# Video forensics based on expression dynamics

Duc-Tien Dang-Nguyen[1], Valentina Conotter[2], Giulia Boato[2], Francesco G. B. De Natale[2]

[1]DIEE - University of Cagliari, Italy
ductien.dangnguyen@diee.unica.it
[2]DISI - University of Trento, Italy
{conotter, boato}@disi.unitn.it, denatale@ing.unitn.it

***Abstract*—Digital graphics tools are nowadays capable of rendering highly photorealistic imagery, which easily puzzle our perception of reality. This poses serious ethical and legal issues, which in turn create the need for further technologies able to ensure the trustworthiness of digital media as a true representation of reality, especially when depicting humans. In this work, we propose a novel forensic technique to tackle the problem of distinguishing computer generated (CG) from real humans in videos. It exploits the temporal information inherent of a video sequence by analyzing the spatio-temporal appearance of facial expressions in both CG and real humans. Even if rendering facial expression has reached outstanding performances, CG face appearance over time still presents some underlying mechanical properties that greatly differ from the natural muscle movements of real humans. We build an efficient classifier on a set of features describing facial dynamics and spatio-temporal changes during smiling to distinguish CG from human faces. Experimental results demonstrate the effectiveness of the proposed approach.**

## I. Introduction

In November 2013, Sweetie, a 10-year-old Filipina girl, opens a profile in a video chat room and is immediately contacted by thousands of people for sex act online. But Sweetie is not a real baby, she is just an incredibly life-like computer generated (CG) girl set up by a Dutch child protection group to identify network pedophiles (Fig. 1) [1]. A happy ending story, which poses serious doubts about our ability to distinguish between real and synthetic media, wondering if the line between physical and virtual worlds is well traced anymore.

As this, and numerous similar examples illustrate, the tremendous advances in computer graphics, supported by powerful computing facilities, promote the easy creation and fast distribution across the Web of stunning realistic CG contents, often indistinguishable from real imagery at naked eye, thus easily puzzling our perception of reality. Since these technologies continue to evolve, complex ethical and legal issues arise, posing significant challenges for the trustworthiness of digital media as a true representation of reality, especially when depicting humans.

Over the last decade, digital forensics has rapidly emerged as an effective solution to cope with these problematic issues, proposing a plethora of valuable solutions for recognizing CG imagery. State-of-the-art techniques exploited statistical models based on wavelet statistics [2], color compatibility [3], low- to mid-level statistics based on Markov process model [4], and texture features [5]. Based on the assumption that different physical pipelines are followed for generating CG and photographic imagery, several techniques have been proposed, exploiting the noise introduced by the recording camera [6], traces of demosaicing [7], chromatic aberration [8] and traces of camera image processing such as gamma correction [9]. A hybrid classifier based on wavelet features and pattern noise statistics has been proposed in [10]. Recently, few forensics methods for recognizing CG humans have been proposed, based on face asymmetry [11] and repetitive pattern of facial expressions in [12].

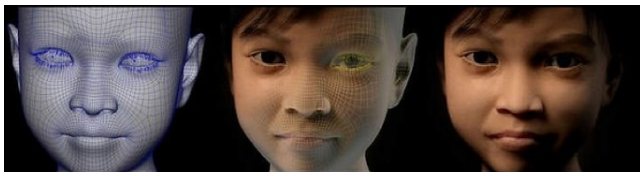

Fig. 1: Sweetie, the stunning realistic avatar created by 'Terre des Hommes Netherlands', a Dutch service that fights for child rights. [image courtesy of 'Terre des Hommes Netherlands'].

Despite their valuable contribution, almost all previous techniques are focused to work only on static images. Even if a video could be thought as a succession of independent still images, the application of image-based forensics technique would be complex and time consuming and would not exploit the rich multi-dimensional information of a video. The method proposed in [12] analyses also video sequences, but the time evolution of faces is not taken into account and multiple instances of the same facial expression have to be present in the video sequence. Recently, two forensic methods have been proposed to exploit the multi-dimensional information of a video for recognizing CG characters, but they do not focus on the analysis of facial expression dynamics [13][14].

In this work we propose a novel forensic technique that takes into account the temporal dimension of a video for discriminating between CG and real human characters. The key idea is that facial expressions in CG characters present intrinsic mechanical properties in terms of motion over time that greatly differ from the natural facial muscle movements

*WIFS'2014, Deecember 3-5, 2014, Atlanta, Georgia.*
*ISBN 978-1-4673-5593-3* *©2014 IEEE.*

responsible for real characters expressions. In order to detect such differences, we extract a set of characteristic features describing facial dynamics and spatio-temporal appearance [15] over smile expression (just a single expression is required). Such features are then used to train a Support Vector Machine (SVM) [16], building a classifier able be successfully distinguish between real and CG characters.

The structure of the paper is the following: in Section II we present the proposed forensic technique, describing in details the set of features used to feed the classifier for discriminating between CG and real characters, while in Section III we present the extensive experimental analysis. Finally in Section IV some concluding remarks are drawn.

## II. Proposed approach

In order to discriminate between CG and real humans in video, we propose a novel forensic technique based on the spatio-temporal evolution of facial expressions. The key idea behind this work is that, exploiting the rich temporal data inherent to video, the variation of human dynamics over time can be modeled: deviations from a smooth and spontaneous temporal evolution, typical of natural human beings, will be taken as a clue to recognize CG motion. In particular, facial expressions in real characters represent an intrinsic fingerprint of each individual, being a combination of voluntary and involuntary muscle movements. On the other hand, even if technologies for rendering facial expression has reached outstanding performances, natural human facial dynamics are difficult to be perfectly rendered in CG synthesis, where subtle visual and temporal artifacts may be present due to limited geometric and physical model accuracy. For instance, the variation among multiple instances of the same facial expressions is smaller for CG characters with respect to natural human beings [12].

In this work, we propose to measure such differences in CG and real human expressions evolution (from the neutral to the expressional and back to neutral state) by exploiting dynamics and spatio-temporal features extracted from videos representing smiling persons. To this end, we leverage the set of features proposed in [15] to characterize smiling expressions over time by analyzing each temporal phase of a smile, as described in details in Section II-B. The reasons behind the selection of smile expressions are twofold. Firstly, smile expressions are likely to frequently occur in videos, since they can show many different feelings, such as happiness, embarrassment, politeness, etc. Secondly, smile detection has been largely studied and many approaches allow to automatically recognize and extract video shots of smiling people (e.g., [17]).

A schema of the proposed approach is reported in Fig. 2. Given a standard video sequence containing human or CG characters as input, we start by detecting extensive facial landmarks in the first frame using a multi-level deep neural network, and subsequently track them over the rest of video [17][18]. The tracked features are used to fit a 3D head model in order to compute the 3D motion of the head, together with the individual 3D landmarks movements. The 2D movements

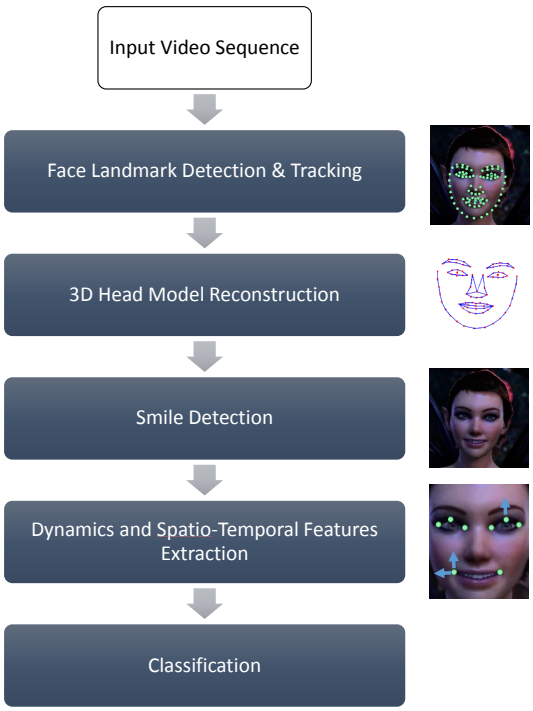


Fig. 2: Schema of the proposed method.

on the face can then be represented as a reprojection of the 3D movements onto the image plane. The tracked landmarks are further normalized with respect to rotation, translation and scale in order to have aligned points suitable for feature extraction and comparison [19]. Details about these first steps of the algorithm are discussed in Section II-A.

Since we are interested in smiling expression, we manually select the frame sequence depicting people with an enjoyment smile; as a future work, we plan to fully automatize this step by using Face++ smile analysis framework [17].

Based on the computed landmarks, we calculate facial expression dynamics on the eyelid and lip corners from each temporal phase: transition from a neutral state, going through an unfolding smile and returning to the initial state (i.e., onset, apex and offset). Besides dynamic features, we extract spatio-temporal features from the eye and lip region by employing Completed Local Binary Pattern from Three Orthogonal Planes (CLBP-TOP) to describe the spatio-temporal changes in the face appearance [15]. Finally, a gaussian kernel SVM [16] is fed with these sets of characteristic features and trained to build a classifier able to distinguish between CG and real human faces. Details about these last steps of the algorithm are discussed in Section II-B.

### A. Facial landmark detection and 3D model reconstruction

In order to be able to extract consistent dynamic and spatio-temporal features from aligned and normalized faces, we firstly have to accurately locate and track extensive facial features. To

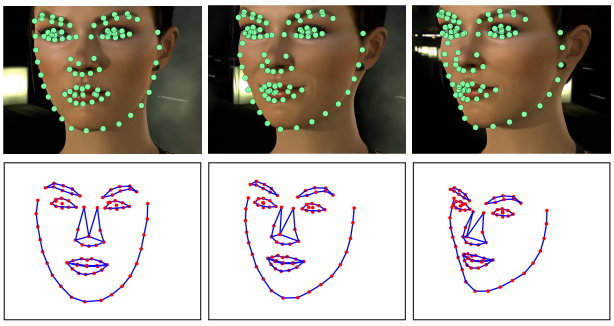

Fig. 3: Shown in the top row are 3 frames of an example video sequence, where the 83 features used for facial tracking have been superimposed; shown in the bottom row are the corresponding reconstructed 3D models.

this end, we use a recent landmark detection algorithm [17], which provides an accurate location of 83 facial keypoints, including eyebrows, eyes, nose and mouth (see top row of Fig. 3). This technology is based on a multi-level deep convolutional network, where each level is trained to locally refine a subset of the facial landmarks, found in previous levels, by applying explicit geometric constraints. We decided to exploit it, since it is able to deal with multiple head poses, complex lighting environment, complex facial expression and multi-angle faces. By using such tracked landmarks, we fit a 3D head model to the face in each frame, so recovering the projection parameters (i.e., a $3 \times 3$ rotation matrix $R$ and $1 \times 3$ translation vector $T$) that relate the 3D model to the observed face in the video frame. To this end, we employ the recent algorithm in [19] that tackles the problem of non-rigid structure from motion by using a spatial, rather than temporal, smoothing constraint. This is done by mapping each 3D point in the model to its input 2D shape in the feature space of a rotation invariant kernel. The learned mapping has the advantage that can be easily evaluated for the 3D reconstruction of any new 2D shape. Shown in the bottom row of Fig. 3 are the reconstructed 3D head models corresponding to the 3 sample frames of a video shown in the top row.

We then exploit the knowledge of the 3D head motion to model the 2D movements of a face by reprojecting the 3D landmarks movements onto the image plane. The projection parameters also allow for the alignment and normalization of faces in each frame, before the feature extraction step. In particular, in order to get a frontal view of the reconstructed 3D model, each 3D point $\ell_i$, with $i \in \{1, \ldots, 83\}$, in the model is transformed by applying the inverse of the rotation matrix $R$, i.e., $\ell_i^f = R^{-1}\ell_i$. Following, a normalization is applied, by setting the origin of the coordinate system to be the middle point between the two eye centers ($c_1$ and $c_2$, as shown Fig. 4a), and their distance from the origin is set to be equal to 100 pixels. In particular, each point $\ell_i^f$ is scaled and translated to get a normalized point, as follows:

$$\ell_i^n = \frac{100}{\rho(c_1, c_2)} \left[ \ell_i^f - \frac{c_1 + c_2}{2} \right]$$

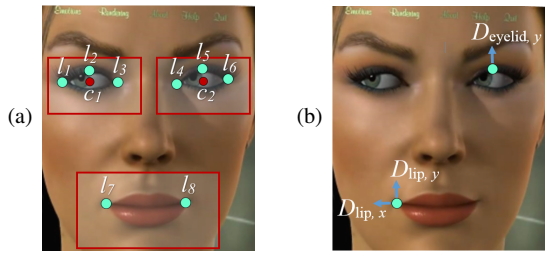


Fig. 4: Panel (a) shows the numbered facial points used in the computation of the smile dynamics, together with the three facial regions used for the spatio-temporal descriptors. Panel (b) depicts the considered facial displacement for dynamics evaluation.

where $c_1$ and $c_2$ are eye centers, $\rho$ denotes the Euclidean distance between two given points. Since the normalized face is frontal, the depth coordinate ($z$) of each point can be ignored; thus, the 2D aligned normalized landmarks will be denoted as $l_i = (x_i, y_i)$.

### B. Features extraction

The set of features selected in this work for discriminating between CG and real humans are extracted from videos representing smiling persons. According to the Facial Action Coding System [20], a smile corresponds to Action Unit 12, defined as an upward movement of the lip corners. It consists of three stages: the *onset*, where the face changes its expression from neutral to smiling state; the *apex*, which is the stable period in which the smiling expression is maintained; and the *offset*, where the face returns to the neutral state. Given the amplitude of a smile $S$ to be the Euclidean distance of the lip corners to the lip center during the smile, we can define the onset as the longest continuous increase in $S$, the offset as the longest continuous decrease and the apex as the phase between the onset and the offset. Notice that, since the faces are normalized and aligned (Section II-A), we can compute the center and the length of the lip only in the initial frame.

Given the tracked landmarks (Section II-A), we extract two sets of features from faces: dynamic features and spatio-temporal features, as in [15].

*Smile dynamics* describe the movements of the tracked facial landmarks over the expression duration. In particular, we compute the vertical displacement of the eyelid $D_{\text{eyelid},y}$ and the horizontal and vertical amplitude signals from lip corners motion $D_{\text{lip},x}$ and $D_{\text{lip},y}$, respectively, as follows:

$$\begin{aligned}
D_{\text{eyelid},y}(t) &= \frac{\frac{y_1^t + y_3^t}{2} - y_2^t}{2\rho(l_1^t, l_3^t)} + \frac{\frac{y_4^t + y_6^t}{2} - y_5^t}{2\rho(l_4^t, l_6^t)} \\
D_{\text{lip},x}(t) &= \frac{|\frac{x_7^1 + x_8^1}{2} - x_7^t| + |\frac{x_7^1 + x_8^1}{2} - x_8^t|}{2\rho(l_7^1, l_8^1)} \\
D_{\text{lip},y}(t) &= \frac{|\frac{y_7^1 + y_8^1}{2} - y_7^t| + |\frac{y_7^1 + y_8^1}{2} - y_8^t|}{2\rho(l_7^1, l_8^1)}
\end{aligned}$$

where $l_i^t$ indicates the location of the $i^{th}$ landmark point in frame $t$ (see Fig. 4b). Subsequently, this set of features is split into the 3 phases of a smile and a total of 15 corresponding

| Feature | Definition |
|---|---|
| Duration | $\left[\frac{\eta(D^+)}{\omega}, \frac{\eta(D^-)}{\omega}, \frac{\eta(D)}{\omega}\right]$ |
| Duration Ratio | $\left[\frac{\eta(D^+)}{\eta(D)}, \frac{\eta(D^-)}{\eta(D)}\right]$ |
| Maximum Amplitude | $\max(D)$ |
| Mean Amplitude | $\frac{\sum(D)}{\eta(D)}$ |
| Maximum Speed | $[\max(V^+), \max(\lvert V^-\rvert)]$ |
| Mean Speed | $\left[\frac{\sum(V^+)}{\eta(V^+)}, \frac{\sum(\lvert V^-\rvert)}{\eta(V^-)}\right]$ |
| Maximum Acceleration | $[\max(A^+), \max(\lvert A^-\rvert)]$ |
| Mean Acceleration | $\left[\frac{\sum(A^+)}{\eta(A^+)}, \frac{\sum(\lvert A^-\rvert)}{\eta(A^-)}\right]$ |

TABLE I: 15 Facial Dynamic Features extracted from a smile expression in video, where $V(t) = \frac{dD}{dt}$ and $A(t) = \frac{d^2D}{dt^2}$ denote speed and acceleration, respectively, $\eta$ defines the number of frames in the analyzed sequence and $\omega$ indicates the frame rate of the video.

characteristics are computed as in Table I, where apices ($^+$) and ($^-$) denote the onset and the offset segments of the related feature, respectively. The selected dynamic features, extracted for the vertical eye aperture and for both the vertical and horizontal amplitude of the movement of the lip corners, will sum up to a vector of dimension $3 \times 15 = 45$. Please note that, in [15] authors suggest to exploit also eyebrow and cheek displacements. However, we chose to exclude them from our set of features, primarily because the detection and tracking of eyebrow and cheek landmarks often results to be a difficult and inaccurate task (mainly due to illumination issues inherent of the examined videos or to the absence of eyebrow in some video sequences).

*Spatio-temporal local texture descriptors*, i.e., the Completed Local Binary Pattern from Three Orthogonal Planes (CLBP-TOP) [21], are used as a second set of features, in order to describe the temporal changes in the texture of a face . They are a natural extension to image sequences of the Completed Local Binary Pattern (CLBP), firstly proposed in [22]. CLBP calculates local pixel differences $d_p = g_p - g_c$, where $g_c$ is the central pixel and $g_p$ , with $p \in \{0, 1, \ldots, P-1\}$, are its $P$ circularly and equally spaced neighbors. The operators used to code the sign (S) and magnitude (M) of the difference $d_p$ and the gray level of the central pixel (C) are :

$$\begin{aligned} CLBPS_{P,R} &= \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \quad s(x) = \begin{cases} 1 & , x \geq 0 \\ 0 & , x < 0 \end{cases} \\ CLBPM_{P,R} &= \sum_{p=0}^{P-1} \tau(m_p - c)2^p, \quad \tau(x, c) = \begin{cases} 1 & , x \geq c \\ 0 & , x < c \end{cases} \\ CLBPC_{P,R} &= \tau(g_c, c_l) \end{aligned} \tag{1}$$

where $R$ is the radius of the neighboring area, $c$ indicates the mean value of the magnitude $m_p = |d_p|$ and $c_l$ stands for the mean gray level of the whole image. The final CLBP histogram is build as the concatenation of histograms in Eq. (1). Since it only uses the $x$ and $y$ dimension of an image, CLBP is purely spatial. CLBP-TOP extends such computation to the 3rd dimension of a video, i.e. time $t$, by computing the CLBP histograms on the three orthogonal planes $xy$, $xt$ and $yt$, which are then concatenated to form the final texture descriptor. In order to reduce the dimensionality of the feature space, thus decreasing the computational complexity of the algorithm, we employ the uniform CLBP-TOP histograms [21]. We extract them from three patches around the eyes and the mouth (see Fig. 4a) only during the onset phase of a smile, since this is the phase showing the change from neutral to smiling status [15]. By setting $P = 8$, the spatio-temporal texture descriptor consists in a histogram of $3 \times 354 = 1062$ bins. Note that before extracting the uniform CLBP-TOP histograms, faces need to be normalized with respect to rotation, cropped and resized by setting inter-ocular distance to 50 pixels. As a result, normalized frames are of size 125 $\times$100 pixels.

Finally, the extracted dynamic features and the spatio-temporal local texture descriptors are concatenated to form a unique feature vector of dimension 1107, which is then used to train a classifier in order to distinguish between CG and real human characters.

## III. Experimental Results

In order to verify the efficiency of the facial expression dynamics and the spatio-temporal texture descriptors for recognizing CG characters in video, we started by constructing a database of both CG and real humans smiling in video sequences. We collected from the Web 35 highly realistic CG characters performing happiness expression, a non-trivial task due to the lack of a publicly available database. On the other hand, videos depicting smiling real humans were collected from two public datasets: Boğaziçi University Head Motion Analysis Project Database (Buhmap) [23], and UvA-Nemo Smile Database (UvA) [24]. In particular, we constructed 2 different databases of real smiling humans to be tested against the CG characters: *(i)* 35 videos randomly selected out of the 440 sequences in the Buhmap database; *(ii)* 35 videos randomly collected out the 1240 sequences in the Uva database. Shown in Fig. 7 and Fig. 8 are 3 frames (representing onset, apex and offset) of 5 real and 5 CG tested videos, respectively[1].

For each test video, we calculated dynamics and spatio-temporal features, as described in Section II, and trained an SVM with RBF kernel. In our experiments, 3-fold cross-validation was used for evaluation and the SVM optimal parameters ($c$,$\gamma$) were found with the "grid-search" method. The two bar plots on the left in Fig. 5 illustrate the achieved results in terms of classification accuracy when discriminating between CG and natural characters in videos. We analyzed both the Buhmap and the UvA dataset, achieving a global accuracy of 95.7% and 97.1%, respectively (gray bars). These results demonstrate the successful nature of the proposed forensic technique in distinguishing CG characters from real

[1]In Fig. 8, from top to bottom, the videos have been downloaded from youtu.be/z0cZin2xDmQ, youtu.be/-wtv4bsLWvw, youtu.be/CvaGd4KqlvQ, youtu.be/o˙7CfWlkqm8, and youtu.be/˙v8I8NJ6Oyc

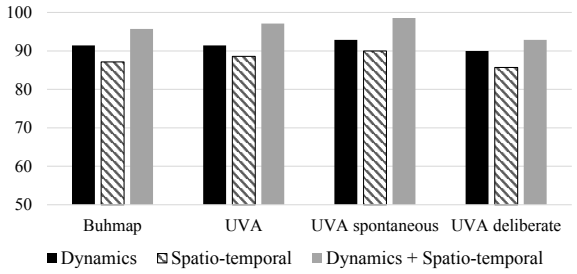


Fig. 5: Evaluation results on different datasets.

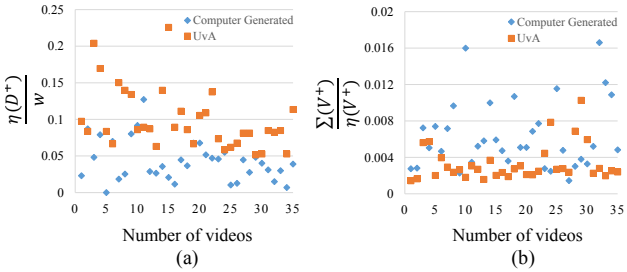


Fig. 6: Plot of two extracted dynamic features over the entire CG and UvA video databases: in panel (a) the duration features extracted from the eyelid displacements, while in panel (b) the mean speed of horizontal component of the lips, both collected on onset phase.

humans based on smile features, confirming that the underlying mechanical properties of a CG smiling character greatly differs from the natural muscle movements of real humans.

Since the UvA database contains both videos of spontaneous (597) and deliberate (597) smiles, we investigated these two categories separately, since posed smile may not show the same natural smooth facial movements as the spontaneous ones. To this end, we randomly selected 35 videos of spontaneously smiling people, and other 35 videos of deliberate smiles. The two bar plots on the right in Fig. 5 show the relative classification accuracy: 98.6% and 92.3% for the spontaneous videos in UvA and the posed expressions in UvA, respectively (gray bars). The slightly lower accuracy achieved for the deliberate smiles database may indicate that the CG and forced smiles share some common properties that differentiate them from a fully natural spontaneous smile.

As a last experiment, we tried to analyze the relative influence of the dynamic and the spatio-temporal features on the global performances of the proposed framework. To this end, we built two different classifiers by considering either the dynamic features or the texture descriptors. In Fig. 5 the achieved accuracy classification for each of the considered dataset is shown. Results demonstrate that the fusion of dynamics and spatio-temporal features (gray bars) allows to reach the best accuracy for CG classification. We may notice that using only dynamic features (black bars) we obtain an accuracy always higher than 90%, while performances decrease when using only the spatio-temporal (diagonal-lines texture bars), achieving an accuracy lower than 90% over all the databases. To further investigate this point, we analyzed some of the dynamics features. Fig. 6 shows the plots of two selected features (vertical axis), namely the duration feature extracted from the eyelid displacements (panel (a)) and the mean speed of horizontal component of the lips (panel (b)), both collected on onset phase, for each tested CG (◊) and UvA (□) videos (horizontal axis). It is clear that these two features are already highly discriminating, even if not self-sufficient for precise classification. We want to point out that in [15] also eyebrow and cheek displacements are exploited. However we chose to exclude them from our set of features, since such landmarks are often hard to detect and track due to either illumination issues inherent of the examined videos or to the possible absence of eyebrow in some video sequences. Anyway, we verified that the selected dynamics allow the trained classifier to reach a very high detection accuracy and demonstrated that the eyelid and lip corners displacements are sufficient, together with the spatio-temporal texture descriptors, to accurately characterize CG and real human smiles.

## IV. Conclusion

In this work, we presented a novel forensics technique for distinguishing CG from real humans in videos, by exploiting the temporal dimension of a video sequence and in particular the spatio-temporal appearance of facial expressions. The main idea behind this contribution is that, even if rendering facial expression has reached outstanding performances, facial expressions in CG characters present intrinsic mechanical properties in terms of dynamics which are very different from the natural facial muscle movements responsible for expressions in real human beings. We demonstrated that such differences in expressions evolution can be accurately detected by analyzing a set of characteristic facial dynamic features and spatio-temporal local texture descriptors. In particular, we focused on the smile expression since this is the most frequently observed and its automatic detection is pretty accurate. Nevertheless, the proposed idea could be modified to work also on all other expressions. Experimental results validated the effectiveness of the proposed approach on different datasets.

## Acknowledgment

This work was partially supported by the ICT COST Action IC1105 - 3D-ConTourNet - 3D Content Creation, Coding and Transmission over Future Media Networks.

## References

[1] BBCnews, "Computer-generated 'Sweetie' catches online predators," www.bbc.com/news/uk-24818769, November 2013, retrieved June 2014.
[2] S. Lyu and H. Farid, "How realistic is photorealistic?" *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 845–850, 2005.
[3] J.-F. Lalonde and A. A. Efros, "Using color compatibility for assessing image realism," in *IEEE Int. Conf. on Computer Vision*, 2007, pp. 1–8.
[4] P. Sutthiwan, X. Cai, Y. Shi, and H. Zhang, "Computer graphics classification based on markov process model and boosting feature selection technique," in *IEEE Int. Conf. on Image Processing*, 2009, pp. 2913–2916.
[5] Z. Li, J. Ye, and Y. Q. Shi, "Distinguishing computer graphics from photographic images using local binary patterns," in *Int. Workshop on Digital Watermarking IWDW*, 2012, pp. 228–241.

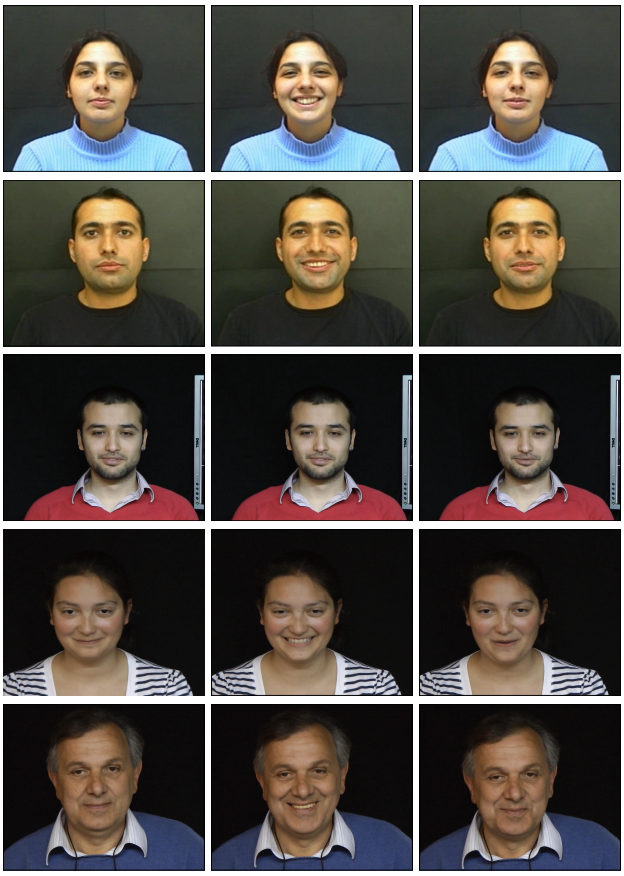

Fig. 7: Five examples of tested videos depicting real humans. The top two rows are from BUHMAP, following by two spontaneous smiles from UvA while the last row shows a deliberate one.

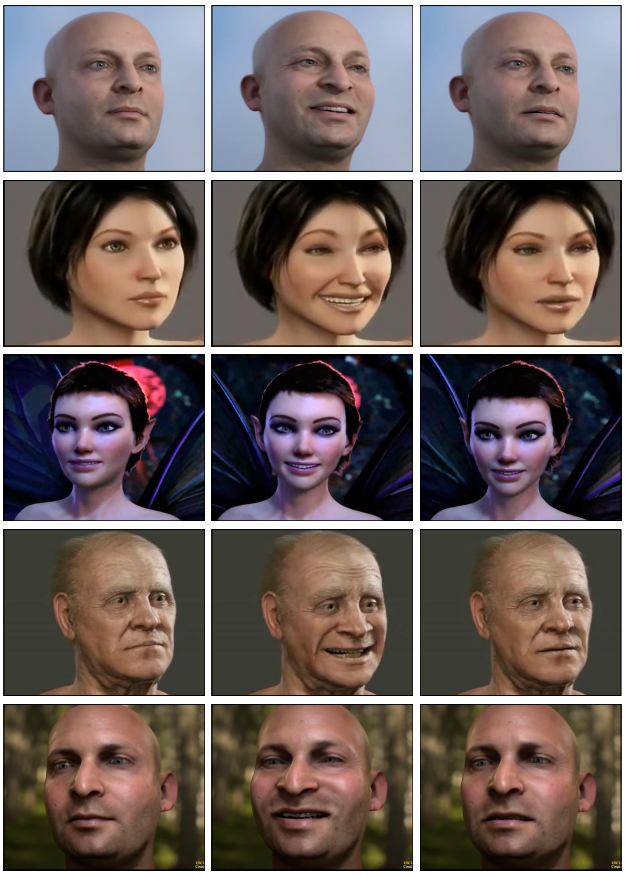

Fig. 8: Five examples of tested CG videos.


[6] N. Khanna, G. T. C. Chiu, J. P. Allebach, and E. J. Delp, "Forensic techniques for classifying scanner, computer generated and digital camera images," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2008, pp. 1653–1656.

[7] A. Dirik, S. Bayram, H. Sencar, and N. Memon, "New features to identify computer generated images," in *IEEE Int. Conf. on Image Processing*, vol. 4, 2007, pp. IV–433 – IV–436.

[8] A. Gallagher and T. Chen, "Image authentication by detecting traces of demosaicing," in *IEEE Computer Vision and Pattern Recognition Workshop*, 2008.

[9] T. T. Ng, S. F. Chang, J. Hsu, L. Xie, and M. P. Tsui, "Physics-motivated features for distinguishing photographic images and computer graphics," in *ACM Multimedia*, 2005, pp. 239–248.

[10] V. Conotter and L. Cordin, "Detecting photographic and computer generated composites," in *SPIE Symposium on Electronic Imaging*, vol. 7870, 2011.

[11] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, "Discrimination between computer generated and natural human faces based on asymmetry information," in *European Signal Processing Conf.*, 2012, pp. 1234–1238.

[12] ——, "Identify computer generated characters by analysing facial expressions variation," in *IEEE Int. Workshop on Information Forensics and Security*, 2012, pp. 252–257, (Best paper award).

[13] V. Conotter, E. Bodnari, G. Boato, and H. Farid, "Physiologically-based detection of computer generated faces in video," in *IEEE Int. Conf. on Image processing*, 2014.

[14] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, "Revealing synthetic facial animations of realistic characters," in *IEEE Int. Conf. on Image processing*, 2014.

[15] H. Dibeklioglu, A. A. Salah, and T. Gevers, "Like father, like son: Facial expression dynamics for kinship verification," *IEEE Int. Conf. on Computer Vision*, pp. 1497–1504, 2013.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at : http://www.csie.ntu.edu.tw/~cjlin/libsvm, Retrieved June 2014.

[17] MegviInc., "Face++ research toolkit," www.faceplusplus.com, 2013.

[18] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional neural network," in *ICCV workshop on 300 Faces in-the-Wild Challenge*, 2013, pp. 386–391.

[19] O. C. Hamsici, P. F. Gotardo, and A. M. Martinez, "Learning spatially-smooth mappings in non-rigid structure from motion," in *12th European Conf. on Computer Vision ECCV*, 2012, pp. 260–273.

[20] P. Ekman and W. Friesen, *The Facial Action coding System: a technique for the measurement of facial movement*. San francisco, CA, USA: Consulting Phsycologists Press Inc., 1978.

[21] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *IEEE Conf. on Computer Vision*, 2011, pp. 868–875.

[22] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," in *IEEE Transactions on Image Processing*, vol. 19, 2010, pp. 1657–1663.

[23] O. Aran, I. Ari, M. A. Güvensan, H. Haberdar, Z. Kurt, H. . Türkmen, A. Uyar, and L. Akarun, "A database of non-manual signs in Turkish sign language," in *Signal Processing and Communications Applications*, 2007, pp. 1–4.

[24] H. Dibeklioglu, A. A. Salah, and T. Gevers, "Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles," in *European Conf. on Computer Vision*, 2012, pp. 525–538.