PAPER

Facial Micro-Expression Detection in Hi-Speed Video Based on Facial Action Coding System (FACS)

Senya POLIKOVSKY[†a)], *Student Member*, Yoshinari KAMEDA[†], *Senior Member*, *and* Yuichi OHTA[†], *Fellow*

SUMMARY Facial micro-expressions are fast and subtle facial motions that are considered as one of the most useful external signs for detecting hidden emotional changes in a person. However, they are not easy to detect and measure as they appear only for a short time, with small muscle contraction in the facial areas where salient features are not available. We propose a new computer vision method for detecting and measuring timing characteristics of facial micro-expressions. The core of this method is based on a descriptor that combines pre-processing masks, histograms and concatenation of spatial-temporal gradient vectors. Presented 3D gradient histogram descriptor is able to detect and measure the timing characteristics of the fast and subtle changes of the facial skin surface. This method is specifically designed for analysis of videos recorded using a hi-speed 200 fps camera. Final classification of micro expressions is done by using a k-mean classifier and a voting procedure. The Facial Action Coding System was utilized to annotate the appearance and dynamics of the expressions in our new hi-speed micro-expressions video database. The efficiency of the proposed approach was validated using our new hi-speed video database.

key words: *facial motion analysis, high speed camera, video descriptor*

1. Introduction

The increase of extreme violent actions around the world requires new technological solutions that are helpful in the detection of hostile intent and prevention of those actions. The combination of computer vision and psychology has the potential for developing such technology.

Facial micro-expressions are brief, involuntary expressions that appear when emotions are concealed or repressed and usually occur in high stakes situations [1]. Two independent research groups, Ekman and al. [1] and Porter and al. [2] found that facial micro-expressions are the most important nonverbal sign of hidden emotions and can be used for lie and danger demeanor detection [3], [4]. Not only the detection but also the timing characteristics of facial muscle motions during micro-expressions are considered as important clues. Those characteristics have high potential to be used for psychological and behavioral analysis [1].

However, there are several technical issues that make micro-expression recognition and characterization difficult. First, they have a short duration that varies from 1/3 to 1/25 seconds, second they are caused by a small muscle contraction inducing little changes in the skin texture, all this making them almost imperceptible. Traditional computer vision approaches such as Active Appearance Model (AAM) or Gabor descriptors would not be suitable for detecting such changes.

The two main targets of the research initiated with this work are to propose a method for both detection and characterization of micro-expressions. Due to the short duration of micro-expressions, attempting to analyze such motions using 25 fps cameras would result in approximately 5 frames for an entire expression. Such number of frames would not be sufficient in order to extract the timing characteristics of the expression.

The solution we propose uses a "flow approach" that was specially designed to analyze videos from 200 fps hi-speed camera. Our flow approach is based on 3D gradients (spatial-temporal) descriptors that measure the changes between following frames. The descriptor proposed in this paper combines pre-processing masks, histograms and concatenation of 3D gradient vectors, providing both classification and motion characterization of micro-expressions. In addition, the time characteristics can be measured directly from the descriptor values. Furthermore, use of gradient descriptors provides an appropriate solution to address the noise induced by the use of hi-speed video.

In addition to our method, we introduce a FACS annotated new hi-speed video database that will be extended in the future for analyzing variety of micro-expression motion characteristics. This database target to provide an important instrument for researchers in the field of psychological and behavioral analysis that are using computer vision technology for analysis automatization.

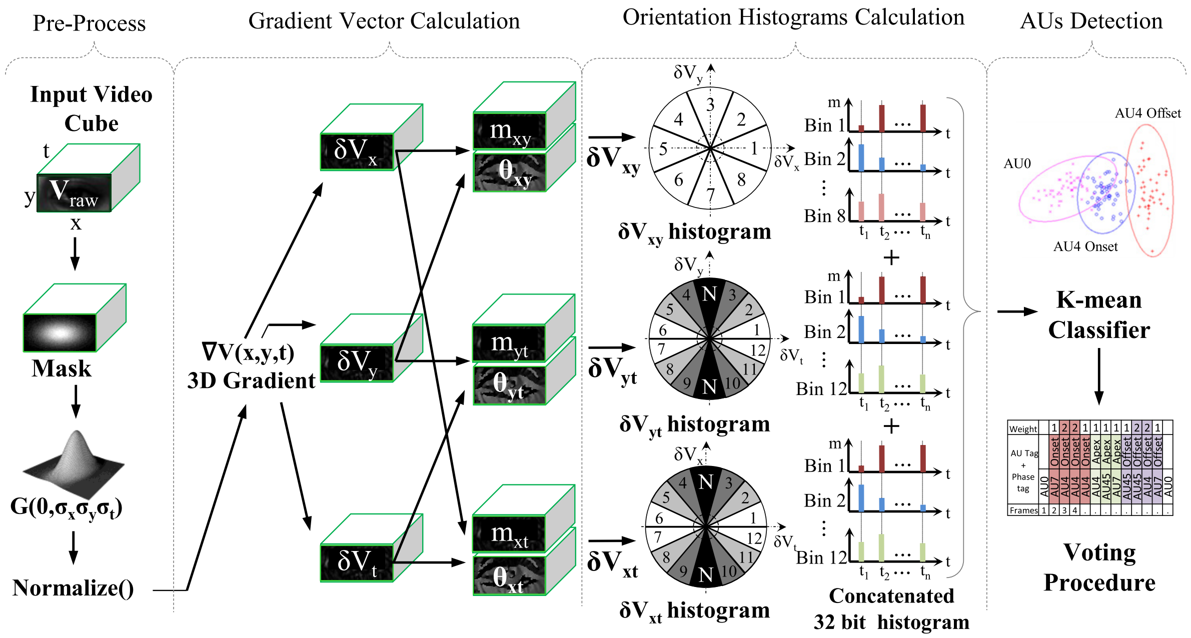
The proposed method contains following steps: First, the face in the video is divided into facial regions and corresponding video cubes are extracted. Facial regions are defined in accordance with the facial action coding system (FACS) [5]. Second, motion in each region is described using a 3D-Gradient orientation histogram descriptor that is based on partial derivative vectors. Figure 1 presents a facial video cube along with partial derivative vectors. Finally, micro-expression recognition is achieved through a k-mean classifier and weighted voting procedure. Figure 2 shows an overview of the proposed method.

The structure of the paper is as follows: In Sect. 2 we discuss related work. Detailed descriptions of our methods in Sect. 3. In Sect. 4 our new database of micro-expressions is presented. Experimental results are discussed in Sect. 5. Conclusions and future work are presented in Sect. 6.

Manuscript received April 19, 2012.
Manuscript revised August 7, 2012.

[†] The authors are with University of Tsukuba, Tsukuba-shi, 305–8573 Japan.
a) E-mail: senya@image.esys.tsukuba.ac.jp

DOI: 10.1587/transinf.E96.D.81



Weight		1	2	2	1	1	1	1	1	2	2	1	
AU Tag + Phase tag		Onset	Onset	Onset	Onset	Apex	Apex	Apex	Offset	Offset	Offset	Offset	
	AU0	AU7	AU4	AU4	AU4	AU4	AU45	AU7	AU45	AU45	AU4	AU7	AU0
Frames	1	2	3	4	.	.	.	.	.	.	.	.	.

Fig. 2 Our micro-expression detection method consists of four main parts: Pre-Process, Gradient Vector Calculation, Orientation Histograms Calculation and AUs Detection. The Pre-Process part consists of video cube masking, Gaussian smoothing and normalization. In the Gradient Vector Calculation part, first δv_x, δv_y, and δv_x gradients cubes and then their magnitudes $m_{xy}(x, y, t)$, $m_{yt}(x, y, t)$, $m_{xt}(x, y, t)$ and orientation $\theta_{xy}(x, y, t)$, $\theta_{yt}(x, y, t)$, $\theta_{xt}(x, y, t)$ cubes are calculated. In the Orientation Histograms Calculation part, first an 8 bins histogram δv_{yx} and two 12 bins histograms δv_{yt} and δv_{xt} are calculated, and then all the histograms are concatenated into one 32 bins description vector. This 32 bins vector represents the motion between every frame in the video $v(x, y, t)$. Finally in the AU Detection part the final micro-expression recognition is made through a k-mean classifier and a weighted voting procedure.

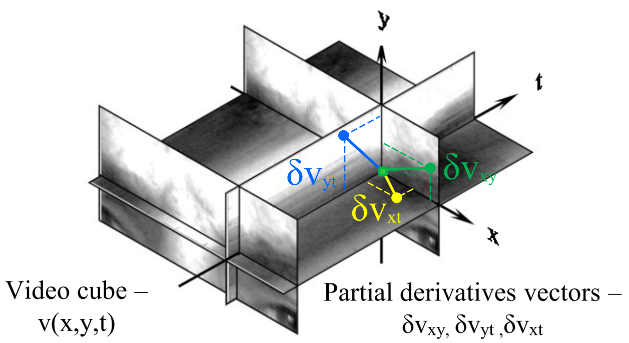


Fig. 1 Visualization of the 3D video cube $v(x, y, t)$ in the *Between Eyes* area and partial derivative vectors of one pixel in the cube δv_{xy}, δv_{yt}, and δv_{xt} corresponding to the xy, xt, and yt surfaces. Our 3D-Gradient orientation histogram descriptor combines all the partial derivatives vectors from all the pixels in each frame.

2. Related Work

In this section we summarize facial expression analysis methods and spatio-temporal descriptors to detect motion in video sequences.

2.1 Automatic Facial Expression Recognition Systems

General automatic facial expressions recognition systems consist of three steps: (1) face acquisition, (2) facial data extraction and representation, and (3) facial expression recognition. Face acquisition consists of automatic detection and tracking of the face in the input video, in some cases extraction of the face direction is part of this step. For facial data extraction and representation for expression analysis, two main approaches exist: (a) geometric feature-based methods and (b) appearance-based methods. (a) Geometric facial features methods represent facial expressions by reconstructing the approximate shape of the face based on the location of facial feature points (such as mouth corners, eyes corners, eyebrows edges and etc...). A feature vector storing these feature points represents geometry the face [2]. For example 20 facial feature points have been directly tracked using a particle filter [6]. This approach gives good results for several facial expressions, however some subtle motions such as micro-expressions that can only be observed by skin surface changes cannot be detected.

In (b) appearance-based methods, image filters, such as Gabor wavelets, are applied to either the entire face or specific regions in the face. This method was applied to spontaneous facial motion analysis [7]. More recently Gabor filters were replaced by haar-like features providing similar recognition rate with less computational load [8]. However, both methods are based on analyzing the video frame by frame, without considering motion between the frames. In addition, applying this approach for facial surface analysis requires large database for training more than 0.6 million Gabor filters.

2.2 Active Appearance Model and Active Shape Model

Recently, advanced research results were reported on Active Appearance Model (AAM) by the Kanade group [9]. AAM combines texture and shape analysis. However, there are two disadvantages of AAM that makes them inapplicable for our research setting. First, this approach requires an extensive database with a large amount of manually tagged points on the face area. Second, the accuracy of facial feature tracking significantly decreases when applied on faces that were not included in the training set [10]. A similar approach is the Active Shape Model (ASM) which combines a point distribution model together with a local appearance pattern for every point. Although it is less accurate than AAM, it is provides more robust results on faces which were not in the training set. Additional detail on AAM and ASM can be found in [10].

While AAM / ASM are proven to have good performance in detecting regular expressions where facial structure has a significant deformation, they are not suitable for detecting a subtle and low muscle contractions such as during micro-expressions. For example a micro-expression smile can be observable only by few wrinkles on the skin surface above the cheeks. Futhermore Marks [11] proved that in cases the texture changes are more significant than changes in shape, flow approaches have better performance than template approaches such as AAM and ASM. Therefore 3D descriptor approach can be considered more suitable for micro-expression detection.

2.3 3D Descriptors

2.3.1 Spatio-Temporal Local Descriptors

Concept of motion descriptors were introduced by Laptev for automatic event detection in videos [12]. Dollar [13] compared action recognition in videos using local descriptors, such as normalized pixel values, brightness gradients, and windowed optical flow. Experiments on three datasets: facial expressions, mouse behavior, and human activities, showed best results for gradient descriptors. In this method, the gradient descriptor was computed by concatenating all gradient vectors in a region, leading to rough descriptor. As a result more complicated and subtle motions such as micro-expressions can not be descried.

A more advanced descriptor was proposed by Scovannere [14], who used an extension of the SIFT descriptor for 3D data. After the spatio-temporal gradient vectors are computed for each pixel, their orientation quantization in polar coordinates is done. This leads to singularities at the poles since bins get progressively smaller closer to them (like the longitude and latitude grid of the globe), which leads to non uniform quantization. Polyhedrons were proposed as a possible solution to the singularity problem and the gradient vector being computed for sub-blocks and not for every pixel [15].

Both previous methods have the benefits of a general descriptor, but are not suitable for precise measurements of time characteristics due to the lack of connection between the values of the gradient histogram and the actual facial movements making. Review of algorithms for motion classification in video, using 3D gradient descriptors can be found in [15].

2.3.2 3D Descriptors for Facial Expression Analyis

Pantic group proposed the use of dynamic descriptors for analysis of facial texture changes in videos [16], [17]. They also compared Motion History Images (MHI) and Free-Form Deformations (FFDs) descriptors, GenleBoost being used as a classifier [16]. In addition, this work defines an approach for comparing results in expression phase detection. In [17] the family of local binary pattern descriptors for FACS AU detection are compared.

The main focus in these works is the detection of the regular facial expressions, where the algorithms and the capturing systems were based on regular speed cameras of 25 fps. However, the analysis of micro-expression requires adaption of the descriptor to subtle and imperceptible motions as well as much higher sampling rate.

The descriptor we propose is specifically adapted to facial movement analysis of micro-expressions and is able to describe separately the surface of the face and the facial motion. In addition it allows the examination of the facial movement's timing characteristics though direct observation of the descriptor values on the gradient's histogram. The use of 3D gradient histograms descriptors using hi-speed camera for micro-expression detection as well as the creation of a new hi-speed video database for validation is novel. In this work 200 fps presents a suitable tradeoff between noise, accuracy and camera price. Use of hi-speed camera poses some difficulties that we propose to solve through our descriptor.

3. 3D Gradient Histogram Descriptor

In this section we present the procedure used to calculate the 3D gradient histogram descriptor for micro-expression detection. In general, micro-expressions are directional deformation of certain facial skin textures, starts from non-expression (known as neutral state) to pick of the expression (contraction) and back to neutral state in continuous way. By detecting the skin changes in particular direction across time we can detect and identify the corresponding expression. In this manner, the 3D gradient histogram are used to describe the spacial-temporal changes of the facial skin. The calculated descriptor values serve to classify the changes acroses AU tags.

The proposed algorithm contains three main phases: First, extraction of twelve facial video cubes (Sect. 3.1).

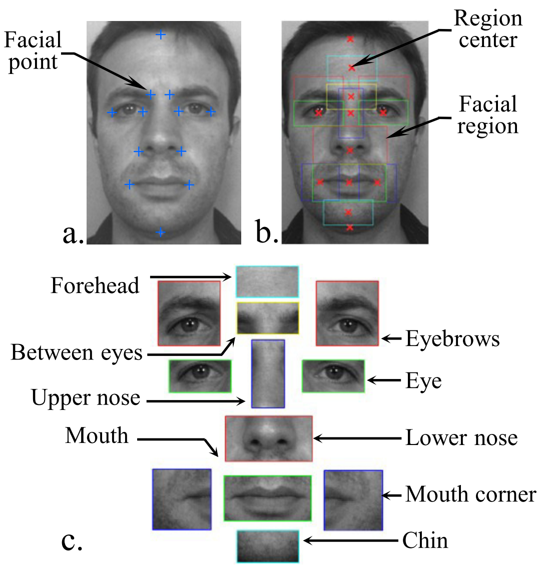


Fig. 3 Facial region selection. a.) Facial points marked on the first frame of the video. b.) Calculation of the facial regions' size and position. c.) 12 selected facial regions that will be used in further analysis.

Second, a pre-processing step that includes cube normalization, smoothing and masking of uninformative pixels (Sect. 3.2). Third, computation of the 3D gradient orientation histogram descriptor for each video cube (Sect. 3.3).

In Sect. 3.4 the classification process is described, in which all the frames in the cube are classified into one of of AU tags, based on their descriptors. The final identification of the micro-expressions is made using a voting procedure. Finally, in Sect. 3.5 we present parameters of our descriptor.

3.1 Facial Region Selection

Estimation of whole facial muscles structure and movement simultaneously is difficult and high computational task. In addition, as the appearance of micro-expressions can be spotted by observing certain facial regions, we separate the analysis of the face into 12 facial regions.

Following the facial action coding system (FACS) [5] that decomposes facial expressions in terms of 46 component movements called Action Units (AUs), we select the most representative facial regions in terms of micro-expressions.

In order to estimate each face muscle motion precisely, each facial regions is defined so that its appearance is affected only by a limited number of muscles. Hence, during the classification step, every region will hold a limited number of classes to distinguish. The regions are selected manually using the following procedure:

First, 12 facial points are marked on the first frame of the input video as shown in Fig. 3 a. The positions of the facial points were defined by based on [9] and [6] works.

Second, the average size of the eyes and the center location for each region are calculated based on the selected facial point position as shown in Fig. 3 b. Based on the portrait drawing guidelines [18], eye size has an important proportional value in the human face, as the size of facial features and their position can be regularized in proportion to the eye size. More information on locations and proportions of facial regions is given in [19]. The size of the regions is defined as slightly bigger than the actual region to make sure that, in spite of small face movements and rotations, the important features will stay inside the region.

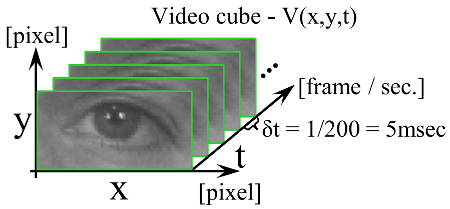


Fig. 4 Video cube of the "Eye" region extracted from the 200 fps video sequence.

Finally, 3D facial cubes are extracted for all the facial regions (Fig. 3 c). The units for the x, y and t axes of the cube are "pixels", "pixels" and "frames" respectively (Fig. 4).

3.2 Pre-Processing

Our descriptor includes three pre-processing procedures as follows: First, as expressed in Eq. (1), each video cube $v_{raw}(x, y, t)$ is multiplied by a weight mask $M(x, y, t)$ that remove uninformative pixels from the descriptor calculations. There are several ways to obtain the mask cube $M(x, y, t)$. One, is to learn the weights for the mask based on statistical information using a training set of videos from a database.

$$M(x, y, t) = \begin{cases} w & \text{informative pixels} \\ 0 & \text{uninformative pixels} \end{cases} \tag{1}$$
$$v_{mask}(x, y, t) = M(x, y, t) \circ v_{raw}(x, y, t)$$

$\circ$ being the Hadamard product, which is an element wise multiplication.

$$v(x, y, t) = G_{3D}(x, y, t, \sigma_x, \sigma_y, \sigma_t) * v_{mask}(x, y, t) \tag{2}$$

Another option is based on face tracking algorithms in which the weight mask is updated dynamically in correspondence with the location of the face in the video cube. A third option is to use a static mask specially designed for every cube. This option was implemented in this paper due to it simplicity and the fact that the face position doesn't significant change in our video database. Binary masks were used to set all uninformative pixels to "0" and all informative pixels to "1" (see Fig. 5 a).

After the masking, the video cube is smoothed in the spatial and temporal directions using a 3D Gaussian $G_{3D}()$ kernel with $\sigma_x, \sigma_y, \sigma_t$ parameters (see Eq. (2)). One of the reasons for applying the Gaussian filter is the fact that the use of a hi-speed camera produces higher level of noise compared to regular cameras. Sigma values were selected in accordance with the lighting conditions of the input videos. The last step of the pre-processing is a cube normalization

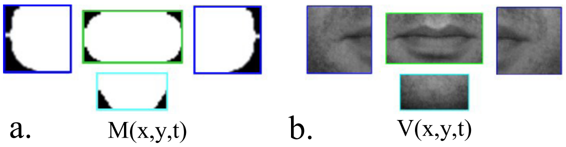


Fig. 5 a.) Weight mask images that were defined for every video cube, “Black” pixels stand for weight value ‘0’ and “White” pixels for ‘1’. b.) Corresponding video cubes for the masks.

that brings all the pixel values between 0 and 1 Fig. 2 illustrates all the steps in the Pre-Processing section.

3.3 3D Orientation Gradients Histogram

In this section we describe the 3D histogram descriptor calculation. First lets consider the 2D case, where $I(x, y)$ is the image and $\delta I_x(x, y)$, $\delta I_y(x, y)$ are the image's partial derivatives. The gradient magnitude and the orientation of each pixel is then defined by Eq. (3).

$$\begin{aligned} m_{x,y}(x, y) &= \sqrt{\delta I_x(x, y)^2 + \delta I_y(x, y)^2} \\ \theta_{x,y}(x, y) &= \tan^{-1}(\delta I_x(x, y)^2 / \delta I_y(x, y)^2) \end{aligned} \tag{3}$$

In the 3D case of video $V(x, y, t)$, where the third dimension is time, we selected the gradient representation specifically adapted to facial micro-expression movement analysis.

The first step in the 3D descriptor calculation given a video cube $v(x, y, t)$, is the calculation of the partial derivatives along the x, y, and t axes. Then, for each couple of partial derivatives $(\delta v_x, \delta v_y)$, $(\delta v_y, \delta v_t)$, and $(\delta v_x, \delta v_t)$, corresponding magnitude $m_{xy}(x, y, t)$, $m_{yt}(x, y, t)$, $m_{xt}(x, y, t)$ and orientation $\theta_{xy}(x, y, t)$, $\theta_{yt}(x, y, t)$, $\theta_{xt}(x, y, t)$ cubes are computed using Eq. (4). A diagram of the gradient vector calculation can be seen in Fig. 2.

$$\begin{aligned} m_{xy}(x, y, t) &= \sqrt{\delta v_x(x, y, t)^2 + \delta v_y(x, y, t)^2} \\ \theta_{xy}(x, y, t) &= \tan^{-1}\left(\frac{\delta v_y(x, y, t)^2}{\delta v_x(x, y, t)^2}\right) \\ m_{yt}(x, y, t) &= \sqrt{\delta v_y(x, y, t)^2 + \delta v_t(x, y, t)^2} \\ \theta_{yt}(x, y, t) &= \tan^{-1}\left(\frac{\delta v_y(x, y, t)^2}{\delta v_t(x, y, t)^2}\right) \\ m_{xt}(x, y, t) &= \sqrt{\delta v_x(x, y, t)^2 + \delta v_y(x, y, t)^2} \\ \theta_{xt}(x, y, t) &= \tan^{-1}\left(\frac{\delta v_x(x, y, t)^2}{\delta v_t(x, y, t)^2}\right) \end{aligned} \tag{4}$$

In this work $\delta v_{xy} = \{m_{xy}, \theta_{xy}\}$ represents the shape of the surface, $\delta v_{yt} = \{m_{yt}, \theta_{yt}\}$ represents the vertical changes, and $\delta v_{xt} = \{m_{xt}, \theta_{xt}\}$ represents the horizontal changes.

The gradient orientation histogram for every frame in the δv_{xy}, δv_{yt} and δv_{xt} cubes are then computed.

The gradient orientation histogram of the shape of the surface δv_{xy} contains 8 bins (Fig. 6 a). All the vectors that belongs to the same bin are added, except for vectors of very small magnitude which can be considered as informative. By setting 8 bins we can characterize the appearance of skin wrinkles and other texture changes in 8 directions (up, up-left, left, etc...). Each bin is wide enough to give the flexibility to deal with small rotations of the face.

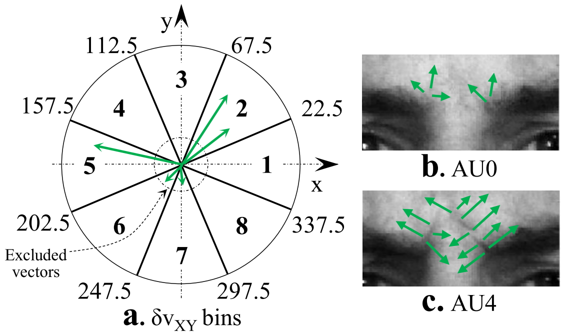


Fig. 6 a.) Shape of the surface δv_{xy} gradient orientation histogram bins. The vectors on the histogram illustrate the bin quantization process. All the vectors that appear inside the threshold (dotted line circle) will be excluded from the histogram summation. b.) Eyebrows during neutral expression and c.) Eyebrows during micro-expression AU4 activation.

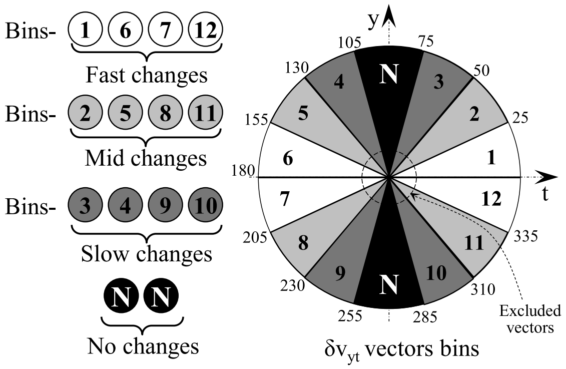


Fig. 7 δv_{yt} and δv_{xt} orientations histogram are split to 12 equal size bins and additional 2 “No changes” bins. “No changes” bins are not used in the descriptor calculation. All the vectors that appear inside the threshold (dotted line circle) will be excluded from the histogram summation.

The changes of the skin surface can be observed by examining the changes in the gradient vectors' density, magnitude and orientation. The computation of vectors histogram provide stable quantization values for our descriptor.

Illustration of our descriptor is presented in the following example, the eyebrow lowering expression is described as Action-Unit 4 (AU4) by FACS [20]). In Fig. 6 b. we can see the eyebrow area during a neutral expression and Fig. 6. shows the activation of AU4. Figure 6 b and Fig. 6 c illustrate the changes of the gradient magnitudes and orientation of δv_{yx} during the micro-expressions.

The gradient orientation histograms for δv_{yt} (vertical changes) and δv_{xt} (horizontal changes) is shown in Fig. 7. The figure also illustrates the bins' separation between the subgroups that are used only for explanatory purpose. As the separation between δv_{xt} and δv_{yt} surfaces are identical, the following explanation applies to both cases.

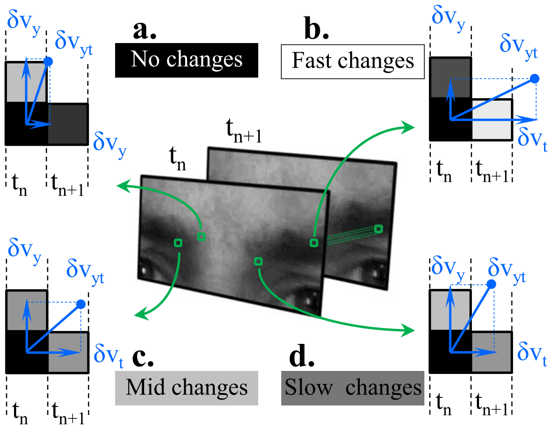


Fig. 8 Example of four gradient vectors in yt surface that are associated to subgroups a. "No changes", b. "fast changes", c. "mid changes" and d. "slow changes".

Radial segmentation is done as follows: Vectors corresponding to no or small changes between frames are eliminated from descriptor calculation. These vectors point vertically to a positive or negative direction, resulting into two "no change" bins. Changes in the x or y direction and in the time direction t may correspond to a positive or a negative direction, resulting in four quadrants. These four quadrants are further split into three parts of equal size, resulting in 12 additional bins, for a total of 14 bins.

Each bin can be further described: "No changes" regions contain all the gradient vectors whose change rate in direction t is small, and they indicate no change between the frames in the corresponding pixel. In addition we can say that δv_y (or δv_x)) part of gradient vector in the "No changes" regions was already included in the δv_{xy} histogram, so we did not include them in the δv_{yt} and δv_{xt} histograms (see Fig. 8 a).

We define the "fast changes" subgroup as the group containing gradient vectors which have small changes in adjacent pixels along the y (or x)direction but have a significant changes in t. This means that the corresponding pixels have a significant change only in their intensity between frames (see Fig. 8 b).

The "mid changes" subgroup contains gradient vectors with similar change rate in y (or x)) and t directions. (see Fig. 8 c).

The "slow changes" subgroup contains gradient vectors that indicate a strong change in y (or x)) direction and relatively small change in t (see Fig. 8 d).

In summery, we consider that vectors in the "no changes" regions indicate no or little motions in the face between the frames. Vectors inside "fast changes" subgroup indicate big change in pixel intensity between the frames and can represent motions such as blinking and eye movements. The "mid changes" and "slow changes" subgroups vectors describe motions such as the appearances or disappearances of skin folds on the face surface. Similarly as in the δv_{xy} histogram, all vectors whose size is smaller than a threshold value are excluded from the histogram summation.

The final step is the concatenation of the δv_{yx}, δv_{yt} and δv_{xt} histograms from the same frame. Consequently the motion between every frame in the video $v(x, y, t)$ is described by a 32-dimensional descriptor vector. Each value in the descriptor vector is the sum of the magnitudes of the gradient vectors in the corresponding bin. A scheme of the orientation histogram calculation can be seen in Fig. 2.

We can note that δv_{xt} and δv_{yt} correspond to change of intensity of pixels in both time and spacial direction. As we work under the assumption of constant lighting and temperature conditions, and as head and other movements happen at a speed negligible compared to those of micro-expression, we can consider that changes in intensity of pixels in subsequent frames correspond to changes in the facial skin surface caused by the motion of facial features.

3.4 Descriptor Classification and AU Detection

The detection of AUs is done in two steps, first, a descriptor classification and second, a voting procedure.

First, after calculating descriptors for every frame in the video cube $v(x, y, t)$, all the frames are classified based on the k-mean algorithm.

The descriptor classification procedure was as follows: Based on leave-one-subject-out cross-validation approach we cluster the training data using the k-mean algorithm, the number of clusters is determined by the number of AU that appeared in each one of the video cube types. Next, the label for each cluster is extracted based on the closest 50% of the points to the cluster center. Finally, the descriptor is classified according to its distances to the clusters calculated from the training data. Euclidean distance is used during training and classification steps.

The k-mean algorithm clusters both different AUs of the video cube, and *Onset*, *Apex*, and *Offset* of each AU. *Onset*, *Apex*, and *Offset* are phases of the micro-expressions. *Onset* is the phase where muscles go from their neutral state until they reach expressions' *Apex*. In the *Apex* phase the muscles sustain their contraction until the *Offset* phase. During the *Offset* phase the muscles relax and go back to there neutral state.

In spite of of the simplicity of the presented approach it provides adequate classification results. In addition, in general in case of partial labeled data the use of semi-supervised k-mean will be more suitable in comparison to other standard supervised classifiers.

Second, it has been shown in [19] that the presents of some AU, compared to AU0 (neutral face), can be detected with high accuracy. Moreover, frames that belong to the *Onset* and *Offset* phases have higher accuracy classification, in comparison to frames corresponds to *Apex*. We can assume that frames during a full AU section (video sequence containing *Onset*, *Apex* and *Offset* of the AU) belong to the same AU or to the same AU combination. Based on this, we attribute "phase weights" w to each phase in the voting

Weight (w)			1	2	2	2	2	2	1	1	1	1	1	1	1	1	1	2	2	2	2	2	1	
AU Tag + Phase tag			Onset	Onset	Onset	Onset	Onset	Onset	Onset	Apex	Apex	Apex	Apex	Apex	Apex	Apex	Offset	Offset	Offset	Offset	Offset	Offset	Offset	
		AU0	AU7	AU4	AU4	AU4	AU4	AU4	AU4	AU4	AU7	AU45	AU45	AU45	AU7	AU7	AU45	AU45	AU4	AU4	AU4	AU4	AU7	AU0
Frames	.	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	.

$$\begin{aligned}\text{AU4} &= 9 \cdot 2 + 1 \cdot 1 = \boxed{19} \leftarrow \text{Winner}\\ \text{AU7} &= 5 \cdot 1 = 7\\ \text{AU45} &= 1 \cdot 2 + 4 = 7\end{aligned}$$

Fig. 9 The final step of the voting procedure for an AU. The table represents the sequence of frames of AU4 classified using k-mean. The upper row in the table presents weights that are associated with each frame. The calculation underneath the table represents the voting procedure.

procedure.

After the first and last frames of the AU section were detected and each frame in the section was classified, the voting procedure is performed. First, each frame is associated with a weight value w based on its location in the AU section. Then, we add the votes of the frames with the same AU tag, finally the AU that gets the highest vote value is associated to the section. This process is illustrated in Fig. 9.

In order to set the values of the weight for the voting procedure, first we observed in the frame-to-frame classification results (presented in [19]) that the descriptors corresponding to *Onset* and *Offset* phases were more reliable than descriptors corresponding to the *Apex* phases. Therefore we considered that the related weight of the *Onset* and *Offset* phases should be higher. Furthermore, we observed that the maximum ratio between duration of *Onset*, *Apex*, *Offset* are of the order of 1/4, 1/2, 1/4 respectively. Therefore setting a weight of value "2" to *Onset* and *Offset* frames and of value "1" to *Apex* ensures that, in case of maximum ratio and complete miss-classification of *Apex* frame, AU will still be classified correctly. Different strategies for setting the weight values were tested, however no significant improvements in the classification results were observed.

In addition, in some cases during micro-expressions, multiple *Apex* phases may appear, for example *Neutral* → *Onset* → *Apex* → *Onset* → *Apex* → *Offset* → *Neutral*. Our voting procedure is suitable for recognizing such cases.

3.5 Descriptor Parameters

This section summarises and discusses the tuning process of the three groups of parameters of the proposed descriptor.

The first group of parameters is the the threshold values used for eliminating vectors of small magnitude from the calculation of histograms corresponding to the δv_{xy}, δv_{xt}, δv_{yt} plans. The source of these vectors are highly grainy textures such as skin with facial hairs, wrinkles, or freckles. Due to the summation process during histograms calculation, a large amount of these vectors can influence the value of the descriptor. The threshold values were determined by maximizing, in cubes with highly grainy textures, the distances between descriptors of the frames corresponding to the still face and descriptors of the frames corresponding to facial expressions.

The second group is composed by the parameters of the Gaussian smoothing function that is essential for dealing with the high level of white noise produced by the high speed camera. In addition, the smoothing process reduces the influence of the strong edges produced due to the depth of certain facial features such as eyes and nose. The smoothing factor has a strong influence on the descriptor value and therefore should be carefully determined. Based on the training data for each type of cube the parameters were tuned so that the distances between the descriptors inside the clusters were minimized the and the distances between the centers of the clusters were maximized.

The third group of parameters is the number and spread of the histogram bins in the δv_{xy}, δv_{xt}, δv_{yt} plans. Various bins spread configurations were tested during the development process. The bin configuration resulting in the best descriptive results was selected.

4. Database

In this section we present the creation and extended annotation of a new micro-expression hi-speed video database.

4.1 Available Databases

Facial expressions analysis research suffers from lack of extensive databases for training, validation and comparison between different approaches [21].

A popular database of facial spontaneous expressions is RU-FACS [22]. This database was created using a 'false opinion' paradigm and contains 100 subjects. The subject's faces were captured by four synchronized Dragon cameras by Point Grey, whose maximum frame rates with 640×480 resolution are 30 fps. Until now only 33 subjects have been FACS-coded. There are two more, relatively small, facial spontaneous expressions databases [23], [24]. However, it is very difficult to estimate the quality of the tags in small scale data. Another group consists of databases containing only pseudo-spontaneous expressions. Cohen and Kanade's DFAT-504 [25] contains videos of 100 university students. The emphasis was put on regular facial expressions (not micro-expressions). MMI Facial expression [6] is another fast growing database, that contains 300 manually coded frame-by-frame annotations. In 2010, these two databases started providing phase tags. In [26] the author presents an overview of some additional databases, including ones that were captured by IR cameras. To our knowledge, no database dedicated to hi-speed video of facial expression has been proposed yet.

4.2 Database Creation

Two basic approaches can be used for setting the ground truth for facial expressions in the videos. The first is done by using descriptive tags such as FACS as it was done in previously presented databases. The other is performed by mea-

Fig. 10 Example frames of our hi-speed micro-expression database. Database contains 10 university student subjects (5 Asian, 4 Caucasian, 1 Indian).

suring the displacement of predefined feature points on the face and reconstruction of 3D shape of the face [11]. First, IR markers are attached to the face (ther markers are invisible to regular cameras). Then a calibrated set of 3 near-infrared cameras together with visible light cameras captured the face motion. In a last step the 3D location of each marker is triangulated based on near-IR cameras and this information is used as ground truth. In this research FACS seems to be more suitable as it can subsequently be easily related to emotional states. In addition, the use of strong lights in our system (for minimizing the noise from the hi-speed camera) made it impossible to replicate the data collection procedure based on IR pen markers from [11].

The lack of a micro-expression database captured by hi-speed camera motivated us to create our own video database.

We separated the database creation task in two stages. The first stage targets videos of posed facial expressions that will allow us a evaluation of our algorithm. The second stage includes more realistic videos, that will be acquired in the near future in collaboration with psychologists.

This paper focuses on the results of the first stage of the database that contains posed micro-expressions. For video capturing a Grasshopper camera by Point Grey was used. Camera settings are: 480×640 resolution, 200 fps, RAW8 mode (in this mode minimum internal signal processing is done allows it to reach 200 fps).

McCabe's [27] recommendations for mugshot and facial image filming were used as the guidelines for video face recording. Three lights were used for shadow cancelation, left, right and hight lights with diffusion sheets to minimise hot spots on the facial image. Uniform background approximately 18% gray was used and the camera was rotated 90 degrees (640×480) to maximize the amount of pixel on the face region. The database contains video of 10 university student subjects (5 Asians, 4 Caucasians, 1 Indian, with average age of 25 years old with standard deviation of 4), (see Fig. 10). The participates were trained to perform mild and subtle facial expression. After manually cutting each one of the expression to separate sections, expression that were similar to micro-expression were added to the database. The extracted video cubes from the section correspond to one AU, starting from neutral expression, going through all three phases of AU and going back to neutral. The average length of the section is 0.51 sec. with standard deviation of 0.2 sec.

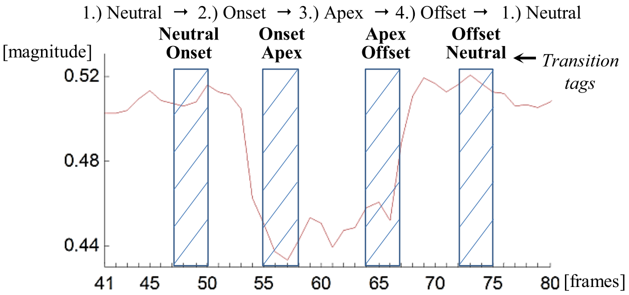


Fig. 11 Magnitude changes of one of the representative bins in our descriptor during micro-expression. The text represents a standard phase tags described by FACS, starting from 1.) Neutral, followed by 2.) Onset, 3.) Apex, 4.) Offset and back to 1.) Neutral. In bold are the new transition tags, they correspond to boundary frames for which it is ambiguous to which phase of micro-expression they belong. The frames with transition tags are represented within a rectangle.

4.3 AU's Boundary Phase Tags

Recently, FACS tagged databases [6], [25] started providing additional AU phase tags such as *Onset*, *Apex* and *Offset*. For that type of tagging every AU is split into 5 sections: starting from 1.) *Neutral*, followed by 2.) *Onset*, 3.) *Apex*, 4.) *Offset*, and back to 1.) *Neutral*. (see Fig. 11 Bottom row).

When using 200 fps camera, transition between the AU phases happens over several frames. Selection of a single frame representing the sharp border between phases during the tagging process would introduce an arbitrary element in the ground truth. In order to avoid this issue, we introduce additional transition tags: *Neutral-Onset*, *Onset-Apex*, *Apex-Offset*, and *Offset-Neutral* (see Fig. 11 upper row). During the classification step, the frames corresponding to transitional tags are considered classified correctly when classified to either of its neighboring phase tags. This approaches allows to limit issues related to arbitrary elements inherent to manual tagging processes and leads to more reliable results.

Next we shortly explain the process for determining the transition frames in the video. First, the tagger view the video section and tagged the last frame that contained a certain AU phase of the expression. Next the tagger was asked to jump approximately 15 frames forward, to the next phase of the expression and was asked to view the video in reverse and to tag the last frame that belonged to that AU phase. The frames between the first and the second tag during this process were defined as transition frames.

5. Experiment Results

In this section we first present the analysis of our 3D gradient histogram descriptor based on frame-by-frame classification results of the k-mean algorithm, followed by a comprehensive analysis of miss-classified frames. Next, the results of our AU recognition are presented. Finally, an example of micro-expression timing characteristics measure-

Table 1 Frame by frame classification results.

Facial Cubes	AU	Onset	Apex	Offset	Neutral
a) Forehead	AU2	0.93	0.95	0.91	0.95
b) Eyebrows	AU4	0.84	0.83	0.9	0.93
	AU5	0.86	0.83	0.85	
c) Eyes	AU4	0.84	0.81	0.84	0.92
	AU7	0.86	0.8	0.81	
	AU43	0.85	0.85	0.84	
d) Between the eyes	AU4	0.93	0.9	0.84	0.9
e) Lower nose	AU10	0.95	0.93	0.95	0.94
f) Mouth	AU12	0.81	0.85	0.85	0.88
	AU24	0.81	0.79	0.79	
	AU26	0.83	0.67	0.8	
g) Mouth corners	AU13	0.85	0.81	0.89	0.83
h) Chin	AU17	0.83	0.83	0.84	0.89

ments based on the proposed descriptors is given.

Current version of the algorithm was implemented in Matlab using "Piotr's Image & Video Toolbox" [28] without special emphasis on performance. However, the parallel structure of the 3D gradient histogram descriptor benefits from implementation on a parallel architecture hardware such as GPGPU.

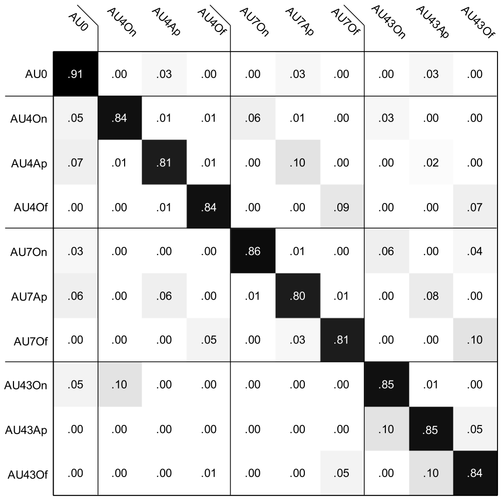
5.1 Frame-by-Frame Classification

First, facial video cubes from our database were divided into 8 groups: (a) forehead, (b) left and right eyebrows, (c) left and right eyes, (d) between the eyes, (e) lower nose, (f) mouth, (g) left and right mouth corner, (h) chin. Only one expression appears across every video cube, and in total we obtained 15000 frames from all the cubes. The 3D-Gradient historian descriptor with k-mean algorithm (without the voting procedure) was used and then all the frames were classified and compared against the ground-truth.

In Table 1 we report the classification results for each group. The results indicate good classification precision in cubes of the (a) forehead, (d) between the eyes, (e) lower nose. This is due to the small number of AU classes and the fact that they differ greatly from each other. The group (h) showed lower rates due to the beards on two faces on paticipants in the database. The mouth cube (f) shows the lowest recognition rate, which is consistent with many other works on facial expression recognition that report that mouth movements are the most challenging for classification.

By analyzing the results in Table 1 we can see that *Onset* and *Offset* phases, in most of the AUs, have higher classification precision than the *Apex* phase. This indicates that the proposed descriptor is more suited for motion recognition and segmentation than for classification of the static frames such as during the *Apex* phase. In addition, the AU0 (frames with *Neutral* expression) tags are detected with high accuracy.

The confusion matrix presents the classification of the frames in video cubes of the *Eyes* (see Fig. 12). It is important to note that for most of the AU's that were miss-classified during the *Apex* phase, there phase was detected correctly. Similar classification behavior can be seen in the *Onset* and *Offset* phases.



	AU0	AU4On	AU4Ap	AU4Of	AU7On	AU7Ap	AU7Of	AU43On	AU43Ap	AU43Of
AU0	.91	.00	.03	.00	.00	.03	.00	.00	.03	.00
AU4On	.05	.84	.01	.01	.06	.01	.00	.03	.00	.00
AU4Ap	.07	.01	.81	.01	.00	.10	.00	.00	.02	.00
AU4Of	.00	.00	.01	.84	.00	.00	.09	.00	.00	.07
AU7On	.03	.00	.00	.00	.86	.01	.00	.06	.00	.04
AU7Ap	.06	.00	.06	.00	.01	.80	.01	.00	.08	.00
AU7Of	.00	.00	.00	.05	.00	.03	.81	.00	.00	.10
AU43On	.05	.10	.00	.00	.00	.00	.00	.85	.01	.00
AU43Ap	.00	.00	.00	.00	.00	.00	.00	.10	.85	.05
AU43Of	.00	.00	.00	.01	.00	.00	.05	.00	.10	.84

Fig. 12 Confusion matrix of classification results of AUs *Onset*, *Apex*, *Offset* and neutral phases related to eyes movements: AU4 (brow lowerer), AU7 (lid tightener), AU43 (eyes closed) and AU0 (neutral). Cells get darker as the corresponding rate gets higher. The classification was done using the leave-one-subject-out cross-validation approach. In the figure 'On' refers to *Onset*, 'Ap' to *Apex* and 'Of' to *Offset*.

Table 2 Miss-classified frames across 15000 frames.

	Onset	Apex	Offset
Total frames	4000	7000	4000
Miss-classified frames	875	2216	821
Classification rate	78.13%	68.34%	79.48%
Miss-classified frames (TT)	799	2131	726
Classification rate(TT)	80.02%	70.99%	81.85%

TT - Transition Tags were added

In conclusion, the high accuracy in AU0 detection and correct detections AU phase in miss-tagged frames led us to introduce the voting procedure that will result in higher final classification rates (see Sect. 5.3).

5.2 Miss-Classified Frames Analysis

This section presents the correlation between miss-classified frames and subjects, and between miss-classified frames and AU phase.

First, based on our 15000 video cube frames that were classified frame by frame using 34 different tags (AU0 + 11 AUs each with 3 phases), we present the number of miss-classified frames during each phase in Table 2. Additionally we present the classification results by use of Transition Tags (TT), this tags helps to remove the errors introduced by manual tagging process.

Next, the same miss-classified frames are shown corresponding to subject identity. By doing so the stability of the algorithm across different subjects is checked. In Fig. 13, for each subject we present miss-classified frames during

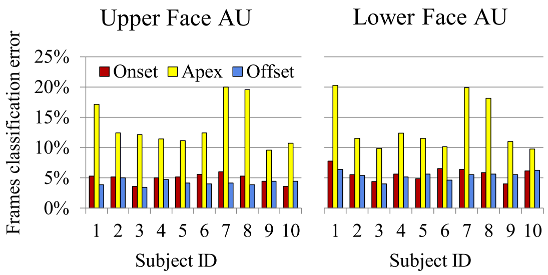


Fig. 13 Miss-classified frames across subject ID, upper face AUs on the right side and lower face AUs on the left side. *Onset* and *Offset* phases have low correlation to the subjects, contrary to *Apex* phase which error influenced by the subject.

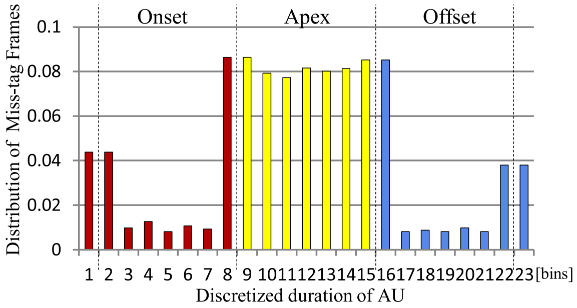


Fig. 14 AUs happen over a different number of frames. To represent the distribution of miss-classified frames over the duration of the AU, all AU video sequences were discretized into 23 bins. One bin for the beginning of the AU time sequence, seven bins per phase and and one at the end. Results are discussed in Sect. 5.2.

the *Onset*, *Apex*, and *Offset* phases, upper and lower AUs are represented separately. We can see that the miss-classified frames during the *Onset* and *Offset* phases have low correlation to the subjects. In contrast the errors during the *Apex* phase are influenced by the subject. However, the errors during *Apex* phase don't have a strong influence on the final recognition results, see Sect. 3.4.

In the second analysis we investigated how the miss-classification error was distributed over the AU phases. This clarified the reliability of our descriptor across the different AU phases. The distribution is presented in Fig. 14. First, we could see that the classifications during the *Onset* and *Offset* phases were more reliable than in the *Apex* phase. Second, there were more miss-classified frames in the transition stage between the phases (bins 2, 8, 16, and 22), indicating that classification of frames in the transition stage in a frame by frame fashion remained a hard task. Third, the miss-classified frames during the *Apex* phase were spread homogeneously, suggesting that miss-classification of AU happens across all *Apex* phases.

The difference of distribution between phases is that contrary to the *Onset* and *Offset* phases for which all the 32 bins in our descriptor are giving information about the frame, the *Apex* phase the descriptor relies on only 8 bins corresponding to the shape of the surface δv_{yx} histogram. The values of rest of the bins from δv_{yt} and δv_{xt} histograms will be zero due to lack of motions. All these results are consistent with the behavior of our descriptor.

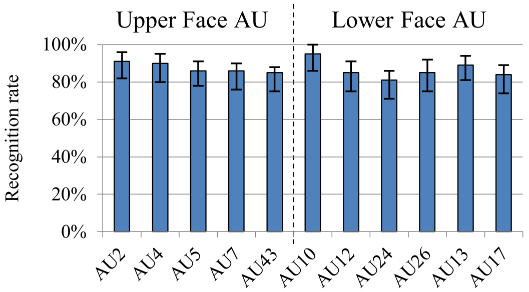


Fig. 15 Recognition rate of 11 AUs after the voting step, 5 upper face and 6 lower face.

5.3 AUs Recognition Rate

Calculation of the final AUs recognition rate is based on a leave-one-subject-out cross-validation approach, and the presented results are rate averaged over all the trials. The presented classification is the output of a voting procedure applied on frames previously classified by k-mean algorithm. (see Sect. 3.4). The recognition rate of 11 AUs is presented in Fig. 15.

The recognition rate of the 11 AUs for micro expressions analyzed by us are similar and sometimes better than the recognition for the same AUs for full expression using other state-of-the-art approaches [7], [29] and comparable to more recent works such [16], [17], [30].

The comparison was possible due to the similarity of the captured scenes (frontal view of the face without significant head motions), between our hi-speed video database and databases such as [25], [6]'s with posed expressions, and [22] with spontaneous expressions commonly used in AU detection evaluation.

5.4 Measurement of AU's Timing Characteristics

The changes of the bins' values in the 3D gradient oriented histogram reflect the changes and the motion accelerations of facial movements. Therefore, timing characteristics of micro-expressions can be measured directly from the descriptors' values. Figure 16 shows an example of the change in magnitude of the representative bins over time during a micro-expression. In this example, the magnitude of YT-bins (1 and 8) and XY-bin (1 and 5) of our descriptor during AU4 provide a quantitative measurement of the movement of the eyebrow over time. After classification it become clear that the duration of *Onset* is 29 frames (approximatively 0.145 seconds), the duration of *Apex* is 23 frames (approximatively 0.115 seconds), and the duration of *Offset* is 25 frames (approximatively 0.125 seconds). A possible use for such a magnitude/time profile is the distinction and classification between posed and spontaneous micro-expressions. Such a use remains to be further investigated.

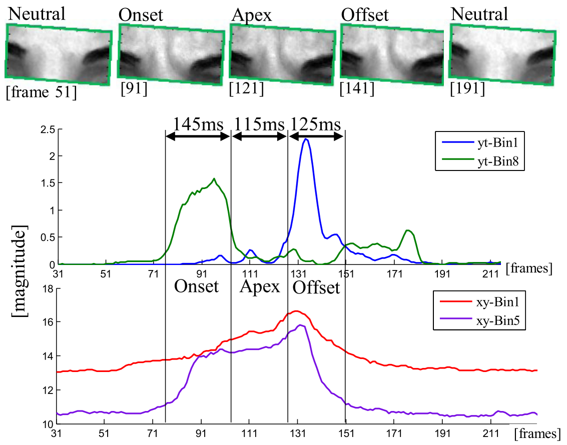


Fig. 16 Example of descriptor magnitude/time profile reflecting movements of a micro-expression. Magnitude of YT-bins (1 and 8) and XY-bin (1 and 5) of our descriptor during AU4 provide a quantitative measurement of the movement of the eyebrow over time. The top row shows frames representative of each phase during the activation and release of AU4.

6. Conclusions and Future Work

In this paper, we presented a novel approach for facial micro-expressions detection using 3D gradient histogram descriptor applied on hi-speed video captured by 200 fps camera. Experimental results indicate its particular usefulness in detecting the facial skin surface motion in hi-speed videos. Due to similarity between the scenes in our new hi-speed video database and in both Cohn-Kanade [25] and MMI [6] databases (that are commonly used in AU detection evaluation), we compared the recognition rate results for 11 AUs. Obtained recognition rate produce similar and in some cases better results compared with the state-of-the-art approaches and is targeting specifically micro-expressions.

Timing characteristics of facial expression and micro-expressions were found to be significant for psychological and behavioral analysis but have so far not been taken into account due to lack of suitable technology. Until recently state-of-the-art algorithms for facial expression detection were unable to measure such characteristics and mostly were applied on regular speed videos. The proposed descriptor is specially adapted to hi-speed video analysis and our method has the potential to provide finer description of the timing characteristics of micro-expressions. More importantly, correlation between the different bins values of our descriptors to the corresponding physical motion can provide information concerning the velocity and the acceleration of the expression.

To explore these possibilities, we are currently working in extending the database described in this paper: it will include videos captured by hi-speed, hi-resolution and infrared synchronized cameras as well as data from other other sensor for automatic ground truth extraction. Benefits of this database will be: 1) As the database will be annotated using FACS with additional transitional tags, it will allow to evaluate the performance of algorithms for time characteristics extraction 2) Synchronized 25 fps and 200 fps videos will allow comparison of our method adapted to hi-speed video with other approaches.

Also a face tracking technique based on infrared camera is under development and will complete our procedure by adding a face tracking step. In addition to face tracking, IR images will be used for masking uninformative parts of the face (see Sect. 3.2). For example the face boundaries and area of an open mouth can be easily detected in IR images.

Finally, other standard classification algorithms for AU detection will be combined with our descriptor.

References

[1] P. Ekman, "Facial expressions of emotion: An old controversy and new findings," Philosophical Transactions of The Royal Society of London, Series B Biological Sciences, vol.335, no.1273, pp.63–69, 1992.

[2] S.Z. Li and A.K. Jain, Handbook of Face Recognition, Springer, 2005.

[3] P. Ekman, Telling Lies (second edition), NORTON, 2009.

[4] S. Porter and L.T. Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," Psychological Science, vol.19, no.5, pp.508–514, 2008.

[5] P. Ekman and W.V. Friesen, "Facial action coding system: A technique for the measurement of facial movement, consulting psychologists press," 1978.

[6] M. Pantic and I. Patras, "Detecting facial actions and their temporal segments in nearly frontal-view face image sequences," IEEE Int'l Conf. on Systems, Man and Cybernetics 2005, pp.3358–3363, 2005.

[7] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, and J.R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," J. Multimedia, vol.1, no.6, pp.22–35, 2006.

[8] P. Yang, Q. Liu, and D. Metaxas, "Boosting encoded dynamic features for facial expression recognition," Pattern Recognit. Lett., vol.30, no.2, pp.132–139, 2009.

[9] S. Lucey, A.B. Ashraf, and J.F. Cohn, "Recognition through aam representations of the face," in Handbook of Face Recognition, pp.275–286, Springer, 2005.

[10] K. Daijin, Automated Face Analysis: Emerging Technologies and Research, Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2009.

[11] T.K. Marks, J.R. Hershey, and J.R. Movellan, "Tracking motion, deformation, and texture using conditionally Gaussian processes," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.2, pp.348–363, 2010.

[12] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," Comput. Vis. Image Understand., vol.108, no.3, pp.207–229, 2007.

[13] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.65–72, 2005.

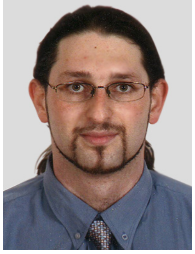
[14] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," Proc. 15th International Conference on Multimedia MULTIMEDIA 07, p.357, 2007.

[15] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," British Machine Vision Conference, pp.995–1004, Citeseer, 2008.

[16] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.11, pp.1940–1954, 2010.

[17] B. Jiang, M.F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes,"

Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), pp.314–321, Santa Barbara, CA, USA, March 2011.
[18] D. Renee, "Face proportions (front view)." http://dryggirl.com/teaching/art1/faces/faces.html, July 2012.
[19] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," pp.P16–P16, 2009.
[20] C.M.U. Robotics Institute, "Facs - facial action coding system table (2002 revision is here)," Aug. 2011.
[21] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," Face Recognit., pp.377–416, June 2007.
[22] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, and J.R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," J. Multimedia, vol.1, no.6, pp.22–35, 2006.
[23] L. Max, "Multiple aspects of discourse research lab." http://madresearchlab.org, Nov. 2009.
[24] R. Cowie and M. Schroeder, "The description of naturally occurring emotional speech," System, pp.2877–2880, 2003.
[25] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," Robotics, vol.4, pp.46–53, 2000.
[26] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, and F. Chen, "A natural visible and infrared facial expression database for expression recognition and emotion inference," Design, vol.12, no.7, pp.682–691, 2010.
[27] M. McCabe, "Best practice recommendation for the capture of mugshots." http://www.itl.nist.gov/iaui/894.03/face/bprmug3.htm, 2009.
[28] P. Dollar and V. Rabaud, "Piotr's image video toolbox for matlab." http://vision.ucsd.edu/ pdollar/toolbox/doc/, 2009.
[29] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.10, pp.1683–1699, 2007.
[30] F. Tsalakanidou, "Real-time 2d+3d facial action and expression recognition," Pattern Recognit., vol.43, no.5, pp.1763–1775, 2010.



Senya Polikovsky received a B.Sc. degree in Electrical Engineering from Holon Institute of Technology (Israel) and a M.Sc. degree from University of Tsukuba (Japan) in 2005 and 2009, respectively. He is currently a Ph.D. candidate at the Computer Vision and Image Media Laboratory. In 2011, he received a scholarship and research grant from Japan Society for the Promotion of Science. He has more than 10 years of industrial research and development experience.

Yoshinari Kameda received B., M., and Ph.D from Kyoto University in 1991, 1993, and 1999. He started working at Kyoto University at 1996. While he was at Kyoto University, he was also a visiting scholar in MIT in 2001. In 2003, He became an assistant professor of University of Tsukuba, and then an associate professor in University of Tsukuba in 2004. A member of IPSJ and IEEE. His research interest covers computer vision for human interface, augmented and mixed reality, multimedia handling.

Yuichi Ohta is a professor of the Faculty of Engineering, Information and Systems, and the Center for Computational Sciences, University of Tsukuba. He received his Doctoral degree in Information Science from Kyoto University in 1980. After holding a faculty position in the Department of Information Science, Kyoto University, he joined University of Tsukuba in 1981. In the early 1990s, Dr. Ohta started the pioneering studies on the 3D image media as a promising application field of computer vision. He was the President-elect (2008) and the President (2009) of the Information and Systems Society, IEICE. He has been elected as a Fellow of IAPR and IPSJ, in 2004, 2004, and 2007.