subsection III-A. These faces are parameterized following the MPEG-4 Facial Animation standard [9].

Our prototype (2) allows us to load a single face or use a database specification script, which loads a sequence of faces from the database. The user can interact through the graphics user interface, setting lighting and camera parameters. Here it is possible to load and play an animation script (as described in subsection III-B). As the faces are parameterized, the same animation can be played by different faces. The database specification script can also indicate a sequence of animation scripts to be loaded, producing rich and varied animation/ground truth.

Output from the prototype is in the form of ground truth files (see subsection III-B) and the corresponding synthetic videos, composing VHuF (Virtual Human Faces) database. The data can then be used to evaluate CV algorithms. Alternatively, the prototype also allows interactive visual analysis, as the CV algorithms can be directly integrated into it (i.e. directly using the screen rendered image).

The following sections describe in details the two main processes of the workflow, and result in a prototype that automatically generates facial video sequences with ground truth.

A. Generation of 3D Faces

Our 3D face models were generated in the *FaceGen Modeller*[7], and the eyes, body and hair were modeled by artists. *FaceGen* uses a technique called *morphable model* [3] which

[7] Singular Inversions – http://www.facegen.com/modeller.htm

transforms a generic face model according to attributes extracted from a photo, or simply by the combination of facial attributes and texture parameters in its graphical interface. All faces (without eyes and hair) have the same number of vertices, which is an important issue for our methodology, because we use the vertex correspondence among the 3D models to provide the feature points from which the ground truth is generated. Furthermore, we can provide display of more than one face in the same frame and also our prototype is not limited to visualize only faces, since it can import a human body model as well.

When building the faces, some issues were considered:

- the 3D eye model has the eyeball, iris and pupil modeled as separated objects (i.e. not just texture features), as we want to easily extract the geometric information of these attributes. The pupils also have one vertex positioned exactly in their center, which allows us to simply extract the center of pupil by its vertex index;
- most of our face models are a reconstruction of real people. We opted for this to ensure variety in geometry and textures compatible with the reality. Our goal is to produce faces representing different ethnicity and ages. Figure 3 illustrates some models from our database.
- the quality of the textures depends directly of the photo used in reconstruction. At this stage, we decided to use photos with different resolutions (from webcam images to high resolution stills from digital cameras), in order to simulate different levels of detail in the face generation. Figure 4 shows some of our textures.

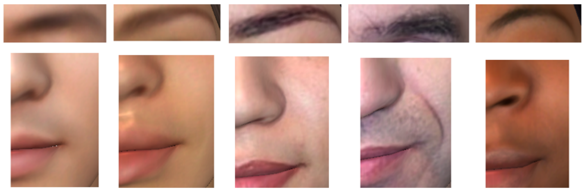


Figure 4. A close-up of some of our face skin textures. The level of detail is dependent of the quality of the photo used for generating the mesh.

B. Generation of the Facial Animation Videos with Ground Truth

Facial animation platform is implemented based on the facial animation framework described in details by [11]. The framework follows the MPEG-4 Facial Animation (FA) [9] standard for parameterization of face and animation. The animations are described as high level facial actions in a scripting language called FDL (Facial Description Language). In the current stage of our research, the FDL files can describe a sequence of three types of high-level face actions: speech, facial expressions and eye behaviors.

To generate animations, our prototype 5 receives as input a FDL file containing the description of one or more face actions. Then it interprets these actions and generates the corresponding animation in the MPEG-4 facial animation format, producing what is called a FAP file. Once a FAP file is generated, it can be used in different 3D faces, which perform the desired animations. The Facial Description Language is described in more detail in [10].

The database videos can be generated by two ways:

- directly by the manipulation of the prototype interface, where the user can manually load FDL/FAP files, the desired 3D models and adjust background, light and camera conditions; and
- by the generation and loading of a script with random scene conditions for the various 3D models of the database. In this mode, the user can opt to generate an entire sequence of images (all the animation frames) or just a few selected images corresponding to some frames of the animation in batch mode.

The user interface allows the user to interactively generate the image sequences with the desired scene configurations. It allows the selection of face models, animation scripts or the scripts for the batch mode generation. Figure 1 shows a sample of images generated by our prototype, after running a script file in batch mode. It also offers the manipulation of lights (color, position and intensity) and camera parameters (position, orientation and viewing angle), as well as the background image.

When the user chooses to record a video sequence, the prototype saves all frames in a directory and generates a ground truth file that currently contains the position (in pixels) of the feature points in each frame. In addition, we provide the horizontal gaze shift (rotation angle) that accurately provide data for evaluation of algorithms that detect the gaze direction. The scheme of the ground truth file is shown in Figure 6, considering a video sequence of n frames.

The framework is also integrated with the OpenCV[8] library. Hence it can be extended to provide the interactive visualization of CV algorithms. In order to simulate real camera conditions, we also apply camera noise to the images, as proposed in [8]. An important factor to be considering when evaluating computer vision algorithms is the degradation caused by sensor noise. In our approach, there is a simple routine to capture the noise produced by the user camera, which is inserted into the synthetic video sequence to improve the realism of the scene.

IV. Evaluation of Computer Vision Algorithms

In this section we present the performed evaluation of VHuF and BioID [9] databases. Castrillon-Santana et al. [12] discusses that since the publication of Viola-Jones work [15], an increasing number of applications have been proposed,

[8] http://www.opencv.org

[9] http://www.bioid.com/downloads/facedb/index.php

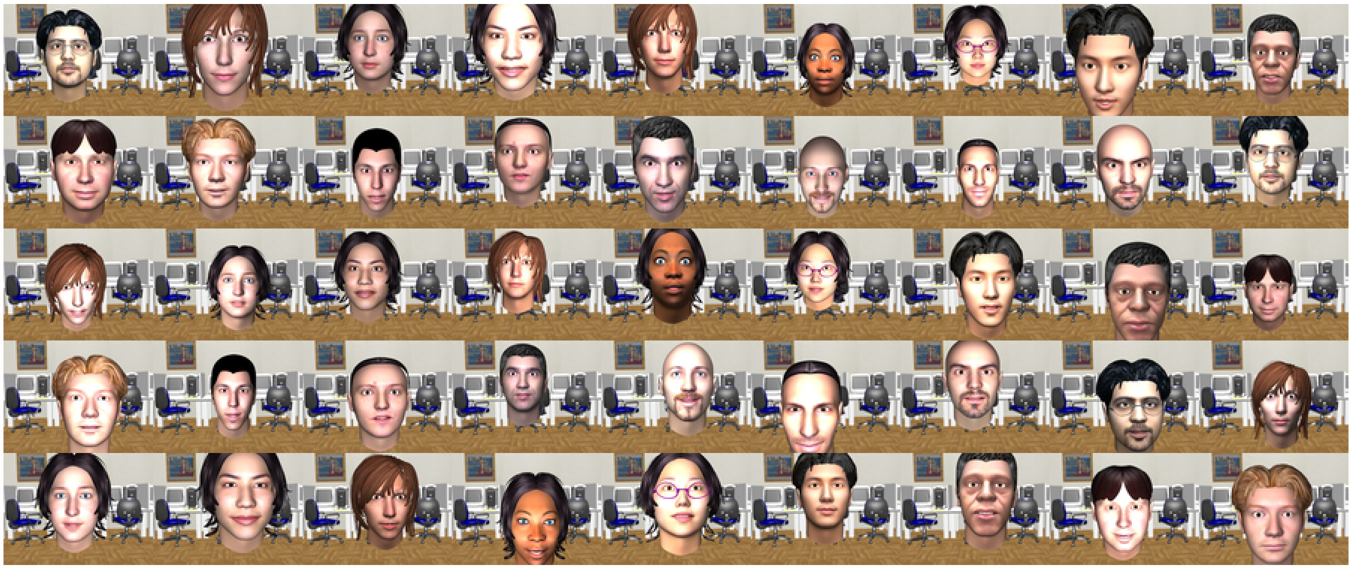
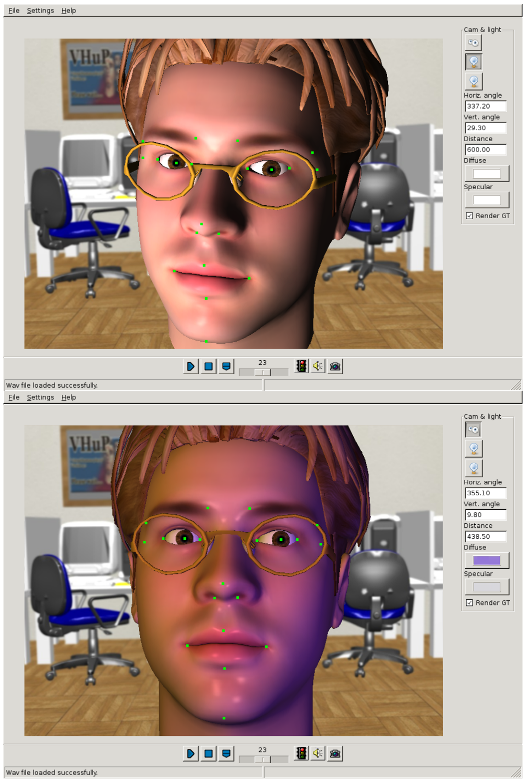


Figure 3. A sample of 3D models from the database.



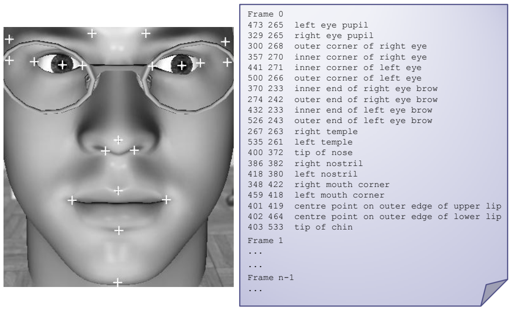
Horiz. angle 337.20
Vert. angle 29.30
Distance 600.00
Diffuse
Specular
☑ Render GT

Wav file loaded successfully.

Horiz. angle 355.10
Vert. angle 9.80
Distance 438.50
Diffuse
Specular
☑ Render GT

Wav file loaded successfully.

Figure 5. Two screenshots of the interface, showing the use of different light and camera positions. Ground truth data can be seen as square dots on top of the model.



```
Frame 0
473 265  left eye pupil
329 265  right eye pupil
300 268  outer corner of right eye
357 270  inner corner of right eye
441 271  inner corner of left eye
500 266  outer corner of left eye
370 233  inner end of right eye brow
274 242  outer end of right eye brow
432 233  inner end of left eye brow
526 243  outer end of left eye brow
267 263  right temple
535 261  left temple
400 372  tip of nose
386 382  right nostril
418 380  left nostril
348 422  right mouth corner
459 418  left mouth corner
401 419  centre point on outer edge of upper lip
402 464  centre point on outer edge of lower lip
403 533  tip of chin
Frame 1
...
...
Frame n-1
...
```

Figure 6. Feature points and an example of output ground truth file.

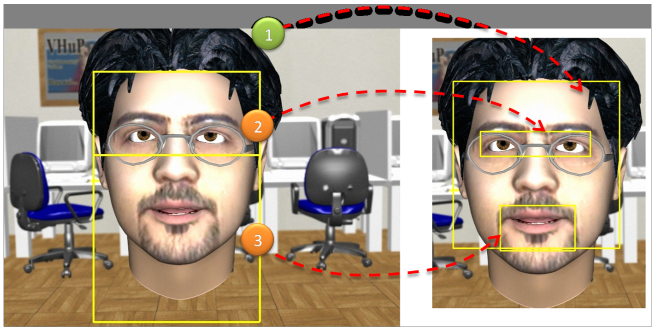


Figure 7. Example of face, eye pair and mouth detection on the VHuF database. Left: image regions sent as input for the face (1), eye pairs (2) and mouth (3) detectors; right: detection results.

mainly in the context of facial processing. Consequently, the OpenCV community shares a collection of public domain classifiers. Hence authors present individual performance of several public classifiers, which can be useful to define a baseline for other approaches.

In this work, we have chosen CV methods that can be used to compare their performance in both CG and real databases, in order to show that synthetic images can be used as automatic ground truth to evaluate CV algorithms. We use the Viola-Jones method [15] to detect faces, and the method of Castrillon-Santana [4] to detect eye pair and mouths. Basically we run the face, eye pair and mouth detectors in 3 different regions, illustrated in Figure 7 (on the left):

1) for the face detector, we use the whole image;
2) for the eye detector, we use the upper half area of the detected face;
3) for the mouth detector we use the lower half area increased by the face radius (this is needed as sometimes the mouth is close to the bottom face edge).

The algorithms [4], [15] return a rectangular region containing the detected feature (Figure 7, on the right). As the ground truth data supplied by both databases are points, we consider success in detection when the points (see Figure 6) of each feature belong to the respective region (Figure 7, right), as following specified:

- for the eyes, if the both pupils are within the eyes region;
- for the mouth, if the four mouth points are within the mouth region; and
- for the face, if eye and mouth points are within the face region.

Table II
95% CONFIDENCE INTERVAL OF SUCCESS IN FEATURES DETECTION.

Feature	BioId	VHuF
Faces	$(96.12 \pm 0.97)\%$	$(96.54 \pm 0.97)\%$
Eye Pair	$(90.56 \pm 1.50)\%$	$(89.34 \pm 1.67)\%$
Mouths	$(87.96 \pm 1.67)\%$	$(86.89 \pm 1.74)\%$

Table II shows the confidence interval of the percentage of success obtained, where the percentage of detection for faces, eyes and mouth on BioId and VHuF datasets are quite similar. The 95% achieved confidence intervals indicate that the CG dataset ground truth information of these features can be used to evaluate CV algorithms as well as the real image dataset ground thuth can be used for the same purposes.

Other datasets were also tested, however we have chosen BioId for comparison, since it presents a great diversity of images.

V. Final Remarks

This work presented a methodology for generation of facial ground truth with synthetic faces. We developed a prototype in which we can generate facial animation videos with 3D MPEG-4 parameterized face models, controlling face actions, illumination and camera position. The facial animation platform allows to generate animations with speech, facial expressions and eye motion, in order to approach realistic human behavior. In addition, our model provides the ground truth of a set of facial feature points at each frame. As result, we are able to build a video database of synthetic human faces with ground truth, that can be used for the evaluation of several algorithms of feature tracking and/or detection.

The prototype can also be considered a framework for the generation of other kinds of face features, and allows the integration and interactive visualization of CV algorithms, changing parameters while the algorithm runs. It is an advantage that we can obtain only (or at least in an easier way) using CG.

Generally speaking, we can say that the proposed methodology is a valid contribution in the sense of it is one of the pioneer efforts to explore the use of CG faces to evaluate CV tracking algorithms.

Empirically, we can cite some factors that should contribute to improve the realism of our faces and videos, which in turn could help to generate a better database. For instance, we could improve the rendering of the face and eyes, in general. The eyes rendered by our project are opaque hence they do not exhibit natural properties of the human eye, such as reflection and refraction of the lens. Another improvement would be to enhance the presentation of eyeglasses - they also exhibit refraction, which usually impacts in the results of CV algorithms.

The described method can be applied to test/evaluate CV algorithms focused on many applications. In particular, we are interested on improving the methodologies for simulating human facial behavior as realistic as possible.

Acknowledgements

This work was developed in cooperation with Hewlett-Packard Brasil Ltda, using incentives of Brazilian Informatics Law (Law n 8.248 of 1991). Thanks to the people who provided their pictures for face reconstruction.

References

[1] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2d and 3d face recognition: A survey. *Pattern Recogn. Lett.*, 28(14):1885–1906, 2007.

[2] E. L. Andrade and R. B. Fisher. Simulation of crowd problems for computer vision. *In: First International Workshop on Crowd Simulation*, 2005.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[4] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández. Encara2: Real-time detection of multiple faces at different resolutions in video streams. *J. Vis. Comun. Image Represent.*, 18(2):130–140, 2007.

[5] R. Gross. Face databases. In *Handbook of Facial Recognition*, pages 301–327, New York, NY, USA, 2005. Springer.

[6] Y. Hu, Z. Zhang, X. Xu, Y. Fu, and T. Huang. Building large scale 3d face database for face analysis. pages 343–350. 2007.

[7] T. List, J. Bins, J. Vazquez, and R. B. Fisher. Performance evaluating the evaluator. In *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 129–136, Washington, DC, USA, 2005. IEEE Computer Society.

[8] S. R. Musse, R. Rodrigues, M. Paravisi, J. C. S. Jacques. Junior, and C. R. Jung. Using synthetic ground truth data to evaluate computer vision techniques. In *IEEE Workshop on Performance Evaluation of Tracking Systems (in conjunction with ICCV 07)*, pages 25–32, 2007.

[9] I. S. Pandzic and R. Forchheimer, editors. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. John Wiley & Sons, Inc., New York, NY, USA, 2003.

[10] R. B. Queiroz, M. Cohen, and S. R. Musse. A facial animation interactive framework with facial expressions, lip synchronization and eye behavior. In *SBGames 2008*, 2008.

[11] R. B. Queiroz, M. Cohen, and S. R. Musse. An Extensible Framework for Interactive Facial Animation with Facial Expressions, Lip Synchronization and Eye Behavior. In *ACM Computers in Entertainment (CIE)*, Volume 7, Issue 4, December 2009, New York, NY, USA.

[12] M. C. Santana, O. Déniz-Suárez, L. Antón-Canalís, and J. Lorenzo-Navarro. Face and facial feature detection evaluation - performance evaluation of public domain haar detectors for face and facial feature detection. In *VISAPP (2)*, pages 167–172, 2008.

[13] K. Song and C. Chlen. Visual tracking of a moving person for a home robot. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 219(4):259–269, 2005.

[14] G. R. Taylor, A. J. Chosak, and P. C. Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[15] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[16] A. Woodward, P. Leclercq, P. Delmas, and G. Gimel'farb. Generation of an accurate facial ground truth for stereo algorithm evaluation. In *Computer Vision and Graphics*, pages 534–539, 2006. Springer Netherlands.