Macro- and Micro-Expression Spotting in Long Videos Using Spatio-temporal Strain

Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar

Abstract— **We propose a method for the automatic spotting (temporal segmentation) of facial expressions in long videos comprising of macro- and micro-expressions. The method utilizes the strain impacted on the facial skin due to the non-rigid motion caused during expressions. The strain magnitude is calculated using the central difference method over the robust and dense optical flow field observed in several regions (chin, mouth, cheek, forehead) on each subject's face. This new approach is able to successfully detect and distinguish between large expressions (macro) and rapid and localized expressions (micro). Extensive testing was completed on a dataset containing 181 macro-expressions and 124 micro-expressions. The dataset consists of 56 videos collected at USF, 6 videos from the Canal-9 political debates, and 3 low quality videos found on the internet. A spotting accuracy of 85% was achieved for macro-expressions and 74% of all micro-expressions were spotted.**

I. INTRODUCTION

When analysing the facial dynamics of an individual, many algorithms rely on the manual segmentation of expressions or other non-rigid facial movements. For example, a large number of facial expression recognition algorithms have been proposed that successfully identify several facial expressions, but do not address the critical stage of initially spotting them.

In this paper, we propose a unified method that automatically spots expressions in long videos using spatio-temporal strain. Since strain represents the deformation incurred during non-rigid motion, it directly corresponds to the deformation of facial skin during an expression. This gives it several main strengths: (i) since it is directly related to the local gradient of motion, and not global motion, it is robust to moderate amounts of head translations; (ii) strain has been shown in [8] [12] to be robust to adverse lighting conditions and heavy make-up, demonstrating its suitability for real-world application.

In our previous work on automatic expression spotting [11], we proposed two different algorithms for detecting macro- and micro-expressions, based on thresholding the strain magnitude calculated over the entire image frame; however, due to a lack of face tracking, we had reported low performance on videos containing subjects with large head movements. The dataset used for micro-expressions was also small – just 7. In this work, we significantly expand our study and present an unified approach to detect both macro- and micro- expressions. We show results on much larger datasets containing longer videos with 181 macro-expressions and 124 micro-expressions.

The proposed algorithm has four main steps: (i) facial landmarks are located automatically and the face is spatially segmented into regions; (ii) the strain magnitude is calculated for each of these regions over time; (iii) a global threshold is used on the strain magnitude to determine macro-expressions, and removes them from the sequence; (iv) a local thresholding method based on both the temporal duration and localization of the strain magnitude is used to spot micro-expressions over remainder of the sequence.

Throughout literature, algorithms that analyse facial expressions often begin after facial expression have been manually segmented. Methods that do approach the problem of automatically spotting macro-expressions typically fall into three categories [10]. The first category consists of techniques [9] that use a selective point-model representation of the face, and then analyzes and segments based on the inter-dynamics of these points over a video sequence. Poor lighting and other adverse facial conditions often cause problems for algorithms in this category, since they rely on consistent and accurate detection of several key features. The second category consists of more holistic approaches [2] that model the face in its entirety, i.e., a dense representation of all points on the face. The third category consists of methods that combine both these approaches. Our approach fits in this category, since we derive our model from the optical flow fields covering the entire face, after initially locating the face and extracting each eye coordinate.

Our approach has several advantages: (i) first, we have not discovered any other methods that attempt to spot both macro- and micro-expressions; (ii) we have restricted the limitations of the first category to just the head and eyes, for which there have been several successful head and eye trackers developed [14] [3]; (iii) the reliability of optical strain has been demonstrated in [8] and [12], even under adverse illumination and heavy make-up; (iv) lastly, optical strain can be quickly and accurately calculated from optical flow fields, so in combination with specialized optical flow hardware, it may be viable for real-time application.

II. EXPRESSIONS

Expressions typically convey the emotional state of the individual, although they may be feigned, forced, or suppressed. In general, there are two types of expressions: i) macro-expressions, which usually occur over multiple regions of the face and are easily observed, and ii) micro expressions, which are rapid and occur in a small regions

All authors are with the Department of Computer Science and Engineering, University of South Florida, Tampa, Florida 33625, USA
[mshreve, sgodavar, goldgof, sarkar]@cse.usf.edu

of the face. In the following section, we further distinguish these two types of expressions.

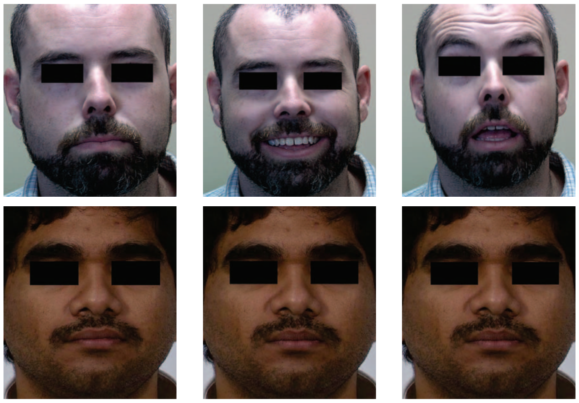

Fig. 1: Example expressions. Eyes masked for privacy concerns. The first row contains the macro-expressions smile and surprise. The second row are 3 frames containing an example type 1 micro-expression (slightly raised left cheek).

A. Macro Expressions

Macro-expressions typically last 3/4th of a second to 2 seconds. There are 6 universal expressions: happy, sad, fear, surprise, anger, and disgust. Spatially, macro-expressions can occur over multiple or single regions of the face, depending on the expressions. For instance, the surprise expressions generally causes motion around the eyes, forehead, cheeks, and mouth, whereas the expression for sadness typically generates motion only near the mouth and cheek region.

B. Micro Expressions

In general, a micro-expressions is described as an involuntary pattern of the human body that is significant enough to be observable, but is too brief to convey an emotion [5] [7]. Micro-expressions occurring on the face are rapid and are often missed during casual observation, or even sometimes extremely hard to observe. Lasting between1/25th to 1/5th of a second [6], micro-expressions can be classified, based on how an expression is modified, into three types [5]:

- Type 1. Simulated Expressions: When a micro-expressions is not accompanied by a genuine expression.
- Type 2. Neutralized expressions: When a genuine expression is suppressed and the face remains neutral.
- Type 3. Masked Expressions: When a genuine expression is completely masked by a falsified expression.

Type 2 micro-expressions are not observable and type 3 micro-expressions may be completely eclipsed by a falsified expression. In this paper, we focus on type 1 micro-expressions, i.e., micro-expressions that correspond to rapid, but observable and non-suppressed motion on the face.

III. Background

There are two main approaches for calculating optical strain: (i) integrate the strain definition into the optical flow equations, or (ii) derive strain directly from the flow vectors. The first approach requires the calculation of high order derivatives, hence is sensitive to image noise. The second approach allows us to post-process the flow vectors before calculating strain, possibly reducing the effects any errors incurred during the optical flow estimation. We use the second approach in this paper.

A. Optical Flow

Optical flow is a well-known motion estimation technique that is based on the brightness conservation principle [1]. In general, it assumes (i) constant intensity at each point over a pair (sequence) of frames, and (ii) smooth pixel displacement within a small image region. It is typically represented by the following equation:

$$(\nabla I)^T \mathbf{p} + I_t = 0 \tag{1}$$

where $I(x, y, t)$ represents the temporal image intensity function at point x and y at time t, and ∇I represents the spatial and temporal gradient. The horizontal and vertical motion vectors are represented by $\mathbf{p} = [p = dx/dt, q = dy/dt]^T$.

Since large intervals over a single expression can often cause failure in tracking (due to the smoothness constraint), we implemented a vector linking (or stitching) process that combines small, local pairs of small intervals (1-3 frames) into larger pairs to expand over the entire sequence of frames.

B. Optical Strain

The projected 2-D displacement of any deformable object can be expressed by a vector $\mathbf{u} = [u, v]^T$. If the motion is small enough, then the corresponding finite strain tensor is defined as:

$$\varepsilon = \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T], \tag{2}$$

which can be expanded to the form:

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) \\ \varepsilon_{yx} = \frac{1}{2}(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix} \tag{3}$$

where (ε_{xx},ε_{yy}) are normal strain components and (ε_{xy},ε_{yx}) are shear strain components.

Since each of these strain components are a function of displacement vectors (u,v) over a continuous space, each strain component is approximated using the discrete optical flow data (p,q):

$$p = \frac{\delta x}{\delta t} \doteq \frac{\Delta x}{\Delta t} = \frac{u}{\Delta t}, u = p\Delta t, \tag{4}$$

$$q = \frac{\delta y}{\delta t} \doteq \frac{\Delta y}{\Delta t} = \frac{v}{\Delta t}, v = q\Delta t \tag{5}$$

where Δt is the change in time between two image frames. Setting Δt to a fixed interval length, we can estimate the partial derivatives of (4) and (5):

$$\frac{\partial u}{\partial x} = \frac{\partial p}{\partial x}\Delta t, \frac{\partial u}{\partial y} = \frac{\partial p}{\partial y}\Delta t, \tag{6}$$

$$\frac{\partial v}{\partial x} = \frac{\partial q}{\partial x}\Delta t, \frac{\partial v}{\partial y} = \frac{\partial q}{\partial y}\Delta t, \tag{7}$$

The second order derivatives are calculated using the central difference method. Hence,

$$\frac{\partial u}{\partial x} = \frac{u(x+\Delta x) - u(x-\Delta x)}{2\Delta x} \doteq \frac{p(x+\Delta x) - p(x-\Delta x)}{2\Delta x} \tag{8}$$

$$\frac{\partial v}{\partial y} = \frac{v(y+\Delta y) - v(y-\Delta y)}{2\Delta y} \doteq \frac{q(y+\Delta y) - q(y-\Delta y)}{2\Delta y} \tag{9}$$

where (Δx,Δy) $\approx$ 2-3 pixels.

Finally, each of these values corresponding to low and large elastic moduli are summed to generate the strain magnitude. Each value can also be normalized to 0-255 for a visual representation (strain map). Figures 2 and 3 give illustrations of the optical flow fields and normalized strain values obtained during both a macro- and micro-expression.

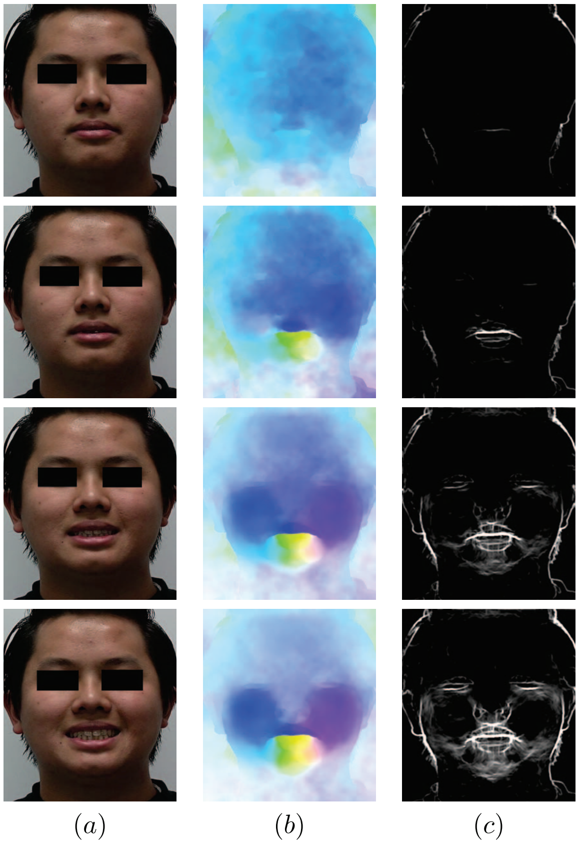

Fig. 2: (a) Example smile expression. Eyes are masked for privacy concerns. Column (b) contains the linked optical flow fields (color denotes direction and the magnitude is represented by the intensity [1]). Column (c) contains the corresponding strain maps.

IV. Algorithm

A. Facial Landmark Detection and Alignment

To avoid error due to large head translations and to increase the reliability of optical flow estimation, we use the Viola-Jones face detector [14] to detect and crop out faces from the whole frames (the final size of the cropped face is the average size observed over all frames). This also has additional benefit of speeding up the optical flow calculations due to the smaller image frame. After the face images are obtained, we perform the following steps:

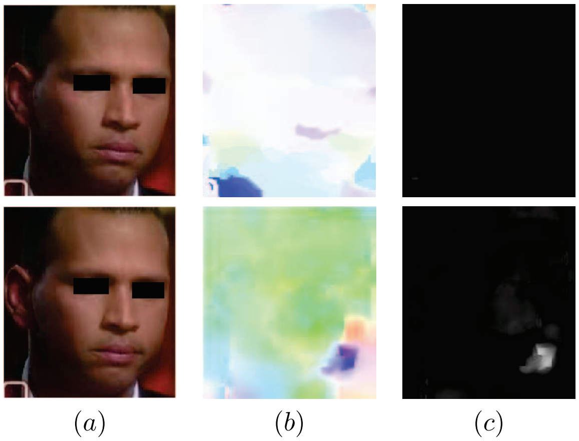

Fig. 3: (a) Example ”scorn” micro-expression, which occurs when the corner of the lip is raised slightly. Column (b) contains the linked optical flow fields (color denotes direction and the magnitude is represented by the intensity [1]). Column (c) contains the corresponding strain maps.

- Locate eyes using the OpenCV Haar classifier [3].
- Calculate centroid of the eye locations.
- Construct the line joining the centroids of the two detected eyes.
- Register all frames to starting frame by aligning each pair of lines.
- Match top left skin pixel, for increased stability [4].

The first step uses the location of the eyes as an alignment axis for removing large head rotations. Skin pixel alignment is performed to remove the slight ”jittering” caused by the 2-3 pixel difference in eye-tracking coordinates between each frame.

B. Spatial Segmentation

After aligning each face image, we divide the face into eight regions: forehead, left and right of eye, left and right of cheek, left and right of mouth, and chin (see Fig. 4.b). The eye regions of each image are masked due to the noise caused from eye saccades and blinking. The nose and mouth regions are also removed, since (i) the nose typically rigid and (ii) opening of the mouth violates the smoothness assumption in optical flow equations, leading to erroneous tracking results. Fig. 4 shows the final segmentation template.

To align the mask for segmentation, we expanded or contracted the generic mask given in Fig. 4(b) to align with the respective boundary points for each eye (the left and right halves on top of the bricked ’T’). The region of the face image above the eyes contains the forehead region. For the left and right halves of the face, we divided the remaining distance from the bottom of the eyes to the bottom of the face into two main halves (cheek region and mouth region). For region between these halves we segment the chin region starting at 3/4th of the remaining distance below the bottom of the eyes, to the end of the face.

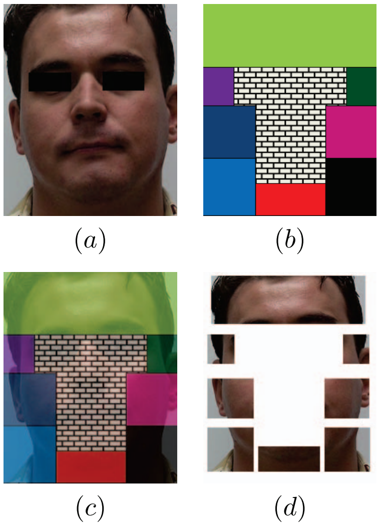

Fig. 4: Facial Segments. (a) Sample subject. (b) Each color contains a different region of the face (forehead, left and right of eye, left and right of cheek, left and right of mouth, and chin). The 'bricked' region represents the removed areas. (c) Mask alignment is performed by fitting the top of the 'T' bricked region to the rectangular boundaries of the detected eyes. (d) Final segments.

C. Optical Strain, Thresholding, and Detection

For each region of the face given in Fig. 4, the magnitude (the total sum of the strain) is calculated. The strain plots obtained for each region are then subjected to thresholding in order to spot expressions. The difference in selection criteria for macro- and micro-expressions will now be discussed.

1) Macro-expression: Since macro-expressions can occur over multiple regions of the face simultaneously, all strain values contained in all regions are summed to generate an overall strain magnitude. Next, a global threshold is chosen using the following steps:

- Fit curve (2D) to the sequence of total strain magnitude using least squares method.
- At each point, subtract this curve from the sequence.
- Fit a line (1D) to the result of previous step using least squares method.
- Intersections of this line initially correspond to expression boundaries.

To remove accumulative error due to vector linking, a 2-dimensional curve is first fit to the sequence, and then subtracted from it. Next, a line is fit to the resulting function. Time points at which the strain magnitude rises above the intersecting points are considered frames containing macro-expressions, if the interval contains at least 10 frames (approx. 1/3 of a second in duration). Fig. 5 shows the result of this procedure.

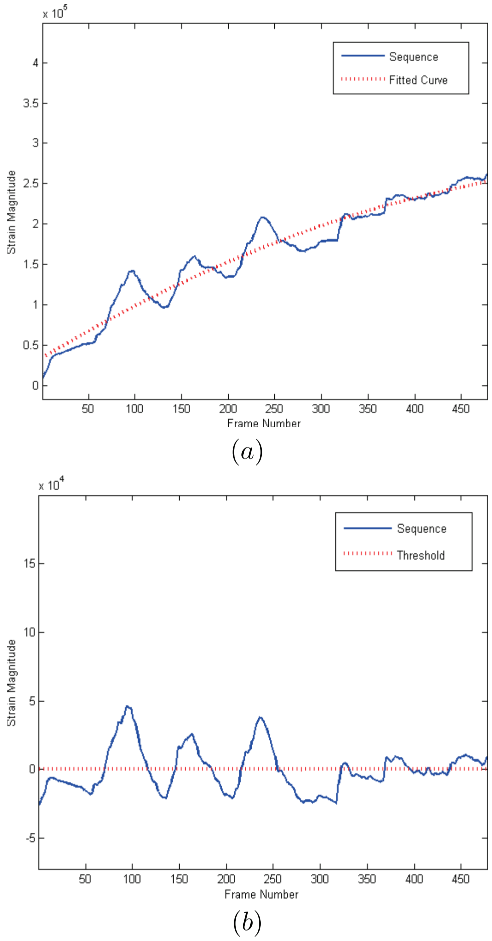

Fig. 5: Macro-expression thresholding. (a) Curve is fit to sequence of strain magnitude over all frames using least squares method. (b) This curve is subtracted from each point in sequence, and then a global threshold (line) is calculated again using least squares method.

2) Micro-expression: The thresholding technique used for micro-expressions is more constrained than that of macro-expressions. Because of their rapid and spatially localized characteristics, a local thresholding technique is needed [7]. First, segments of the video containing macro-expressions are removed using the algorithm in the previous section, in order to minimize the effects of dominate motions. Next, two additional criteria added: (i) the strain magnitude must be significantly larger than the surrounding regions and (ii) the duration of this increased strain can only be at most 1/5th of a second. The first criteria ensures the spatial property of micro-expressions in that they can only occur in one or two regions of the face, while the second enforces the maximum duration of a type 1 micro-expression given in the literature [5].

First for each of the eight regions R, the average strain $(\mu_1, \mu_2, ..., \mu_8)$ is calculated, where

$$\mu_R = \frac{1}{N}\sum_{f=1}^{N}(S_f^R) \tag{10}$$

where N is the total number of frames and S_f^R is the strain magnitude calculated at frame f and region R. Next, local peaks are detected every n frames ($n = 9$ in our experiments, or roughly 1/3 of a second). A micro-expression peak P_f^R is detected if

$$P_f^R > 2\mu_R \tag{11}$$

and over all ±4 frames around the peak frame,

$$S_f^R > \alpha \times P_f^R \tag{12}$$

and

$$S_f^R > \mu_R \tag{13}$$

where $\alpha \in (0, 1)$. Note that (12) ensures that the threshold is not set at the base of the peak, resulting in a missed micro-expression, and (13) ensures the strain magnitude is large enough. Optimal experimental results were obtained with $\alpha = .35$. Additionally, only one other region may have a peak within the same interval of n frames. After it is determined if the frames do / do not contain a micro-expression, all n frames are removed from the sequence. Then μ_R is re-calculated for all regions. Fig. 6 shows an example of a detected micro-expression.

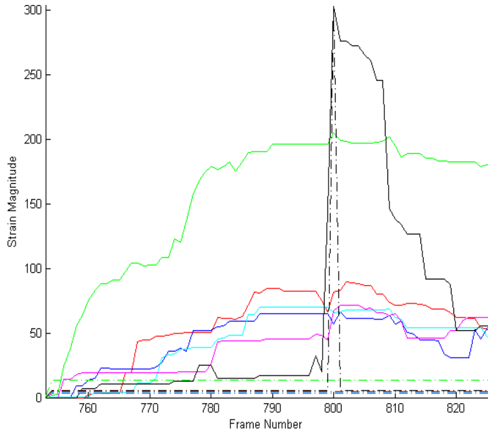

Fig. 6: Example spotted micro-expression at frame 800 (note: color is needed for this visualization). Each color corresponds to the regions of the same color given in Fig. 4. A strain peak occurred in the region to the right of the mouth, indicative of the scorn micro-expression.

V. Experiment

Experiments were performed on several datasets. We now discuss each dataset, and then report the results of macro- and micro-expression spotting.

A. Datasets

USF-HD: This database consists of 47 sequences and contains 181 macro-expressions (smile, surprise, anger, sad) and 100 micro-expressions. Videos were collected either by a JVC-HD100 or a Panasonic AG-HMC40 camcorder at a resolution of 720 x 1280 and frame-rate of 29.7 fps. The length of each video is on average approximately 1 minute in length (the longest is near 2 minutes and the shortest 20 seconds). Subjects were recorded under normal lighting conditions and asked to perform both macro- and micro-expressions. For micro-expressions, subjects were shown some example videos containing micro-expressions prior to being recorded. The subject was then asked to mimic them in any order and to avoid out-of-plane head motion.

Canal-9: The Canal-9 database [13] contains 72 political debates recorded by the Canal 9 local TV station and was broadcast in Valais, Switzerland. The videos are recorded in HD format. There are up to five participants in each debate and were often asked yes / no questions on political issues. Twenty four sequences containing a micro-expression were selected (each around 6 seconds). Several angles and rotations appear in the videos, however the clips chosen for our experiments contain no change in pose over the sequence.

Found Video: This collection contains examples of micro-expressions used by Ekman [5] that we obtained from the internet. They include the English spy Kim Philby's last public interview and Alex Rodriguez's interview with Katie Couric where he denies taking drugs. Overall, this dataset consists of 4 micro-expressions and each clip lasts about two seconds.

There two things to note about the Found Video and Canal-9 datasets: (i) ground-truthing was performed by two annotators who have studied micro-expressions but are not formally trained at spotting them, and (ii) the facial segmentation process given in section IV (b) has to be assisted manually in segments that contain large out-of-plane alignments. Our goal is to make this an automated process in future work.

Training

For macro-expressions, training was not needed since our threshold is calculated from the polynomials found using least squares distance. For micro-expressions spotting, training was needed to determine the threshold (α) given in equation 12. Training was performed on 20% of the micro-expressions in the USF-HD dataset to arrive at the final value given in section IV.

B. Results

The results for both macro- and micro-expression spotting can be found in Table I. Overall the results are positive for all datasets, with peak results achieved on the USF-HD dataset, where 154 (85%) macro-expressions in 35 sequences and 77 (80%) micro-expressions in 12 sequences were successfully spotted. This is mainly attributed to the controlled environment in which subjects were recorded (even lighting, only moderate head movement). Also, each sequence collected in this dataset only contains expressions. Fig. 7 contains an example of a spotted macro-expressions, and Fig. 8 contains an example of spotted micro-expressions.

A total of 12 (50%) of genuine micro-expression were found in the Canal-9 political debates, however at the expense of a higher false alarm rate (also 50%). The lower micro-expression spotting accuracy in this dataset can be mainly attributed to numerous head movements, and talking

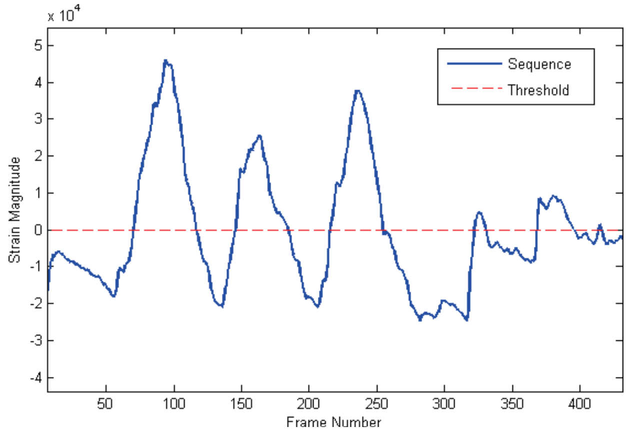

Fig. 7: Example results for macro-expression spotting. Expression intervals are defined by the boundary intersection points at threshold. Five expressions were accurately segmented with one false alarm at interval (420,424).

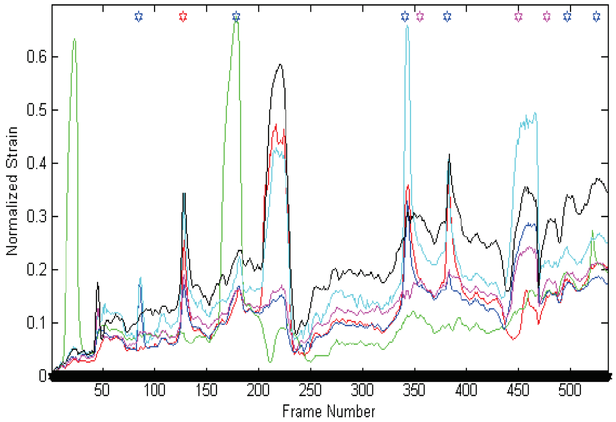

Fig. 8: Example results for micro-expression spotting (note: color is needed for this visualization). Hexagons indicate spotted micro-expression, in the region corresponding to the color given by the mask in Fig. 4. False positives appear at the hexagons located at frames 73, 358. A missed spotting occurs at frames 24 and 220. Strain magnitude values were normalized for this illustration. Detection is shown only for the first detected region.

which was often misclassified as a micro-expression. For the Found Videos, three out of the four micro-expressions were detected, at a 0% false alarm rate.

VI. Conclusions

In this paper, we proposed a method for the automatic spotting of facial expressions videos comprised of numerous facial expressions. The method relies on the magnitude of the strain observed on the facial skin as a subject performs an expression. This new approach is able to successfully detect and distinguish between regular, universal macro-expressions and rapid micro-expressions. An accuracy of 85% was achieved for spotting macro-expressions, and 74% of all micro-expressions were successfully spotted. Testing was performed on a variety of videos including our own collected data, Canal-9 political debates, and found videos.

TABLE I: Expression spotting results

Dataset	# Seq.	# Macro / Micro	TP	Missed	FP
USF-HD	44	181 (Macro)	154	27	7
USF-HD	12	96 (Micro)	77	19	36
Canal-9	6	24 (Micro)	12	12	18
Found Videos	3	4 (Micro)	3	1	0
Total (Mac)	44	181	85%	15%	4%
Total (Mic)	25	124	74%	26%	44%

References

[1] M. J. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. In *Computer Vision and Image Understanding*, volume 63, pages 75–104, New York, NY, USA, 1996. Elsevier Science Inc.

[2] M. J. Black and Y. Yacoob. Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion. In *Computer Vision*, volume 25, pages 23–48, 1997.

[3] G. Bradski and A. Kaehler. Learning OpenCV. O'Reilly Media, 2008.

[4] C. Conaire, N. O'Connor, and A. Smeaton. Detector adaptation by maximising agreement between independent data sources. In *Computer Vision and Pattern Recognition*, pages 1–6. CVPR, 2007.

[5] P. Ekman. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised and Updated Edition). In *W.W. Norton and Company*, 2001.

[6] P. Ekman, E. T. Rolls, D. I. Perrett, and H. D. Ellis. Facial Expressions of Emotion: An Old Controversy and New Findings [and Discussion]. In *Philosophical Transactions: Biological Sciences*, volume 335, pages 63–69, 1992.

[7] S. Godavarthy. Micro-Expression Spotting in Video Using Optical Strain. Masters Thesis. University of South Florida, 2010.

[8] V. Manohar, D.B. Goldgof, S. Sarkar, and Y. Zhang. Facial strain pattern as a soft forensic evidence. In *Workshop on Applications of Computer Vision*, page 42, Austin, TX, USA, 2007. IEEE Computer Society.

[9] C. Padgett and G.W. Cottrell. Representing Face Images for Emotion Classification. In *Advances in Neural Information Processing Systems*, pages 894–900, 1996.

[10] M. Pantic and L. Rothkrantz. Pattern Analysis and Machine Intelligence, IEEE Transactions on. In *Automatic analysis of facial expressions: the state of the art*, volume 22, pages 1424 –1445, 2000.

[11] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar. Towards macro- and micro-expression spotting in video using strain patterns. In *Workshop on Applications of Computer Vision*, pages 1–6, 2009.

[12] M. Shreve, V. Manohar, D. Goldgof, and S. Sarkar. Face recognition under camouflage and adverse illumination. In *Biometrics: Theory, Applications, and Systems, First IEEE International Conference on*, 9 2010.

[13] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops*, pages 1–4. ACII, 2009.

[14] P. Viola and M. Jones. Robust Real-Time Face Detection. In *International Journal of Computer Vision*, volume 57, pages 137–154, 2004.