JISKa, Vol. 5, No. 1, MEI, 2020, Pp. 23 – 35
ISSN : 2527 – 5836 (print) | 2528 – 0074 (online)

Penentuan Emosi pada Video dengan *Convolutional Neural Network*

Daru Prasetyawan [(1)], Shofwatul 'Uyun [(2)]
Magister Informatika UIN Sunan Kalijaga Yogyakarta
Jl. Marsda Adi Sucipto Yogyakarta
e-mail : daru.prasetyawan@uin-suka.ac.id [(1)], shofwatul.uyun@uin-suka.ac.id [(2)]

Abstract

Human emotions can be shown by facial expressions. Human facial expressions can change dynamically without they realize. This paper proposes a method to determine human emotions by recognizing human facial expressions and recording changes of facial expressions. We classify the 6 basic human facial expressions such as anger, fear, disgust, happiness, sadness, surprise plus a neutral expression using the Convolutional Neural Network (CNN). Data distribution equalization is applied to improve the performance of the model. This paper presents a classification model that can be applied to find out emotions in a video. The model was tested using separate data from the training data and evaluated using a confusion matrix. The results of the evaluation show that the classification model obtained an accuracy percentage of 74,07%, an average precision of 75,07%, and an average recall of 74,18%. At the end of this paper, we conducted an experiment by applying the classification model to some videos that represent human expressions. Every change in expression will be recorded and analysed so the most dominant emotion is found.

Keywords : CNN, Facial Expression, Confusion Matrix, Cross Validation

Abstrak

Emosi seseorang dapat ditunjukan melalui ekspresi wajah. Ekspresi wajah manusia dapat berubah-ubah secara dinamis tanpa disadari oleh orang tersebut. Penelitian ini melakukan penentuan emosi dengan melakukan pengenalan ekspresi wajah manusia dan melakukan perekaman untuk setiap perubahan ekspresi wajah tersebut. Metode dalam penelitian ini adalah dengan melakukan klasifikasi terhadap 6 ekspresi dasar wajah manusia ditambah ekspresi netral dengan *Convolutional Neural Network* (CNN). Pemerataan distribusi data dilakukan untuk meningkatkan kinerja model. Dari pemodelan tersebut, dihasilkan model klasifikasi yang dapat diterapkan pada sebuah video. Model tersebut diuji menggunakan data yang terpisah dari data latih dan dievaluasi menggunakan confusion matrix. Sebagai hasil evaluasi, diperoleh akurasi 74,17%, rata-rata presisi 74,07%, dan rata-rata recall 74,08%. Di akhir artikel ini, penulis melakukan percobaan dengan menerapkan model klasifikasi tersebut pada beberapa video yang mewakili ekspresi seseorang di dalam video tersebut. Setiap perubahan ekspresi akan direkam dan dianalisis sehingga ditemukan emosi yang paling dominan.

Kata Kunci : ***CNN, Facial Expression, Confusion Matrix, Cross Validation***

1. PENDAHULUAN

Ekspresi wajah adalah salah satu cara komunikasi non-verbal untuk mengungkapkan segala macam emosi baik yang negatif maupun yang positif (Prawitasari, 1995). Ekspresi wajah adalah perubahan wajah dalam menanggapi keadaan emosi, niat, atau komunikasi sosial seseorang (Tian, Kanade, & Cohn, 2011). Seseorang dapat memiliki ekspresi wajah yang dapat dikontrol oleh dirinya sendiri secara sengaja, tetapi pada umumnya ekspresi wajah dapat timbul secara alami akibat perasaan atau emosi orang tersebut. Ekspresi yang muncul secara tidak disengaja inilah yang dapat menggambarkan perasaan atau emosi pada saat itu. Ekspresi wajah sangat menarik untuk diteliti karena merupakan salah satu komunikasi non-verbal yang dapat digunakan untuk menyampaikan pesan sosial dalam kehidupan manusia serta menggambarkan keadaan emosi seseorang. Wajah manusia dapat menggambarkan perasaan manusia saat itu. Mengenali ekspresi wajah merupakan cara penting untuk mengetahui apa yang dirasakan seseorang.

Seseorang terkadang ingin menyembunyikan perasaan atau emosinya, tetapi biasanya hal ini sangat sulit dilakukan karena wajah mereka biasanya akan menunjukan perasaan yang

sebenarnya. Misalnya, seseorang ingin menyembunyikan perasaan bencinya terhadap orang lain, tetapi pada saat tertentu tanpa sengaja akan menunjukkan perasaannya tersebut di wajahnya, walaupun orang tersebut sangat pandai menyembunyikan perasaan tersebut. Sebaliknya, banyak orang yang salah membaca emosi sesorang karena hanya melihat sesaat saja. Hal ini tentunya memerlukan pengamatan terus menerus terhadap perubahan ekspresi wajah seseorang. Namun hal ini tidak dapat dilakukan oleh manusia secara langsung karena pada saat tertentu akan mengalami kejenuhan yang mengakibatkan ketidaktelitian. Penentuan emosi pada video merupakan salah satu cara untuk mengatasi permasalahan tersebut.

Deteksi wajah merupakan langkah pertama yang harus dilakukan dalam analisis wajah, termasuk di dalamnya adalah pengenalan ekspresi wajah. Deteksi wajah bertujuan untuk menentukan apakah ada wajah atau tidak di dalam citra, dan jika ada dimana letak wajah tersebut dan ukuran masing-masing wajah pada citra (M.-H. Yang, Kriegman, & Ahuja, 2002). Dalam deteksi wajah terdapat beberapa tantangan seperti posisi wajah, skala wajah, ekspresi wajah, wajah terhalang objek lain, dan kondisi pencahayaan (S. Yang, Luo, Loy, & Tang, 2016). Metode yang biasa digunakan dalam deteksi wajah antara lain *Knowledge-based methods*, *Feature invariant approaches*, *Template matching methods*, dan *Appearance-based methods* (M.-H. Yang et al., 2002). *Knowledge-based method* (Kotropoulos & Pitas, 1997). *Knowledge-based methods* merupakan metode deteksi wajah yang bergantung pada sekumpulan aturan yang dibuat manusia, misalnya sebuah wajah terdiri dari dua mata, satu hidung dan satu mulut dalam jarak tertentu dan posisi relatif terhadap bagian-bagian yang lain (Chauhan, 2014). *Feature invariant approaches* merupakan teknik untuk deteksi wajah dengan mengekstrak fitur struktur wajah (Chauhan, 2014), seperti kulit wajah (Mahmoodi, 2017) dan tekstur wajah (Saito, Lingyu, Liwen, Nagano, & Hao, 2016). *Template matching approaches* menggunakan *template* wajah yang telah ditentukan atau diberi parameter untuk menemukan dan mendeteksi wajah, dengan menghitung nilai korelasi antara *template* dan citra yang diinputkan (Chauhan, 2014). *Appearance-based methods* bergantung pada satu set citra wajah yang dilakukan pelatihan untuk mengetahui model wajah (Chauhan, 2014).

Seiring dengan berkembangnya metode-metode pengenalan wajah dan meningkatnya kemampuan perangkat keras, banyak penelitian-penelitian yang tidak hanya melakukan deteksi wajah tetapi juga melakukan pengenalan wajah termasuk di dalamnya pengenalan ekspresi wajah. Pada umumnya, sistem pengenalan wajah biasanya terdiri dari 4 bagian, yaitu *face detection*, *face alignment*, *feature extraction*, dan *feature maching* (Li & Jain, 2011). Pendeteksi wajah mencari wajah di dalam citra dan mengembalikan kordinat area wajah untuk setiap wajah yang terdeteksi. Sedangkan *face alignment* bertujuan untuk menyelaraskan wajah yang terdeteksi dengan menggunakan satu set titik referensi yang terletak di lokasi tertentu di dalam citra (Trigueros, Meng, & Hartnett, 2018). Kemudian, ekstraksi fitur dilakukan untuk memberikan informasi yang berguna untuk membedakan antara wajah orang yang berbeda yang berhubungan dengan variasi geometris dan fotometrikal (Li & Jain, 2011).

Metode tradisional mengandalkan kemampuan tangan manusia, seperti penegasan garis tepi dan tekstur, dikombinasikan dengan teknik pembelajaran mesin, seperti *Principal Component Analysis (PCA)*, *Linear Discriminant Analysis (LDA)*, atau *Support Vector machine (SVM)* (Trigueros et al., 2018). Pendekatan secara konvensional yang diterapkan dalam pengenalan wajah biasanya melibatkan tahapan akuisisi dan pemrosesan gambar, pengurangan dimensi, ekstraksi fitur, dan klasifikasi secara berurutan (Liew, Khalil-Hani, Ahmad Radzi, & Bakhteri, 2016). Setiap langkah pemrosesan biasanya dilakukan terpisah sehingga kesesuaian setiap motode sangat diperlukan. Kinerja sistem pengenalan wajah juga sangat tergantung pada jenis algoritma klasifikasi yang dipilih. Selanjutnya jenis algoritma klasifikasi juga akan bergantung pada jenis dan metode ekstraksi fitur yang digunakan. Hal ini memerlukan ketelitian dalam memilih jenis ekstrasi fitur dan algoritma klasifikasi yang diterapkan untuk menghasilkan kinerja sistem pengenalan wajah yang optimal. Pada pendekatan tradisional, *Support Vector machine (SVM)* sering digunakan untuk melakukan pengenalan ekspresi. SVM adalah alat prediksi klasifikasi dan regresi yang menggunakan teori pembelajaran mesin untuk memaksimalkan akurasi prediktif (Vasanth & Nataraj, 2015). Pengenalan ekspresi wajah dengan SVM yang dilakukan oleh Vasanth dan Nataraj menggabungkan *Gabor Features* dan *Local Binary Pattern (LBP)* untuk melakukan ekstraksi fitur mata dan mulut, serta memanfaatkan *Principal Component Analysis (PCA)* untuk mereduksi dimensi matriks fitur yang dihasilkan dari proses ekstraksi.

Matriks fitur yang berukuran kecil dapat membantu meningkatkan kecepatan klasifikasi (Vasanth & Nataraj, 2015). Selanjutnya SVM digunakan sebagai pengklasifikasi untuk melakukan pengenalan ekspresi wajah.

Saat ini metode pengenalan wajah tradisional telah digantikan oleh metode pembelajaran mendalam (*deep learning*) seperti *Convolutional Neural Network* (CNN). Salah satu hal terpenting keberhasilan metode tersebut adalah ketersediaan data pelatihan dalam jumlah besar (Parkhi, Vedaldi, & Zisserman, 2015). Metode *deep learning* dapat dilatih dengan data yang sangat besar untuk mempelajari fitur dalam merepresentasikan data (Trigueros et al., 2018). CNN adalah varian jaringan saraf yang terdiri dari sejumlah lapisan konvolusional dan lapisan subsampling dan diakhiri dengan satu atau lebih lapisan standar *multilayer perceptron (MLP)* (Liew et al., 2016). CNN terdiri dari dua bagian dasar, yaitu ekstraksi fitur dan klasifikasi. Ekstraksi fitur mencakup beberapa lapisan konvolusi diikuti oleh lapisan *pooling* dan fungsi aktivasi (Khoshdeli, Cong, & Parvin, 2017). Bagian pengklasifikasi biasanya terdiri dari lapisan MLP biasa (*fully-connected layer*). Keuntungan signifikan dari CNN dibandingkan pendekatan konvensional adalah kemampuannya untuk secara bersamaan mengekstrak fitur, mengurangi dimensi data, dan proses klasifikasi dalam satu struktur jaringan (Liew et al., 2016). Sebelum CNN diperkenalkan, banyak waktu yang dihabiskan untuk pemilihan atau ekstraksi fitur. Kemampuan belajar yang kuat dari CNN sebagian besar disebabkan oleh penggunaan beberapa tahap ekstraksi fitur pada lapisan tersembunyi yang dapat secara otomatis mempelajari representasi dari data (Khan, Sohail, Zahoora, & Qureshi, 2019). CNN melakukan ekstraksi fitur dan klasifikasi dalam satu struktur jaringan melalui pembelajaran pada sampel data (Lecun, Bottou, Bengio, & Ha, 1998). Permasalahan yang sering muncul dalam penggunaan CNN dalam klasifikasi citra tidak terlepas dari kualitas model yang dihasilkan itu sendiri. Peningkatan kualitas biasanya diselesaikan dengan penambahan jumlah dataset. Terkadang model yang dihasilkan dapat memprediksi suatu kelas dengan sangat akurat, tetapi kurang baik untuk kelas lain. Hal ini dapat terjadi karena tidak seimbangnya distribusi data latih pada saat pemodelan sehingga model seperti menghafal kelas-kelas tertentu saja. Permasalahan lainnya yang sering muncul dalam klasifikasi citra adalah terjadinya o*verfitting*, yaitu suatu kondisi dimana model dapat mengenali dengan baik data yang digunakan dalam proses pelatihan, tetapi kurang baik terhadap data yang belum pernah ditemui. Selain itu, apabila terdapat lapisan-lapisan yang sangat rumit dan mendalam, akan mengakibatkan model terlalu besar dan dalam proses pelatihannya juga memakan waktu yang cukup lama.

Untuk mengatasi permasalahan tersebut diperlukan arsitektur model CNN yang sederhana karena dalam proses pengenalan ekspresi wajah di dalam video harus dilakukan dengan cepat. Pemerataan distribusi data diperlukan agar model dapat mempelajari data dengan kelas yang seimbang, sehingga model tidak hanya seperti menghafal kelas-kelas tertentu saja. Selanjutnya permasalan *overfitting* dapat dikurangi dengan regularisasi, yaitu dengan mengatur agar *loss* atau *error* yang diperoleh dari data yang pernah dilihat sebelumnya dengan *loss* atau *error* yang diperoleh dari data yang belum pernah dilihat sebelumnya. Selain itu, penerapan *cross-validation* dilakukan pada saat proses pelatihan, sehingga setiap data dapat terlibat sebagai data latih dan data uji. Penulis mengusulkan metode untuk menentukan emosi manusia di dalam video melalui pengenalan ekspresi wajah menggunakan CNN. Optimalisasi model CNN dilakukan pada saat pra pemrosesan data dengan melakukan pemerataan distribusi data untuk setiap kelas yang digunakan sebagai data latih, sehingga dihasilkan sebuah model CNN yang sederhana tetapi memiliki kinerja yang baik dalam melakukan klasifikasi. Model CNN yang sederhana sangat diperlukan karena proses pengenalan ekspresi wajah pada video dilakukan pada setiap *frame* yang memerlukan pemrosesan yang cepat. Selanjutnya, penentuan emosi manusia dengan menghitung banyaknya ekspresi wajah yang muncul di dalam video pada rentang waktu tertentu.

2. DATA DAN METODE

2.1. Sumber Data

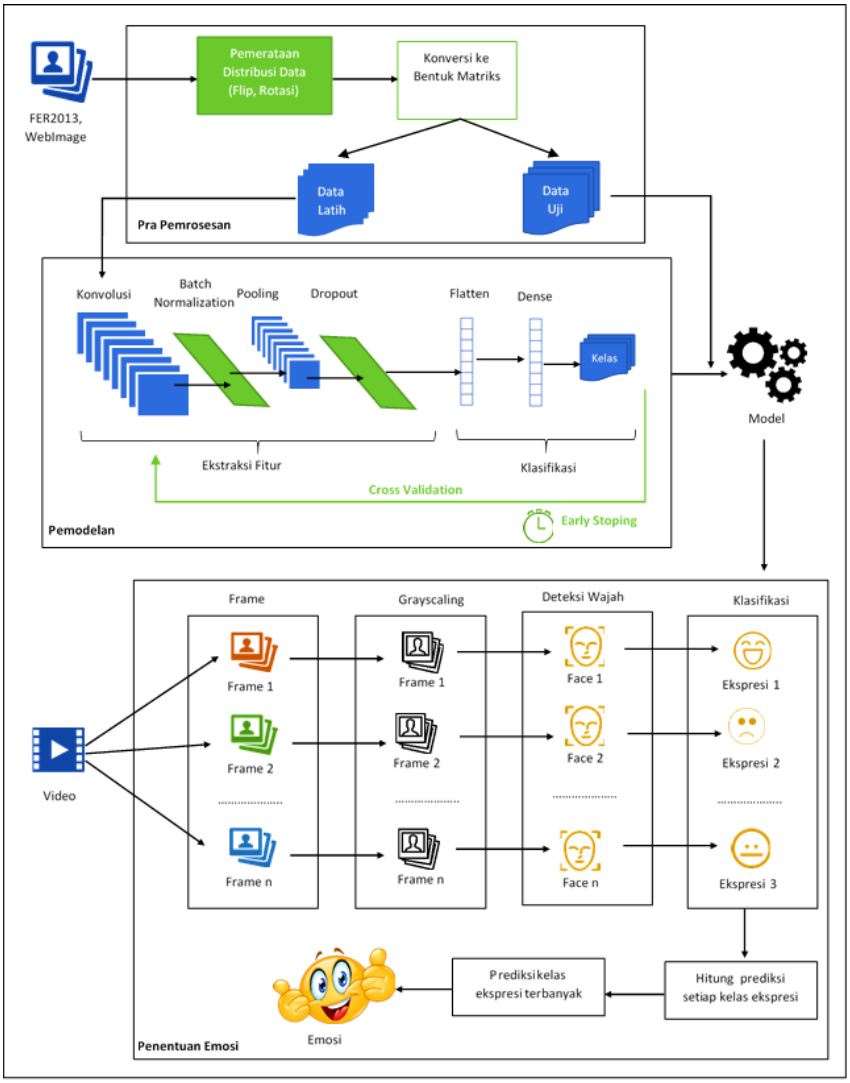
Pengembangan model CNN di dalam artikel ini menggunakan data *The Facial Expression Recognition 2013* (FER-2013) yang dikenalkan pada *International Conference on Machine Learning* (ICML) 2013 (Goodfellow et al., 2013). FER-2013 berisi 35.887 citra *grayscale* wajah berukuran 48x48 yang terdiri dari 7 jenis emosi yang berbeda. Data tersebut sudah dilabeli dan diklasifikasikan menjadi 7 kelas dengan indeks antara 0 sampai dengan 6 seperti pada Tabel 1. Analisis data dilakukan dengan tujuan untuk memperoleh informasi mengenai data yang akan digunakan untuk pemodelan CNN. Tahap ini juga akan memastikan integritas data sehingga tidak menimbulkan masalah pada proses pelatihan. Identifikasi kelas dilakukan dengan tujuan untuk mengetahui kelas dan jumlah anggota setiap kelas yang ada. Integritas data diperlukan untuk menjamin proses pembelajaran berjalan sesuai dengan yang diharapkan. Memastikan integritas data yang dimaksud antara lain memastikan bahwa data-data tersebut seharusnya merupakan data numerik karena data tersebut nantinya digunakan dalam operasi matematika. Selain itu, karena data tersebut berisi piksel-piksel gambar, sehingga harus memenuhi nilai yang sesuai, yaitu antara 0 sampai dengan 255.

Tabel 1. Kelas emosi pada FER2013

Label	Jenis Emosi	Jumlah
0	Marah (*Angry*)	4593
1	Jijik (*Disgust*)	547
2	Takut (*Fear*)	5121
3	Bahagia (*Happy*)	8989
4	Sedih (*Sad*)	6077
5	Terkejut (*Surprise*)	4002
6	Netral (*Neutral*)	6198

2.2. Metode

Penentuan emosi pada video dilakukan melalui 3 tahap, yaitu pra pemrosesan, pemodelan, dan penentuan emosi. Gambar 1 menunjukan tahapan dalam menentukan emosi dalam video dengan CNN.



Gambar 1. Tahapan penentuan emosi pada video

2.2.1. Pra Pemrosesan

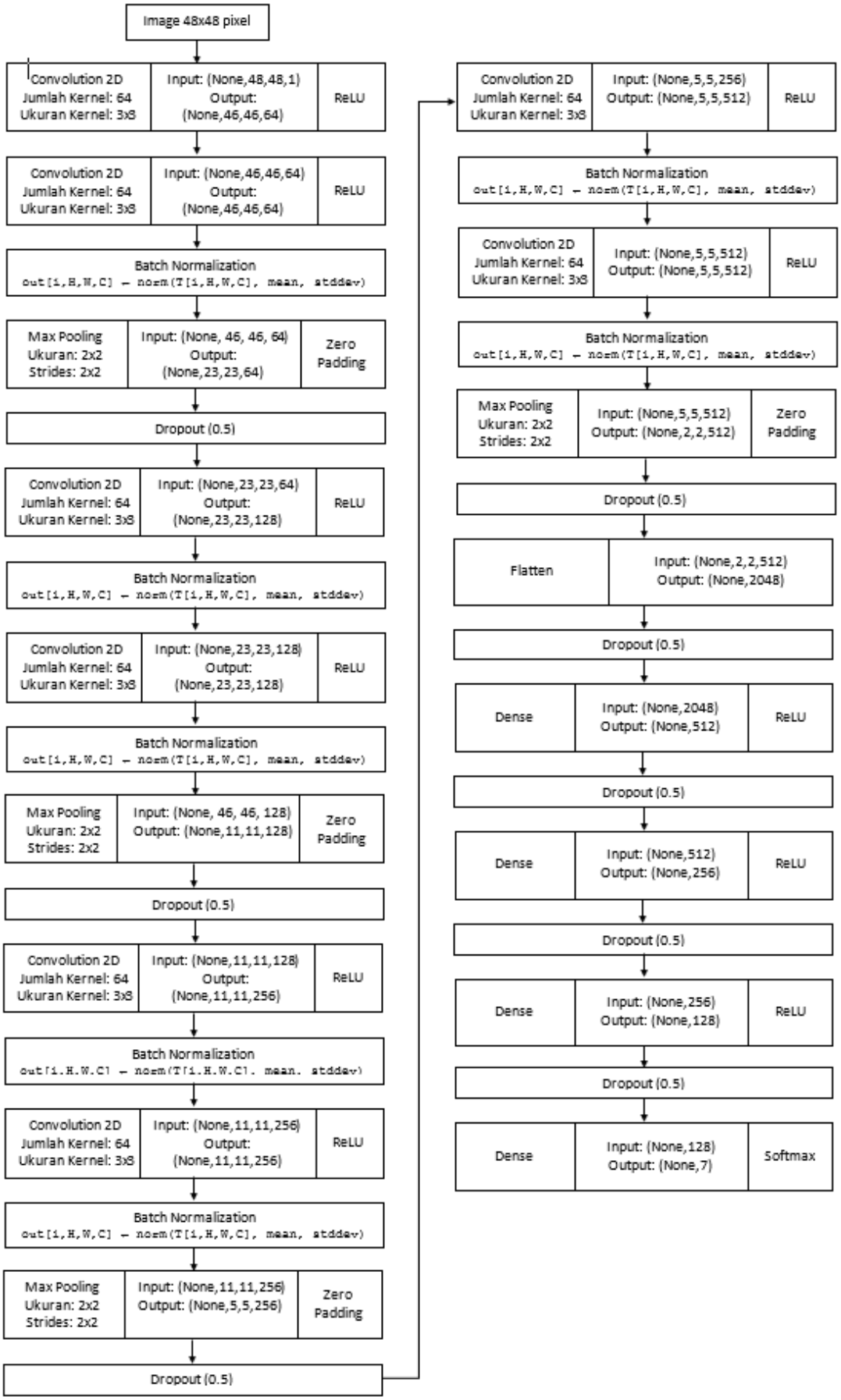
FER2013 memiliki distribusi kelas yang tidak merata, artinya terdapat kelas yang memiliki anggota sangat sedikit, sehingga perlu dilakukan pemerataan distribusi data masing-masing kelas. Untuk itu, penulis menambahkan beberapa data eksternal pada beberapa kelas dengan anggota yang sedikit. Selain itu, penulis juga melakukan penambahan data dengan melakukan flip dan rotasi pada data tersebut. Tujuannya adalah membuat distribusi data pada data latih menjadi lebih merata. Selain itu, data FER-2013 disimpan sebagai file *excel* yang berisi piksel-piksel citra, sehingga perlu dilakukan pra pemrosesan untuk mendapat data yang sesuai untuk pemodelan. Selain itu, pra pemrosesan data juga untuk mengubah label ke dalam matriks.

Di dalam *deep learning*, umumnya data dibagi menjadi 3 yaitu data latih (*training set*), data uji (*testing set*), dan data validasi (*validation set*). Data latih digunakan di dalam *Multilayer Perceptron (MLP)* untuk membentuk sebuah model. Data tersebut digunakan untuk menemukan bobot yang optimal untuk setiap *back-propagation* agar menghasilkan model yang sesuai. Data uji digunakan untuk mengukur kinerja model yang terbentuk. Data uji seharusnya dipisahkan dengan data latih dengan tujuan agar model yang terbentuk memiliki kemampuan generalisasi yang baik dalam melakukan klasifikasi.

2.2.2. Pemodelan

Pemodelan yang dilakukan adalah dengan memanfaatkan pustaka *Keras* dan *Tensorflow* dari Google untuk membantu pemodelan CNN. Tahap pertama dalam pemodelan adalah membuat arsitektur CNN. Arsitektur yang dibagun sangat sederhana dengan tujuan untuk mendapatkan model yang ringan karena akan digunakan pada video yang di dalamnya terdapat banyak ekspresi wajah dan mengalami perubahan yang sangat cepat. Gambar 2 menunjukan arsitektur CNN yang terdiri dari 4 blok ekstraksi fitur dan 1 blok klasifikasi. Blok pertama ektraksi fitur diawali dengan lapisan konvolusi yang terdiri dari 64 kernel dengan ukuran kernel 3x3. Kernel dengan ukuran 3x3 dipilih agar kompleksitas komputasi tidak terlalu besar. Lapisan ini menerima input citra *grayscale* dengan ukuran 48x48 piksel dan menghasilkan output sebuah tensor [*i*,46,48,64] dimana i adalah sebuah citra. Selanjutnya output dari lapisan pertama menjadi input bagi lapisan konvolusi kedua yang terdiri dari 64 kernel dengan ukuran 3x3. Fungsi aktivasi yang digunakan pada lapisan konvolusi adalah *ReLU*. Output dari lapisan konvolusi akan diterima oleh lapisan normalisasi *batch*. Normalisasi *batch* merupakan sebuah metode yang dapat digunakan untuk menormalkan input dari setiap lapisan, dengan tujuan untuk mengatasi masalah pergeseran kovariat internal (*internal covariate shift*). Pergeseran kovariat internal didefinisikan sebagai perubahan distribusi aktivasi jaringan karena perubahan parameter jaringan selama pelatihan (Ioffe & Szegedy, 2015). Penggunaan lapisan *pooling* akan mengurangi dimensi dari *feature map*, sehingga proses komputasi akan semakin cepat karena parameter yang harus di update semakin. Jenis *pooling* yang digunakan dalam pemodelan ini adalah *Max Pooling*, yaitu mengambil nilai tertinggi untuk setiap filter. Ukuran filter pada lapisan *pooling* yang digunakan digunakan untuk semua blok ekstraksi fitur adalah 2x2 dengan strides 2x2 dan *padding* yang digunakan adalah *zero padding*. Lapisan terakhir pada setiap blok ekstraksi fitur adalah *dropout*. Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2015) adalah pendekatan untuk regularisasi dalam jaringan saraf yang membantu mengurangi ketergantungan di antara neuron. Teknik regularisasi ini merupakan cara yang sangat efisien untuk mengurangi *overfitting* di dalam jaringan saraf dengan mencegah co-adaptasi yang kompleks pada data latih. Output dari blok ektraksi fitur pertama adalah sebuah tensor [i,23,23,64].

Pada blok ekstraksi fitur yang kedua terdiri dari 2 lapisan konvolusi, 2 lapisan normalisasi *batch*, 1 lapisan *pooling*, dan 1 lapisan *dropout*. Lapisan konvolusi pada blok ekstraksi fitur yang kedua terdiri dari 128 kernel dengan ukuran 3x3 dan fungsi aktivasi ReLU. Output dari blok ekstraksi fitur yang kedua adalah sebuah tensor [i,11,11,128]. Blok ekstraksi fitur yang ketiga terdiri dari 2 lapisan konvolusi, 2 lapisan normalisasi *batch*, 1 lapisan *pooling*, dan 1 lapisan *dropout*. Lapisan konvolusi pada blok ekstraksi fitur yang ketiga terdiri dari 256 kernel dengan ukuran 3x3 dan fungsi aktivasi ReLU. Output dari blok ekstraksi fitur yang ketiga adalah sebuah tensor [i,5,5,256]. Blok ekstraksi fitur yang keempat terdiri dari 2 lapisan konvolusi, 2 lapisan normalisasi *batch*, 1 lapisan *pooling*, dan 1 lapisan *dropout*. Lapisan konvolusi pada blok ekstraksi fitur yang keempat terdiri dari 512 kernel dengan ukuran 3x3 dan fungsi aktivasi ReLU. Output dari blok ekstraksi fitur yang keempat adalah sebuah tensor [i,2,2,512].



Gambar 2. Arsitektur CNN

Bagian terakhir dalam arsitektur CNN yang dibangun adalah blok klasifikasi yang terdiri dari lapisan *flatten*, 4 lapisan *dense*, dan beberapa lapisan dropout 0.5. Lapisan *flatten* mengubah data *array* multi-dimensi menjadi 1-dimensi sebagai input untuk lapisan berikutnya. Input dari lapisan ini berupa tensor [i,2,2,512] dan menghasilkan output sebuah *array* dengan panjang 2048. Lapisan *dense* yang pertama memiliki filter sebanyak 512 filter dengan fungsi aktivasi *ReLU*, lapisan *dense* kedua memiliki filter sebanyak 256 filter dengan fungsi aktivasi *ReLU*, lapisan *dense* ketiga memiliki filter sebanyak 128 filter dengan fungsi aktivasi *ReLU*, dan lapisan *dense* terakhir memiliki filter sebanyak jumlah kelas klasifikasi, yaitu 7 filter dengan fungsi aktivasi *softmax*. Proses pelatihan dilakukanndengan mekanisme *cross-validation* sebanyak 8 *fold*. *Dataset* akan dibagi menjadi sebanyak 8 bagian, kemudian dilakukan sejumlah 8 percobaan, dimana masing-masing percobaan menggunakan 7 bagian sebagai data latih dan 1 bagian sebagai data uji secara bergantian.

Salah satu masalah yang sering terjadi dalam pelatihan jaringan syaraf adalah pemilihan *epoch* yang akan digunakan. Terlalu banyak *epoch* yang digunakan akan terjadi *overfitting*, sedangkan terlalu sedikit *epoch* yang digunakan akan mengakibatkan *underfitting*. Untuk itu, proses pelatiahan model akan menerapkan teknik *early-stopping,* yaitu sebuah teknik regularisasi dengan menghentikan pelatihan sebelum model menyelesaikan proses pelatihan sejumlah *epoch* yang ditentukan, dengan melakukan monitor terhadap proses pelatihan. Apabila dalam beberapa epoch sebuah model tidak mengalami perbaikan, maka proses pelatihan akan dihentikan.

Selanjutnya untuk mengukur kinerja model yang dihasilkan adalah dengan melakukan pengujian dan evaluasi. Pengujian dilakaukan dengan data uji yang telah disiapkan. Di dalam model klasifikasi, kinerja model yang dihasilkan menggambarkan sejauh mana model tersebut dapat mengklasifikasikan suatu data ke dalam kelas-kelas tertentu. Salah satu metode yang dapat digunakan untuk mengukur kinerja model tersebut adalah *confusion matrix*. Akurasi merupakan suatu cara yang biasa digunakan untuk mengukur kinerja sistem klasifikasi. Perhitungan akurasi bertujuan untuk memperkirakan seberapa efektif algoritma tersebut dengan menunjukkan probabilitas nilai sebenarnya (*actual*) dan keseluruhan label kelas. dengan kata lain akurasi menilai keefektifan algoritma secara keseluruhan (Sokolova, Japkowicz, & Szpakowicz, 2006). Akurasi didefinisikan melalui persamaan 1.

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \qquad (1)$$

dimana:
tp adalah *true positive*, yaitu jumlah data positif yang diklasifikasikan benar.
tn adalah *true negative*, yaitu jumlah data negatif yang diklasifikasikan benar.
fp adalah *false positive*, yaitu jumlah data positif yang diklasifikasikan salah.
fn adalah *false negative*, yaitu jumlah data negatif yang diklasifikasikan salah.

Selain akurasi, presisi dan *recall* juga sering digunakan untuk mengukur kinerja model klasifikasi. Presisi merupakan presentase hasil klasifikasi yang relevan, sedangkan *recall* merupakan presentase total hasil yang relevan yang diklasifikasikan dengan tepat. *Precision* atau *Positive prediction value* merupakan tingkat ketepatan sistem klasifikasi dalam memberikan nilai suatu prediksi. Presisi dihitung dengan membagi jumlah data positif yang terklasifikasi benar dengan jumlah keseluruhan data yang positif, sehingga dapat didefinisikan seperti pada persamaan 2.

$$Precision = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fp_i)} \qquad (2)$$

Recall dapat didefinisikan sebagai rasio dari jumlah total sampel positif yang terklasifikasi benar dibagi dengan jumlah total sampel positif, sebagaimana dalam persamaan.

$$Recall = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fn_i)} \qquad (3)$$

2.2.3. Penentuan Emosi

Penentuan emosi pada video dilakukan dengan memecah video tersebut menjadi beberapa *frame* yang berisi citra digital. Kemudian citra tersebut diubah menjadi citra *grayscale* untuk mempermudah pemetaan warna. Deteksi wajah perlu dilakukan sebelum proses pengenalan ekspresi dengan menggunakan metode *Haar* (Viola & Michael, 2004). Selanjutnya model CNN akan melakukan klasifikasi terhadap citra wajah tersebut. Hasil klasifikasi akan dihitung dan dikelompokan berdasarkan kelas ekspresi. Penentuan emosi akan mengacu pada kelas dengan jumlah anggota terbanyak pada video tersebut.

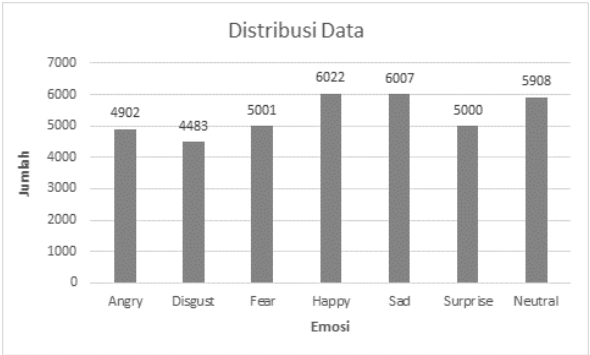
3. HASIL DAN PEMBAHASAN

3.1. Pemerataan Distribusi Data

Pemerataan distribusi data dilakukan dengan menambahkan data pada kelas yang memiliki angggota sedikit dan mengurangi data pada kelas yang memiliki anggota banyak. Data yang ditambahkan berasal dari eksternal data set dan dengan melakukan flip dan rotasi pada pada citra yang ada. Hasil flip dan rotasi data dapat dilihat pada Gambar 3.

Gambar 3. Hasil flip dan rotasi data

Hasil pemerataaan distribusi data dapat dilihat pada Gambar 4.



Gambar 4. Distribusi data setelah dilakukan pemerataanPemodelan

3.2. Pelatihan Model

Proses pelatihan model dilakukan dengan menerapkan *cross-validation*, artinya melakukan validasi terhadap model yang dihasilkan dengan membagi data latih dan data validasi menjadi sejumlah k *fold* (partisi), kemudian dilakukan percobaan pelatihan sebanyak k menggunakan data latih dan data validasi tersebut. Hal ini bertujuan untuk melatih model agar dapat mengenali data baru dengan baik. Data dibagi menjadi 8 partisi atau 12.5 % untuk setiap partisinya, 7 partisi sebagai subset pembelajaran, dan sisanya sebagai validasi. Proses pembagian antara subset pelatihan dan subset validasi akan terus dilakukan sehingga semua data akan berperan sebagai subset pelatihan dan subset validasi secara bergantian. Dari hasil pelatihan model dengan

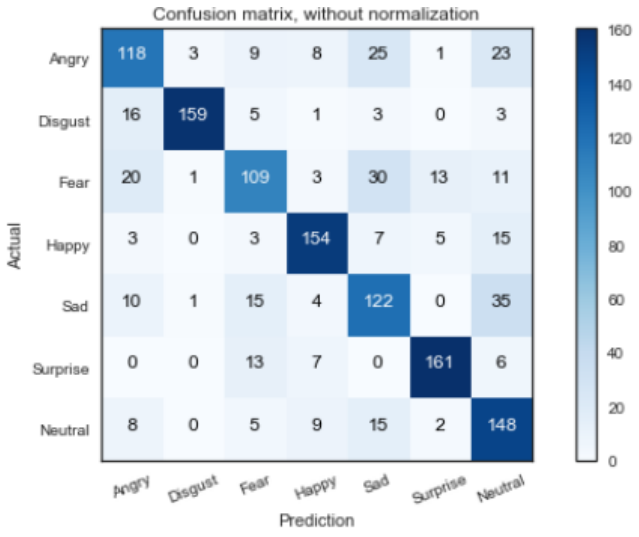
menggunakan *cross-validation* sebanyak 8 fold, *loss* dan akurasi yang dihasilakan disajikan dalam Tabel 3.

Tabel 2. Hasil pelatihan menggunakan Cross Validation

Fold	Loss	Accuracy
1	0.7987	0.74
2	0.7884	0.7495
3	0.8337	0.718
4	0.7625	0.7458
5	0.8488	0.7126
6	0.7968	0.7415
7	0.7607	0.7411
8	0.7869	0.7338
Average	**0.7971**	**0.7353**

3.3. Evaluasi

Karena jumlah populasi tidak diketahui, maka dalam menentukan jumlah sampel uji, penulis menggunakan rumus perhitungan jumlah sampel *Paul Leedy* dengan *confident level* 97% dan tingkat kesalahan tidak lebih dari 3%. Berdasarkan rumus tersebut, diperoleh jumlah sampel minimal sebanyak 1309 sampel. Penulis menggunkan 1309 data sebagai data uji yang terbagi menjadi 7 kelas, yaitu *Angry* (187 data), *Disgust* (187 data), *Fear* (187 data), *Happy* (187 data), *Sad* (187 data), *Surprise* (187 data), dan *Neutral* (187 data). Hasil *confusion matrix* dari model yang dihasilkan disajikan dalam Gambar 5.



Gambar 5. Hasil confusion matrix

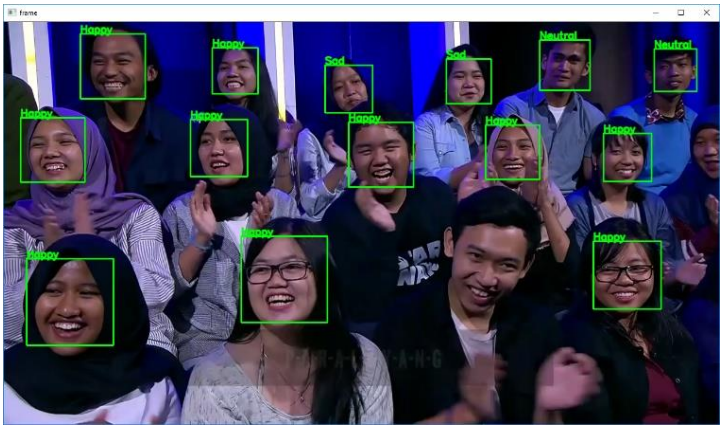
Berdasarkan persamaan perhitungan akurasi, presisi, dan *recall*, diperoleh akurasi sebesar 74,17%, presisi sebesar 74,07%, dan *recall* sebesar 74,18%. Semakin tinggi nilai recall menunjukan bahwa kelas tersebut dapat diprediksi dengan benar. Apabila nilai *recall* tinggi tetapi nilai presisi rendah berarti banyak sampel positif yang diberikan dapat dikenali dengan benar tetapi banyak juga hasil positif yang palsu (seharusnya bukan positif). Sebaliknya, jika nilai *recall* rendah tetapi presisi tinggi mengindikasikan bahwa banyak sampel positif yang tidak dapat dikenali dengan benar tetapi menunjukan bahwa sampel positif tersebut memang benar positif.

3.4. Percobaan

Model yang terbentuk diimplementasikan pada video dengan objek wajah tunggal dan objek wajah banyak. Pada objek wajah tunggal penulis menerapkan model CNN untuk mengklasifikasikan perubahan ekspresi pada setiap frame dan mencatatnya, kemudian disimpulkan ekspresi yang dominan pada wajah tersebut. Pada kasus pertama, penulis mengambil contoh video dari Presiden RI ke-6 yang berjudul “Tanggapan SBY soal Tuduhan Terkait Sakitnya Ibu Ani Yudoyono”. Di dalam video tersebut SBY mengklarifikasi tuduhan terkait sakitnya Ibu Ani dengan ekspresi marah. Model CNN akan melakukan klasifikasi terhadap setiap perubahan ekspresi wajah yang terjadi. Video yang berdurasi 3 menit 17 tersebut terdeterksi 2193 ekspresi wajah dengan ekspresi marah yang paling dominan (2131 atau 97,17%). Proses dan hasil penentuan emosi pada video dengan objek tunggal disajikan pada Gambar 6.

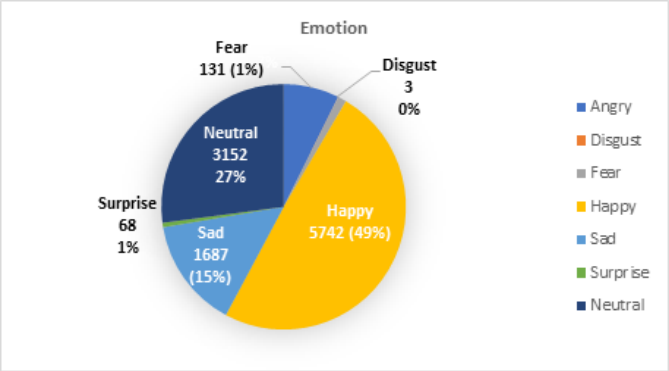
Gambar 6. Proses dan hasil penentuan emosi pada video dengan objek tunggal

Pada kasus kedua (video dengan objek banyak), penulis menggunakan sampel video acara komedi televisi. Acara tersebut menggambarkan suasana bahagia dimana pemain dan penonton pada acara tersebut banyak menunjukan ekspresi tertawa ceria. Gambar 7 menunjukan implementasi model klasifikasi CNN pada video dengan objek wajah banyak.



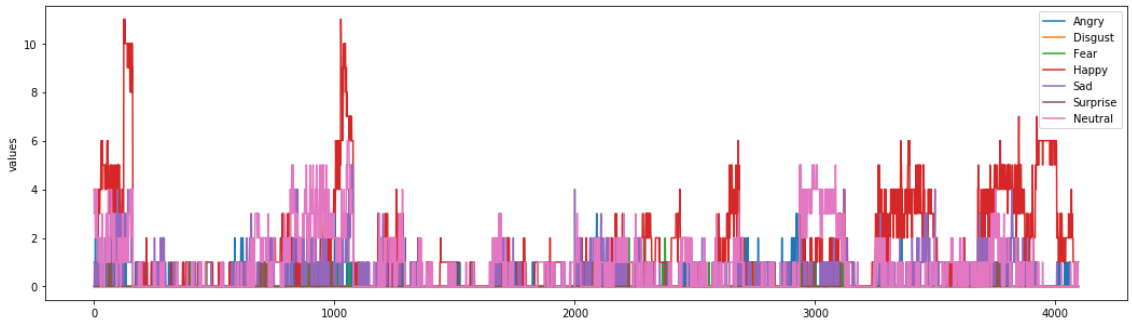
Gambar 7. Implementasi model klasisfikasi CNN pada video dengan objek wajah banyak

Pada kasus video kedua, terdeteksi ekspresi wajah sebanyak 11.640 dengan ekspresi bahagia yang paling dominan dengan 5.742 ekspresi wajah (49%). Gambar 8 menunjukan persentase emosi pada kasus video kedua (objek banyak).



Gambar 8. Presentase emosi pada video dengan objek banyak

Di dalam video dengan objek wajah banyak, jumlah wajah dengan ekspresi tertentu dapat ditentukan pada setiap *frame*. Jumlah ekspresi yang terdeksi pada setiap *frame* menggambarkan suasana pada saat itu. Gambar 9 menunjukan jumlah ekspresi pada setiap *frame* secara bersamaan. Dalam kasus ini jumlah ekspresi bahagia di saat bersamaan berada pada *frame* ke-123 dan 1034 dengan 11 ekspresi bahagia secara bersamaan.



Gambar 9. Jumlah ekspresi pada setiap frame

4. KESIMPULAN

Emosi manusia dapat ditemukan dengan melakukan klasifikasi terhadap setiap perubahan ekspresi manusia dengan memanfaatkan teknologi kecerdasan buatan, terutama *machine learning*. Model klasifikasi menggunakan *Convolution Neural Network (CNN)* menghasilkan akurasi sebesar 74,17%, presisi sebesar 74,07%, dan rata-rata recall 74,18%. Pemerataan distribusi data latih dapat meningkatkan kinerja model. Artikel ini menghasilkan sistem klasifikasi yang mampu mengenali setiap perubahan ekspresi wajah kemudian menghitung jumlah setiap jenis ekspresi yang dapat digunakan untuk menentukan emosi seseorang.

DAFTAR PUSTAKA

Chauhan, M. (2014). Study & Analysis of Different Face Detection Techniques, *5*(2), 1615–1618.

Goodfellow, I. J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., … Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, *64*, 59–63. https://doi.org/10.1016/j.neunet.2014.09.005

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*, 448–456.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2019). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *ArXiv*, 1–67.

Khoshdeli, M., Cong, R., & Parvin, B. (2017). Detection of Nuclei in H & E Stained Sections Using

Convolutional Neural Networks. *Conference: IEEE International Conference on Biomedical Health Informatics*, (February).

Kotropoulos, C., & Pitas, I. (1997). Rule-based face detection in frontal views. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *4*, 2537–2540. https://doi.org/10.1109/icassp.1997.595305

Lecun, Y., Bottou, L., Bengio, Y., & Ha, P. (1998). Gradient-Based Learning Applied to Document Recognition, (November), 1–46.

Li, S. Z., & Jain, A. K. (2011). *Handbook of Face Recognition*. (S. Z. Li1 & A. K. Jain, Eds.), *Handbook of Face Recognition* (2nd ed.). Springer. https://doi.org/10.2990/29_1_103

Liew, S. S., Khalil-Hani, M., Ahmad Radzi, S., & Bakhteri, R. (2016). Gender classification: A convolutional neural network approach. *Turkish Journal of Electrical Engineering and Computer Sciences*, *24*(3), 1248–1264. https://doi.org/10.3906/elk-1311-58

Mahmoodi, M. R. (2017). Fast and Efficient Skin Detection for Facial Detection. *ArXiv*.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *Conference: British Machine Vision Conference 2015*, (Section 3), 41.1-41.12. https://doi.org/10.5244/c.29.41

Saito, S., Lingyu, W., Liwen, H., Nagano, K., & Hao, L. (2016). Photorealistic facial texture inference using deep neural networks. *ArXiv*, 14 pp.-14 pp.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *Australasian Joint Conference on Artificial Intelligence*, *WS-06-06*(c), 24–29.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2015). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Tian, Y., Kanade, T., & Cohn, J. F. (2011). Facial Expression Recognition. In *Handbook of Face Recognition* (pp. 487–519). https://doi.org/10.1007/978-0-85729-932-1

Trigueros, D. S., Meng, L., & Hartnett, M. (2018). Face Recognition: From Traditional to Deep Learning Methods. *ArXiv*, (October 2018).

Vasanth, P. ., & Nataraj, K. . (2015). Facial Expression Recognition Using SVM Classifier. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, *3*(1), 16–20.

Viola, P., & Michael, J. (2004). Robust Real-Time Face Detection. *Nternational Journal of ComputerVision*, *57*, 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb.

Yang, M.-H., Kriegman, D. J., & Ahuja, N. (2002). Detecting Faces In Image : A Survey - Presentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(1), 1–25.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A face detection benchmark. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 5525–5533. https://doi.org/10.1109/CVPR.2016.596