Facial Expression Recognition Using Robust Local Directional Strength Pattern Features and Recurrent Neural Network

Anders Skibeli Rokkones
Huddly AS
Oslo, Norway
E-mail:anders.r@huddly.com

Md Zia Uddin
Department of Informatics
University of Oslo
Oslo, Norway
E-mail:mdzu@ifi.uio.no

Jim Torresen
Department of Informatics
University of Oslo
Oslo, Norway
E-mail:jimtoer@ifi.uio.no

***Abstract*—This work proposes a novel facial expression recognition approach to contribute to better human-machine interactions. To do that, edge features in facial expression images are combined with a recurrent neural network (RNN) to classify different facial expressions. Robust edge features are first obtained by using Local Directional Strength Pattern (LDSP) and applied with RNN. This LDSP-RNN approach achieves superior recognition performance than other conventional approaches on a randomly distributed training and testing datasets obtained from a public dataset. The proposed approach should be useful for various practical applications such as a robot analyzing and understanding different human emotions from facial expressions based on robotic vision.**

I. Introduction

Humans have multiple ways of showing their feelings, which differ from person to person and culture to culture. Facial expressions have a significant role when individuals communicate with each other and sometimes can play a part to resolve misunderstandings while conveying a message. Human-machine interactions have attracted a large number of attentions from many researchers all around the world due to their practical applications in ubiquitous systems. For example, adapting a robot for emotion recognition in a ubiquitous smart system can contribute to improve the system by identifying the users' expressions and react accordingly [1], [9]. For emotional healthcare systems, facial expression analysis utilizing cameras are getting more and more attentions by computer vision researchers these days [2], [8].

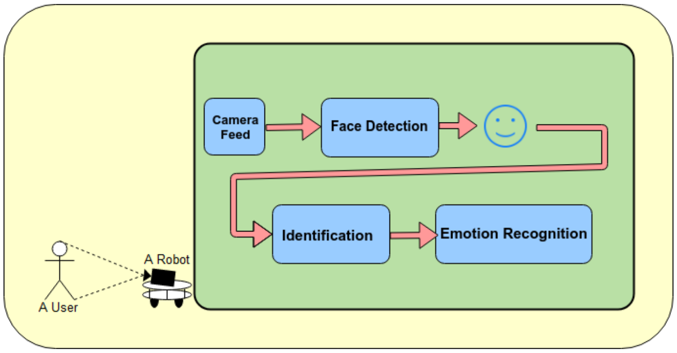


Fig. 1: Illustration of the information flow of a typical human-robot interaction system based on facial expressions.

With the advancements of technologies such as autonomous robots which interact with people, a good understanding between robots and humans can be one of the most important factors to be considered to enable further accomplishments in this field. In the last couple of decades, machine learning and artificial intelligence have been the center of attention in computer science, resulting in a better environment for significant progress in the human-robot interaction domain. There are a lot of different aspects of this domain, such as natural behavior and following basic norms.

The human face is a very complicated part of the body, consisting of multiple tiny muscles that together can form a wide range of expressions. Enabling machines (e.g., robots) to use sensors (e.g., microphones and cameras) to capture and analyze human faces and expressions, can contribute to better mental healthcare. Fig. 1 shows a schematic structure of a typical emotion recognition process from facial expressions based on the images acquired by a camera installed on a robot.

In this paper, a novel approach is proposed for emotion recognition from facial expressions using typical color cameras. The idea is to identify emotions through facial expressions using robust edge features in image sequences rather than a single RGB-image, which is a more common approach [1], [7] when dealing with images and deep neural networks. A Recurrent Neural Network (RNN) is trained on different expressions from image sequences obtained from the public dataset CK+ [6]. Then, the trained network is used for the expression recognition task later. Based on the robustness of Local Directional Strength Pattern (LDSP) for facial expression description, it is adopted in this work to apply with RNN instead of Convolutional Neural Netowrk (CNN) [9]. The proposed approach is named as LDSP-RNN.

II. Methods

This section represents the description of the feature extraction approach and learning the neural network expression model used in this work. Fig. 2 shows the overall methodology of the proposed approach, which consists of two parts: training and recognition of the emotions.

978-1-7281-2745-3/19/$31.00 ©2019 IEEE

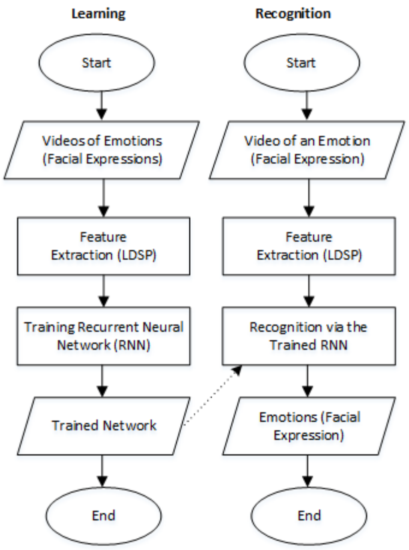

Learning
Recognition
Start
Start
Videos of Emotions (Facial Expressions)
Video of an Emotion (Facial Expression)
Feature Extraction (LDSP)
Feature Extraction (LDSP)
Training Recurrent Neural Network (RNN)
Recognition via the Trained RNN
Trained Network
Emotions (Facial Expression)
End
End

Fig. 2: Flow chart of the proposed emotion recognition system.

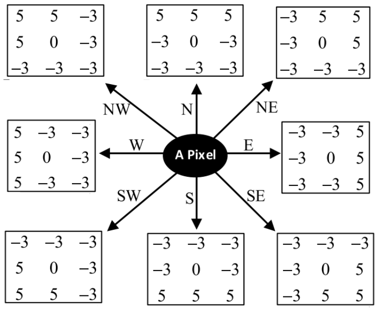


Fig. 3: Kirsch edge masks in eight directions [9].

A. Kirsch Edge Detection

Kirsch edge detection is a mask that considers edge responses in all eight directions around a single pixel. It is done by applying eight separate filters with values specifically to highlight edges oriented in the specified orientation [4] as shown in Fig. 3.

B. Local Directional Strength Pattern (LDSP)

Local Directional Strength Pattern (LDSP) is an improved version of Local Directional Pattern (LDP) [4] and it generates a 6-bit representation for a pixel by focusing on its strengths to eight different surrounding directions. Unlike LDP that considers the top n absolute values of the surrounding directions, LDSP looks at the directions representing the maximum and minimum values instead. LDSP was first proposed in [9] as a feature descriptor of emotions in depth images. Since conventional LDP considers only absolute values of edge strengths of a pixel, it can result in the generation of the same pattern for two different types of edge pixels. However, LDSP can overcome this and produce more robust patterns than LDP.

$$LDSP(x_c, y_c) = \sum_{n=0}^{7} L_n \times 2^n \tag{1}$$

$$L = binary(Arg(h)) \,||\, binary(Arg(l)) \tag{2}$$

h is the highest value of the surrounding n neighborhood pixels and l is the lowest. Their location *Arg()* is then converted into a binary representation and combined to form a 6-bit equivalent of their location(*Arg*) around the center pixel, where h is the three left bits and l is the three right bits, essentially making the highest neighboring pixel the most significant bits. The extraction method is described in Fig. 4. In Figs. 5 and 6, it is shown that LDSP produces separate edge representations for two different edge pixels. On the contrary, conventional LDP produces the same edge representation for both of them.

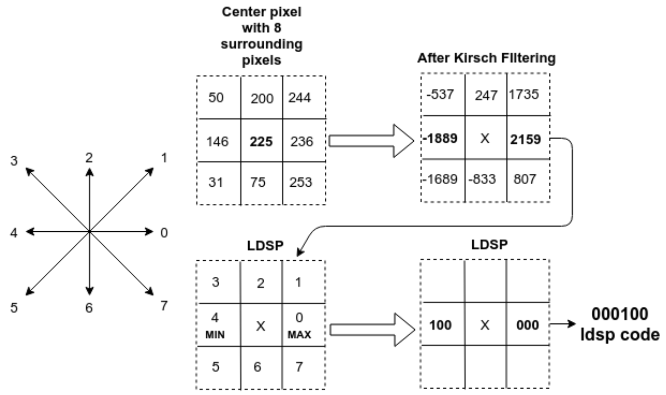


Fig. 4: Local Directional Strength Pattern (LDSP) on a 3x3 pixel neighborhood in an image. The resulting LDSP code is for the pixel at center.

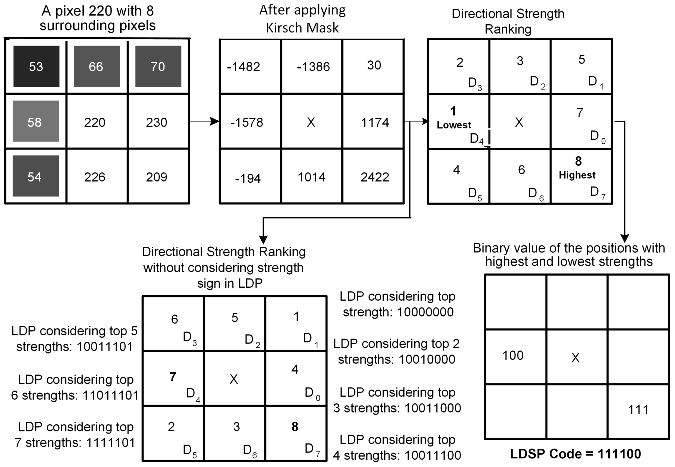

A pixel 220 with 8 surrounding pixels
53 66 70
58 220 230
54 226 209
After applying Kirsch Mask
-1482 -1386 30
-1578 X 1174
-194 1014 2422
Directional Strength Ranking
2 D3
3 D2
5 D1
1 Lowest D4
X
7 D0
4 D5
6 D6
8 Highest D7
Directional Strength Ranking without considering strength sign in LDP
6 D3
5 D2
1 D1
7 D4
X
4 D0
2 D5
3 D6
8 D7
LDP considering top 5 strengths: 10011101
LDP considering top 6 strengths: 11011101
LDP considering top 7 strengths: 1111101
LDP considering top strength: 10000000
LDP considering top 2 strengths: 10010000
LDP considering top 3 strengths: 10011000
LDP considering top 4 strengths: 10011100
Binary value of the positions with highest and lowest strengths
100
X
111
LDSP Code = 111100

Fig. 5: LDSP vs. LDP for a sample pixel with the value of 220.

C. Modeling Expressions

Recurrent Neural Network (RNN) is designed to handle sequences of data to make decisions based on data point

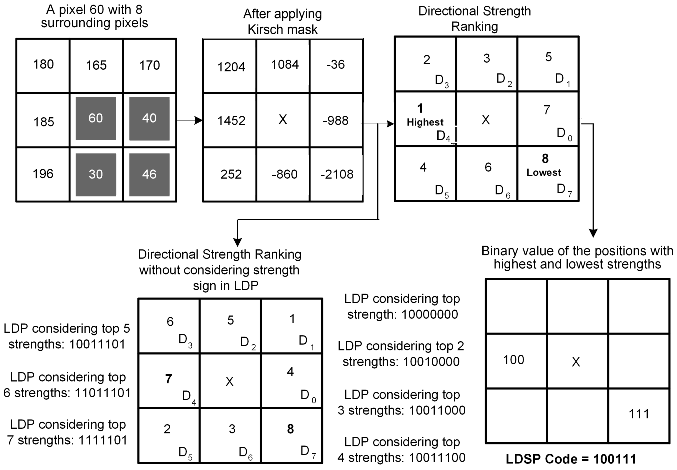

A pixel 60 with 8 surrounding pixels
180
165
170
185
60
40
196
30
46
After applying Kirsch mask
1204
1084
-36
1452
X
-988
252
-860
-2108
Directional Strength Ranking
2 D3
3 D2
5 D1
1 Highest D4
X
7 D0
4 D5
6 D6
8 Lowest D7
Directional Strength Ranking without considering strength sign in LDP
6 D3
5 D2
1 D1
7 D4
X
4 D0
2 D5
3 D6
8 D7
LDP considering top 5 strengths: 10011101
LDP considering top 6 strengths: 11011101
LDP considering top 7 strengths: 1111101
LDP considering top strength: 10000000
LDP considering top 2 strengths: 10010000
LDP considering top 3 strengths: 10011000
LDP considering top 4 strengths: 10011100
Binary value of the positions with highest and lowest strengths
100
X
111
LDSP Code = 100111

Fig. 6: LDSP vs. LDP for a sample pixel with the value of 60.

occurrences over time. Examples of this can be text (i.e., sentences) or videos. By looking at how specific observations develop over some time, it can predict the next step in a sequence.

RNNs process a sample for each time step from the sequence of one-dimensional vectors. The hidden state is computed based on the previous hidden state(h_{t-1}) at time(t-1) and the current input(x_t) at time(t):

$$h_t = \sigma_h(W_i x_t + W_h h_{t-1}) \tag{3}$$

The input and previously hidden state is multiplied with W_i, which is the input weight matrix, and W_h which is the recurrent matrix. The sum of these two results is then fed through a hidden activation function σ_h, usually Tanh-activation:

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{4}$$

which represents the activation ranges from -1 to 1.

$$y_t = \sigma_y(W_o h_t) \tag{5}$$

For every time step, an output (i.e., y_t) is computed where W_o is the output matrix and σ_y the output activation function. Thus, we can get the output at the time step that we prefer as illustrated in Fig. 7.

In [3], the authors presented an approach that includes using an RNN to classify facial expressions in videos. Their work is very much related to the proposed implementation, but they make use of a CNN to extract features, whereas the approach explained in this paper experiments with LDSP as features and RNN. The authors also utilized a variant of RNN called IRNN. Usually, an RNN has a hyperbolic tangent function (i.e., Tanh) as an activation function that consists of the problem of vanishing gradients. Hence, an alternative was proposed where Rectified Linear Units (ReLU) was used with a recurrent matrix that was initialized with scaled variations of the identity matrix, as described in [3], [5]. A standard RNN architecture is used in this work for emotion modeling and recognition.

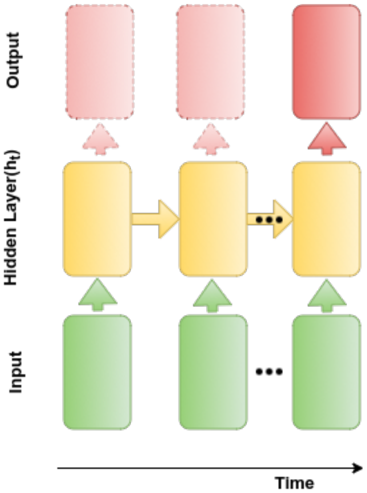


Fig. 7: Simple RNN structure used in this work.

III. Dataset, Experiments, and Results

This section represents the description of the dataset and outcomes of the experiments on that dataset using the proposed approach. Hence, we start here with describing the dataset. Then, the image pre-processing procedure is discussed followed by the visualization of the robust features. Then, we focus on describing the training and recognition of the features. Finally, we show the results using the proposed and other traditional approaches.

A. Dataset

The first stage of training and testing an expression model is preparing the datasets. RNNs, as stated earlier, needs a sequence of data over time. For example, in this case, facial expressions evolving over a sequence of images. If the input sequence is too long, the RNN will forget during the processing of the images, resulting in poor performance. The key is then to find a sequence length, which is not too long but still maximizes the RNNs information extraction. Given that the shortest length of a sequence in the benchmark dataset [6] is 5 images, the number of input samples is then set to 5. Fig. 8 represents sample sequences of three images per sequence progressing from a small degree of that expression to a full display. The expressions shown in the figure are disgust, surprise, and anger, respectively from the top to bottom row.

The selected image sequences are then randomly divided into three subsets: training, validation, and test set. Thus, 50%, 25%, and 25% of the dataset are considered for training, validation, and testing, respectively. In the CK+ dataset [6], each emotion image sequence starts from neutral to the specific expression. However, five sequential images are extracted from each original image sequence of the dataset. The assumption of choosing the images in the sequences is that the highest gradient of the emotions lies in the middle of each sequence of the dataset. Fig. 9 shows the distribution of the expression image sequences in different classes.

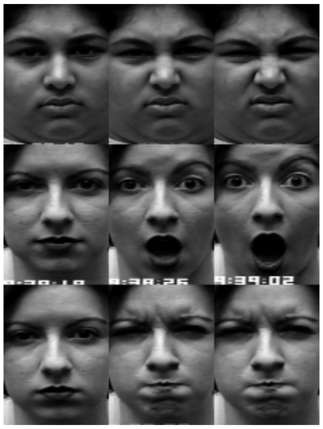


Fig. 8: Emotions evolving over time: disgust (top row), surprise (middle row), and anger (bottom row).

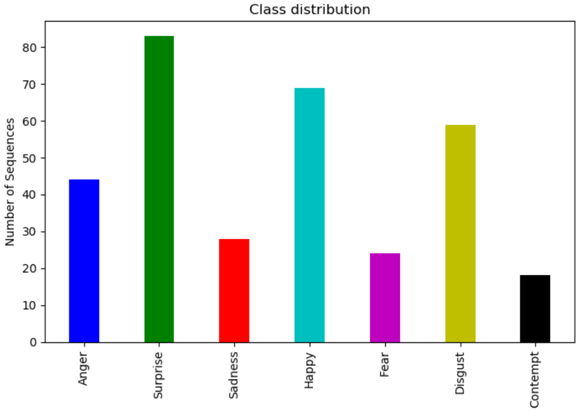


Fig. 9: Class distribution based on the expression image sequences.

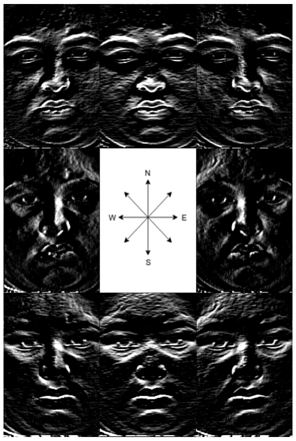
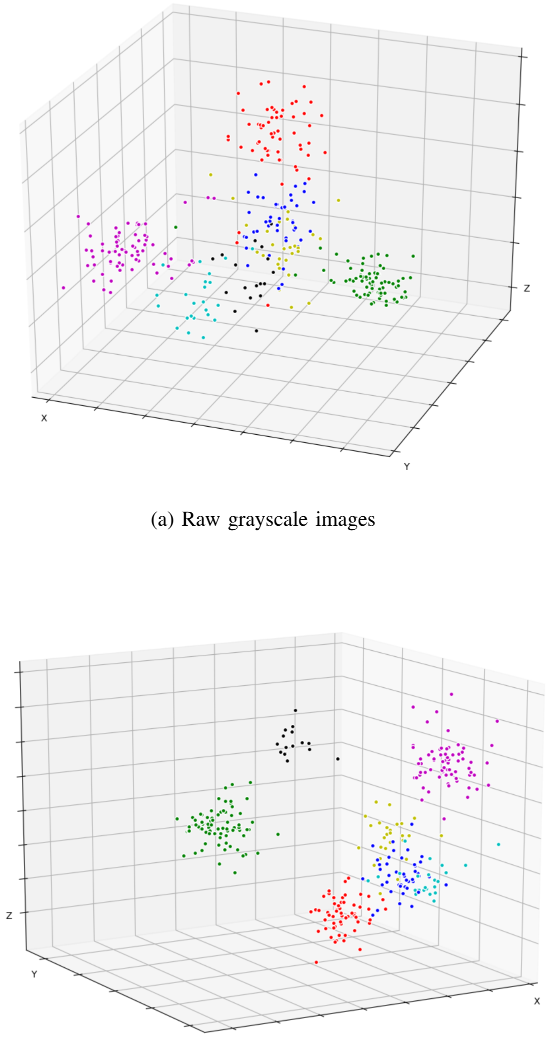


Fig. 10: Kirsch mask applied in all eight directions on a single grayscale image.



(a) Raw grayscale images

(b) LDSP features

Fig. 11: 3D plot of (a) raw grayscale images and (b) LDSP features after using LDA on the expression images. The expressions are anger(blue), contempt(black), disgust(red), fear(cyan), happy(pink), sad(yellow), and surprise(green).

B. Image Pre-processing

Before applying the RNN model, feature extraction from the image sequences is needed to enhance the accuracy of the classification process [9]. The CK+ dataset used in this work is of grayscale images from seven different expressions. As the first step of edge feature extraction, Kirsch mask is used for each pixel to obtain the edge strengths in its eight surrounding directions. Fig. 10 shows the edge response images in the eight different directions for the first face on the top-left in the Fig. 8. The figure indicates the importance of adopting the Kirsch mask to represent the important pixel patterns in the image. In our experiments, LDSP is directly applied to RNN without typical histogram representations.

C. Visualization of Features

To visualize the strength of the LDSP and traditional grayscale image features, Linear Discriminant Analysis (LDA) has been applied to the expression images, as shown in the 3D plots of the Fig. 11. LDA chooses the axes based on their ability to separate the different classes. In this regard, defining a decision line between the classes is usually a challenging problem to solve. The figure shows that the LDSP features have the strength to represent the better separability of different emotions than typical raw pixel features of grayscale images.

D. Results

From the dataset, 325 random samples were used in total where 50% was used for training, 25% for testing, and the rest 25% for validation. Table I depicts the result of 10 training iterations with 50 epochs per iteration. The LDSP features with RNN obtained a mean recognition rate of 73.6%. The model has been trained over ten different runs. The RNN with raw images as input (raw grayscale image-RNN) produced less mean recognition rate than LDSP-RNN's, indicating that with the use of LDSP features it produces more robust results in the form of a lower standard deviation and a higher mean value (i.e., Table I). Fig. 12 shows the best training model obtained using the proposed approach.

Approach	**Mean**	**Std**
RNN w/ Raw garyscale image	0.728	0.086
RNN w/ LDSP features	0.736	0.023
SVM w/ Raw garyscale image	0.72	0
SVM w/ LDSP features	0.72	0

TABLE I: Test results from different approaches with the mean and standard deviation.

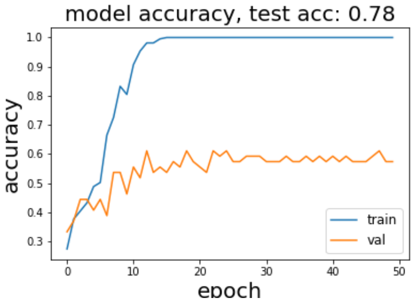
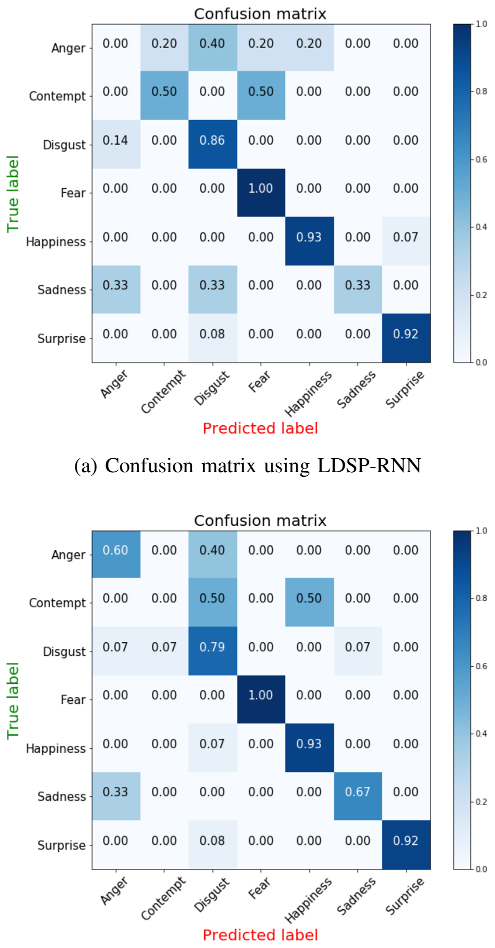


Fig. 12: Best training run of LDSP-RNN model.

As for comparing RNN to a standard image classifier, SVM is trained and tested on the same input data (i.e., raw grayscale image-SVM and LDSP-SVM) with a gamma parameter value of 0.01. In both cases of LDSP and raw grayscale images as input to SVM, they produced the same recognition rate of 72%, lower than the proposed approach.

Since deep neural networks are stochastic, a different result can be obtained for each training run. Hence, multiple runs are usually applied to deep neural networks to get statistically viable results. The confusion matrices from testing of LDSP-RNN and raw grayscale images with RNN are presented in Fig. 13, highlighting the emotions that both approaches struggle to classify.



(a) Confusion matrix using LDSP-RNN

(b) Confusion matrix using raw grayscale images with RNN

Fig. 13: Confusion matrices from RNN emotion recognition on the test set, showing recall for each emotion.

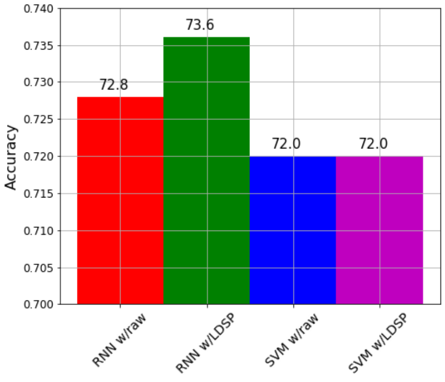


Fig. 14: Model performances on raw image features and LDSP features.

Contempt, anger, and sadness are the classes that show the most prominent variations between the two models. Samples from anger and sadness seem to be mixed with each other and difficult to be clustered, as indicated in Fig. 11. Contempt is most likely misclassified due to the low-class representation in the training set, as shown in Fig. 9. In Fig. 14, the average recognition rate from each approach on seven classes is represented where LDSP-RNN shows the best outcome.

IV. Concluding Remarks

In this paper, a novel human emotion recognition approach is proposed by analyzing facial expressions from videos captured by a typical color camera. A robust feature extraction process LDSP is applied to the time-sequential images of the videos and combined with RNN for robust expression recognition. With a standard RNN architecture on LDSP features as input, the model managed to get a superior recognition performance than others, indicating the proposed approach as a suitable option for classifying facial expressions for better human machine interaction. For the experiments, a public dataset was used here from which the training, validation, and testing datasets were obtained based on a random procedure. The proposed approach can be further analyzed to develop real-time human emotion recognition systems in real-life environments for different practical applications such as mobile robots for the elderly to help them to prolong their independent life.

Acknowledgments

This work is supported by the Research Council of Norway as a part of the Multimodal Elderly Care Systems (MECS) project, under grant agreement 247697. The authors would also like to thank Huddly AS for its support to finish the paper.

References

[1] S. Alizadeh and A. Fazel. Convolutional Neural Networks for Facial Expression Recognition. *arXiv preprint arXiv:1704.06756*, 2017. doi: 10.1007/978-3-319-47952-1

[2] P. Baxter and J. G. Trafton. Cognitive architectures for human-robot interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*. ACM Press, 2014. doi: 10.1145/2559636.2560026

[3] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent Neural Networks for Emotion Recognition in Video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pp. 467–474. ACM Press, New York, New York, USA, 2015. doi: 10.1145/2818346.2830596

[4] T. Jabid, M. H. Kabir, and O. Chae. Local Directional Pattern (LDP); A Robust Image Descriptor for Object Recognition. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 482–487. IEEE, 8 2010. doi: 10.1109/AVSS.2010.17

[5] Q. V. Le, N. Jaitly, and G. E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015.

[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, June 2010. doi: 10.1109/CVPRW.2010.5543262

[7] R. Ranjan, V. M. Patel, and R. Chellappa. HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019.

[8] S. Tadeusz. Application of vision information to planning trajectories of adept six-300 robot. In *2016 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, aug 2016. doi: 10.1109/icarm.2016.7606900

[9] M. Z. Uddin, W. Khaksar, and J. Torresen. Facial expression recognition using salient features and convolutional neural network. *IEEE Access*, 5:26146–26161, 2017. doi: 10.1109/ACCESS.2017.2777003