AUTOMATIC PARENTAL GUIDE SCENE CLASSIFICATION MENGGUNAKAN METODE DEEP CONVOLUTIONAL NEURAL NETWORK DAN LSTM

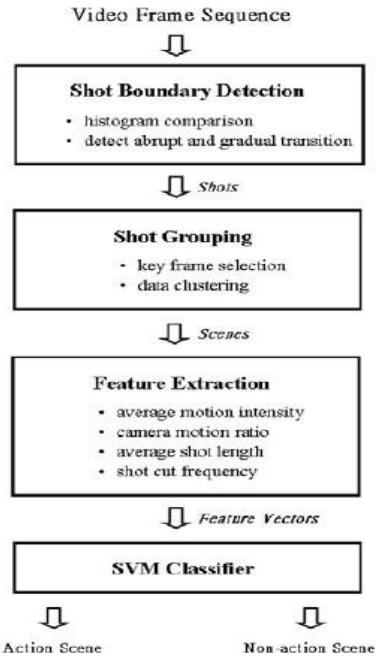
Riko Gunawan, *Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya*,
Yosi Kristian, *Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya*

***Abstrak*— Menonton film merupakan salah satu hobi yang paling digemari oleh berbagai kalangan. Seiring dengan semakin bertambahnya film yang beredar di pasaran, semakin banyak pula konten tidak pantas pada film-film tersebutu. Oleh karena itu, dibutuhkan sebuah metode untuk mengklasifikasikan film agar konten yang ditonton sesuai dengan usia penonton. Konten film yang kurang cocok untuk pengguna di bawah umur yang akan diklasifikasikan pada penelitian ini antara lain: kekerasan, pronografi, kata-kata kasar, minuman keras, penggunaan obat-obatan terlarang, merokok, adegan mengerikan (horror) dan intens. Metode klasifikasi yang digunakan berupa modifikasi dari convolutional neural network dan LSTM. Gabungan kedua metode ini dapat mengakomodasi data training dalam jumlah yang kecil, serta dapat melakukan multi klasifikasi berdasarkan video, audio, dan subtitle film. Penggunaan multi klasifikasi ini dikarenakan sebuah film selalu memiliki lebih dari satu klasifikasi. Dalam proses training dan testing pada penelitian ini digunakan sebanyak 1000 data untuk klasifikasi video, 600 data klasifikasi audio, dan 400 data klasifikasi subtitle yang didapatkan dari internet. Dari hasil percobaan dihasilkan tingkat akurasi yang diukur dengan menggunakan F1-Score sebesar 0.922 untuk klasifikasi video, 0.741 untuk klasifikasi audio, dan 0.844 untuk klasifikasi subtitle dengan rata-rata akurasi sebesar 0.835. Pada penelitian berikutnya akan dicoba dengan menggunakan metode Deep Convolutional Neural Network yang lain serta dengan memperbanyak jumlah dan variasi dari data testing.**

***Kata Kunci*—Convolutional Neural Network, Deep Learning, LSTM, Movie Classification**

I. PENDAHULUAN

Pesatnya perkembangan dalam dunia perfilman menciptakan sebuah tantangan baru bagi penikmat film, khususnya mereka yang telah berkeluarga dan memiliki anak. Hal ini dikarenakan dalam proses pembuatannya, film yang dibuat dapat memiliki beragam konen yang kurang sesuai untuk dinikmati oleh semua kalangan [1]–[3]. Konten yang dimaksud dalam hal ini dapat berupa adegan kekerasan [4], [5], horror, ponorgrafi, dan lain-lain. Studi lebih lanjut menunjukan, bahwa anak-anak berumur 8 tahun yang terpapar tontonan dengan unsur kekerasan memiliki kecenderungan perilaku agresif ketika mereka mencapai umur 18 tahun [1]. Selain konten kekerasan, konten horror juga memiliki dampak negatif khususnya secara psikilogis. Semakin sering seorang anak terpapar dengan konten horror, maka anak akan lebih mudah memiliki phobia [6]. Pada penelitian ini, peneliti bermaksud untuk memberikan sebuah metode klasifikasi konten film, yang lebih lanjut akan disebut sebagai scene. Klasifikasi scene ini diharapkan dapat bermanfaat bagi orang tua untuk membantu dalam membatasi paparan scene film, khususnya yang kurang sesuai bagi anak-anak. Pendekatan klasifikasi yang akan digunakan berupa visual,audio, dan teks [7] dalam scene film. Jenis klasifikasi yang digunakan adalah: Violence & Gore [8], [9], Sex & Nudity [10], Profanity [11], [12], Alcohol, Drugs, and Smoking, dan Frightening & Intense.



Gambar. 1. Alur klasifikasi action scene menggunakan SVM

II. TINJAUAN PUSTAKA

A. *Action Scene Detection with Support Vector Machines*

Pada paper ini dilakukan pendekatan metode berdasarkan konstruksi dari struktur video. Dalam hal ini penulis hanya menggunakan fitur visual dari video sebagai input dari SVM. SVM digunakan oleh penulis sebagai metode klasifikasi dikarenakan SVM memiliki tingkat akurasi yang cukup tinggi dengan training sample yang relatif kecil. Hasil klasifikasi dari input video akan berupa action scenes atau bukan. Alur klasifikasi action scene dapat dilihat pada Gambar 1.

Tahap pertama yang dilakukan oleh penulis dalam melakukan klasifikasi ini adalah dengan membagi video input menjadi shots atau frame, kemudian akan dipilih beberapa key frame dari hasil segmentasi tersebut. Key frame yang dimaksud adalah sebuah frame yang dianggap

dapat mewakili konten secara utuh dari sebuah video. Untuk mendapatkan keyframe ini, penulis menggunakan metode clustering.

Setelah didapatkan keyframe yang sesuai, maka hasil akan diteruskan kedalam SVM. Penulis memilih metode SVM ini dikarenakan memiliki structural risk minimalization principle, sehingga dapat mengurangi jumlah kesalahan pada training set meskipun data yang digunakan tidak terlalu banyak. Hasil dari penelitian ini akan ditunjukan pada tabel berikut, dimana penulis menggunakan 4 buah film sebagai test video.

TABEL 1
TINGKAT AKURASI KLASIFIKASI TEST VIDEO

Movie ID	No of Action Scene	Correct Detection	*Missed Detection*	*False Detection*
1	21	17	4	6
2	16	14	2	3
3	14	13	1	3
4	19	16	3	5

B. Violence Detection in Video Using Computer Vision Techniques

Tujuan dari penelitian ini adalah agar dapat menemukan sebuah action recognition yang dapat digunakan untuk mendeteksi perkelahian atau tindak kekerasan, baik dalam film maupun hasil dari video surveillance. Selain itu peneliti juga bertujuan untuk menghasilkan sebuah dataset video tindak kekerasan dengan menggunakan metode action recognition STIP dan MoSIFT.

TABEL 2
TINGKAT AKURASI DETECTION MENGGUNAKAN CLIP HOCKEY

Vocabulary	STIP (HOG) + HIK	STIP (HOF) + HIK	*MoSIFT + HIK*
50	87.8%	83.5%	87.5%
100	89.1%	84.3%	89.4%
150	89.7%	85.9%	89.5%
200	89.4%	87.5%	90.4%
300	90.8%	87.2%	90.4%
500	91.4%	87.4%	90.5%
1000	91.7%	88.6%	90.0%

Agar dapat mencapai tujuan di atas, peneliti mengumpulkan 1000 video clip dari olahraga hockey. Setiap video clip berisi 50 frames dengan ukuran 420x576 pixels dan proses pemilahan apakah video termasuk kekerasan atau tidak dilakukan secara manual oleh peneliti. Video clip yang telah dikumpulkan akan diteruskan kedalam metode STIP dan MoSIFT sebagai model dari action recognition. Selain menggunakan kedua metode di atas, peneliti juga menggunakan Bag of Words yang diadaptasi sehingga menghasilkan cluster yang didapat dari metode k-means clustering. Dalam penelitian ini cluster yang digunakan sebagai test sebanyak 50, 100, 150, 200, 300, 500 dan 1000 cluster center.

TABEL 3
TINGKAT AKURASI DETECTION MENGGUNAKAN CLIP FILM

Vocabulary	STIP (HOG) + HIK	STIP (HOF) + HIK	*MoSIFT + HIK*
50	44.5%	51.2%	76.0%
100	45.0%	56.5%	79.5%
150	49.0%	59.0%	80.0%
200	46.5%	53.5%	80.0%
300	44.5%	52.5%	87.5%
500	44.5%	50.5%	89.5%
1000	38.5%	52.5%	89.0%

Dari kedua tabel di atas, peneliti mengambil kesimpulan bahwa metode MoSIFT memiliki tingkat akurasi yang lebih baik jika dibandingkan dengan metode STIP. Peneliti juga menyimpulkan bahwa penambahan jumlah vocabulary tidak menghasilkan peningkatan akurasi yang berarti.

C. Horror Video Scene Recognition via Multiple Instance Learning

Penelitian yang dilakukan dalam paper ini akan melalui tiga tahapan utama, yaitu video segmentation, feature extraction, dan multiple instance learning method. Dalam proses video segmentation, peneliti menggunakan metode mutual information (MI) berdasarkan dari [13]. Metode ini menggunakan video transition dan fades sebagai dasar acuan.

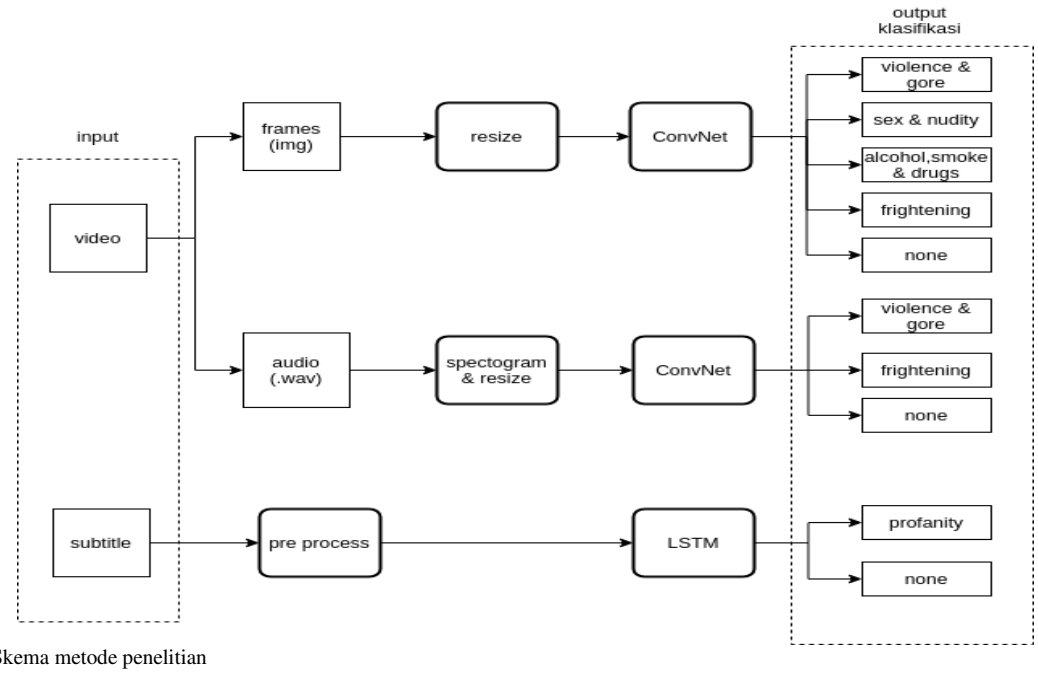
Ketika proses segmentasi telah selesai, maka hasil dari setiap segmen akan diambil sebuah frame utama yang akan dijadikan input. Visual feature extraction akan dilakukan pada tiap keyframe dengan mencari variasi warna dengan rumus covariance L (Luv coor frame) sebagai berikut:

$$\rho = \begin{bmatrix} \sigma_L^2 & \sigma_{Lu}^2 & \sigma_{Lv}^2 \\ \sigma_{Lu}^2 & \sigma_L^2 & \sigma_{uv}^2 \\ \sigma_{Lv}^2 & \sigma_{uv}^2 & \sigma_v^2 \end{bmatrix} \quad (1)$$

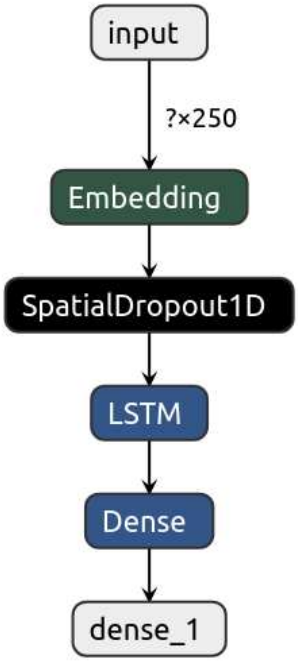
Selain feature extraction pada keyframe, peneliti juga melakukan feature extraction pada audio dan color emotional features sebelum akhirnya akan diteruskan kedalam tahapan multiple instance learning. Multiple-Instace Learning (MIL)[6] merupakan sebuah metode supervised classification dimana setiap training class memiliki asosiasi terhadap sebuah pola. Pada kasus ini, sebuah film akan diklasifikasikan sebagai film horror jika salah satu segmentasi terklasifikasi sebagai film horror. Sebagai perbandingan, peneliti juga menggunakan metode Continous K-Nearest Neighbour (CKNN), EM-DD, serta SI.

III. METODE PENELITIAN

Proses yang dilakukan dalam penelitian ini terbagi menjadi 2 bagian yaitu pengumpulan data dan pembuatan model Deep Convolutional Neural Network dan LSTM sebagai penentu klasifikasi. Skema proses ini lebih jelas dapat dilihat pada Gambar 2.



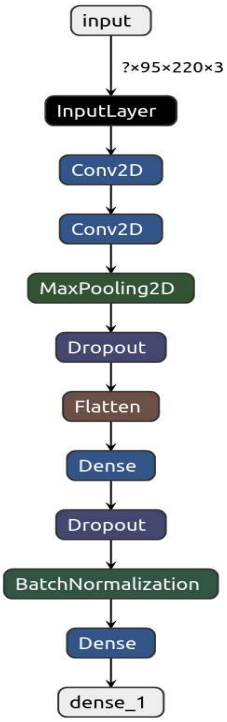
Gambar. 2. Skema metode penelitian



Gambar. 3. Model arsitektur LSTM untuk klasifikasi subtitle

A. *Proses Pengumpulan Data*

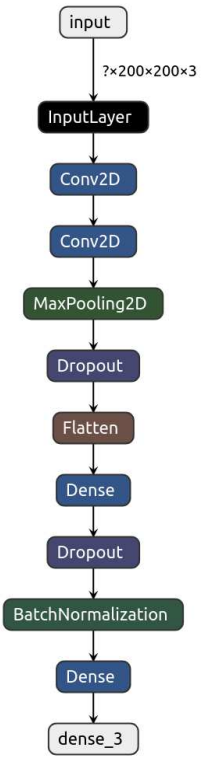
Data yang dikumpulkan oleh peneliti berupa video clip didapatkan melalui beberapa sumber, baik diunduh melalui situ youtube maupun pemotongan film secara manual dari file film digital. Film yang dikumpulkan oleh peneliti merupakan film mancanegara dengan subtitle berbahasa inggris dengan berbagai genre film kecuali animasi. Data film yang berhasil dikumpulkan oleh peneliti lebih lanjut dipotong sehingga setiap video clip akan memiliki durasi sepanjang 10 detik beserta dengan pemotongan subtitle yang sesuai dengan durasi video. Potongan video dan subtitle kemudian diberi label secara manual dengan 6 klasifikasi yaitu violence and gore, sex and nudity, frightening, alcohol, smoke and drugs, profanity, dan none. Klasifikasi ini didasarkan pada klasifikasi parental guidance dari imdb.



Gambar. 4. Model Arsitektur ConvNet untuk klasifikasi audio

Setelah setiap potongan video dan subtitle berhasil dikumpulkan dan diberi label, maka dilakukan proses lebih lanjut. Potongan video kemudian dikonversi menjadi image dan spectogram. Proses konversi dari potongan video menjadi image adalah dengan mengambil 1 frame untuk setiap detik dari video, maka setiap video menghasilkan 10 image yang kemudian diresize ulang menjadi berukuran

200x200 pixel. Sedangkan spectogram didapat dari mengkonversi video menjadi file .wav yang kemudian dikonversi menjadi gambar spectogram dengan jenis amplitude. Hasil gambar spectogram ini akan diresize menjadi berukuran 95x220 pixel.



Gambar. 5. Model Arsitektur ConvNet untuk klasifikasi video

Sedangkan subtitle dari video akan melalui proses normalisasi, case folding dan tokenisasi. Proses ini bertujuan agar didapatkan data subtitle yang bebas dari noise seperti simbol khusus dan angka yang terdapat pada subtitle. Dimana subtitle yang dikumpulkan oleh peneliti berupa subtitle film dengan bahasa inggris dan subtitle yang masih belum disensor agar klasifikasi yang didapat lebih akurat

B. Pembuatan Model

Model arsitektur yang digunakan oleh peneliti merupakan model custom yang dibuat oleh peneliti berdasarkan dari simple CNN dan simple LSTM. Model arsitektur untuk setiap tahapan dapat dilihat pada Gambar 3,4, dan 5. Dalam pembuatan model ini juga diberikan dropout layer sebesar 0.3 agar tidak terjadi overfitting pada data pelatihan. Layer ini bertujuan untuk memilih neuron secara acak untuk tidak dipakai dalam proses training. Selain itu peneliti juga menambahkan layer batch normalization hal ini dilakukan untuk meningkatkan akurasi dari model serta mengurangi kebutuhan data training yang besar. Namun dropout layer ini tidak digunakan pada model LSTM, hanya pada model ConvNet.

IV. HASIL DAN PEMBAHASAN

Dalam bagian ini dijelaskan tentang hasil dari uji coba yang dilakukan peneliti untuk mendapatkan akurasi terbaik dari model yang dibuat. Testing yang dilakukan oleh peneliti adalah dengan melakukan variasi pada jumlah cluster data testing, alat ukur tingkat akurasi yang digunakan oleh peneliti adalah F1 score dari hasil model yang digunakan. Data yang digunakan terbagi menjadi data training dan data testing dengan 90% sebagai training data dan 10% sebagai testing dikarenakan terbatasnya jumlah data yang dikumpulkan oleh peneliti.

Jumlah variasi cluster data yang digunakan oleh peneliti adalah dengan variasi 30,100,150, dan 200 cluster data untuk setiap kategori tag yang bersangkutan pada setiap model. Sebagai contoh pada model klasifikasi subtilte yang menghasilkan 2 label klasifikasi profanity dan none, dengan menggunakan cluster data sebesar 30 maka data yang akan digunakan adalah 30 data dengan label profanity dan 30 data dengan label none.

TABEL 4
TINGKAT AKURASI DETECTION DENGAN BATCH NORMALIZATION

Vocabulary per Label	Subtitle	Video	*Audio*	*Rata-rata*
30	0.667	0.848	0.303	0.606
100	0.649	0.881	0.707	0.745
150	0.756	0.877	0.733	0.788
200	0.844	0.922	0.741	0.835

Tingkat akurasi klasifikasi test video, akurasi detection menggunakan clip hockey, akurasi menggunakan clip film, dan akurasi dengan batch normalization dapat dilihat pada Tabel 1, 2, 3, dan 4. Dari hasil percobaan yang dilakukan terhadap model ditemukan bahwa penggunaan batch normalization layer dapat meningkatkan akurasi dari data training, khususnya pada data spectogram audio yang memiliki total data yang jauh lebih kecil jika dibandingkan dengan data video. Selain itu peningkatan vocabulary pada model juga memiliki dampak yang cukup signifikan pada peningkatan akurasi secara keseluruhan dalam mengklasifikasikan jenis label pada cuplikan film.

V. KESIMPULAN

Dengan menggunakan gabungan dari model yang telah dibuat oleh peneliti khususnya dengan menggunakan batch normalization layer dapat memberikan nilai akurasi yang tinggi dengan rata-rata sebesar 0.835 meskipun dengan jumlah data training yang relatif kecil. Namun pada penelitian mendatang akan dicoba dengan menggunakan metode Deep Convolutional Neural Network yang lain serta dengan memperbanyak jumlah dan variasi dari data testing.

DAFTAR PUSTAKA

[1] L. H. Chen, H. W. Hsu, L. Y. Wang, and C. W. Su, “Violence detection in movies,” 2011, doi: 10.1109/CGIV.2011.14.
[2] U. A. Khan, N. Ejaz, M. A. Martinez-Del-Amor, and H. Sparenberg, “Movies tags extraction using deep learning,” 2017, doi: 10.1109/AVSS.2017.8078459.
[3] Q. Dai *et al.*, “Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning,” in *CEUR Workshop Proceedings*, 2015, vol. 1436.
[4] L. H. Chen, C. W. Su, C. F. Weng, and H. Y. M. Liao, “Action scene detection with support vector machines,” *J. Multimed.*, vol.

4, no. 4, 2009, doi: 10.4304/jmm.4.4.248-253.

[5] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems*, 2014, vol. 1, no. January.

[6] J. Wang, B. Li, W. Hu, and O. Wu, “Horror video scene recognition via multiple-instance learning,” 2011, doi: 10.1109/ICASSP.2011.5946656.

[7] L. Li, L. Xiao, N. Wang, G. Yang, and J. Zhang, “Text classification method based on convolution neural network,” in *2017 3rd IEEE International Conference on Computer and Communications, ICCC 2017*, 2018, vol. 2018-January, doi: 10.1109/CompComm.2017.8322884.

[8] E. Acar, M. Irrgang, D. Maniry, and F. Hopfgartner, “Detecting violent content in hollywood movies and user-generated videos,” *Adv. Comput. Vis. Pattern Recognit.*, vol. 66, 2015, doi: 10.1007/978-3-319-14178-7_11.

[9] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, “Violence detection in video using computer vision techniques,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6855 LNCS, no. PART 2, doi: 10.1007/978-3-642-23678-5_39.

[10] M. Moustafa, “Applying deep learning to classify pornographic images and videos,” *arXiv Prepr. arXiv1511.08899*, 2015.

[11] H. Yenala, A. Jhanwar, M. K. Chinnakotla, and J. Goyal, “Deep learning for detecting inappropriate content in text,” *Int. J. Data Sci. Anal.*, vol. 6, no. 4, 2018, doi: 10.1007/s41060-017-0088-4.

[12] N. Nikhil, R. Pahwa, M. K. Nirala, and R. Khilnani, “Lstms with attention for aggression detection,” in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 2018, pp. 52–57.

[13] Z. Cernekova, C. Kotropoulos, N. Nikolaidis, and I. Pitas, “Video shot segmentation using fusion of SVD and mutual information features,” in *2005 IEEE International Symposium on Circuits and Systems*, 2005, pp. 3849–3852.