EVA: expressive multipart virtual agent performing gestures and emotions

[1]IZIDOR MLAKAR, [2]MATEJ ROJC
[1]Roboti c.s. d.o.o, [2]Faculty of Electrical Engineering and Computer Science, University of Maribor
[1]Tržaška cesta 23, [2]Smetanova ulica 17
SLOVENIA

Abstract**—** Embodied Conversational Agents (ECAs) play an important role in the development of personalized and expressive human-machine interaction, allowing users to interact with a system over several communication channels, such as: natural speech, facial expression, and different body gestures. This paper presents a novel approach to the generation of ECAs for multimodal interfaces, by using the proprietary EVA framework. EVA's articulated 3D model is mesh-based and built on the multipart concept. Each of its 3D sub-models (body-parts) supports both bone and morph target-based animation, in order to simulate natural human movement. Each body movement's structural characteristics can be described by the composite movement of one or more elementary units (bones and/or morphs), and its temporal characteristics by the durations of each of the movement's stages (expose, present, dissipate). EVA scripts provide a means of defining and fine-tuning body motion in the form of predefined gestures, or complex behavioural events (provided by external behaviour modelling sources). Since behavioural events can also be described as a combination of tuned predefined gestures and the movements of elementary units, a small number of predefined gestures can form infinite sets of gestures that ECA can perform. ECA EVA, as presented in this paper, provides both: a personalization of its behaviour (gesture level), and a personalization of its outlook.

Keywords**—** bone and morph-based animation, distributive, expressive ECA, mesh-based articulated model.

I. Introduction

The research goal of multimodal interaction research has moved from user-friendly interfaces towards more advanced platforms that simulate human-like interaction, ranging from simple speech-embedded applications to complex applications, such as: intelligent-environment interfaces [1], e-commerce applications [2] etc. Virtual characters, more often referred to as embodied conversational agents (ECAs) [3] play central roles in such interfaces. The ECAs are embedded into multimodal human-machine interfaces as animated talking heads [4],[5], or fully-functional conversational agents [1],[6]. Most current research incorporates the input processing stage and the combined behaviour planning and realization stage. For instance, MAXINE [1] incorporates several inputs, input processing, and behavioural modelling within a programmable interfaced structure.

Similarly, most speech and facial emotion-synthesis oriented frameworks (most popular concepts within ECA context) suggest incorporating one or more input processing techniques (e.g. speech recognition for visual speech synthesis, behaviour modelling, etc.). The system for generating facial animation (including emotional speech synthesis) [7] is based on statistical modelling, and HUGE architecture [8]. A more elaborated segmentation of the interaction processes is presented in [9] and [10]. In this case the framework follows the principle of three stage based interaction architecture as described in [11]. The concept described in [12] (and [6]) presents an approach for animating expressive speech, non-speech related facial and head gestures (e.g. gaze), and more complex emotions. The descriptions of movement are specified by using Affective Presentation Mark-up Language APML [13], XML-based abstract language, and definitions for facial expressions based on MPEG-4 FAP. The AMPL-based description of movement is transversed into low level facial expression that can be stored as animation files. In [14] the authors present a novel approach to the generation of coordinated multimodal behaviour, by using the behavioural mark-up language (BML), and the low level animation language (EMBRScript).

These above-stated approaches can, in general, be described as mesh-based [15] and use morphed shapes and baked animations (predefined animation sequences) as templates for animating the desired behaviour. Animation that uses the morphed target-blending concept is very popular and is used in several speech-centred agents. The mesh-based models and morphed target animations are relatively easy to handle. In contrast to the appearance based-models [16], mesh-based models also require reasonable amounts of system resources. On the other hand, the skeleton-based (bone based) animation is seldom used for generating speech centric interfaces. Since most of the speech centric interfaces usually incorporate only the head, morphed targets are sufficient for the natural representation of finite sets of facial gestures. An example of human motion analysis and simulation using skeletal chains is presented e.g. in [17].

This paper describes the embodied conversational avatar EVA, created within a modular environment (named EVA framework), used for the development of personalised and expressive human-machine interaction systems (e.g. including facial and body gesture synthesis, audio/visual speech synthesis, etc.). The EVA framework is based on the concepts of DATA framework [18], and Panda 3D [19]. Panda 3D is used as the core of the animation engine within the EVA framework. In order to achieve believable and as natural gestures and facial animations as possible, the mesh-based animation procedures of the EVA framework support both skeletal and morph-based animations. Behaviour can, therefore, be described as a set of morphed target translations and bone movements, moving in parallel or at sequential intervals. ECA EVA is driven by EVA scripts that describe the desired body motion. These scripts are XML-based descriptions, similar to BML, that provide information about the temporal, spatial, and repetitive characteristics of human motion. EVA scripts are also optimized in order to be easily transformed into low level sequences of animation parameters. EVA scripts support the specification of human motion as a combination of predefined behaviour, and a movement description of the elementary control units. In this way, human motion can be specified manually, or by some external process used for behaviour modelling, and by systems of human action and motion recognition (as e.g. in [20]).

II. Articulated model

The Human body consists of a rigid skeleton. This skeleton is an articulated object with joints and rigid elements. When animating the human body and face, it can be assumed that by applying motion to skeletal chains, the rest of the body animation (layers, such as muscles, skin, hair, clothes) will follow accordingly [21]. Several approaches have already been proposed for realistically modelling the human body and its movement. These approaches apply translation, and orientation to joints and joint chains, apply muscle dependant volume to a body that is created by muscle systems, etc. The second part of human body modelling is movement. Applying kinematics to simulate human like movement can be a complex task, since the influences are not localized but can also be generated by different environmental conditions, and interactions. A realistic human presentation should, in the context of animation, provide synchronized limb movement, realistic skin deformations, realistic facial expression, and continuous, synchronized visual speech synthesis, etc. Additionally, a human model should also take into account the fluidity of movement, and its randomness, in order to look natural. The human modelling approaches can be divided into:

- *Stick figure models:* are models based on sets of rigid elements, and connected to joint chains. Motion is specified as a set of hierarchical transformations, controlled by joint-constrains. The stick figure models are results of the earliest studies in human animation. The studies of Korein [22] and Thalmanns [23] further explain the concept of these models.

- *Surface models (mesh-based models):* represent an upgrade of stick figure models. In this case, a polygonal mesh-layer (skin) is applied on the skeleton chains. Each joint in the skeleton chain (backbone) influences an area of the skeleton. Therefore, the skin-deformations are directly related to the movements of underlying skeleton chains. The skinning (full skinning) process allows each vertex to be influenced by one or more joints, preventing some of the unnatural movements, and allowing simulation of natural human movement (for a instance, the hand is not only influenced by the wrist-joint, but also by both the knuckle and forearm joints). Motion within these models is presented by the transition of vertices, and can be modelled either by joint rotations (Euler's angles [24]), or by direct transformations on the mesh model (morphing).

- *Volumetric models:* use simple volumetric primitives such as spheres, cylinders and ellipsoids, in order to construct the body shape. Volumetric models can perform a more realistic presentation of human movement, but are relatively hard to handle. Today such models are commonly used when handling collision. Since the base primitives of volumetric models are naturally smooth, they are also used when presenting organic forms.

- *Multilayered models (muscle-based models):* present anatomically-correct models. The animator of such models introduces different kinds of constraints to the relationship between layers. The basic layers of such models are: skeletal, muscle and skin. Muscle-based models tend to be ellipsoid, since ellipsoids are a good approximation of the muscle appearance.

A more detailed explanation of different articulated models and animation is given in [25] and [21]. The following section presents an articulated model of ECA EVA. EVA is a surface-based ECA. Although surface models may not always perform best in the context of realism, mesh-based models are widely used in the field of natural human-machine interaction. Mesh-based models are relatively easy to construct, personalize, and model. Additionally, most of the existing animation engines support animation described as a movement of vertices (e.g. OpenGL, D3D) and, therefore, most mesh-based models can be animated at interactive speed by using common hardware setups.

A. Multipart mesh-based articulated model

The articulated model of the EVA embodied conversational agent can be described as a set of independent polygonal meshes applied to skeletal chains (sets of 3D sub-models). The 3D sub model of body hosts the main skeletal chain. Each one of the other sub-models (eyes, hair, lower and upper teeth, tongue, accessories, and dress) shares one joint with the main skeletal chain (usually the first joint in the sub models chain). The multi-part actor is automatically built within the EVA framework's animation engine environment by interconnecting the sub-chains with the main skeletal chain and, subsequently extending the forward kinematics from the main joint-chain to each of the sub-chains. Each sub-chain retains influence on its corresponding mesh, and no additional direct influences from the other joint chains (attached to the main joint chain) are given to any set of the vertices. Fig. 1 presents the taxonomy of the articulated 3D model of ECA EVA.

The ECA EVA's articulated model is, as mentioned at the beginning, a set of 3D sub-models with individual skeletal chains, all connected to the main skeletal chain. Each 3D sub-model has its own polygonal mesh, UV connection map, and its own textures. The body-skeleton is usually selected as a main (base) skeletal chain (e.g. spine), and since it shares a common bone with other 3D sub-models, an attaching process can be implemented. In addition to the joints, ECA EVA also uses morphed shapes for synthesizing the desired behaviour. The main and 3D sub-models are usually modelled within a certain external 3D modelling environment (e.g. Maya3D, Blender 3D, Daz3D etc.), and then stored as an exchangeable 3D format (e.g. *X* format for use with direct 3D).

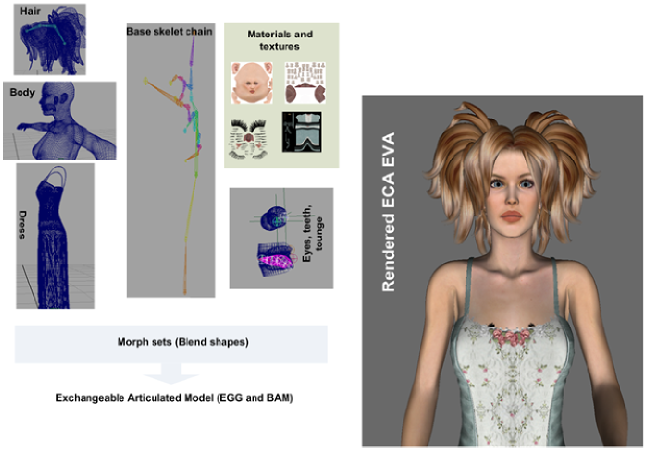


Figure 1: Taxonomy of the ECA EVA's 3D model

Since the core of EVA's framework animation engine is the Panda 3D game engine, the articulated model of ECA EVA is stored in the .egg or .bam formats (binary form of .egg). The process of connecting 3D sub-models is implemented automatically within the animation engine, and the obtained result is presented in Figure 1 (as rendered ECA EVA). The multi-part actor concept enables us to easily change and update each of the 3D models separately, as long as those joints common to the main bone-chain are specified. The process of changing the appearance can even be implemented on-line, and can serve as an important indicator of ECA EVA's personalization and mood. The EVA's articulated model was first modelled within Maya 3D, and Daz3D modelling environments. The DAZ3D was used to obtain polygonal meshes of each of the 3D sub models, and their corresponding textures whereas the model rigging process (e.g. setting up the skeletal chains, and creating morphs) was implemented within Maya 3D. The facial expression-related morphs were defined in the respective to MPEG-4 FAPs and the skeletal-chains support the HPR based 3D rotations. By using the Panda3D's native Maya to .egg exporter, the 3D sub-models were exported into Panda3D's understandable format (exchangeable articulated models). The 3D sub models were then processed, and connected within the animation engine automatically. The EVA framework's animation system and animation engine running the ECA EVA will be presented in the following section.

III. EVA's ANIMATION ENGINE

ECA EVA provides both gesture and speech-related interactive techniques. The speech-related techniques are provided by visual synthesis of a text (process similar to [26]), whereas the gesture related techniques encapsulate facial, head and hand gestures. The animation techniques define how, in the context of conversation, a control point (bone or morph) should be moved in order to present the desired body movement (e.g. what facial control points to move, and how to move them to display facial expression; or what body control points to move, and how to move them to display some pointing-gestures, etc.). Fig. 2 presents the general architecture of ECA EVA's animation engine.

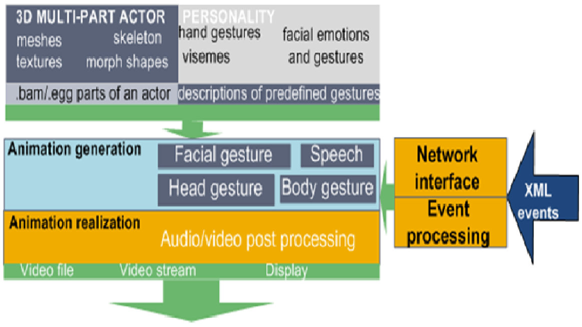
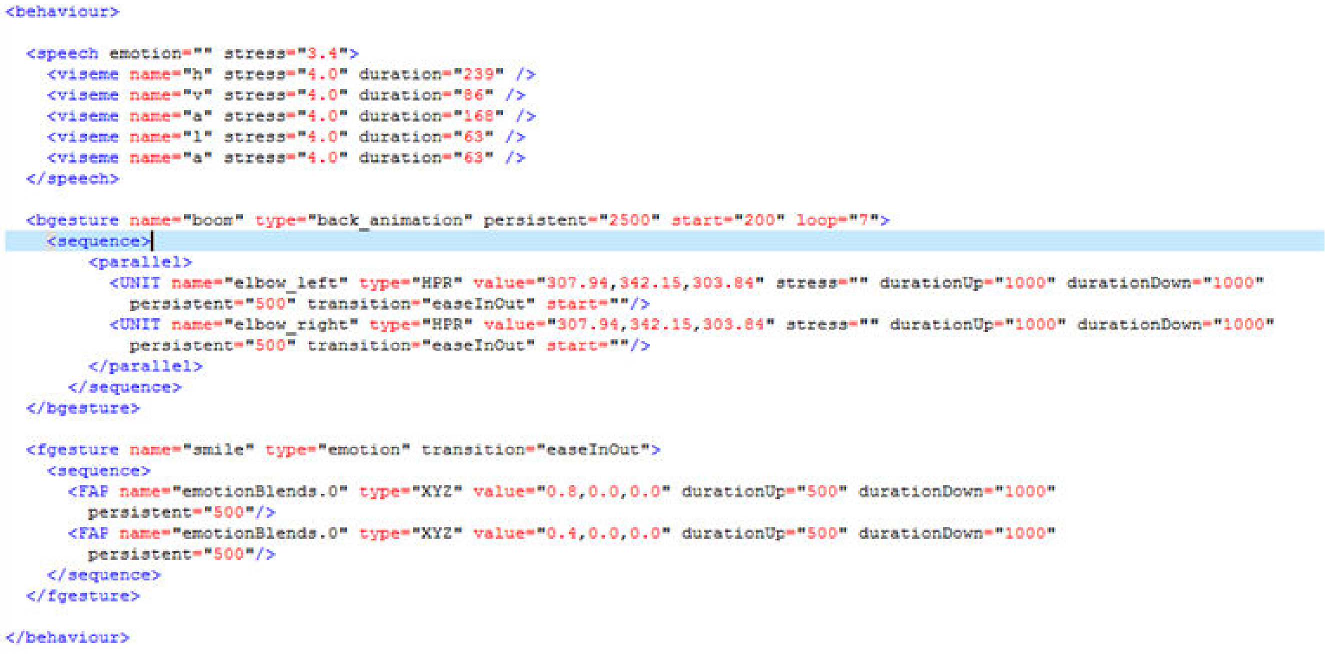


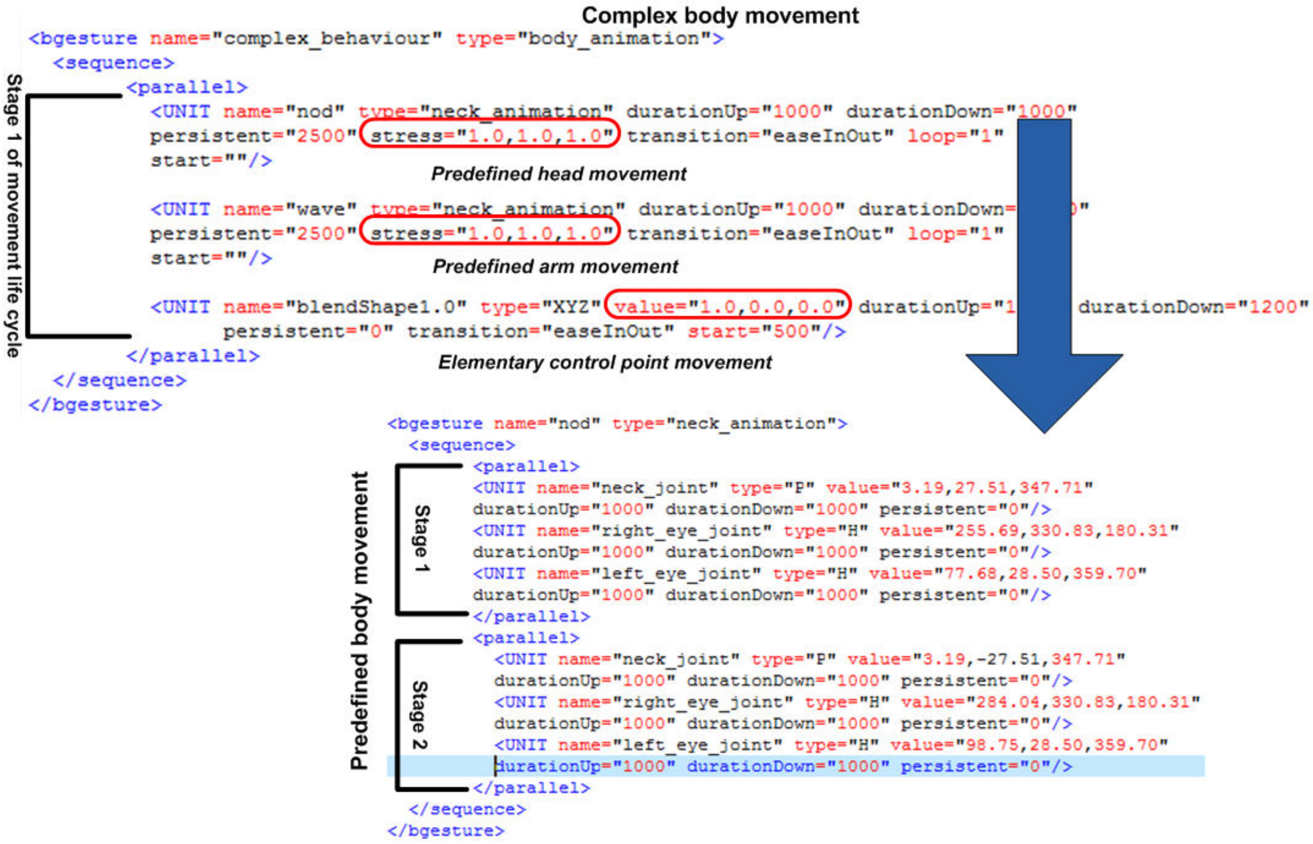
Figure 2: General architecture of ECA EVA's animation engine

When simulating human like body movements, the system assumes that each body movement can be described as an event, and that each event is generated by an external process, as shown in Figure 2. Within the current development stage of the EVA framework, we foresee two important services that are necessary when implementing advanced human-machine interaction systems: text-to-speech based service using ECA EVA (in our systems we use the proprietary PLATTOS TTS system [27]), and modelling service using ECA EVA (converting between high level abstract behavioural descriptions, such as: AMPL, ALMA [28], MURML [29] etc., and intermediate XML descriptions, EVA scripts). EVA's animation engine inputs are, therefore, XML packets describing behaviour, exchangeable articulated models, and personality based descriptions (using EVA scripts' syntax) of predefined gesture concepts (body movement such as hand-wave, head-nod, etc.). EVA framework's animation engine itself is Panda 3D based, and supports the Python programming language and all native animation techniques from Panda 3D, such as: forward kinematics, particles, animation blending, etc. The animation engine separates behaviour generation (e.g. behaviour and emotion modelling), and animation-realization (from behavioural description to animation).



```
<behaviour>

  <speech emotion="" stress="3.4">
    <viseme name="h" stress="4.0" duration="239" />
    <viseme name="v" stress="4.0" duration="86" />
    <viseme name="a" stress="4.0" duration="168" />
    <viseme name="l" stress="4.0" duration="63" />
    <viseme name="a" stress="4.0" duration="63" />
  </speech>

  <bgesture name="boom" type="back_animation" persistent="2500" start="200" loop="7">
    <sequence>
        <parallel>
          <UNIT name="elbow_left" type="HPR" value="307.94,342.15,303.84" stress="" durationUp="1000" durationDown="1000"
            persistent="500" transition="easeInOut" start=""/>
          <UNIT name="elbow_right" type="HPR" value="307.94,342.15,303.84" stress="" durationUp="1000" durationDown="1000"
            persistent="500" transition="easeInOut" start=""/>
        </parallel>
      </sequence>
  </bgesture>

  <fgesture name="smile" type="emotion" transition="easeInOut">
    <sequence>
      <FAP name="emotionBlends.0" type="XYZ" value="0.8,0.0,0.0" durationUp="500" durationDown="1000"
        persistent="500"/>
      <FAP name="emotionBlends.0" type="XYZ" value="0.4,0.0,0.0" durationUp="500" durationDown="1000"
        persistent="500"/>
    </sequence>
  </fgesture>

</behaviour>
```

Figure 3: Behaviour event description by using ECA EVA’s script syntax



```
<bgesture name="complex_behaviour" type="body_animation">
  <sequence>
     <parallel>
       <UNIT name="nod" type="neck_animation" durationUp="1000" durationDown="1000"
       persistent="2500" stress="1.0,1.0,1.0" transition="easeInOut" loop="1"
       start=""/>

       <UNIT name="wave" type="neck_animation" durationUp="1000" durationDown=    "
       persistent="2500" stress="1.0,1.0,1.0" transition="easeInOut" loop="1"
       start=""/>

       <UNIT name="blendShape1.0" type="XYZ" value="1.0,0.0,0.0" durationUp="1     durationDown="1200"
             persistent="0" transition="easeInOut" start="500"/>
     </parallel>
  </sequence>
</bgesture>
```

```
<bgesture name="nod" type="neck_animation">
  <sequence>
      <parallel>
      <UNIT name="neck_joint" type="P" value="3.19,27.51,347.71"
      durationUp="1000" durationDown="1000" persistent="0"/>
      <UNIT name="right_eye_joint" type="H" value="255.69,330.83,180.31"
      durationUp="1000" durationDown="1000" persistent="0"/>
      <UNIT name="left_eye_joint" type="H" value="77.68,28.50,359.70"
      durationUp="1000" durationDown="1000" persistent="0"/>
      </parallel>
      <parallel>
        <UNIT name="neck_joint" type="P" value="3.19,-27.51,347.71"
        durationUp="1000" durationDown="1000" persistent="0"/>
        <UNIT name="right_eye_joint" type="H" value="284.04,330.83,180.31"
        durationUp="1000" durationDown="1000" persistent="0"/>
        <UNIT name="left_eye_joint" type="H" value="98.75,28.50,359.70"
        durationUp="1000" durationDown="1000" persistent="0"/>
      </parallel>
  </sequence>
</bgesture>
```

Figure 4: Describing complex body movement within EVA scripts

The animation generation layer provides the means and interfaces for animating control points described within the exchangeable articulated models. The animation engine's primary interfaces are: Network interface (implements communication with remote behaviour description service such as TTS), and Event processing interface that is used for converting intermediate XML event descriptions into low level animation parameter sets. ECA EVA's personality concept is best described as a set of person dependent predefined body movements (gestures) that ECA uses whilst interacting, and can be fine-tuned to express ECA's emotional state, mood, etc.

The realization layer of EVA's animation engine then transforms these animation parameters sets into sequences of animated movements. By using VLC open source audio/video processing libraries, the animated sequences can be further processed, and transmitted as several video streams (e.g. HTTP or RTP based), or just stored as video files. The audio/video post-processing stage also plays an essential role, when implementing an animation engine as a network-based distributive service. In such cases, the animation engine acts as a server. Since ''heavy'' processing must be transferred from the user to the service side, the multimedia based network streams are an ideal solution for presenting the synthesis of body movement to the user. The following section will present in more detail the event processing interface, and the concept of behaviour as set of events.

A. EVA scripts for describing animation

Several researchers have already studied human motion (e.g. [30], [31]) in order to detect, track, and recognize human motion and, thus, in general interpret human behaviour. These researchers have encoded different motions into categories, such as: slow/fast, pleasant/unpleasant, weak/energetic, and small/large [26]. In order to describe (model) both the structural and physical characteristics of human movements, an XML-based concept for describing human movement was used for ECA EVA's scripts. The XML-based ECA EVA's scripts are written in an intermediate, human understandable language that specifies how to connect the abstract-high level descriptions of human motion (such as AMPL, or EML), and low level animation parameters. On the basis of human motion recognition and visual motion synthesis research (including concepts, as described e.g. in [32], [17], etc.), we have defined three main groups (regions) of human motion: speech, facial gestures with emotions, and body gestures. Circumstantially, *<speech>*, *<fgesture>* and *<bgesture>* XML tags were defined in the context of ECA EVA's scripts. All three main groups of human motion are also used in the EVA script shown in Fig. 3. The structural characteristics of human motion are described by the contents of the three main XML tags, whereas the physical characteristics of body movement are described by the attributes of the three main tags, and their child tags.

Fig. 3(A) describes a part of a speech event. The speech event is defined by a viseme sequence, where each viseme is placed on the lip motion timeline (by using duration attribute). The stress attribute in the viseme sequence describes the degree of articulation (physical characteristic) for each viseme, and the name attribute indirectly describes what control points will be moved, when animating a specific viseme. Each viseme that ECA EVA can animate is defined by one or more control points. The minimal definition contains a morphed-shape of the viseme, whereas additionally defined points are based on the viseme's influence on the other mouth regions (e.g. position of lower/upper teeth, and tongue). The viseme tags within the speech XML tag are assumed to be part of a viseme sequence, and will, therefore, be animated sequentially (one after another). In addition to the viseme sequence, a speech event can also describes the facial emotion (and its corresponding degree of influence on the facial region) relating to the viseme sequence, simply by defining emotion attribute. Since emotion as defined within a speech tag cannot be fine-tuned (neither in structure, nor in its temporal component can change), it will always consume 1/3 of overall speech sequence time for each of its animation stages (stages of formation, presentation, and disappearance). ECA EVA's scripts however incorporate fully tuneable facial gestures for animating both speech and non-speech related facial gestures (Fig. 3(C)).

Body gestures (Fig. 3(B)) represent movements that relate to the head, hand, and other movements not contained within speech sequences, or facial gestures. The definition of body gesture is similar to that of facial gestures, and is specified within the *<bgesture>* XML tag. The structure of each body gesture is described by its child UNIT tags. The relationship between "UNIT" tags can, similar to the context of facial gestures, be defined as parallel or sequential. The type attribute of each unit defines the type of movement to be used in order to animate the described movement. Types "X", "Y", "Z", and "XYZ" describe translations, whereas types "H", "P", "R", and "HPR" describe animations generated by rotating the selected bone (defined by name attribute).

Fig. 3(C) shows how a facial gesture can be defined. Facial gestures present an additional information channel in human-machine interaction. Each facial gesture is defined by its type (e.g. speech or emotion related), name (e.g. smile), duration (durationUp, durationDown, and persistence), and stress level (to what extent the gesture will be displayed). Since each facial gesture can be described in terms of the elementary units used to animate it, the *<fgesture>* tags can contain a set of FAP child tags (a set of elementary units). The relationship between FAPs is defined as parallel (if the FAPs animate facial segments simultaneously), or as sequential (if the FAPs are displayed one after another). Additionally, the EVA script also allows for defining the facial gesture as a combination of different sequential and parallel movements. For instance (Fig. 3(C)), a smile can be defined as a parallel lip and jaw movement. Since most facial gestures are closely related to MPEG-4 FAP, the elementary morphed shapes of the EVA framework were generated based on MPEG-4 FAP. By utilising the powers of bone and morph based animation, the animation blending and internal synchronization process ECA EVA can form and synthesize both simple and complex gestures.

In addition to describing the body movement as a set of elementary units' movements, the ECA EVA's script also permits the description of body movement in its abstract form.

For such descriptions, ECA EVA's script defines those so-called predefined body movements (gestures) that also represents an important part of EVA's personalization, and will be addressed in the following section.

B. Predefined body movement

Predefined body movement (or predefined gestures) are ECA EVA's script based descriptions of a human movement. Similarly, as in the case of body-movement events, predefined body-movement describes the life-cycle of each movement. Each stage during a movement's life-cycle is presented as a sequential entry within the type dependent gesture tag. These entries are encapsulated either by UNIT/FAP, or parallel XML tags. All child-tags contained within each parallel tag are regarded as part of one sequence, or stage in the movement's life-cycle. The concept of describing a predefined human movement is similar to event-based descriptions of body movement. The life-cycle of each movement is contained within movement type dependent <*bgesture*>, or <*fgesture*> XML tags. The same principal of sequence/parallel concept is applied in order to provide the structural characteristics of the movement. The movement's physical characteristics are also described within elementary control point encapsulations (elementary units UNIT/FAP). The basis of the temporal and special fine tuning of body movement events can also be applied to predefined body gestures. The main differences between predefined body movement and body movement events are:

- *A predefined body gesture can only contain elementary control point encapsulations, whereas, an event can contain both elementary control point encapsulations, and predefined body gesture.*

- *Within a predefined body gesture all elementary units contain the attribute ''value" in order to describe the movement. The control units use within a body movement event use "value" attribute to describe the movement of the elementary control-points and "stress" attribute(circled attributes in Fig. 4).*

Predefined body movement is defined in order to ease the process of generating a gesture, to enable implementation of complex gestures, such as the one presented in Fig. 4, and to support ECA learning-ability in the future. Complex gestures are formed by combining different combinations of predefined body-movements and elementary units.

The movement in Figure 4 is defined as a parallel movement of predefined body gestures nod (head movement), wave (arm movement), and the elementary control point blendShape1.0 (pointing hand gesture). A more-detailed description of the formation, and animation progression described in Figure 4, is presented in the chapter Animating body movement. In short, the entire set of elementary units will move at the same time, each one within its own specified temporal characteristics. ECA EVA will move its head, right arm and, at the same time, perform a pointing-gesture. This complex movement, as presented in Figure 4, has only 1 stage in its life cycle (due to the used parallel container). Figure 4 also shows an example of the predefined body movement's description. This description is implemented over two stages, and presents movement of head and a correlated movement of the eyes (gaze).

By combining predefined body-gestures and elementary control-points, a complex gesture is formed involving the animation of different body segments at the same time. We claim that this approach, as implemented in EVA scripts, can eventually lead to a natural and personalized finite base-set of predefined body gestures for an ECA (similarly, to FAP being defined for facial expression, and gestures). By using combined predefined gestures from such a base-set, ECA will be able to synthesize most of the complex gestures that cover natural human movement. The small set of gestures to be modelled enables the animators to give each gesture more attention, which usually results in a more natural ECA behaviour.

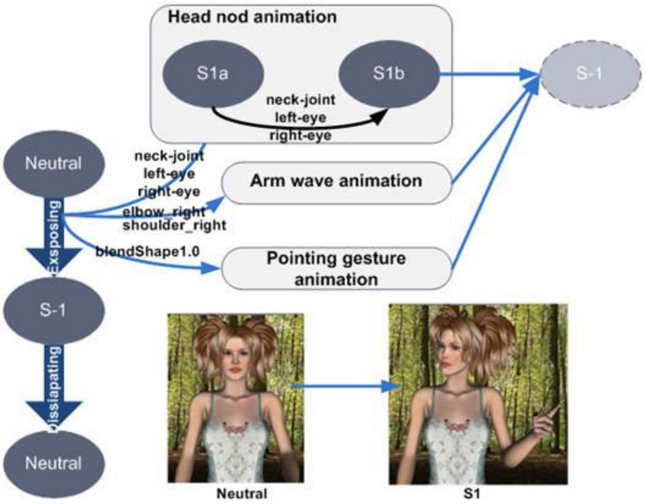


Figure 5: Animation generation and synchronization within EVA's animation engine

IV. Animating body movement (from EVA script to animation)

Each bone/morphed target in a 3D model presents a control joint (control point) that can move or influence a certain amount of vertices within the polygonal mesh. A continuous sequence of different influences can be regarded as a human-gesture or movement. When generating animated sequences, different animation segments (represented by animation states within the animation graph) must transit between each other, without transition to a neutral state. During transition between two animation states, the corresponding control-point moves to the desired position. Therefore, when animating behaviour, the animation-engine simply acts as a finite-state machine (FSM). Fig. 5 presents the animation generation concept, and also internal synchronization within EVA's animation engine. The animation description is provided by body-movement events, as already presented in Fig. 4.

The basic concepts of generating animation from EVA's scripts and the concept of gesture synchronization, state that animation is performed in the form of a finite-state machine (FSM). The control units change physical characteristics (e.g. transition or rotation) only on those transitions between different FSM states. During each transition between states, all joints from the previous state (e.g. A_i) are compared against the joints of the state in the process of being exposed (e.g. A_{i+1}), and properly transited to the positions according to the state A_{i+1}. This concept, therefore, forms the following three sets of control units:

Exposed control units [StiR]: control units exclusively described by state A_{i+1}. Exposed control units transit from their neutral state to state Ai+1, and can be specified as (1):

$$a \in \mathrm{StiR} \Leftrightarrow (a \in A_{i+1}) \Lambda (a \notin A_i) \qquad (1)$$

Dissipated control units [StiL]: control units exclusively described by state A_i. Dissipated control units are set in the A_i state of the animation, and are no longer required in the animation state A_{i+1}, therefore, all units in the set StiL will move to their neutral position. The StiL set can be specified as (2):

$$d \in \mathrm{StiL} \Leftrightarrow (d \in A_i) \Lambda (d \notin A_{i+1}) \qquad (2)$$

Transiting control units [StiT]: control units described by states A_i and A_{i+1}. Such control units will transit from the position described by state Ai, to the position described by state A_{i+1}. The StiT set can be specified as (3):

$$\mathrm{t} \in StiT \Leftrightarrow (t \in A_i) \Lambda (t \in A_{i+1}) \quad (3)$$

The animation sequences can be defined as parallel movements of control units belonging to sets StiR, StiL, and StiT. For instance, the animation of body-movement as presented in Fig. 5, transits following control-points: *neck_joint*, *left_eye, right_eye, elbow_right, shoulder_right*, and *blendShape1.0*. Since the first animation state is neutral (Ai is empty), all the control points will belong to the StiR set, and will therefore move from their neutral position to the exposed position (S1 in Fig. 5). Since the animation sequence is complex, and uses predefined gesture nod (Fig. 4), the transition period for animating nod can be extended to include internal states S1a and S1b (equivalent to S-1). In this way, animation of the gesture nod will be performed in two stages: Neutral → S1a → S1b. On the transition from Neutral to S1a, *neck_joint*, *left_eye*, and *right_eye* control-points will transit from the neutral position into the position described at Stage 1 of the predefined body-movement (Fig. 4), and the StiL and StiT sets will be empty. On transition from animation state S1a to S1b (Stage 2 of predefined body movement in Fig. 4), the *neck_joint, left_eye*, and *right_eye* control points will transit from S1a into the position described by S1b. The control-points will now be contained in the StiT set, and both StiL and StiR sets will be empty. In the context of overall animation, the transitions Neutral → S1a → S1b describe how the control-points will act within the exposed animation period. The Dissipating transition-phase of animation (from state S1 back to the Neutral state) is used, due to the assumption that each body movement originates from, and always returns to its Neutral state. The Dissipating transition phase, therefore, describes the movement of all control points (that are not already in the Neutral state) from the excited state (e.g. S1) into the Neutral state. In our case, the StiL set of the dissipating transition phase will contain all the control points (*neck_joint, left_eye, right_eye, elbow_right, shoulder_right and blendShape1.0.*), and sets StiR and StiT will be empty. By decomposing any complex animation (described within EVA script) into different states that could also contain different sub-states (related to predefined behaviour), any complex description of body-movement can be animated fluidly, without unexpected abrupt movement (e.g. unnecessary transition to neutral state, “jerky” movement, fast jumps, etc.).

V. RESULTS

ECA EVA and the underlying EVA framework present a novel engine for generating multimodal human-machine interfaces, supporting more-personalized and more-expressive human-machine interactions. The EVA framework enables synthesis of both verbal and non-verbal behaviour. It is based on a distributive concept, and physically separates the behavioural generation, and behavioural realization (animation generation and realization) phases. Therefore, it enables the usage of different types of sources that can provide animation parameters in the form of behavioural events. An example of complex-behaviour animation is presented in Fig. 6. The images in Fig. 6 are snaps taken from the video sequence of such animated behaviour.

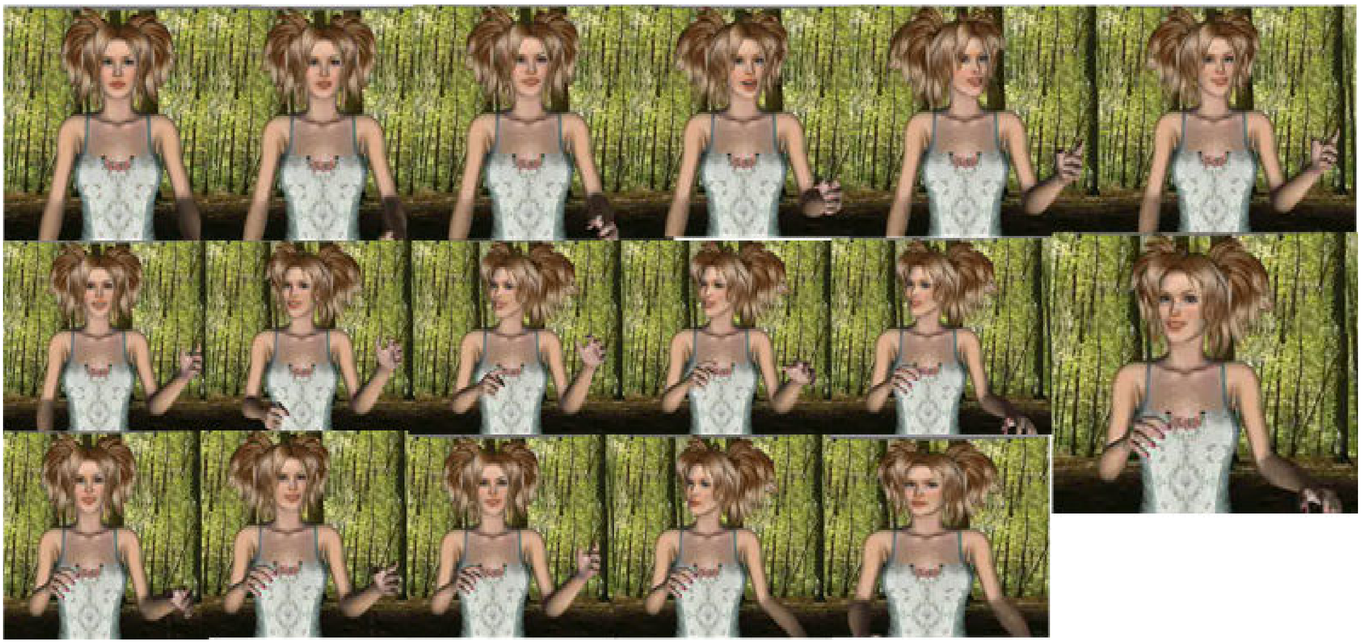


Figure 6: Animation of complex behaviour by using EVA framework.

The images in Fig. 6 present a sequence of animated behaviour. The animation merges both bone and morph-based animation and animation blending technique is used on different body regions in order to properly present different contexts of behaviour (e.g.: animating speech and facial emotion in the mouth region at the same time). Several morphed-shapes (all in compliance with MPEG-4 FAP standard), for animating facial gestures and vizemes were defined within the modelling processes of the ECA EVA's articulated model. All morphed shapes (e.g. *left_outer_eye_brow_up*, *right_inner_eye_brow_up*, *left_mouth_corrner_down*, etc.) were generated by manually analysing facial emotion recognition databases (CohnCadne [33] and MMI [34]) and approximating the morphed shapes to the presentation of different action units (and to different combinations of action units). In addition a skeletal formation was formed in order to perform movement of body parts such as tongue, jaws, eyes, hands and head.

The EVA framework supports use of different input sources that can either provide combined behaviour (e.g. behavioural modelling), or each one its own behavioural events that are processed in parallel into common animated behaviour. The behavioural events provide descriptions of the desired animated behaviour, and the synchronization process then ensures that all sequential segments of an animation (animation states) are always continuous.

By using the animation-blending technique, different animation segments (different body part animations) are combined into smooth and continuous animated-movement. In addition, by using multi-part based 3D models of an ECA, the EVA framework enables on-line changes of the ECA's personalization. Namely, each body part can be modelled and animated independently. ECAs generated by the EVA framework can generate different types of gestures, gaze, and both simple and complex emotion in an expressive, fully adjustable way. The ECA EVA presented in this article is also capable of animating expressive behaviour. Fig. 6 shows an example of such expressive behaviour. The body movement is described by a behavioural event as already described in Figure 4, but with additional description for predefined arm movement, speech, and speech related facial gesture.

VI. CONCLUSION AND FEATURE RESEARCH

The expressivity of an ECA plays a central role in defining its personality, its emotional state, and can further explain the context of the spoken dialogue (e.g. which parts of the dialog are important – emphasis, visualization of the spoken word etc.). The expressivity basically defines ”how” information is presented through physically-based behaviour (movement), and plays a central role in the perception of verbal and non verbal dialogue. Our next steps will be directed towards context oriented behavioural modelling. ECA EVA can express several emotions and gestures at the same time and, as presented in Fig. 6, also complex gestures by animating any part of the ECA's body.

Currently, ECA EVA has no external-behaviour modelling service that would generate EVA's behavioural scripts automatically (with the exception to speech, where TTS service PLATTOS is used to generate speech sequences). The plans for our research in the future, therefore, include research into the behavioural modelling and study of engines, such as: ALMA and EMOTE[35], and their underlying abstract languages, such as: AMPL and EML.

Additionally, a broadening of expressivity regarding ECA EVA is also planned, by defining a finite-set of hand and facial-gestures, either accompanying speech, or non-speech related tasks. Such a set should be defined based on different video databases and will incorporate basic gestures' movement description of arm, hand, neck, etc. By combining basic gestures into complex behaviour, we should then cover most of those gestures generally used in natural multimodal human-machine interactions. Further, extension of the EVA framework with a service that will support online automatic learning of conversational gestures, by processing video input and using different gesture recognition techniques (e.g. facial gesture recognition, facial emotion recognition, hand tracking, etc.) is also planned.

ACKNOWLEDGEMENTS

Operation part financed by the European Union, European Social Fund.

REFERENCES

[1] Sandra Baldassarri , Eva Cerezo , Francisco J. Seron, Chaos and Graphics: Maxine: A platform for embodied animated agents, Computers and Graphics, v.32 n.4, p.430-437, August, 2008

[2] Dimitrios Rigas, Nikolaos Gazepidis, A Further Investigation of Facial Expressions and Body Gestures as metaphors in E-Commerce, Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications August, (2007)

[3] Y Fu, R Li, TS Huang, M Danielsen , Real-Time Multimodal Human–Avatar Interaction, IEEE Transactions On Circuits And Systems For Video Technology, Vol. 18, No. 4, April 2008

[4] E. Cosatto , H. Graf, Sample-Based Synthesis of Photo-Realistic Talking Heads, Proceedings of the Computer Animation, p.103, June 08-10, 1998

[5] POGGI I., PELACHAUD C., DE ROSIS F., CAROFIGLIO V., DE CAROLIS B.: Greta. a believable embodied conversational agent. In Multimodal Intelligent Information Presentation (Text, Speech and Language Technology Vol. 27) (2005)

[6] Erika Chuang , Christoph Bregler, Mood swings: expressive speech animation, ACM Transactions on Graphics (TOG), v.24 n.2, p.331-347, April 2005

[7] Goranka Zoric, Igor S. Pandzic : in Towards Real-time Speech-based Facial Animation Applications built on HUGE architecture, Proceedings of International Conference on Auditory-Visual Speech Processing AVSP 2008

[8] Karlo Smid, Goranka Zoric and Igor S. Pandzic, "[HUGE]: Universal Architecture for Statistically Based HUman GEsturing", Lecture Notes on Artificial Intelligence LNAI 4133, pp. 256-269 (Proceedings of the 6th International Conference on Intelligent Virtual Agents IVA 2006)

[9] Schroder, M.: The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. Advances in Human-Computer Interaction (2010)

[10] A. Heloir and M. Kipp, "EMBR—a realtime animation engine for interactive embodied agents," in Proceedings of the 9th International Conference on Intelligent Virtual Agents (IVA '09), pp. 393–404, Springer, Amsterdam, The Netherlands, 2009.

[11] Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Th_orisson,K.R., van Welbergen, H., van der Werf, R.J.: The Behavior Markup Language: Recent developments and challenges. In: Proc. of IVA-07. (2007)

[12] E Bevacqua,M Mancini,R Niewiadomski,C Pelachaud, An expressive ECA showing complex emotions. Proceedings of the AISB Annual Convention (2007)

[13] DeCarolis B., Pelachaud C., Poggi I, and Steedman M. (2004). APML, a mark-up language for believable behavior generation. In H. Prendinger and M. Ishizuka, editors, Life-like Characters. Tools, Affective Functions and Applications, 65--85. Springer.

[14] Alexis Heloir, Marc Schröder, Patrick Gebhard Realizing Multimodal Behavior Closing the gap between behavior planning and embodied agent presentation

[15] O.Mazany, T. Svoboda Articulated 3D human model and its animation for testing and learning algorithms of multi-camera systems, Master Thesis, Czech Technical University, FEL, CTU-CMP-2007-02, 2007, Prague.

[16] Y. Fu and N. Zheng, "M-face: An appearance-based photorealistic model for multiple facial attributes rendering," IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 7, pp. 830–842, Jul. 2006.

[17] J. Kang, B. Badi, Y. Zhao and D. K. Wright, Human Motion Modeling and Simulation. Proceedings of the 6th WSEAS International Conference on Robotics, Control and Manufacturing Technology (April 2006).

[18] ROJC, Matej, MLAKAR, Izidor. Finite-state machine based distributed framework DATA for intelligent ambience systems. (CIMMACS '09), WSEAS Press, cop. 2009, page 80-85.

[19] Mike Goslin , Mark R. Mine, The Panda3D Graphics Engine, Computer, v.37 n.10, p.112-114, October 2004

[20] Adem Karahoca, Murat Nurullahoglu Human Motion Analysis And Action Recognition, International Conference on Multivariate Analysis and its Application in Science and Engineering (MAASE '08), WSEAS press, cop. 2008

[21] Ugur Güdükbay, Bülent Özgüç, Aydemir Memişoglu and Mehmet Šahin Yeşil in Modeling, Animation, and Rendering of Human Figures. Signals and Communication Technology, 2008, pages 201-238, Springer

[22] Korein, J. and Badler, N., \Techniques for generating the goal directed motions of articulated gures", IEEE Computer Graphics and Applications,Vol. 2, No. 9, pp. 71-74, 1982.

[23] Thalmann, M.N. and Thalmann, D., Computer Animation: Theory and Practice, Springer-Verlag, Berlin, 1985.

[24] Eulers angles:Landau, L.D.; Lifshitz, E. M. (1996), Mechanics (3rd ed.), Oxford: Butterworth-Heinemann

[25] Mehmet Sahin YESIL, REALISTIC RENDERING OF A MULTI-LAYERED HUMAN BODY MODEL, Master of science thesis, August 2003

[26] Mario Malcangi. Soft-computing Methods for Text-to-Speech Driven Avatars, Mathematical methods and applied computing (ACC '09), WSEAS press, cop. 2009

[27] ROJC, Matej, KAČIČ, Zdravko. Time and space-efficient architecture for a corpus-based text-to-speec synthesis system. Speech commun.. [Print ed.], 2007, vol. 49, iss. 3, str. 230-249.

[28] GEBHARD P.: Alma: a layered model of affect. In Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (2005),ACM Press, pp. 29–36.

[29] Alfred Kransted, Stefan Kopp, and Ipke Wachsmuth. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In Andrew Marriott et al., editor, Embodied Conversational Agents: Let's Specify and Compare Them!, Workshop Notes AAMAS, Bologna,Italy, 2002.

[30] G. Ball and J. Breese, 'Emotion and personality in a conversational agent', in Embodied Conversational Characters, eds., S. Prevost J. Cassell, J. Sullivan and E. Churchill, MITpress, Cambridge, MA, (2000).

[31] F. E. Pollick, 'The features people use to recognize human movementstyle', in Gesture-Based Communication in Human-Computer Interaction,eds., Antonio Camurri and Gualtiero Volpe, number 2915 in LNAI, 10–19, Springer, (2004).

[32] CHUN-HONG HUANG1, CHING-SHENG WANG2 and MENG-LIANG YU1, Automatic 3D CBIR on Kinematical Human Motion, Proceedings of the 2007 WSEAS International Conference on Computer Engineering and Applications, WSEAS press, cop. 2007

[33] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), Grenoble, France, 46-53.

[34] M. F. Valstar and M. Pantic, "Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database," in Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION, Malta, 2010, pp. 65-70.

[35] Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. 27th annual conference on Computer graphics and interactive techniques (2000) 173-182