Facial Expression Recognition on partial facial sections

Ryan Melaugh, Nazmul Siddique, Sonya Coleman, Pratheepan Yogarajah
Intelligent Systems Research Centre
Ulster University

***Abstract*—Research by psychologists have shown that subjects had a preference for a side of a face when it was expressing emotions. This paper seeks to find what accuracies can be attained when only a segment of the face is considered. We show that using one side of the face only reduces accuracy by 0.34% but at half the computationally time required. Various other sections of the face are evaluated for similar performance. We demonstrate that using smaller portions of the face have an expected computation reduction but dont suffer the same degree of accuracy loss. For evaluating we train with a Convolutional Neural Network. To test what portions of a facial image are useful, the full face, half face, eyes, single eye, mouth and half of the mouth are chosen. These images come from the JAFFE, CK+ and KDEF datasets.**

***Index Terms*—Facial Expression Recognition, Neural Network, CNN, Occlusion, Hemisphere differences, Image Processing**

I. Introduction

Despite its age, the work Ekman [1] has done on emotions and emotion recognition still holds value as a basic introduction to emotion expression within the face. It is generally considered the basis of psychological work in this area. In particular, the original six emotions act as the basis of all emotions (with others being a combination of these basic six), they are: joy, sadness, anger, disgust, surprise, and fear [2]. It should be noted that when working with many facial image databases such as Japanese Female Facial Expression (JAFFE) [3] and Karolinska Directed Emotional Faces (KDEF) [4], they split the images into these six emotions, plus a seventh non-expressive state called neutral.

Ekman also created the Facial Action Coding System (FACS) [5]. These codes identify and score the movement of muscles and together make up a score for a particular facial expression. These FACS codes are not typically used in Facial Expression Recognition (FER), but they did form the basis for the feature extraction stage. It is assumed by many that when we express our emotions through our facial reactions that what we do is usually the case for both sides of the face and thus most research in FER makes use of the full face image to determine the emotional label [3] [6] [7] [8] [9]. However the face is symmetrical in appearance [10]. This is one aspect explored in this paper, to find what accuracy using only one side of the symmetrical face can achieve.

Psychological research by Blackburn and Schirillo [11] showed that there is a bias to one side of the face for expressing emotions. Subjects were asked to rate facial images using one side or the other. This was correlated with their pupil dilation which is associated with pleasantness [11]. Kowner provided a theory [12] that the right-hemisphere, which controls the left hemiface, has higher specialisation in the perception and expression of emotions. Blackburn and Schirillo found that subjects provided higher ratings for the emotion when the left portion of the face (as in the image split vertically over the bridge of the nose) was displayed. To establish this preference, subjects pupil dilation was tracked and correlated to their results. If humans use and prefer one side of the face, this does hint that taking one side could produce enough information to successfully classify an emotion.

Since it is an open question as to what accuracies are possible with only select regions of the face, this paper seeks to find those results. However instead of limiting tests to only one side of the face, we explore accuracies across the major landmarks used in Expression. These are the eyes, eyebrows, nose and mouth. Eyes and eyebrows can be grouped by the closeness of the muscles, as too can nose and mouth. Both of these regions can split into the left and right parts to assess whether sides differ in accuracy.

Prior research was more fragmented on machine learning usage though many used and still use Support Vector Machines (SVM). Shan, et al. [13], took Local Binary Pattern (LBP) data of facial expressions and used it with linear SVM to achieve 87.2%. Castillio et al. [14] used Local Sign Directional Pattern with a Polynomial SVM for 95.1%. Others made use of other methods such as Dapogny and Bailly [15], who used Pairwaise Conditional Random Forest for a 76.1% on BU-4DFE. Current research in FER has been moving towards Convolutional Neural Networks (CNN). CNNs are well suited to image tasks due to its effective feature extraction stage and flexibility in designing a model which allows it to scale well to larger datasets. Xie [8] used a CNN on the Cohn-Kanade+ database, and achieved an

accuracy of 92.06%, while Lopes et al. [9] used the JAFFE database and achieved an accuracy of 84.48%. Lopes stated that the time it took them 20 minutes to train their CNN model on CK+ (98.92%), and was performed on a NVIDIA GTX 660 GPU.

This paper investigates regions of the face using a shallow learning CNN model. The shallow learning CNN model is better suited to the datasets used in this case as they are much smaller than datasets often trained on CNNs such as CIFAR10 (60,000 images) [16] or MNIST (70,000) [17]. Not only does this paper compare the whole face versus either side of the face for FER, it also makes use of different portions of the face, such as the eyes, both separate and individually, as well as the full mouth and the left and right hemispheres of it as well.

In section II, the methodology is described, detailing the databases used for training and testing in II-A. In II-B the CNN model which is used for training and testing is respectively described. Section III is split into III-A and III-B. In III-A, the accuracy performance for the tests are mentioned along with the comparison between portions of the face. In III-B, computational times are covered where a comparison is again made between portions of the face. Section IV finishes this paper with a conclusion with an interpretation of the results as well as future research paths.

II. Methodology

A. Image selection and preprocessing

Images are loaded into memory from which the region of the face can be extracted. Viola-Jones Haar cascaders [18] is used as a face detector to locat the face in the image. This face region is then cropped and resized to 60×60 pixels. The image is then cropped to the specific region that is being tested, i.e. mouth region. These are then fed into the CNN. To crop one side of the face, the image can be split in half on the x-axis, which for a 60×60 pixel image the split line at the 30th pixel. This returns a 30×60 pixel image. For right side of the face images this is all data in the XY dimension before 30th pixel on x-axis, and vice versa for the left side images. For the eye region, we define it as between the mid-point of the forehead to the mid-point of the bridge of the nose. This exposes the eyebrows and eyes, and the range of positions that the muscles of the eyebrows move to during expressions. The mid-point of the forehead is found around the 20th pixel on the y-axis. The mid-point of the nose is around the 34th pixel on the y-axis. This returns a cropped image of 60×14, or 30×14 when using only a single eye. The last region used is the mouth. This is defined as from the same mid-point on the bridge of the nose to just below the chin. Contained in this region is the bottom of the nose, and mouth region. The nose and check regions were kept due to the tendency of muscles of the mouth to push and pull, forming laugh lines for example. The point just below the chin is around the 55rd pixel on the y-axis. With these settings, the mouth region image is 60×13 pixels or 30×13 in the case of using one side of the mouth. These settings amount to an estimate of the region as facial structures differ between subjects but will still be within a reasonable range (i.e. around the mid-point of the forehead for example when using 20th pixel on y-axis).

B. CNN

The use of the CNN is based on the model commonly used [19] [20]. It is a simplified version that uses a single layer of filters and pooling rather than three filter layers. This was because tests showed minimal difference in accuracy but with much higher computation time when using three layers. Using three layers (each accompanied by max pooling) returned an accuracy of 89.76% with total computation time of 936.96 seconds. The proposed model achieved 89.4% which took only 244.28 seconds. The neural network portion uses shallow learning with a fully connected dense layer as opposed to multi-layer Deep Learning CNNs. The reduction in layers is to account for the smaller size of the databases being used, as larger models are better suited to more data [21]. Figure 1 shows an illustration of the model in use.

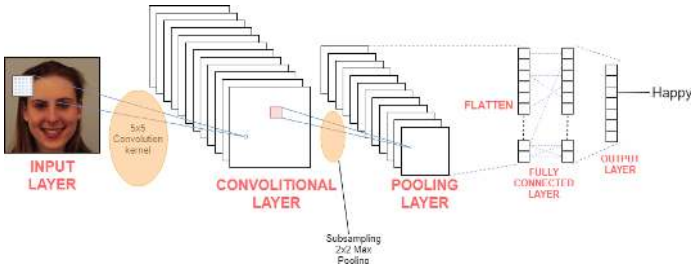

Fig. 1. Illustration of the CNN model in use.

In the CNN architecture the single convolutional layer consists of a 3D convolutional filter with a kernel size of $5 \times 5 \times 3$, consisting of width, height and depth, followed by a 2×2 Max pooling layer which is flattened and fed into a Dense Connected layer of 128 neurons with a dropout rate of 75%. The network finally ends in a dense output layer with a softmax function. Images are batched in groups, usually of 50 (with exception to using JAFFE), for an input shape of $50 \times 60 \times 60 \times 1$.

III. Results

Images came from three datasets, JAFFE [3], KDEF [4] and CK+ (Extended Cohn-Kanade) [22] datasets. KDEF and JAFFE included seven labels (Happy, Sad, Angry, Afraid, Disgust, Surprise and Neutral) and an eighth label (contempt), which was only included in CK+. Only front facing images are used from KDEF which gives 980 usable images. CK+ are images taken from frames of a video sequence which start with a neutral face expression and then changes emotional expression. The last three frames, showing the epxression, are used plus the first frame is kept as a neutral image. This gives a total of 469 images for CK+. All JAFFE images were used which gives 213 images.

A. Accuracy

Results are evaluated with respect to accuracy and computational run-time. By how much if any does one side impact accuracy and by how much can this reduce time taken to compute? Results are broken up into the sections of the face that were used, that is the half view, eyes and mouth with the full face as a benchmark. Computational run-time was measured from the point images were read to the end of evaluation of CNN. For the full face, it was able to achieve a benchmark accuracy score of 89.4% at a total time of 82.929 seconds (82.149 seconds without testing) to compute on the KDEF database. This time includes training and testing with 50 images per batch. All computation was on an Intel i7-6700 CPU with a clock rate of 3.4GHz. Images were resized to 60×60, lower images sizes reduced accuracy while higher sizes retained same accuracy but higher computation time. CK+ had an accuracy of 87.32% which took a total time of 75.539 seconds. While on the JAFFE database it benchmarked an accuracy score of 76.56% at 19.42 seconds, with a batch of 10 images to account for the smaller number of images. Discrepancy in accuracy is likely related to the relative size of each database, with JAFFE only containing 213 images, and 10 subjects. It has been shown in [21] that more data are beneficial to the accuracy in CNN.

The first test was similar to that performed by Blackburn [11]. The face is separated vertically down the bridge of the nose as in Figure 2 and only one side is used during full iteration of the classification. This gives us equal symmetry for the face. In this case, half of the face is used as seen in Figure 2 with each side tested and compared.

Fig. 2. Example of a face split into two sections. Each section was trained independently and tested. Image from KDEF.

The results are shown in Figure 3. For KDEF, the difference between the right side and the left is 0.71% but this was much higher for the CK+ dataset. Here the difference was 6.33% but JAFFE shows no difference at all. Importantly the difference between using all of the information from the face wasnt significantly higher that using just a single side. For

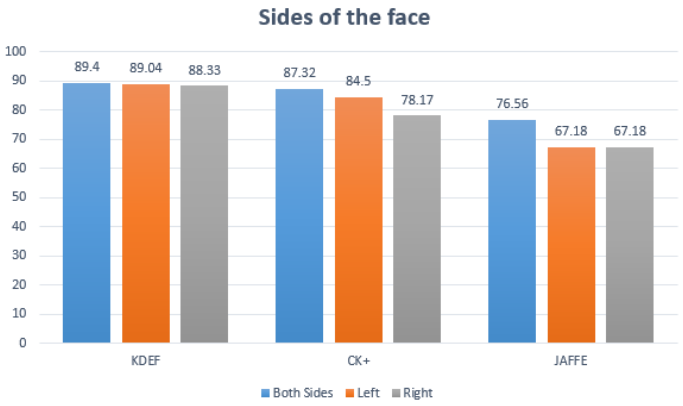

Fig. 3. Accuracy for symmetrically divided face.

KDEF this difference was only 0.34%. That being said, as the size of the dataset decreases, the difference between using the full face compared to using one side increases. This indicates that using more data is more important than the proportion of the original image used.

Fig. 4. Example of a two hemisphere mouth (Both Sides). Image from JAFFE.

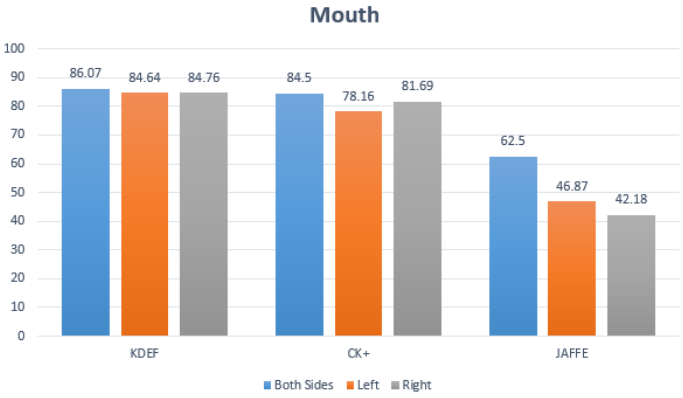

Fig. 5. Accuracy for symmetrically divided mouth region.

The next region used was the mouth. Results are shown in figure 5. For KDEF and CK+, there wasnt a large drop in accuracy from using the whole face or the mouth region. KDEF had an accuracy of 86.07% and CK+ 84.5%. However JAFFE had the largest drop to 62.5%. The results for JAFFE had the largest reduction in accuracy between the mouth region and symmetrically divided sections. Using the Left side, this accuracy jumped by 15.63%, which shows again the variance that smaller datasets suffer from when using selected regions. The left/right dichotomy is reversed from the full face and side tests. Both KDEF and CK+ exhibited a higher accuracy for the

right although for KDEF this was only a difference of 0.12%. The right side of the mouth for CK+ images displayed a 3.53% accuracy difference over the left. JAFFE was the only dataset to have a higher accuracy for the left side by 4.69%. These increases of differences from the larger dataset to the smallest suggest that these effects are exaggerated by reduction of data.

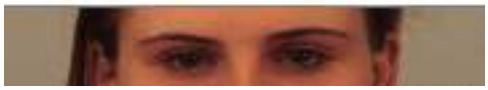

Fig. 6. Example of a two hemisphere eyes (Both Sides). Image from KDEF.

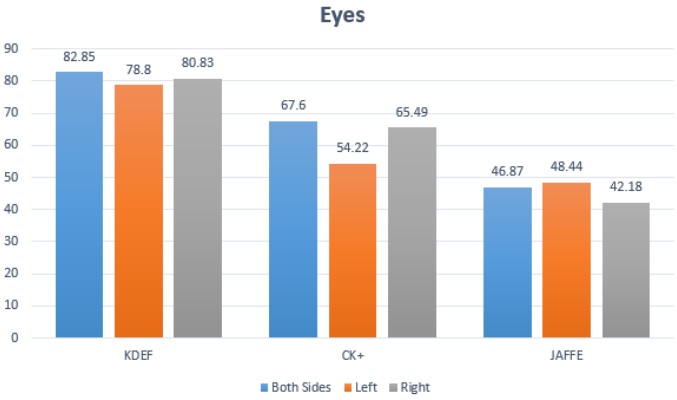

Fig. 7. Accuracy for symmetrically divided eye region.

The final test was for the eye region. KDEF accuracy was reduced from the mouth region but still only 6.55% decreased from using the full face. As the sizes of the datasets decrease, the ability of the CNN to keep up accuracies is reduced. CK+ managed 67.6% and JAFFE 46.87%. In terms of the difference between the left and right side, the results showed an overall preference for the right. KDEF had an increase of 2.03% but for CK+ the right was more dominant with an 11.27% difference. This was the only situation where CK+ demonstrated a higher difference between the sides than JAFFE. JAFFE found a higher accuracy for the left side by 6.26%. While not as large a variance as CK+, this still followed the trend of small datasets exhibiting large differences between the sides. JAFFE contains 213 images which is likely too small a representative that it skews the results and creates a larger margin of error.

B. Times

The base time for KDEF using the whole face took 244.277 seconds. Of that time, training took 237.25 seconds and evaluation 0.78 seconds, the remainder of time was from read in of images. When using the left side, this train time was cut to 109.89 seconds (evaluation 0.44s). Times follow along this general trend of train time cut by the proportion of the image used. So, using half, was able to train by approximately half the time. This can be seen from figures 9, 10 and 11. However accuracy wasnt affected by the same proportion as shown in figure 8. This table plots the times and accuracy in a scatter plot, with an exponential trend line. The highest accuracy was attained with using the full face however the time required to train using the full face is disproportionate to the accuracy it is able to achieve. The most effective region is using the mouth area. This was able to achieve 86.07% and took a total time of 65.35 seconds. Other regions and sides had a much more reduced accuracy for little extra benefit.

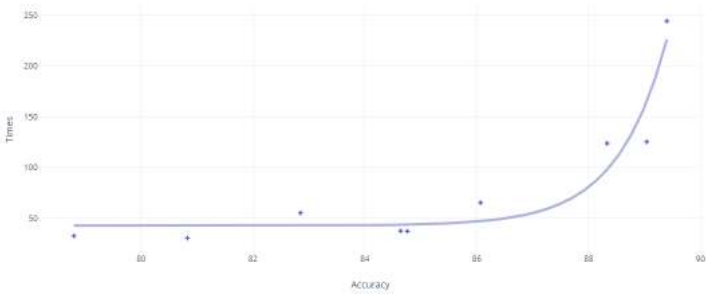

Fig. 8. Accuracy by times for KDEF.

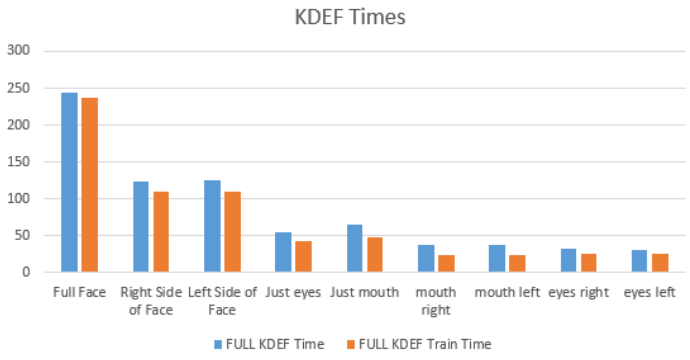

Fig. 9. Times for KDEF.

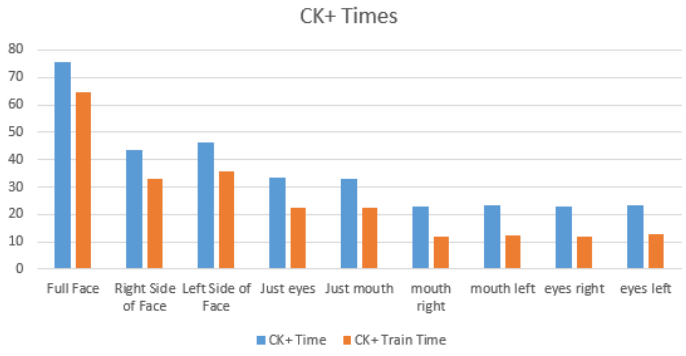

Fig. 10. Times for CK+.

C. Evaluation of Results

With only a difference of 0.36% when using half of the face compared to the full face, there is very little need to ever use the entirety of the face as the input. Especially since

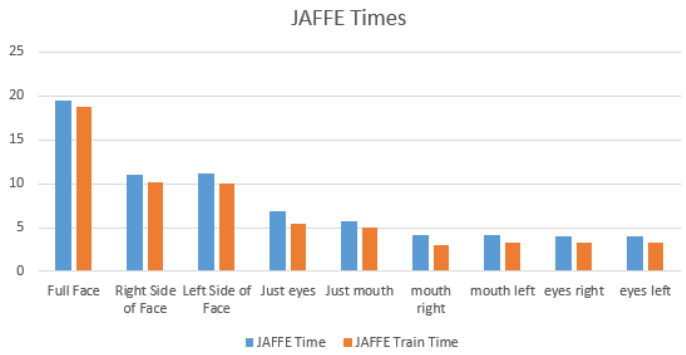

Fig. 11. Times for JAFFE.

taking half of the face leads to half the computation time, an ineffective use of training time. When this was broken down into further smaller regions, there is a degradation to accuracy, particularly if only using one side, for example one eye. However this reduction isnt proportional to the amount of information that is reduced. Taking the example of CNN on KDEF, the full face takes 82.929 seconds to compute. A single eye is approximately 30.5 seconds. That is 156% faster to compute than the full face. Whereas the accuracy for the full face was 89.4%, and this was reduced to 80.83% when using the left eye. That is only an 8.5% reduction in accuracy.

The overall face tests with the symmetrically divided sides show a slight preference with the left side which is in line with research by Blackburn [11]. However when broken into regions, this dichotomy mostly flips with a preference to the right side. This flip is with a caveat that there wasnt much consistency in the results. KDEF would never show much variance compared to the others, for example. The smaller the dataset used, the greater the variance between sides were. Whether the discrepancy between sides is due to localized physiological response in specific regions of the face or simply due to the size of the datasets cant be conclusively drawn. There is observed differences in accuracy but further data would be needed. That said, if one side was required, the right would be preferable due to the strengths it showed in the results at least for the smaller regions of the face.

IV. Conclusion

Bringing it back to the original premise, is taking a portion of the face enough to classify emotion? Our research shows that the answer is yes, at least when using the one side of the face. A secondary consideration was if there was a detectable difference in accuracy between the two sides of the face. This would be in line with human preference for the left side when expressing emotions. There is variation but the larger the dataset the less variance actually appear between the two sides. Further research would be useful in this area, such as person identification and whether one side or section can be favoured for recognition. Furthermore it would be interesting to find what, if any, emotion is favoured by the one side, and whether the intensity of the expression is correlated to the type of emotion.

References

[1] P. Ekman, “Ekman group - website.” [Online; accessed 13-January-2019].

[2] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion.,” *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[3] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, Apr 1998.

[4] E. Goeleven, R. D. Raedt, L. Leyman, and B. Verschuere, “The karolinska directed emotional faces: A validation study,” *Cognition & Emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.

[5] P. Ekman, *Basic Emotions*, pp. 45–60. John Wiley & Sons, 1999.

[6] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, “Local features based facial expression recognition with face registration errors,” *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008.

[7] C. A. Corneanu, M. O. Simn, J. F. Cohn, and S. E. Guerrero, “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1548–1568, Aug 2016.

[8] S. Xie and H. Hu, “Facial expression recognition with frr-cnn,” *Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.

[9] A. T. Lopes, E. D. Aguiar, A. F. D. 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, p. 610628, 2017.

[10] B. Fink, N. Neave, J. T. Manning, and K. Grammer, “Facial symmetry and the big-five personality factors,” *Personality and Individual Differences*, vol. 39, no. 3, p. 523529, 2005.

[11] K. Blackburn and J. Schirillo, “Emotive hemispheric differences measured in real-life portraits using pupil diameter and subjective aesthetic preferences,” *Experimental Brain Research*, vol. 219, no. 4, p. 447455, 2012.

[12] R. Kowner, “Laterality in facial expressions and its effect on attributions of emotion and personality: A reconsideration,” *Neuropsychologia*, vol. 33, no. 5, pp. 539–559, 1995.

[13] Caifeng Shan, Shaogang Gong, and P. W. McOwan, “Robust facial expression recognition using local binary patterns,” *IEEE International Conference on Image Processing 2005*, vol. 2, pp. II–370, 2005.

[14] J. A. Rojas Castillo, A. Ramirez Rivera, and O. Chae, “Facial expression recognition based on local sign directional pattern,” *2012 19th IEEE International Conference on Image Processing*, pp. 2613–2616, 2012.

[15] A. Dapogny, K. Bailly, and S. Dubuisson, “Pairwise conditional random forests for facial expression recognition,” in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[16] A. Krizhevsky, “Learning multiple layers of features from tiny images,” 2009.

[17] Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010.

[18] P. Viola and M. J. Jones, “Robust real-time face detection,” *Int. J. Comput. Vision*, vol. 57, pp. 137–154, May 2004.

[19] M. ejmo, M. Kowal, J. Korbicz, and R. Monczak, “Nuclei recognition using convolutional neural network and hough transform,” pp. 316–327, 01 2018.

[20] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, p. 12851298, 2016.

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[22] 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,” *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010.