A Model for Facial Activity Recognition using Metarepresentation: a Concept

Boris Knyazev and Yuri Gapanyuk
Faculty of Informatics and Control Systems
Bauman Moscow State Technical University
Moscow, Russia
emails: {bknyazev@bmstu.ru, gapyu@bmstu.ru}

***Abstract*—Recognition of the facial visual properties (physiognomy) and its static and dynamic behavioral patterns (action units) has proved to be an important part in many multimedia retrieval and analysis applications. Apart from the previous studies, where methods to extract part of the action units from an image or video have been developed, in this ongoing research project we work on a model for more accurate and detailed facial activity semantic description adaptable to new behavioral patterns and real conditions. In this paper, we address challenges of building this model and suggest its basic multilevel concept. On the low level, we propose using wavelet-based multiresolution representation of video data. On the middle level, several multiclass classifiers are being examined for the purpose of attribute learning, and a custom multiple metric is provided. On the high level, facial elements, behavioral patterns and their attributes can be connected and further extended using the ontologically-compliant architecture of this model. On the abstraction layer, all three levels of this model are seamlessly integrated via graph-based hierarchies of metavertices, metaedges and their mappings. Having this structure, the proposed model can be trained and employed to solve the problems of human behavior retrieval and human-computer multimodal interaction more efficiently. Current results, however, reveal that to be reliable, this model requires further research studies and their comprehensive experimental evaluation.**

Keywords-facial behavrior recognition; semantic annotation; video multiresolution representation; metagraph modeling

I. INTRODUCTION

Human appearance and nonverbal behavior, particularly facial visual properties (physiognomy) and its static and dynamic behavioral patterns (action units), convey a lot of overt and covert data [1, 2]. Mining of these data is inherent to tackle the problems of human intelligent monitoring and facial expression analysis more efficiently. It has also proved to be useful in psychophysiological and neurological diagnosing, e.g., autism, schizophrenia and other disorders [13], in synthesis of virtual agents [28], examining correlation between face asymmetry and brain disharmony [14] and enhancing human-computer multimodal interaction as a whole. If this mining is automatic, accurate and detailed, then its results – objective data or ground-truth – could be supplied to an expert, clinician or some logical rule-based model for the purpose of making important real-life decisions, e.g., preventing car crashes caused by drowsiness, as well as for entertainment.

The automatic behavior recognition engine could also be a core component of either general-purpose [15] or more specific [16, 17] multimedia annotators. The primary intent of this type of software is to provide means, usually via graphical user interface, to spatiotemporally bind annotations with other modalities, like audio, and with context, which may have a huge impact on interpreting behavior and making a more reliable decision [2].

This research is aimed to further develop previous studies on the automation of facial behavioral patterns recognition (e.g., [3-5]) and is inspired by the works on wavelet multiresolution decomposition [6], Gabor and Dual-Tree complex wavelets [7-9], on graph-based models [10, 11], attribute learning [12] and body segmentation [27].

The contribution of this paper is a model recognizing more action units (AUs) based on the Facial Action Coding System (FACS) [21, 23] and able to extract the new ones, including body AUs [24] and those defined by an expert. Additionally, this model is integral on the abstraction layer and expected to provide easier learning procedures, return semantic annotations improving expert-computer interaction and produce more precise results in more real conditions.

The goals of this paper are to address challenges of the development of this model, suggest its basic concept, give its basic experimental results and propose the directions of its further development.

In Section 2 of this paper, the four-level structure of the proposed model is introduced. In Section 3, we discuss the alphabet of human behavioral patterns and facial action units in particular, and suggest extracting only a specific set of attributes of these actions. Section 4 is devoted to the low level of our model, in which general properties of a metagraph, its vertices and edges are presented. In Section 5, the ways of training our model are shortly overviewed and basic evaluation results are given and discussed.

II. MODEL OVERVIEW

At this stage of research, we do not concern context-based video retrieval and real-time requirements, as well as cluttered environments and multiple persons, so that this model accepts videos (or image sequences) with only one human on a simple background as in the databases [18, 19, 20]. But, this model should be adaptable to real application

 ISBN: 978-1-61208-320-9

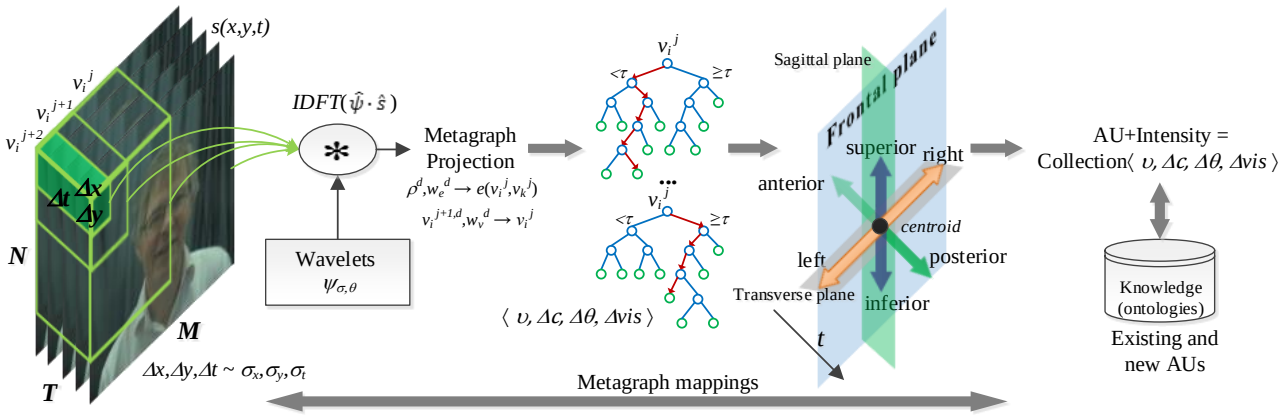

Figure 1. Model overview. Metagraph multilevel representation and relationship between human kinematics and video representation. The WT is faster calculated in the frequency domain applying the convolution theorem and the inverse Discrete Fourier Transform (IDFT); ˆ denotes the Fourier transform.

environments. Our model's structure is a three-level pipeline usually implemented in a visual recognition and understanding system plus a fourth, integration level (Fig. 1).

(1) On the low level, it maps an input video signal to a finite combination of 3D (or 2D) complex wavelets, a naturally multiresolution way to describe an N-dimensional non-stationary signal, such as a video. In addition to capturing changes in texture on different scales and orientations, a number of geometric and color properties can be computed.

(2) On the mid-level, we investigate several multiclass machine learning methods, including the unsupervised ones, such as k-means, and supervised, such as Random Forests (exemplified in Fig. 1), and suggest a multiple weighted metric to separate classes. The classes to be learned are facial primitives, attributes of the AUs as well as more abstract entities.

(3) On the high level, we propose describing facial primitives, behavioral patterns and their attributes merging existing human ontologies with our own being created on the basis of FACS, a broadly accepted facial coding scheme.

(4) To provide integrity on the abstraction layer, that is to encapsulate the three above levels making a model more flexible and scalable, metagraph-based representation of these levels is introduced.

Graphs and their extensions allow intuitive describing of hierarchical data and processes. Furthermore, graph-based models often give competitive results in vision applications, e.g., [11], [22]. One of their extensions is a metagraph, proposed in [10], which is a universal structure to describe properties (attributes), logical relations and complex mappings and, therefore, may be effectively employed to represent humans and their behaviors on the different levels.

One of the benefits of this approach is that since patterns recognition and description are two tightly related processes, detailed comprehensive description should boost the recognition of the patterns and vice versa. Even though the problem of signal reconstruction is not directly related to this study, it might be extremely important for further developments.

On the other hand, among the weaknesses of this model is its computational overload, and it is mostly the low level where optimization techniques should be applied. In spite of this cost, our experience has demonstrated that the benefits of accurate automatic description significantly overweight this negative side effect. In fact, unreliable recognition results mean double work for an expert-annotator: checking plus editing. In any case, the dependency of the accuracy of the results on the extent of detailing should be evaluated.

III. CLASSIFICATION OF BEHAVIORAL PATTERNS

Classification of human nonverbal patterns is challenging to be complete because there are many specialized versions, e.g., [32]. Nevertheless, there is a quite reliable coding system for a face, FACS, and a more recent development for a body, The Body Action and Posture Coding System (BAP) [24], which represent behavioral patterns as combinations of simple AUs and their intensities (A-E) in case of a face. These AUs can be further combined to extract more complex patterns, such as facial expressions in [5].

Training a classifier for each action unit separately is time consuming and is not straightforwardly adaptable to new action units. To remedy this, based on existing coding systems and human anatomy and kinematics, we can define all action codes using a more primitive set(s): $\langle \upsilon, \Delta c, \Delta\theta, \Delta vis \rangle$, where υ – a facial primitive, Δc – change of its centroid (translation), or other measure of spatial relocation, $\Delta\theta$ – change of its spatial orientation (rotation) or movement direction, Δvis – change of its visibility (as in AUs 43, 45, 46 and others). The latter argument must also capture appearance and disappearance of texture features like furrows, e.g., for AU 4 (Brow Lowerer) in the glabella area. To reduce biases in these changes, their values must be normalized and compared to a reference state. This state should include the current orientation and position of an upper level node (in our case, it is a face or a head), person's individual features and context.

Movement directions $\Delta\theta$ can be quantized in the human anatomic planes: left-right (intersection of the F and T planes), superior-inferior (intersection of the F and S planes)

 ISBN: 978-1-61208-320-9

and anterior-posterior (intersection of the *T* and *S* planes), where *F*, *T*, *S* – frontal (coronal), transverse (horizontal) and sagittal (medial) planes respectively (Fig. 1).

The intensity of an action unit is a measure of how distant is a certain facial primitive from a reference position, which may be weighted in Δc and how much texture is changed, which may be weighted in Δvis.

To quantize υ, facial primitives should be divided into smaller ones and include teeth, tongue and other elements involved in some facial activity [21]. To provide more flexibility, though, in addition to verbal description the facial primitives should be also defined as a collection of geometric and texture attributes.

Although, some complex action units, e.g., AU 9 – Nose Wrinkle, AU 23 – Lip Tightener, AU 28 – Lip Suck, AU 32 – Bite, AU 37 – Lip Wipe, etc., and their intensities (A-E) are laborious to be expressed in this way, this representation is more complete from a physical point of view and it is still possible, even though one of the hardest, yet tractable, challenges seems to be quantization (sampling) of the parameters υ, Δc, $\Delta\theta$ and Δvis.

IV. METAGRAPH-BASED MODEL

A. General Model

So far, metagraphs have no unified theory and in this study we adhere to the definition close to [10]:

$$MG = \langle V, E\rangle,\ v_i \in V,\ e_k \in E, \tag{1}$$

where *MG* – a metagraph, *V* – a set of vertices (metavertices), *E* – a set of edges (metaedges), v_i – a vertex of the metagraph and e_k – its edge. In contrast to simple graphs, the vertices are defined as $v_i = \langle\{v_m\},\{e_k\}\rangle$, $v_m \in V$, and can be in turn considered to be a metagraph, so these two terms are interchangeable. We should also distinguish metagraphs from other graph extensions. Compared to hypergraphs, for example, vertices of a metagraph can include both vertices and edges, forming a logical pyramid (a hierarchy). Next, metagraphs can have their own application specific properties. In our study, this pyramid is limited to a video pixel on the one side and by a spatiotemporal voxel $\langle M, N, T\rangle$ with facial activity on the other side. In other words, each metavertex v_i^j resembles an abstract basic type, instances of which include, but are not limited to, a single video pixel; group of interest pixels joined in time, space or other domains (superpixels); input video as a whole, where *j* – is a level of a metavertex v_i^j, and:

$$v_i^j = \bigcup_{m,k} v_m^{j-1}, e_k\ . \tag{2}$$

This abstraction is very convenient because one can work with metavertices in the same fashion as with base abstract types in programming frameworks, i.e., manage objects being unaware of their exact content and values of properties.

$$V_0 \underset{n\to-\infty}{\subset \ldots \subset} V_2{}^j \subset V_2{}^{j+1} = V_2{}^j \oplus O_2{}^j \underset{n\to+\infty}{\subset \ldots \subset} V_2{}^{j+n}$$

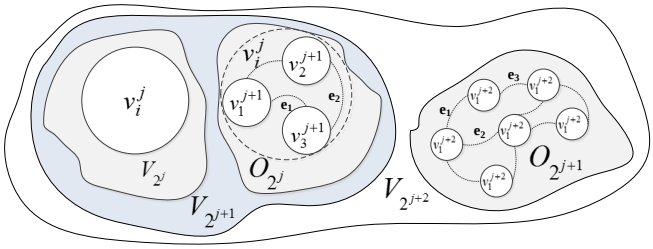

Figure 2. Connection between metagraph-based and wavelet multiresolution reprsentation. For simplicity parental relationships and all edges are not visualized.

In this research, first, a video signal *s(x,y,t)* defined on $\langle M, N, T\rangle$ is represented as a set of abstract metavertices, then mappings to several domains are iteratively constructed in a partially supervised way (see Section 5). The next subsections of this section contain details about the vertices and edges of the proposed metagraph and their properties.

B. Metavertex

Each metavertex has values in at most four domains: spatial, wavelet, color and semantic. These values are essential for further classification of υ, Δc, $\Delta\theta$ and Δvis.

1) Spatial domain (S)

From the kinematics point of view, a human, as well as a human face, can be approximated to a set of objects, the coordinates of their centers of mass and three angles in space. Instead of mass, we can only compute a geometric center (centroid). Important is that the centroid of a more complex object is a linear combination (an average) of its lower level elements.

2) Wavelet (W)

The value of a metavertex in the wavelet domain must reflect behavioral features in the spatial, temporal and frequency domains, where raw video data are difficult to be analyzed. An input signal is transformed to a higher dimensional space, in which video features are more discriminative.

There is no certain algorithm to choose a transform, also referred to as an image/video descriptor or visual words, since there are a couple tradeoffs to consider.

On the one side we want a smooth (continuous) and shift, rotation and scale invariant transform with optimal signal's time and frequency resolutions (limited due to the uncertainty principle) and perfectly reversible, on the other – we want a fast, non-redundant transform requiring less computational power. As was stated above, more accurate results are more valuable for our purpose, and, moreover, we provide optimization techniques. Thus, we suggest to be focused on the wavelet representation of a video signal for a couple of reasons. First, the wavelet transform (WT) is, generally, a type of a complete, i.e., reversible, transform [7] and does not lose information about facial behavior. Second, the wavelet theory provides methods to analyze a signal, such as a video or an image, at different scales, called

multiresolution analysis [6], which is a crucial ability for our metagraph-based model.

Numerous equations of 1D, 2D and 3D wavelet mother functions (wavelets), their Fourier transforms (FTs) and the wavelet criteria can be found in various forms in [6-9, 25], and are omitted here due to their lengthiness. A fast FT (FFT) and its inverse (IDFT) allow computing the WT faster applying the convolution theorem (Fig. 1). Gaussian-modulated complex exponentials, such as the Morlet and Gabor functions, are preferable to be employed as a series of wavelets $\psi_{\sigma,\theta}$, since they are continuous complex wavelets with optimal temporal and frequency resolutions [7], as well as some extensions of B-splines. We suggest binding each level j of the metagraph pyramid with a wavelet scaling (dilation) factor σ (in case of 3D it is σ_x, σ_y, σ_t) by definition included into $\psi_{\sigma,\theta}$. Together with (2) we obtain:

$$v_i^j, \psi_{\sigma,\theta} \in V_2^j, O_2^{j-1}, \quad (3)$$

where V_2^j – a space of approximate mappings, O_2^{j-1} – a complementary to V_2^{j-1} space of detailed mappings; a power of 2 means a dyadic WT [6]. Thus, the pyramidal structure of our metagraph is fixed, but the values of its vertices are assigned in respective spaces (Fig. 2).

In other words, a video volume $\langle M, N, T \rangle$ with facial activity is expressed as a sum of a "blurred" facial video plus detailed facial parts plus more detailed features of the facial parts and so forth up to single pixel values. More general approximate areas around the facial features (areas around forehead, nose, eyes, lips, etc.) and coarse movements (head shaking) can be detected and described on the high scales σ (low temporal and spatial frequencies and level j), while smaller elements (eyes, iris, mouth corners, etc.) and subtle movements (lip twitching, tics, eyes movements) can be detected and described more precisely on the lower scales σ (higher frequencies and level j). In practice, though, we do not need to build both V_2^j and O_2^j spaces on each level j, and as a result, can make scaling of the WT adaptive:

1. Apply the WT with high σ values to the whole video (or one frame) and scale down until a face low-frequency pattern is not found on the video.
2. Analogously apply the WT with lower σ values only to the facial video voxel (or image block) and scale σ down until distinct facial elements are not classified.
3. Recursively repeat step 2 with lower σ values only to specific video voxel and further scale down adaptively until all details are not extracted.

The exact σ values depend on the facial primitive and should be empirically estimated. For instance, they can be calculated on the basis of the entropy-based information gain, similar to [27]. Orientations θ of the wavelets might also vary in a similar sense.

To keep such strengths of the complex WT (CWT) as approximate shift invariance and directional selectivity, while acquiring the ability of perfect reconstruction of the real-valued discrete WT (DWT), the Dual Tree Complex Wavelet transform (DC CWT) could be employed at no extra computational cost compared to the CWT [8]. For both the CWT and DC CWT redundancy is 4:1 for 2D and 8:1 for 3D, whereas the DWT has no redundancy.

3) Color domain (V)

Data in the color domain are useful for facial segmentation. If color details are disregarded in wavelet coefficients, a separate color scheme must be kept, e.g., color histogram. Otherwise, it must be derived from the wavelet coefficients computed for each color channel independently.

4) Semantic domain (S)

The semantic structure of our model should mirror the metagraph pyramidal structure except the semantically meaningless, abstract metavertices, such as some facial regions. Semantic (verbal) terms are necessary for a more natural interaction between a clinician and lower level parts of the model. We suggest integrating existing ontologies, such as Virtual Human Ontology, Foundational Model of Anatomy Ontology and Mental Functioning Ontology to define facial AUs and more complex patterns based on a primitive set defined in Section 3.

C. Metaedge

A metaedge is an attributed multiple edge wrapping distances between two metavertices v_m^j and v_k^j of the same level j in respective domains:

$$e(v_i^j, v_\kappa^j) = \langle \rho^S, \rho^W, \rho^V, \rho^O \rangle, \quad (4)$$

where $e(v_i^j, v_k^j)$ must satisfy the three distance axioms, described in [26]. In addition, $e(v_i^j, v_k^p) = \varnothing$ for $j \neq p$, which means that a metaedge can connect only vertices within one level of hierarchy (one scale).

The distance in the spatial domain ρ^S is the Euclidean distance between the centroids of two vertices in the (x,y,t) space. The distance in the wavelet and color (visual) domains ρ^W, ρ^V can be one of the distance learning metrics or similar to ρ^S, because the values both in the wavelet and visual domains are already in the feature space. However, compared to the spatial one, in the case of nonlinear wavelets it is less trivial to compute the distance between metavertices of a lower level j (e.g., face) as a combination of the distances between the ones of a higher level j (e.g., eye, lips, nose). The distance in the semantic domain ρ^O measures the difference between the entropies of two vertices, as proposed in [26].

D. Metagraph Projection

Metagraph projection can be perceived as convolution to a metagraph of a lower dimensional space, where the values of its projected metavertices are computed separately for each of the four domains either on the basis of its children vertices or independently:

$$v_i^{j,d} = f^d(v_i^{j+1,d}), \quad (5)$$

where d – is one of the four domains S, W, V and O. The cumulative value of a metavertex is a weighted sum of the children values in the four domains:

$$v_i^j = \sum_d w_v^d f^d(v_k^{j+1,d}). \quad (6)$$

Similarly, the cumulative value of an edge between two metavertices is a weighted sum of the edges in the four domains:

$$e(v_m^j, v_k^j) = \sum_d w_e^d \rho^d(v_m^{j,d}, v_k^{j,d}). \quad (7)$$

The weights w_v^d and w_e^d control the impact of a value and distance (edge) in a certain domain *d* on the overall result.

E. Metagraph construction

In this work, construction of a metagraph, which represents our model, is conducted in a frame-by-frame way, however, there are no limitations to implement a voxel-by-voxel way.

First, the frame is divided into 2-4 square blocks depending on the frame size. These blocks automatically become the lowest level (highest in terms of a hierarchy) blocks. For each such block we then apply the WT adaptive algorithm (see above). After each its step the blocks are further divided into 2-4 blocks together with lowering σ. In result, we obtain the metagraph MG_1, in which some branches of its hierarchy become deeper, whereas for some of them this algorithm interrupts after two-three iterations. Simultaneously, for each metavertex independently on its level we calculate: in the spatial domain, its centroid relatively to the lower (upper in terms of a hierarchy) level metavertex and positions of the local maxima of the responses to the wavelet filters $\psi_{\sigma,\theta}$; in the wavelet domain, a distribution of the sums of these responses for different θ; in the color domain, a color distribution in the HSV color space. Currently, values in the semantic domain we keep blank ($w_v^O = 0$) and to evaluate preliminary results of our model we also assign constant values to the weights in other domains: $\{w_e^S, w_e^W, w_e^V, w_e^O\} = \{1, 1, 0.5, 0\}$. Clearly, their assignment requires more investigation, and in experimental studies various influences of each domain on the correct result depending on a metavertex and its level have been observed.

The next frame is processed analogously in order to construct the metagraph MG_2. In addition, for each its vertex we must determine whether it matches to the vertex at the same position in MG_1 or does not. In the latter case, we calculate the difference in terms of the transformations in respective value domains. Translation (Δc) is calculated by the shift of the local maxima of the responses, rotation ($\Delta\theta$) – by the difference in sums distributions, Δvis – by appearance (disappearance) of new (old) strong responses. Scale change is a change of the metavertex level in the metagraph hierarchy. However, scaling less than two times is not detected due to the dyadic WT we applied, whereas in practice, the scaling varied in a broader range (although, mostly in the range of 1.1 – 2 times), so update of our WT's power base should be considered.

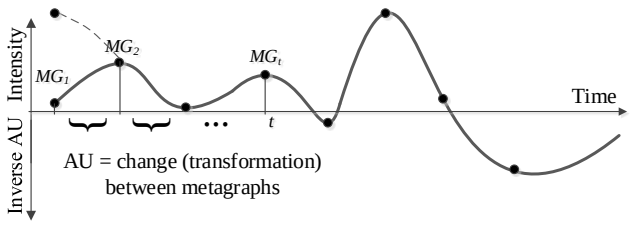

Figure 3. An action unit in terms of metagraphs. Dashed is a possible transformation to the same metagraph (MG_2).

Thus, for each frame, we construct a metagraph. All the metagraphs, excluding the first one are temporal, which means that when the difference between the first (MG_1) and the other (MG_t) metagraphs is found (see (8) below), the transformations (Δc, $\Delta\theta$ and Δvis) to get this difference are clustered to appropriate metavertices of (MG_1).

For each of the vertices of the metagraphs for the following frames we try to find the transformations from either the previous frame only, or from some combination of the previous frames.

In this sense, an action unit is a transformation of a certain metavertex corresponding to a facial element (Fig. 3). Formally, having only one frame in the middle of the frame sequence we cannot certainly infer the current AU, previous states must be known to avoid ambiguities.

Facial activity at some timestamp t is then a set of action units' intensities (a set of transformations) at this timestamp.

V. Training and Evaluation of the Model

Training of our model implies solving the following optimization problem:

$$\arg \min_{w_v^d, w_e^d, T_i} e(MG_1, MG_t). \quad (8)$$

where MG_1 – a metagraph of the first frame, MG_t – a metagraph of the frame at the timestamp t, T_i – a transformation (mapping). Thus, we need to find such mappings and weights, which can transform MG_1 to some MG^*_1, so that $e(MG^*_1, MG_t)$ would be a minimum. In general, this is a complex graph matching problem. In this study, to solve it we assume small changes between frames and as a result, metavertices at the same positions do not differ too much and can be matched more confidently.

Our solution of (8) is composed of two trainings: training the model to recognize attributes and to recognize AUs by these attributes.

A. Learning the attributes

The first training is inspired by the works on attribute learning, such as [12], as one of the ways to deal with the zero-shot learning problem emerged in this research. In our case, it means that no training data for new action units is available. To classify a newly defined AU, the model has to be able to multiclassify its attributes: v, Δc, $\Delta\theta$ and Δvis (see Section 3) during Δt, given metavertices and metaedges associated with a video voxel $\langle \Delta x, \Delta y, \Delta t \rangle$ (or, at least, two

frames) (Fig. 1). Consequently, we need to implement either a supervised or unsupervised multiclass learning method; afterwards, feed the model with training attributive data. In both approaches, weights must be first assigned manually or randomly before a training procedure for the cumulative value (6) and/or for the edge measure (7).

1) Supervised mode

The advantage of Random Forests (exemplified in Fig. 1) among supervised multi classifiers is that their trees are hierarchical by definition and might be associated with the hierarchy of our model further integrating it. Additionally to assigning the weights in (6), thresholds τ for every node of each tree must be also assigned, e.g., as in [27], and then respectively compared with projected values. Training assumes iterative changing of these weights and thresholds for every node of each tree until accuracy is increasing, the trees are not too deep and the gain in information is not sufficient.

2) Unsupervised mode

Among unsupervised methods, k-means, self-organizing maps and other classifiers can be trained. In any case, training assumes an iterative grouping of metavertices with smaller distances (7) between each other closer until no improvement (in the sense of some error function) can be reached. The advantage of these methods is a less tedious training process, since no labeling is required; and facial segmentation can be more objective if a metric is properly chosen, because a human expert is less involved.

B. Learning the AUs

There are two ways to solve the second task, i.e., to recognize AUs by their attributes. First, after semantic values for all facial elements are assigned in a supervised way, each AU can be defined as a set of rules in an xml/owl file. These rules must be written in close accordance with the FACS manual. Another prospective way of learning AUs is to recognize the same attributes from videos of the MMI dataset [33], in which a lot of AUs are labeled, and to infer these rules automatically.

At this stage, all attributes and action patterns (APs) were just clustered in an unsupervised way and we can apply either of the methods in the next works. Note, that feature extraction using the family of unsupervised methods can be tuned to be reliable, but, as it will be shown below, unsupervised classification of AUs themselves is not as reliable, because it is difficult to relate output clusters (APs) with the required classes (AUs).

C. Dataset

Our model can be trained and tested using a labeled dataset with real video [18, 19], images sequences [20] or a synthetic one with inherently labeled action units, for instance generated by the means of [29-31]. In [27], real and synthetic datasets complemented each other, which led to high recognition scores of body pose recognition, and therefore, this approach should be adopted in this study in the future studies.

TABLE I. RESULTS OF THIS WORK FOR THE DATASET [18]

Session Id	Subject Id	No. of AUs (TP)	No. of APs (TP)	No. of false APs (FP)	Overall No. of positives
21	2 (Operator)	6	21	15	36
	3 (User)	9	24	12	36
29	3 (Operator)	9	23	12	35
	16 (User)	7	17	19	36
64	7 (Operator)	11	16	8	24
	11 (User)	9	15	13	28

D. Evaluation

This work is in progress and to determine and, perhaps, correct the further direction of its development we collected qualitative results of a demo version of the model presented above for several subjects from the Semaine Database [18], which seemed to be closer to real environment compared to other datasets.

A simple .NET Framework (ver. 4.5) application integrated with the MATLAB API (ver. 2012a) was developed. The first part was used for object oriented metagraph implementation and abstract manipulation, and the second part was used for wavelet decomposition and calculations of transformations and was compiled for .NET using NE Builder.

The number of facial action patterns (No. of APs, Table 1) that our model clustered turned out to be far more than the number of facial AUs from FACS (No. of AUs), even though they overlapped partially, e.g., 6 from 21.

The coincidences mostly occurred when a particular expression and a respective AU was very intensive, whereas during substantial periods of time expressions were unclear, but it does not mean there was no AU. Another set of ambiguities were observed when a person was talking, which resulted in a lot of APs, which we could not always correspond to one of the AUs. Since the database that we used is labeled only using feeltrace annotations, it was difficult to check our results correctly, therefore we measured them categorically. To calculate the categorical error rate we counted the overall number of action unit types present in a video by watching it, and compared it to the number of output clusters returned by our model, which we could attribute to some AU with high confidence, even though the exact AU was unclear.

TABLE II. RESULTS OF SOME PREVIOUS WORKS

AUs	Method	Dataset	CR/F_1/PR, %
15+	Multi-state geometric face model [3]	CK	82-96.7/-/-
27	Free-form Deformations (FFD) + GentleBoost + Hidden Markov Model (HMM) [4]	MMI	94.3/65.1/59.7
18		CK	89.78/72.14/70.25
15	Viola-Jones + ASM + Gabor filters + GentleAdaboost [5]	private	95.9 (agreement rate)/-/-
9+	This work	Semaine	-/-/58.3

 ISBN: 978-1-61208-320-9

We also counted the number of false positive APs (No. of false APs) which included mismatched metavertices or invalid transformations. Having no results about negatives, we were only able to calculate the precision rate (PR), Altogether, the challenges described above led to an indecent number of false positive errors and PR compared to some previous works (Table 2, in which CR is the classification rate, AUs – the number of analyzed AUs), even though we did not measure them frame by frame.

VI. CONCLUSION AND FUTURE WORK

In this paper, a concept of the model for multilevel facial activity recognition and description is presented, and emerged technical and scientific challenges are discussed. Even though the model is not formalized and not explained in detail here, it promises to fulfill the requirements of this research: (1) recognize more FACS-based action units more accurately and in more real conditions compared to the previous studies; (2) be able to recognize the new ones, including body AUs and those defined by an expert using primitive attributes from the ontological network; (3) be integral and scalable for multiple persons. Among the hardest challenges are reliable quantization of classes and attributes and the complexity of classifying posterior-anterior movements, since they are not intrinsic to a video.

The suggested four level model should not be perceived as an overcomplete application of heavy methods. On the contrary, the metagraph model allows representing low-, mid- and high-level methods using mappings of metavertices making the model more homogeneous and flat. Indeed, this is an important theoretical implication, that many models can be represented using metagraphs and their mappings, even though the mappings are not always trivial.

The preliminary evaluation results demonstrated that our model needs further development, tuning and more comprehensive evaluation to solve the problems of human behavior retrieval and human-computer multimodal interaction more efficiently, which must be the focus of further research studies.

ACKNOWLEDGMENT

This research is partially supported by Bauman Moscow State Technical University. Portions of the research in this paper use the Semaine Database collected for the Semaine project (www.semaine-db.eu) [18].

REFERENCES

[1] T. Kanade, “Visual Processing and Understanding of Human Faces and Bodies”, 9th International Conference (ICVS 2013), Jul. 2013, Keynote Talk, URL: http://workshops.acin.tuwien.ac.at/ICVS/downloads/Kanade_ICVS2013.pdf (December 16, 2013)

[2] P. Ekman, “What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System”, New York: Oxford University Press, 2005.

[3] Y. L. Tian , T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, Feb. 2001, pp. 97-115.

[4] S. Koelstra, M. Pantic, and I. Patras, “A Dynamic Texture Based Approach to Recognition of Facial Actions and Their Temporal Models”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Nov. 2010, pp. 1940-1954.

[5] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders", J. Neurosci. Methods, vol. 200, no. 2, Sep. 2011, pp. 237-256.

[6] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 11, Jul. 1989, pp. 674 -693.

[7] T. S. Lee "Image representation using 2D Gabor wavelets", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 18, no. 10, Oct. 1996, pp. 959 -971.

[8] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform", IEEE Signal Process Mag., vol. 22, no. 6, Nov. 2005, pp. 123-151.

[9] B. Solmaz, S. M. Assari, and M. Shah, “Classifying Web Videos using a Global Video Descriptor”, Machine Vision and Applications (MVA), vol. 24, iss. 7, Oct. 2013, pp. 1473-1485.

[10] A. Basu and R. W. Blanning, “Metagraphs and Their Applications”, Integrated Series in Information Systems, Vol. 15, 2007, 172 p.

[11] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, “Face Recognition by Elastic Bunch Graph Matching”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 17, no. 7, Jul. 1997, pp. 775-779.

[12] V. Ferrari and A. Zisserman, “Learning visual attributes”, Advances in Neural Information Processing Systems, Dec. 2007, pp. 433-440.

[13] M. S. Bartlett and J. Whitehill, “Automated facial expression measurement: Recent applications to basic research in human behavior, learning, and education”, In Oxford Handbook of Face Perception, Oxford University Press, 2011, pp. 489-514.

[14] A. N. Anuashvili, “Fundamentals of Objective Psychology”, Moscow-Warsaw, 2005. (in Russian)

[15] A. Heloir, M. Neff, and M. Kipp, “Exploiting Motion Capture for Virtual Human Animation: Data Collection and Annotation Visualization”, Proc. Workshop on "Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality", 2010, URL: http://embots.dfki.de/doc/Heloiretal10.pdf (December 16, 2013)

[16] C. Delgado, R. Garcia, J. I. Navarro, and E. Hinojo, “Functional analysis of challenging behaviours in people with severe intellectual disabilities using The Observer xT 10.0 software”, Proc. Measuring Behavior, Aug. 2012, pp. 365-367.

[17] B. Knyazev, “Human nonverbal behavior multi-sourced ontological annotation”, Proc. International Workshop on Video and Image Ground Truth in Computer Vision Applications (VIGTA '13), Jul. 2013, Article 2 , 8 p.

[18] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, “The Semaine Corpus of Emotionally Coloured Character Interactions,” Proc. IEEE Int'l Conf. Multimedia and Expo (ICME), Jul. 2010, pp. 1079-1084.

[19] X. Zhang et al., “A High-Resolution Spontaneous 3D Dynamic Facial Expression Database”, Proc. 10th IEEE Int'l Conference and Workshops on Automatic Face and Gesture Recognition (FG'13), Apr. 2013, pp. 1-6.

[20] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," Proc. 2010 IEEE Computer Society Conference on CVPR Workshops, Jun. 2010, pp. 94-101.

 ISBN: 978-1-61208-320-9

[21] P. Ekman and W. Friesen, “Facial Action Coding System: A Technique for the Measurements of Facial Movements”, Consulting Psychologists Press, 1978.

[22] C. Y. Chen and K. Grauman, “Efficient activity detection with max-subgraph search”, Proc. CVPR, Jun. 2012, pp. 1274-1281.

[23] C. H. Hjortsjö, “Man's face and mimic language”, Studentlitteratur, 1969.

[24] N. Dael, M. Mortillaro, and K. R. Scherer, “The Body Action and Posture Coding System (BAP): Development and Reliability”. Journal of Nonverbal Behavior, vol. 36, iss. 2, Jun. 2012, pp. 97-121.

[25] R. J. E. Merry and M. Steinbuch, "Wavelet theory and applications", literature study, Eindhoven University of Technology, 2005.

[26] J. Calmet and A. Daemi, “From Entropy to Ontology”, Proc. 4th Int'l Symp. “From Agent Theory to Agent Implementation”, Apr. 2004, URL: http://www.iks.kit.edu/fileadmin/User/calmet/papers/AT2AI4.pdf (December 16, 2013)

[27] J. Shotton et al., “Real-time human pose recognition in parts from single depth images”, Proc. CVPR 2011, Jun. 2011, pp. 1297-1304.

[28] I. A. Essa, “Analysis, Interpretation and Synthesis of Facial Expression”, PhD thesis, MIT, Media Lab, 1995.

[29] Autodesk MotionBuilder, URL: http://www.autodesk.com/products/motionbuilder/overview (December 16, 2013)

[30] Di3D Inc., URL: http://www.di3d.com (December 16, 2013)

[31] E. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. Scherer, "FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units," Journal of Nonverbal Behavior, vol. 35, 2011, pp. 1-16.

[32] The Nonverbal Dictionary - Center for Nonverbal Studies, URL: http://center-for-nonverbal-studies.org/6101.html (December 16, 2013)

[33] M. F. Valstar and M. Pantic, “Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database”, Proc. International Language Resources and Evaluation Conference, Malta, May 2010, pp. 65-70.

 ISBN: 978-1-61208-320-9