EMOTION DETECTION IN INFANTS USING AN ENSEMBLE CLASSIFIER WITH A NOVEL MEMBER SELECTION TECHNIQUE

Hesham Ahmed Fahmy
Sherif Fadel Fahmy

Dept. of Electronics & Communications
Dept. of Computer Engineering
Arab Academy for Science, Technology and
Maritime Transport, 2401 Smart Village
Campus, 12577, Giza, Egypt
{heshamafahmy,fahmy}@aast.edu

Alberto A. Del Barrio García
Guillermo Botella Juan

Dept. of Computer Architecture and Automation
Complutense University of Madrid,
Ciudad Universitaria, C/ Jose Santesmases 9,
28040, Madrid, Spain
{abarriog,gbotella}@ucm.es

ABSTRACT

In this paper, we develop a method for detecting infant emotion from images using machine learning techniques. It is important to develop automatic infant emotion detection algorithms as they can become part of an integrated system that monitors the well-being of infants at home and in institutional settings. In particular, we develop an ensemble classifier that combines the results from diverse individual classifiers to produce better classification results. We present a novel method for selecting members of the ensemble that is based on the Pearson correlation coefficient of the *location* of the errors in the classifiers. The results show that the ensemble classifier outperforms individual classifiers by at least 1.4 times, and outperforms a state-of-the-art open-source algorithm for adult emotion detection by a larger margin. The results also indicate that Pearson Correlation among classifiers at a certain number of classifier count is a good criterion for selecting ensemble members.

Keywords: emotion detection, infants, ensemble, machine learning.

1 INTRODUCTION

Emotion detection from images is an important topic that has received a lot of research attention (Li et al. 2020, Pathak et al. 2020, Abanoz and Cataltepe 2018, Lee et al. 2018, Liu, Zheng, and Lu 2016, Awasthi 2013, Chen et al. 2012, Gilroy et al. 2009). The practical applications of emotion detection abound, including, but not limited to, gauging the sentiment of crowds to determine appropriate law enforcement response, analyzing the response of customers to products in order to fine-tune them and provide the best service possible and, in the case of infants, helping new parents cope with the problems of parenthood by offering assistance in determining infant mood.

Of particular interest to us is the case of infants. Parenting is a very difficult job, one that requires a lot of skills and is of the utmost importance. However, unlike in any other job, parents are not given any training. They are just expected to be able to cope with the many challenges of parenthood without any guidance. This is particularly true in cultures where independence is prized and new parents receive little or no assistance from extended family.

One of the tools that parents use to help them manage their hectic new life with an infant is a baby monitor. This device provides a video feed of the baby and allows their parents to monitor them from afar. We believe that it would be beneficial if the baby monitor can also analyze still images from its footage to determine the mood of the infant. This can offer assistance to a harried parent, by providing hints on what the infant requires.

While this paper focuses on determining infant emotions from face images, the aim of the authors is to provide a full infant monitoring system that uses multiple cues, including but not limited to, body motion (specifically leg and arm movements), sounds made by the infant as well as the history of emotions of the baby. The authors would also like to develop a time series analysis technique that can be used to predict infant needs using the past history of infant states in order to help new parents plan their day around their child's needs.

This paper presents only part of this integrated system – namely detecting the emotions of infants from baby monitor stills. To the best of our knowledge, no paper has tackled the problem of detecting infant emotions from still images. We believe this problem to be more difficult than its equivalent in adults, as the infant face is not yet fully mature and so may not be as expressive as adult faces.

In addition, most previous work in emotion detection, in adults, used a single classifier or ensemble systems trained using traditional methods. The work in this paper presents an ensemble technique that is based on a novel correlation method for selecting ensemble members. The rest of this paper is organized as follows, Section2 reviews the literature, Section3 describes the dataset used in the paper, Section4 provides an overview of the system used, Section5 presents the experiments and results, and Section6 concludes the paper.

2 LITERATURE REVIEW

As previously mentioned, there are many papers that address the issue of emotion detection from images, or from multimodal (Li et al. 2020, Pathak et al. 2020, Abanoz and Cataltepe 2018, Lee et al. 2018, Liu, Zheng, and Lu 2016, Awasthi 2013,Chen et al. 2012, Gilroy et al. 2009). However, to the best of our knowledge, no work has attempted to solve this problem for infants from image data only. One particular implementation of emotion detection from images that is worth considering is EmoPy reviewed in (Gaggioli 2019). As for the other works mentioned in this paragraph, it targets adults, but is the most mature solution with publically released code, so it is the solution we compare with our proposed work.

There are also works where the authors try to determine whether or not neonates are in pain – such as in (Zamzmi et al. 2018), but they do not address the general issue of emotion detection in infants. Pain detection in neonates is an important clinical tool that can help hospital staff manage neonatal care, but emotion detection, as addressed in this paper, is more general. It would provide a system that would aid new parents in asserting the needs of their infants. In addition, as mentioned in the introduction, it can be built upon to create a system that predicts infant needs.

Many of the works in the literature use transfer learning to detect emotions, examples of such papers include (Li et al. 2020, Abanoz and Cataltepe 2018, Lee et al. 2018, Nguyen et al. 2015, Chen et al. 2012). An important benefit of transfer learning is that it can achieve the results of deep learning on relatively small training datasets. This is because the models are pre-trained on very large datasets, which are sometimes unrelated to emotion detection, and are then fine-tuned on the emotion dataset. The rationale behind this is that the deep neural networks will learn features from the large datasets that can then be useful for solving the emotion detection problem.

During transfer learning, the deep neural networks are either used with all their feature extracting layers intact but with a new output network that combines these features in a manner that is relevant to the problem at hand, or some of their feature extracting layers are re-trained to produce features that are more specific to the problem at hand. Transfers learning is an important tool in emotion detection and is used as the driving idea behind some of the members of the ensemble used in this work.

Another approach that is promising in this field is the use of multimodal data for affective detection systems, examples include (Poria et al. 2017, D'mello and Kory 2015, Hussain et al. 2012, Gilroy et al. 2009).

While affective computing is not synonymous with emotion detection, it is similar enough to inspire us to consider the efficacy of multimodal data in emotion detection.

As mentioned in the introduction of this paper, we intent to use multimodal data from videos, and possibly other data sources, to more accurately determine the emotions of infants. However, in this paper, we limit our problem statement to determine the ability of automatic algorithms to detect emotions of infants from still images.

3 DATASET

We used two datasets inthis work, City Infant Dataset (Webb et al. 2018) and TIF (Maack et al. 2017). The original images were augmented using three different rotations and one flip operation resulting in a combined dataset that contains 80 infants and a total of 610 images. The datasets are all frontal face images of infants expressing various different emotions.

4 SYSTEM OVERVIEW

As previously mentioned, the system presented in this work is an ensemble system. The prospective ensemble members include both deep-learning models, and traditional machine learningtechniques. Deep-learning models do not require the extraction of features, and so are fed the raw images and process them as they deem fit. However, traditional techniques require feature extraction and so, for those algorithms, we extracted features representing the eyes and mouths of infants. The individual classifiers that we considered for the ensemble include:

- A Fuzzy Logic Based Classifier
- A K-Nearest Neighbors Classifier
- A Simple Multilayer Perceptron (MLP)
- Deep Learning using Inception Net
- Deep Learning using MobileNet

We then designed a meta classifier unit that combined the results of these individual classifiers in order to produce a unified result. Figure1depicts a block diagram of the system assuming that all classifiers are used as input to the ensemble.

The ensemble classifier algorithm theory states that selecting the individual classifiers that have the least correlation in the location of their errors according to the Pearson correlation coefficient will result in better results compared with each individual classifier (Benesty et al. 2009). This, to the best of our knowledge, is a novel way for selecting a member of an ensemble classifier. The rest of the subsections of this section explain each of the components of our proposed system in greater detail.

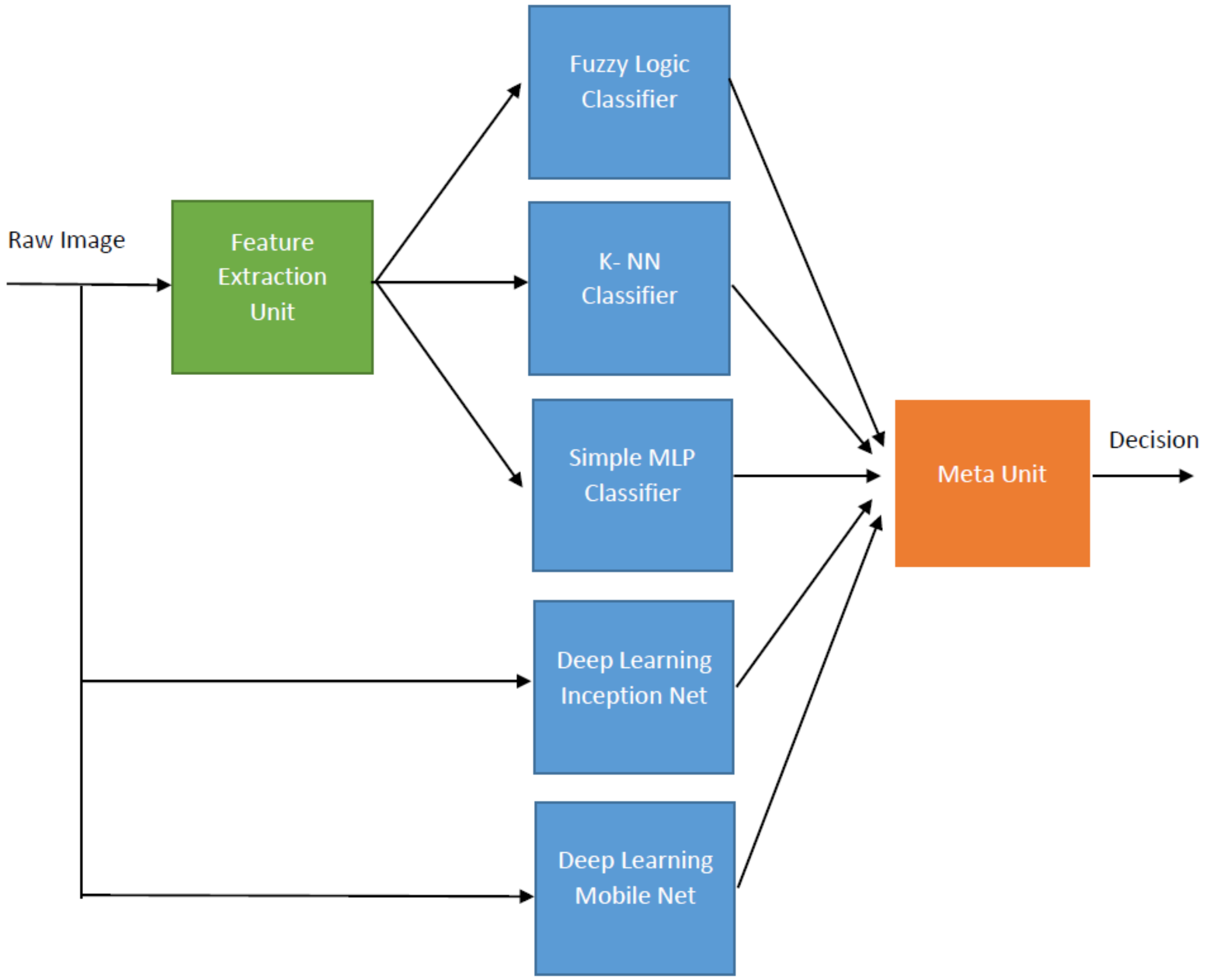

Figure 1: Block Diagram of System.

4.1 Feature Extraction Method

As previously mentioned, the first three machine learning architectures used in this paper are not deep learning architectures. Therefore, we used a number of feature extraction techniques to ensure that the algorithms received data that they could work on to produce good results.

First, the face of the infant is extracted from the image (Ji et al. 2018). Then the eyes and mouth are zoomed on (Ji et al. 2018). From discussions with parents, we learned that they recognized their infant's expression from their eyes and mouth most. Therefore, we extracted the percentage of the area of interest representing the eyes and mouth that actually contained eyes and mouth pixels. This serves as an estimate of how wide the mouth and eyes are open. These features are then used as input to the first three machine learning techniques.

For the deep learning algorithms, no feature extraction is required since the deep neural networks have layers that automatically extract the features that are most relevant to emotion detection.

4.2 The Classifiers

We shall now turn our attention to the classifiers used in the proposed system. We chose five different classifiers to test in our ensemble system. We now discuss the architecture of each of these classifiers with the accuracy of each of them in a greater detail. All the numbers representing accuracy will be reported to the nearest decimal place in this paper.

4.2.1 Fuzzy Logic Based Classifier

The first classifier used is a fuzzy logic system. The reason we considered the use of this classifier is that the problem of emotion detection, especially for infants, seemed particularly used to the fuzzy domain. Experts expressed their opinion in purely subjective terms that seemed best suited for fuzzy inference, for example, they made statements like the following *"If the baby's eyes are very open and his/her mouth is very open then the baby is surprised"*. The nature of these statements is a natural fit for fuzzy classifiers and so this was the first type of classifiers we used.

We used the fuzzy pattern classifier with genetic methods developed in (Stoean et al. 2005) as our fuzzy logic-based classifier. This classifier has an implementation in python's **sklearn** library that we were able to use in our system. On its own, this classifier produced an accuracy of about 54% round to the nearest decimal place.

4.2.2 KNN Classifier

The next classifier we used in the system is a k-nearest neighbor classifier. We used the implementation of KNNs present in python's **sklearn** and set the number of neighbors to five. The default Euclidean metric is used as the distance metric. The classifier, on its own, produced an accuracy of 41%.

4.2.3 MLP

This is a very simple multilayer perceptron that has four layers, each containing ten neurons. The maximum number of iterations was set to one thousand. This classifier produced an accuracy of about 51%.

4.2.4 Inception Net

After implementing the simple classifiers above, we thought it would be a good idea to take advantage of deep neural architectures trained on very large datasets to improve our results. The first deep architecture considered is Inception Net. Specifically, Inception Net V2, presented in this paper (Szegedy et al. 2016).

Naturally, it would not make sense to train a deep neural network on the relatively small dataset we have for our problem, therefore we performed transfer learning to take advantage of the pre-trained network while fine-tuning it to meet the needs of our problem.

Toward that end, we used the Inception Net V2 architecture with weights from the ImageNet dataset (Lab 2020). Then removed the output layer and replaced it with a three-layer feedforward network with 1024, 512, and 1024 neurons each. All neurons in these layers use the *relu* activation function (Agostinelli et al. 2014). The final layer has three neurons, one for each class, that use the *softmax* activation function.

During hyperparameter optimization, we learned that keeping the first thirty layers of the architecture with their original weights and training the remaining layers on the data specific to our problem produced the best result, and thus this is what we did for the results presented in this paper. This classifier produced a result of 41%.

4.2.5 MobileNet

The last architecture used in this paper is MobileNet (Howard et al. 2017), which is a deep network designed to be implemented on mobile devices. Again we used the version trained on ImageNet and performed transfer learning.

The output layer, during hyperparameter optimization, was designed to have four layers of sizes 1024, 2048, 2048 and 2048 respectively. As for the Inception of Net architecture, all neurons used the *relu* activation function. The final output layer contains three neurons, one for each class, each with a *softmax* activation function.

During hyperparameter optimization, we found that the best results are obtained when the first twenty layers retain their weights and the rest are trained on the data specific to our problem, which is what we did in the paper. This classifier produced a result of 37%.

4.3 The Meta Unit

The meta unit of the ensemble classifier (Sagi and Rokach 2018), the network that is responsible for collating the results from the individual classifiers in order to produce a unified result, is designed as a fully connected network with two hidden layers and one output layer. The hidden layers have dimensionality of twenty and twelve neurons respectively. The neurons in these two layers all use the *relu* activation function. The output layer contains one neuron with a linear activation function using the Mean Squared Error loss function and the Adam optimizer. The network was trained for one hundred epochs.

5 RESULTS

In this section of the paper, we present the results of our experiments. The aim of the experiments we conduct in this section of the paper is to see whether or not the ensemble technique we designed is capable of identifying emotions with higher accuracies than individual classifiers and EmoPy (Gaggioli 2019).

The researchers who developed EmoPy report an accuracy in the seventies for their classifier on their dataset, however, they reported that the classifier made mistakes when the emotions being classified were close to each other visually.

Therefore, we chose to test the system on three emotions, that to us, appear visually distinct, namely

- Happy
- Sad
- Neutral

When testing using EmoPy for these three emotions on our combined dataset, we achieved an accuracy of about 33%.

While this is much lower than the accuracy reported by the authors, this is to be expected as infant facial expressions are not fully developed and so may present a greater difficulty for machine learning techniques trained on adult expressions.

Next, we tested the accuracy of different combinations of individual classifiers. We report the best results at each classifier count in table1.

Table 1: Ensemble Results.

Number of classifiers	Members	Accuracy
2	Fuzzy,InceptionNet	66%
3	Fuzzy,MobileNet,InceptionNet	72%
4	Fuzzy,KNN,MobileNet,InceptionNet	70%
5	Fuzzy,MLP,KNN,MobileNet,InceptionNet	71%

By classifier count, we mean the number of individual classifiers used in the ensemble. All results are rounded to one decimal place. As can be seen, the best result obtained was when we used three classifiers. The members of the ensemble in this case were chosen according to the novel selection technique that is based on the Pearson correlation. We selected the classifier with the best result, the Fuzzy Logic classifier, and then selected two classifiers that are least correlated to it and added those to the ensemble. This was done for all cases.

As can be seen, the results of the ensemble outperform individual classifiers trained on infants, and EmoPy. One interesting result that we obtained from the experiments, is that the traditional classifiers that depend on feature selection outperformed deep learning techniques. This was rather unexpected at first, but upon deeper analysis turns out to be logical. Mainly because the size of the dataset available is not sufficient to train the deep network architectures to the desired degree of accuracy even when using transfer learning.

It should be noted, however, that the deep learning architectures made mistakes in different regions of the state space, thus combining them with the traditional techniques in an ensemble yielded better results.

6 CONCLUSION

The work presented in this paper is part of a larger project to accurately classify infant behavior and then predict infant needs using machine learning techniques. The project will use multiple cues and aggregate data from various different sources to do this. The part presented in this paper is emotion detection from infant face images.

We propose a method based on ensemble classification with the members of the ensemble being chosen using the Pearson Correlation Coefficient among the classifiers as the selection criteria. This technique ensures classifier diversity regardless of the way that the classifiers are initially trained. The proposed method shows that it is possible to produce better results with the ensemble than with individual classifiers, with the best individual classifier providing a 54% accuracy, and the best ensemble producing 72% accuracy. The proposed method also performs better than open-source state-of-the-art emotion detection algorithms as shown in the results section.

A couple of important factors became apparent during this work, first, while the Pearson Correlation provides a good indicator for choosing the members of an ensemble at a particular classifier count, more work is needed to determine the optimal classifier count and whether there is some quantitative cut-off we can use to determine this a priori.

Also, it is obvious that emotion detection on infants is a difficult task, and we propose to study the problem in greater detail in order to see whether modifications to existing architectures for emotion detection in adults can be made in order to improve the performance of such techniques on infants. We also hope to be able to collect more data in order to improve the classification of emotions in infants.

Finally, we propose, in the future, to integrate this emotion detection algorithm into a system for infant state detection that uses multiple cues, and that feeds into a time-series analysis algorithm that can be used to predict the temporal needs of infants.

REFERENCES

Abanoz, H., and Z. Cataltepe. 2018. “Emotion recognition on static images using deep transfer learning and ensembling”. In *26th IEEE Signal Processing and Communications Applications Conference, SIU 2018*, pp. 1–4.

Agostinelli, F., M. Hoffman, P. Sadowski, and P. Baldi. 2014. “Learning Activation Functions to Improve Deep Neural Networks”. (2013), pp. 1–9.

Awasthi, A. 2013. “Facial Emotion Recognition Using Deep Learning”. *IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI) Dec. 22, 2017* vol. 1 (September), pp. 9–12.

Benesty, J., J. Chen, Y. Huang, and I. Cohen. 2009. “Pearson correlation coefficient”. In *Noise reduction in speech processing*, pp. 1–4. Springer.

Chen, J., X. Liu, P. Tu, and A. Aragones. 2012. “Person-specific expression recognition with transfer learning”. In *Proceedings - International Conference on Image Processing, ICIP*, pp. 2621–2624.

D’mello, S. K., and J. Kory. 2015. “A Review and Meta-Analysis of Multimodal Affect Detection Systems”. *ACM Computing Surveys* vol. 47 (3), pp. 1—-36.

Gaggioli, A. 2019, may. “Online Emotion Recognition Services Are a Hot Trend”. *Cyberpsychology, Behavior, and Social Networking* vol. 22 (5), pp. 358–359.

Gilroy, S. W., M. Cavazza, M. Niiranen, E. André, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, and M. Billinghurst. 2009. “PAD-based multimodal affective fusion”. In *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*.

Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. *ArXiv* vol. abs/1704.0.

Hussain, M. S., H. Monkaresi, and R. A. Calvo. 2012. “Combining classifiers in multimodal affect detection”. In *Conferences in Research and Practice in Information Technology Series*, Volume 134, pp. 103–108.

Ji, Y., S. Wang, Y. Lu, J. Wei, and Y. Zhao. 2018, feb. “Eye and mouth state detection algorithm based on contour feature extraction”. *Journal of Electronic Imaging* vol. 27 (05), pp. 1.

Lab, Standford Vision 2020. “ImageNet”.

Lee, M. K., D. H. Kim, D. Y. Choi, and B. C. Song. 2018. “Deep Transfer Learning for Emotion Recognition Networks”. In *2018 IEEE International Conference on Consumer Electronics - Asia, ICCE-Asia 2018*.

Li, J., S. Huang, X. Zhang, X. Fu, C. C. Chang, Z. Tang, and Z. Luo. 2020. “Facial Expression Recognition by Transfer Learning for Small Datasets”. In *Advances in Intelligent Systems and Computing*, Volume 895, pp. 756–770.

Liu, W., W. L. Zheng, and B. L. Lu. 2016. “Emotion recognition using multimodal deep learning”. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 9948 LNCS, pp. 521–529.

Maack, J. K., A. Bohne, D. Nordahl, L. Livsdatter, Å. A. Lindahl, M. Øvervoll, C. E. Wang, and G. Pfuhl. 2017. “The Tromso Infant Faces Database (TIF): Development, validation and application to assess parenting experience on clarity and intensity ratings”. *Frontiers in Psychology* vol. 8 (MAR), pp. 1–13.

Ng, H.-W., V. D. Nguyen, V. Vonikakis, and S. Winkler. 2015. “Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning”. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI ’15*, pp. 443–449. New York, New York, USA, ACM Press.

Pathak, A. R., S. Bhalsing, S. Desai, M. Gandhi, and P. Patwardhan. 2020. “Deep Learning Model for Facial Emotion Recognition”. In *Lecture Notes in Electrical Engineering*, Volume 605, pp. 543–558.

Poria, S., E. Cambria, R. Bajpai, and A. Hussain. 2017, sep. “A review of affective computing: From unimodal analysis to multimodal fusion”. *Information Fusion* vol. 37, pp. 98–125.

Sagi, O., and L. Rokach. 2018. “Ensemble learning: A survey”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* vol. 8 (4), pp. 1–18.

Stoean, C., R. Stoean, M. Preuss, and D. Dumitrescu. 2005. “Diabetes Diagnosis through the Means of a Multimodal Evolutionary Algorithm GC represents an evolutionary framework designed for solving problems with multiple”. In *Diabetes*, pp. 9.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. “Rethinking the Inception Architecture for Computer Vision”. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vol. 2016-Decem, pp. 2818–2826.

Webb, R., S. Ayers, and A. Endress. 2018. “The City Infant Faces Database: A validated set of infant facial expressions”. *Behavior Research Methods* vol. 50 (1), pp. 151–159.

Zamzmi, G., D. Goldgof, R. Kasturi, and Y. Sun. 2018, jul. “Neonatal Pain Expression Recognition Using Transfer Learning”. *arXiv preprint arXiv:1807.01631*.

AUTHOR BIOGRAPHIES

HESHAM AHMED FAHMY is a part time PhD student in Complutense University of Madrid and an Assistant lecturer in the Arab Academy for Science, Technology and Maritime Transport (AASTMT), Cairo, Egypt. Received MSc in Electronics & Communications Engineering from AASTMT in 2015. His current research interests include Digital Signal Processing and Design Automation. His email address is hahmed@ucm.es **&** heshamafahmy@aast.edu.

SHERIF FADEL FAHMY received PhD in Computer Engineering in 2010 from Virginia Tech, USA. All certificates were attained with a GPA of 4.0 out of 4.0. His main research interests are distributed systems, real-time systems and operating systems. His email address is fahmy@aast.edu.

ALBERTO A. DEL BARRIO received the Ph.D. degree in Computer Science from the Complutense University of Madrid (UCM), Madrid, Spain, in 2011. Since 2017, he has been an Interim Associate Professor of Computer Science with the Department of Computer Architecture and System Engineering, UCM. His research interests include Design Automation, Arithmetic as well as Video Coding Optimizations. His email address is abarriog@ucm.es.

GUILLERMO BOTELLA received the M.A. Sc. degree in Physics in 1998, the M.A.Sc. degree in Electronic Engineering in 2001 and the Ph.D. degree in 2007, all from the University of Granada, Spain. Currently he is an Associate Professor at the Department of Computer Architecture and Automation of Complutense University of Madrid, Spain. His current research interests include Digital Signal Processing for VLSI, FPGAs, GPUs, HPC and Vision Algorithms. His email address is gbotella@ucm.es.