Target Audience Response Analysis in Out-of-home Advertising Using Computer Vision

Alexandru Costache
Faculty of Automatic Control and Computer Science
University Politehnica of Bucharest
Bucharest, Romania
alexandru.costache0909@gmail.com

Dan Popescu
Faculty of Automatic Control and Computer Science
University Politehnica of Bucharest
Bucharest, Romania
dan.popescu@upb.ro

Stefan Mocanu
Faculty of Automatic Control and Computer Science
University Politehnica of Bucharest
Bucharest, Romania
smocanu@rdslink.ro

Loretta Ichim
Faculty of Automatic Control and Computer Science
University Politehnica of Bucharest
Bucharest, Romania
loretta.ichim@upb.ro

***Abstract*— In this paper we aim to analyze the response of the target audience to out-of-home advertising panels mounted in display windows. We track people passing in front of the panels and, if they approach the panel, looking directly at it, we analyze their facial movements, to see whether the advert appeals to them or not. The approach is based on video data gathering using a static IP camera, storage and analysis, aiming to deliver real-time results. We use neural networks and support vector machines to determine facial microexpression. Results are correlated with information about the adverts displayed to be of use to advertising providers.**

Keywords— emotion analysis, image processing, neural network, object tracking, out-of-home advertising.

I. INTRODUCTION

Out-of-home (OOH) advertising aims to deliver content to a target audience in outdoor (street advertising) and indoor (commercial buildings) environments. Computer vision can aid OOH advertising providers in obtaining the maximum reach for their business, both by optimizing geospatial positioning of their panels, based on detected pedestrian traffic in the area, and by optimizing the adverts shown on the panels, based on the way pedestrians react to them. We aim to aid the providers by offering them precise statistics regarding the number of people passing by their panels and the reactions that those people have when they view certain adverts on those panels. To do so, we intend to use IP cameras for tracking people passing by and for analysing their facial expressions when they approach the panel. This information aggregated with known data about the advert being shown can help the providers understand what the public wants to see. For video processing, neural networks are used.

Identification of human faces and analysis of emotions visually transmitted by them has been an important field of research for more than half a century. Many algorithms were proposed for processing images in order to extract features which would point emotions. A breakthrough in the field was the introduction of neural networks. One of the early works in emotion determination is [1], in which the authors used feature-based recognition networks and achieve good results in gender discrimination, but not so good in emotion recognition due to similarities in the feature descriptors. As technology progressed, researchers were able to identify emotions, with better results achieved for positive emotions (e.g. happiness) which are visually easier to determine [2], [3]. Contemporary researchers propose two important approaches in facial expression recognition (FER). The first, conventional approach follows feature learning, feature selection and classifier construction which are then used with machine learning classifiers (e.g. Support Vector Machines) [4]. The second approach is based on deep learning, usually using neural networks [5]. Researchers also propose various training datasets, containing either static or dynamic images [5], [6]. Two different directions of research are prevalent in the field: some authors study improvements of feature descriptors which are regularly used on static images, as the focus is to obtain good results using few resources [4]; other authors use neural network architectures to achieve fast, reliable results, while sometimes employing pre-processing of input data [7], [8]. Paper [9] presents a modern representation method for facial expression features, modelling spatial-temporal changes in face muscles over a period. A good, modern, real-time solution for FER is given by [10], authors obtaining over 95% accuracy using CNNs trained with full images instead of pre-extracted features.

During our previous work we tried improving OOH advertising business activities by studying the audience [11]. Apart from tracking the persons to see the areas with increased traffic, gathered data to improve advertising panel geospatial position is used [12]. However, there are situations in which we can't further improve the support of advertising, e.g. when the panel is positioned inside a store window. Thus, we must rely on improving the media content presented on the support. We want to do so by determining the viewer's reaction to the displayed content. Most papers regarding emotion recognition and analysis only consider clear, high contrasting emotions, such as happiness, surprise or sadness. For us, such emotions are not that relevant, as it is not expected of an advert to make you laugh or cry. Rather, we need to detect subtle changes in mood as to see if the persons attention was captured by the

978-1-7281-6843-2/20/$31.00 ©2020 IEEE

advert or not. We want to aggregate data about emotions with data regarding the physical presence of the person in front of the panel and with data regarding the content displayed for better understanding the environment.

To identify emotions, first we must extract the face to analyse and then to follow the movement of the face muscles. Many authors who reported good results in emotion cataloguing modelled these emotions as moving patterns for points marked on the studied faces [13], [14]. These points are called facial landmarks and are usually positioned around the eyes, nose and mouth of the subject. Many authors offered their opinion and experience about how many landmarks are necessary to achieve good emotion recognition, depending on the resources available and the response time. A highly regarded algorithm was proposed in paper [15], with 194 landmarks modelling each studied face. We want to use facial landmarks in our research, and we believe that a higher number of landmarks used will help us in detecting even the slightest changes in the person's facial expression.

In this paper we present the conceptual design of a target audience response analysis system in OOH advertising. The target area is surveilled using a static IP camera. The aim is to track any person passing in front of the camera and, if they get close enough, to analyze their response to the advert shown on the panel. Chapter II presents the theoretical and practical details of the proposed system, chapter III contains our experimental results, and chapter IV is dedicated to conclusions and future directions.

II. Methodology, Algorithms and Implementation

Considering a digital advertising panel placed in a display window, we want to see how many pedestrians pass by the panel, how much time they spend in front of it, and, if they approach the panel, we want to study their reactions while viewing the adverts being played. This way, OOH advertising companies can see what type of adverts has the most impact on persons and what kind of reactions different types of adverts produce in the audience. Fig. 1 shows the modules of the proposed system: detection and tracking of moving persons, determining if they had approached the panel, facial and emotion detection when they are viewing the panel, aggregating data about emotions and about displayed adverts, and obtaining business-oriented statistics.

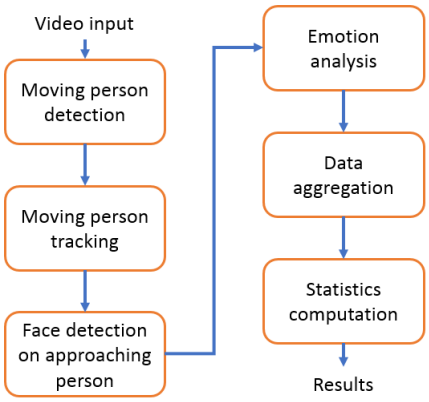

Fig. 1. Proposed system architecture.

The input videos are captured using static cameras mounted on the panel, facing the pedestrian-populated area, with the lens axis perpendicular on the plane determined by the panel's face. 'Moving person detection' is done using background subtraction (BS), with the background mask obtained via mean filtering (MF) on 10 consecutive frames. The moving objects are determined to be persons by a deep neural network (DNN) loaded with the MobileNet Single Shot Detector [16]; only persons predicted with over 80% confidence are considered for the following steps. 'Moving Person Tracking' is done using another DNN loaded with the GoTurn Caffe Model [17]. Since this model is sensitive to overlapping, we retrace persons after they are hidden from view using the future position prediction method described in our earlier work [11]. For efficiency, we attempt 'Face detection on approaching person' only when that person is considered close enough to the camera. The approximation of spatial location inside the camera's field of view is done using the area occupied by the person in the frame [18]. If the person occupies at least 25% of the frame, we crop the area encased by the person's bounding rectangle and use it as input for the Haar-Cascade classifier of OpenCV which gives us another bounding rectangle, this time surrounding the detected face. Fig. 2 offers a still shot from a video containing the results of the first three steps of processing.

Fig. 2. Face detection on approaching person (yellow – moving person far from the panel green – moving person near the panel blue – detected face).

Preprocessing of the input ends with the detected face images being cropped from the original frame and with facial landmarks being determined. The cropped patches of the original image are determined by the bounding rectangles of faces with a 10 pixels buffer to assure high accuracy when placing the landmarks. The landmarks are determined using the DLIB Facial Landmark Detector which gives 68 landmarks for each face, as shown in Fig. 3 [19].

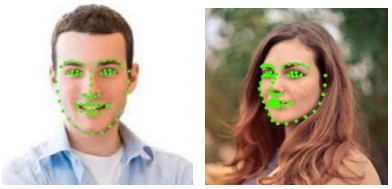

Fig. 3. Facial landmarks.

Most emotion recognition algorithms detect substantial mood changes (e.g. laughing), but it is unlikely such changes may occur when a person is viewing an advertising panel, so we focused our attention towards microexpressions (MEs), i.e. sudden, short-lived, and usually involuntary changes in the facial expression of a person. Researchers define and analyze seven universal emotions, also present in MEs: disgust, anger, feat, sadness, happiness, contempt, and surprise [20]. They also define three groups of MEs: simulated, neutralized (or suppressed) and masked (or

falsified) [21]. While most authors study the seven universal emotions [22], it is unlikely that persons will express, for example, fear while watching an advert. Also, there is no need for persons to suppress or mask their emotions during that time. Thus, our goal is to determine if the advert appeals to the persons or not. In doing so, we study simulated MEs and group them into our proposed 'Approval' and 'Disapproval' classes.

The 'Emotion analysis' module takes the coordinates of the identified facial landmarks of consecutive frames and models the facial movements, i.e. the MEs, then uses support vector machines (SVMs) to distinguish 'Approval', 'Disapproval' or 'Neural' reactions towards the advert. To train the SVMs for our proposed classes, we constructed a dataset recording 10 persons while having them watch 20 different videos. We chose 10 videos that we thought would determine positive ('Approval') reactions, 4 videos that would determine negative ('Disapproval') reactions and 6 videos that would determine neutral reactions, these proportions being chosen based on the apparent ease of determining each reactions (e.g. if the video is unpleasant, the person would probably just walk away from the panel). The 10 persons were 3 females and 7 males, aged between 20 and 55. The persons did not know what videos they were going to watch and were unaware that they were being recorded. While videos from which the training set was created were acquired under controlled conditions, with persons standing or sitting mostly upright in front of the camera, looking directly at the panel, real world condition may have persons watching adverts from a variety of postures. To address this, we used a property of the facial landmark detector that always gives us each landmark with the same identifier (e.g. the tip of the chin is always landmark no. 9). We employed a rotation algorithm that uses the isosceles triangle determined by the tip of the chin and the top of the eyebrows as reference. While the results were correct, the computation time increased greatly, as the rotation algorithm is slow. Fig. 4 shows the rotation result on a face cropped from a frame.

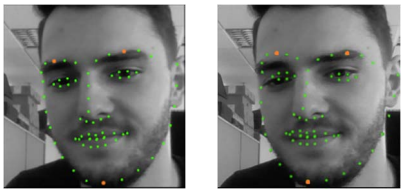

Fig. 4. Face rotation correction (left – original image, with selected reference landmarks marked orange, right – rotated landmark mask over same image).

Each emotion is identified using dedicated SVMs, which again brings an increased computational expense to our algorithm. For identification, descriptors extracted from 10 consecutive frames are used as we study the movement of each facial landmark from one frame to another. Fig. 5 shows subsets of 4 frames for 'Neural', 'Disapproval' and 'Approval' reactions. In this example, for 'Disapproval' it is very easy to perceive the movement of the facial landmarks, even if they are not shown in the images.

Fig. 5. Facial movement displaying emotions (top row – neutral / indifference, middle row – negative / disapproval, bottom row – positive / approval).

We use the tip of the chin (landmark no. 9) as reference for facial movement modelling, determining the relative positions of each other landmark in the image. In a series of 10 frames, the first frame is reference for movement calculation. For each following frame, for each landmark n, we consider Δx_n and Δy_n the number of pixels that landmarks move between the frames, horizontally and vertically, as against the reference landmark. For a single frame, the movement descriptor has the shape of equation (1), where t is the number of the frame. Of course, Δx_9^t and Δy_9^t are always 0. Modelling the movement over 10 frames, we obtain the feature descriptor defined by equation (2), as an array of length 9 (the number of movement frames) times 67 (the number of landmarks except the reference landmark, as it will always have no movement as against itself) times 2 (the movement directions) i.e. 1206 values contained by each feature descriptor used by the SVM.

$$(\Delta x_1^t, \Delta y_1^t, \Delta x_2^t, \Delta y_2^t, ..., \Delta x_{68}^t, \Delta y_{68}^t) \quad (1)$$

$$(\Delta x_1^1, \Delta y_1^1, \Delta x_2^1, ..., \Delta y_{68}^1, \Delta x_1^2, ..., \Delta x_{68}^9, \Delta y_{68}^9) \quad (2)$$

'Data aggregation' combines the results regarding observed emotions from the persons with known information about the playlist of adverts shown by the panel. Such, we can monitor reactions of our audience to the content we show them. The information are the following:

- the time spent by each person in front of the panel (the period from when the person was first detected in the camera's field of view until he leaves it);
- the time spent by each person viewing adverts on the panel (the period during which the person's face was detected and analyzed);
- the type and duration of each emotion or ME identified while the person is facing the camera;
- the exact date and time of each identification of person presence, face presence and emotion;
- the advert playing at each of the date and time instances previously mentioned.

Finally, 'Statistics computation' lets us not only determine the opportunity of the panel placement but also understand the appeal that each advert has for the target audience. This provides valuable information to the OOH

providers, as now they can offer the media providers reliable data about their marketing product, helping them improve their business. Of course, there may be cases in which the person situated in front of the panel is not watching the adverts, but we consider the percentage of these cases is too low to decisively influence our results. To achieve a more detailed model of the environment, we also considered adding a tracking algorithm to observe persons moving from one camera to another, to determine which are the shops with the largest audience as well as the types of shops which spark people interest the most. Such an algorithm we previously described in paper [12].

We have written the software for the system Java and Python, with the help of software libraries such as OpenCV, DLIB and NumPy. The MobileNet Single Shot Detector and the GoTurn Caffe Model were loaded to DNNs before the start of processing. The SVMs used for emotion recognition were trained offline. Training and test videos were acquired with Arecont Vision AV5115 cameras, at HD resolution (1280x720 pixels). Input videos can be captured and processed at any resolution, but inputs for the MobileNet Single Shot Detector must be resized to its working resolution of 300x300 pixels [16]. For optimal results, we recommend cameras positioned with the lens axes perpendicular on the plane determined by the panel's face, at a height of 150-190 cm, as to avoid perspective distortion of analyzed faces. While training and testing was done in an indoor laboratory environment, the system can perform under public and outdoor environments if the lighting conditions are good. Computation was done on an Intel Core i7 processor. Video frames were acquired once every 100 milliseconds and were saved in a FIFO list from which they were extracted and processed. This happens because the algorithm proposed in 'Emotion analysis' module is slow. However, this is not a setback in our chosen field of application, OOH advertising, where real-time statistics are rarely a must.

III. Experimental Results and Discussions

We evaluate our approach using the model from Fig. 6, in which an overhead view of a real-world use situation is represented. The camera can be mounted on a panel in a display window facing towards a sidewalk. The camera can also be oriented towards a hallway in a mall or can be placed on an outside panel (less likely as they are usually taller than the recommended height for the system to perform). We chose 5 adverts which are running in a loop on the advertising panel, each loop lasting 1 minute and 45 seconds:

- pet-shop advert, chosen for determining 'Approval', if the system performed correctly.
- tourism agency advert, chosen for 'Approval'.
- food advert, chosen for 'Neutral'.
- car advert, chosen for 'Disapproval' due to the grim nature of its content.
- Christmas time advert, chosen for 'Approval'.

At any time, a maximum of 2 persons are watching the panel from a near distance so that their emotions are being analyzed. We tested our method on 15 minutes of footage captured by the camera, extracting a frame every 100 milliseconds for a total of 9000 test frames.

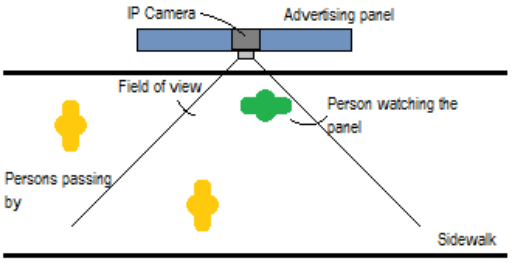

Fig. 6. Environment setting used for testing.

For our implementation we have constructed 3 types of tests, which are detailed below. All target values were determined by human observation. As we mentioned earlier, due to the complexity of the methods used, results are obtained slowly. Processing the test video footage took us 37 minutes and 42 seconds. We thought about increasing the interval at which frames are extracted to reduce time, but then we wouldn't have enough data for the ME identification we want to achieve; the MEs, usually having a very short duration (i.e. under 250 milliseconds), may happen entirely between two frame extractions.

A. *Person Detection, Person Tracking And Adaptive Face Detection*

Our first set of tests verified the correct detection of persons in video inputs and the correct detection of faces for approaching persons. Also, in this stage, we determined the time spent by each person in the camera's field of view, through tracking, and the time the person spent looking at the camera, as the number of frames in which the person's face was successfully detected. Table I gives the first set of statistics determined during testing.

TABLE I. Object Recognition Statistics

Items counted	Target no.	Obtained no.
Person passing by the camera	26	29
Persons near the camera	17	18
Faces detected	17	15

Because the video footage was recorded in a lab environment, there are actually a few humans walking in front of the camera, so they are detected as persons multiple times. We are not keeping track of persons exiting the field of view so with any new entry they will be counted again. Table II gives temporal statistics for the 10 persons who spent most time in front of the panel. As we can see, 8 of the 10 persons who spent most time inside the frame looked at the panel and 4 of them watched the panel for more than 10 seconds, which means they received the intended message of at least one advert. Fig. 7 shows a captured frame in which two persons were identified as watching adverts at the same times.

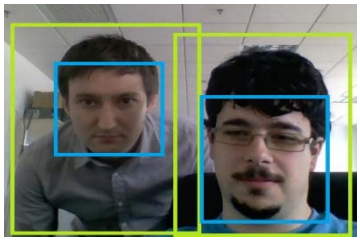

Fig. 7. Adaptive face detection.

Since we imposed that any person must occupy at least 25% of the frame to perform face detection, it means expressions from a maximum of 4 people will be analyzed at any given time. Even if now we process frames asynchronously from a FIFO list, considering there is a maximum amount of data which can be extracted from each frame, future developments could achieve real-time results.

TABLE II. PERSON TRACKING RESULTS

Person no.	Time spent in frame (s)	Face detected	Time spent watching panel (s)
1	119	X	108
2	98	X	82
3	82	-	-
4	80	X	76
5	67	X	48
6	25	X	15
7	23	-	-
8	21	X	13
9	18	X	9
10	14	X	5

B. Person Emotion Analysis

The second set of tests targets the main challenge of this paper, the identification of facial MEs, related to emotions triggered in persons by viewing video adverts. Table III gives general statistics for each of the three sought reactions: approval, disapproval and neutral. The target value was determined by outside observation. The error percentages are large because many neutral reactions were incorrectly classified as 'Approval' or 'Disapproval'.

TABLE III. REACTIONS STATISTICS

Reaction	Target value	Predicted value	Error percentage
Approval	3368	2942	12.64
Neutral	2481	3866	55.82
Disapproval	3934	2975	24.37

We asked the subjects who spent over 1 minute viewing the panel, in our case three persons, what was the reaction (approval, disapproval or neutral) they had watching the adverts. We compared the result of the ME extraction algorithm to the answers. The results are shown in Table IV. This grid can give us two sets of information: first, an important error percentage is visible for person no. 3. We have determined that is because the person had a cap on while watching the adverts, so the facial landmarks were not correctly placed, leading to misinterpretation of MEs, and second, two out of three predictions for advert no. 2 were erroneous, so the SVMs should be trained further.

TABLE IV. PERSON REACTION VARIATION

Person no.	Advert no.	Detected reaction	Declared reaction
1	5	Neutral	Neutral
1	1	Neutral	Neutral
1	2	Approval	Neutral
1	3	Approval	Approval
2	1	Approval	Approval
2	2	Neutral	Neutral
2	3	Neutral	Neutral
2	4	Disapproval	Disapproval
2	5	Approval	Approval
2	1	Disapproval	Approval
3	2	Neutral	Approval
3	3	Neutral	Neutral
3	4	Neutral	Disapproval

Fig. 8 contains a graphical representation of MEs identified in the person (Person 1 in Table IV) who watched the panel the longest (1 minute and 48 seconds). The horizontal axis shows the seconds elapsed since the person has started watching the panel. On the vertical axis, the detection starts from 'Neutral' (marked N) and varies over time as the person is watching adverts, going towards either 'Disapproval' (marked D) or 'Approval' (marked A).

C. Data Aggregation And Statistics Computation

Table V shows each reaction (ME) associated to each of the 5 adverts as the number of frames in which the ME was determined, regardless of the person watching. Ideally, an advert should have absolute majority of 'Approval' reactions, so all of them may be improved. Advert no. 4 should be replaced, as it has mostly 'Disapproval' reactions.

TABLE V. DATA AGGREGATION

Advert no.	Approval	Neutral	Disapproval
1	714	613	579
2	503	1118	468
3	534	1047	475
4	572	477	941
5	619	611	512

Such information may be used by OOH providers in collaboration with their clients to improve content and better understand their target viewers. Table VI offers other business-oriented statistics determined during testing.

TABLE VI. STATISTICS

Label	Value
No. of persons passing by the camera	29
No. of persons looking at the panel	15
No. of persons with at least one 'Approval' reaction	9
No. of persons with at least one 'Disapproval' reaction	8
Average time spent by a person viewing adverts	44 s
Maximum time spent by a person viewing adverts	108 s
Minimum time spent by a person viewing adverts	5 s
Maximum no. of persons watching adverts simultaneously	2

IV. CONCLUSIONS

We have described a system for analyzing target audience responses to adverts placed in display windows. Persons are identified and tracked throughout the video and their emotions are analyzed if they come close enough to the advertising panel. We offer statistics for advertising providers to create better and more targeted content.

Currently, the system performs on previously saved video files. We plan on improving it to real-time usage so that the business can make instant decisions based on audience feedback (e.g. real-time control of video adverts).

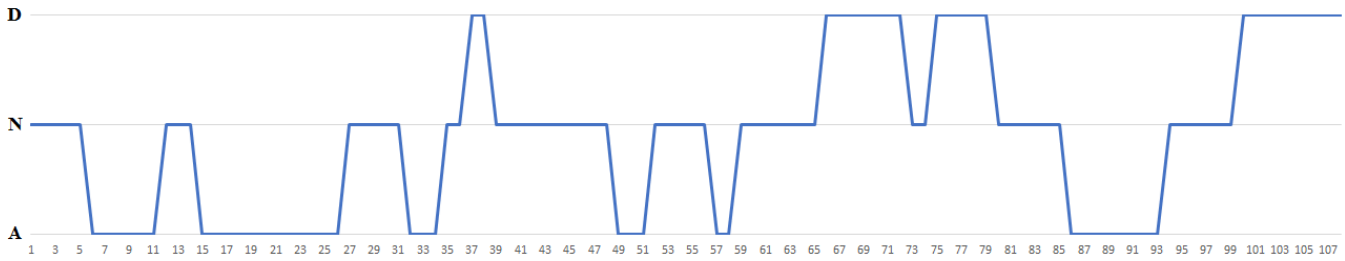

Fig. 8. Detected microexpression variation over time for a single person.

REFERENCES

[1] G. Cottrell and J. Metcalfe, "EMPATH: Face, emotion and gender recognition using holons," Proceedings of the 3rd International Conference on Neural Information Processing Systems, pp. 564–571, 1990.

[2] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network", Neural Networks, vol. 16, no. 5-6, pp. 555–559, June-July 2003.

[3] F. Wallhoff, W. Schuller, M. Hawellek, and G. Rigoll, "Efficient recognition of authentic dynamic facial expressions on the feedtum database," IEEE International Conference on Multimedia and Expo (ICME), pp. 493–496, July 2006.

[4] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[5] B. Ko, "A brief review of facial emotion recognition based on visual information," Sensors, vol 18, no. 2, 401, pp. 1–20, January 2018.

[6] E. Krumhuber, L. Skora, D. Kuster, and L. Fou, "A review of dynamic datasets for facial expression research," Emotion Review, pp. 1–13, October 2016.

[7] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI), pp. 435–442, November 2015.

[8] S. Alizadeh and A. Fazel, "Convolutional neural networks for facial expression recognition," arXiv:1704.06756, pp. 1–8, April 2017.

[9] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4321–4328, June 2014.

[10] A. Texeira Lopes, E. de Aguiar, A. de Souza and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," Pattern Recognition, vol. 61, pp. 610–628, January 2017.

[11] A. Costache, D. Popescu, C. Popa, and S. Mocanu, "Efficient video monitoring of areas of interest," 26th Telecommunications Forum (TELFOR), Belgrade, Serbia, 2018.

[12] A. Costache, D. Popescu, C. Popa, and S. Mocanu, "Multi-camera video surveillance," CSCS22, Bucharest, Romania, 2019.

[13] M. Kostinger, P. Wohlart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization," IEEE International Conference on Computer Vision Workshops (ICCV), pp. 1–8, 2011.

[14] Y. Wu and Q. Ji, "Facial landmark detection: a literature survey," International Journal of Computer Vision, vol. 127, no. 2, pp 115–142, February 2019.

[15] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2014.

[16] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, pp. 1–9, April 2017.

[17] D. Held, S. Thrun, S. Savarese, "Learning to track at 100 FPS with deep regression networks," European Conference on Computer Vision (ECCV 2016), pp 749–765, September 2016.

[18] A. Rosenbrock, "Find distance from camera to object/marker using Python and OpenCV", 2015. [Online]. Available: pyimagesearch.com/2015/01/19/find-distance-camera-objectmarker-using-python-opencv/ [Accessed: 10-Dec-2019].

[19] A. Rosenbrock, "Facial landmarks with dlib, OpenCV, and Python", 2017. [Online]. Available: pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/ [Accessed: 15-Dec-2019].

[20] P. Ekman, "Facial expressions of emotion: an old controversy and new findings," Philos Trans R Soc Lond B Biol Sci., vol. 335, pp. 63–69, January 1992.

[21] M. Shreve, S. Godavarthy, V. Manohar, D. B. Goldgof, and S. Sarkar, "Towards macro- and micro-expression spotting in video using strain patterns," Workshop on Applications of Computer Vision (WACV), pp. 1–6, 2009.

[22] W. Merghani, A. Davison, and M. H. Yap, "A review on facial micro-expressions analysis: datasets, features and metrics," arXiv:1805.02397, pp. 1–19, May 2018