A DEEP REINFORCEMENT LEARNING FRAMEWORK FOR IDENTIFYING FUNNY SCENES IN MOVIES

Haoqi Li[1], *Naveen Kumar*[2], *Ruxin Chen*[2], *Panayiotis Georgiou*[1]

[1]University of Southern California, Los Angeles, CA, USA
[2]SONY Interactive Entertainment LLC, San Mateo, CA, USA

ABSTRACT

This paper presents a novel deep Reinforcement Learning (RL) framework for classifying movie scenes based on affect using the face images detected in the video stream as input. Extracting affective information from the video is a challenging task modulating complex visual and temporal representations intertwined with the complex aspects of human perception and information integration. This also makes it difficult to collect a large annotated corpus restricting the use of supervised learning methods. We present an alternative learning framework based on RL that is tolerant to label sparsity and can easily make use of any available ground truth in an online fashion. We employ this modified RL model for the binary classification of whether a scene is funny or not on a dataset of movie scene clips. The results show that our model correctly predicts 72.95% of the time on the 2-3 minute long movie scenes while on shorter scenes the accuracy obtained is 84.13% .

***Index Terms*—** reinforcement learning, movie scene, facial expression, classification

1. INTRODUCTION

With the widespread video-on-demand mobile phone applications and websites (e.g. youtube.com), people can access unprecedented amount of video from diverse online resources. Because of that, extracting video metadata (e.g. movie scene annotations) is becoming increasingly more important in facilitating efficient search and recommendations. For example, in a movie recommendation system, the system needs to identify the affective information of a movie scene and reference it during subsequent user requests according to explicit queries or user interests.

However, movie scene affective labels, such as "funny" are complex constructs that can often be implicit and covert. Hence most reliable movie annotations are still manually generated. When human annotators try to identify the affective label of a movie scene, they need to take longer context into consideration. This is a complex non-linear process that encompasses the challenges of data representation and human perception. In addition, affective label prediction from movie data is a much more challenging and complex task compared to the prediction on well-prepared standard emotion database; it contains nuances not necessarily present in emotion corpora and more realistic capture conditions. Moreover, scene level affective labels convey cumulative information from multiple modalities, such as the visual, acoustic and lexical channels, that may not always be congruent. For instance, a "funny" label may be due to the actors' facial expression but it may be because of the words they said, or just due to the mood imposed by background music. Despite the large amount of available movie data, the amount of accurately labeled data is severely limited due to copyright and cost of annotation. Thus, affective movie scene classification is still a complicated and challenging task in terms of both algorithmic design and training data.

This work was supported by Sony Interactive Entertainment LLC.

A lot of related research work focuses on movie affective content analysis. For example, Wang *et al.*[1] gave a detailed survey on video affective content analysis, and Irie *et al.*[2] provided a movie affective scene classification algorithm with designed audio-visual features. Chen *et al.*[3] were trying to detect violent scenes from a movie, and related work [4, 5] tried to utilize multimodal features to predict dynamic emotion in movies. Recently, with the use of deep learning methods, Acar *et al.*[6, 7] used convolutional neural networks (CNNs) to extract a mid-level representation and applied it to the affective content analysis of video. However, most of the related research work has been along the route of supervised machine learning approaches.

One of the alternatives to redesigning the system under constraints of data scarcity would be to train one that can understand a scene's affective content through a process of trial and error not unlike how humans themselves first learn to understand and reciprocate emotions. In this paper, we use reinforcement learning for this purpose of movie scene affective label prediction, in an attempt to bridge the gap between these two areas which might otherwise be considered unrelated. In using a RL system, we also try to make a case for online learning for a complex task such as affective label prediction with intermittent human supervision. Consider a scenario where humans may label any instance in a long video clip. We would like to exploit the sparse human annotations available in the training process. We thus want to create a machine learning system that can explore the decision space irrespective of whether label is present, but is simultaneously also able to exploit the human input when available. This mechanism of introduction of human-in-the-loop is similar to training by Reinforcement Learning, in which information from action and feedback reward can be incorporated in the training loop.

Reinforcement learning (RL) is a framework for an experience-driven autonomous learning method [8]. With the use of deep neural networks within a RL framework the field of "deep reinforcement learning" (DRL) is playing a revolutionary role in advancing Artificial Intelligence. The development of DRL has shown great improvement. For example, AlphaGo, defeated the human champion in the game of Go [9], and even showed the real self-learning success without any human knowledge [10].

The rise of DRL mainly comes from two aspects: 1) The powerful function approximation and representation learning properties of deep neural networks (DNNs) [11, 12]. With DNNs, the low-dimensional feature representations can be obtained from high-dimensional data (e.g., images, audio, and text), which makes DRL solve decision-making problems with high-dimensional state and action space. 2) RL's "exploration and exploitation" property. Exploration is related to gathering more information. It means that the system will explore different possible trials to see if they are better than what has been tried before. Exploitation ensures that the system makes the best decision given current information, which means the system memorizes the strategy that has worked best in the past. These two advantages show the priority compared to many other traditional supervised learning methods.

In this paper, we tried to employ these properties of DRL to work on movie scene affective label identification. However, the traditional DRL framework cannot be applied to our task directly. The affective information at each movie frame is related to environment state and the agent uses the environment state to make the classification. In this case, each frame's affective label decision that is generated from the agent cannot interact with the environment, since the movie scene data sequence is determined already. To employ DRL into our work, we proposed a modified DRL framework based on our task to make the interaction between agent and environment available.

The rest of paper is organized as follows: Section 2 describes in detail our proposed DRL framework to predict affective label for movie scenes. It begins with the introduction of traditional DRL, then describes our proposed new framework and use of transfer learning to improve our system. Section 3 provides a brief description of the movie database used in our paper, after which we describe data processing, the pre-training model used in transfer learning and experiments settings. After that, we discussion our results in section 4. Finally, we conclude and discuss future work in the section 5.

2. METHODOLOGY

In this section, we will describe our proposed Reinforcement Learning framework on movie scene affective label identification.

2.1. A brief introduction to Reinforcement Learning

Before illustrating our proposed modification to RL, we first describe a traditional RL application in brief. The main idea in RL, which belongs to a class of experience-driven autonomous learning methods, can be regarded as the typical artificial intelligence idea – improving the system over time through trial and error [8]. A traditional RL algorithm can be modeled within perception-action-learning loops. Within each iteration during training, denoted using the time step t, the agent observes the state $s(t)$ from the environment. The agent then uses its policy to choose an action $A(t)$ based on the state $s(t)$, and once this action is executed, the environment is updated to a new state $s(t+1)$. Most importantly, this transaction provides a feedback in the form of a reward $R(t)$. With these round of updates, we observe the state transitions and the associated reward, which can be denoted in the form of a 4-tuple $(s(t), A(t), s(t+1), R(t))$ associated with time step t. A RL algorithm aims to learn and improve its decision policy using an online update strategy by using the reward as a feedback. For more details of RL, one can refer to [8, 13, 14].

The key point of standard RL application is that the agent cannot observe all state transition of dynamic. Each interaction with the environment will generate new information, which the agent employs to update its knowledge[8], and the environment state will also be changed during the interaction process.

The movie scene clip can be regarded as the sequence of data samples along different modalities. For example, from the speech modality, it consists of audio signals, from the visual modality, there is a sequence of image frames within each movie scene clip. For the affective label prediction task, human annotators need to process dynamic temporal affective information to obtain the correct label. For RL application, the agent's action decision also contains strong temporal correlations, and current action decision plus the reward also depends on previous steps. Thus, we try to use RL to predict the affective label for movie clips.

2.2. Challenges in RL on movie affective label identification

Applying traditional RL framework to our task is a challenging task due to the following:

- Interaction between agent and environment.
- Complexity of movie data.

We will illustrate these challenges briefly. As described above, in the RL framework, usually, the environment will be changed according to the response of agent's action. For example, when playing the Pong game [15, 16] with RL and the agent will give the action to move the pad one step up or down, and further change the video game screen image, which is regarded as environment state. However, in our case, the frame sequence is fixed within each movie scene, agent's action (e.g. the prediction of affective label at current frame) cannot change the movie scene sequence data. So, we cannot use only the movie itself as the environment state, we need to add more components, and make the agent action (e.g. current frame affective label prediction) interactive with the environment state.

Another challenging part is related to the complexity of movie dataset. Affective information identification on movie data is much more complicated compared to other standard test datasets. For example, on the audio channel, most of the human conversation is mixed with background music, which makes it inaccurate and challenging to extract speech prosodic features from movie audio channel directly. On the image channel, the most straight forward idea is to extract affective information from human faces, which contain most affective information. This is the reason that face recognition from movie scene and face expression recognition also received widespread attention in the research community [17].

In this work, we mainly focus on employing RL on a sequence of images, more accurately, we try to identify affective movie scene labels based on actor or actress's face at each movie frame.

2.3. Framework of our proposed method

As mentioned above, we need to design a new RL framework such that the agent action can interact with the environment state. Our proposed RL framework is shown in Fig 1. The movie scene clip is considered to be a sequence of movie frames. At each time step t, the environment state has two components, the original movie face embedding vector ϵ and the predicted affective information α at time step t. By considering this joint information, the agent using the Deep Q-network (DQN) [15], makes a corresponding decision on the action. The action output is, in our case indicative of the affective label prediction at the frame level at time step t. Then the predicted affective information input $\alpha\ (t+1)$ at time $t+1$ is using either the action output at time t or human annotations at time t if available. $\alpha\ (t)$ is then used as part of environment state input to the DQN, and used by the agent to generate the next action at time $t+1$. This is clearly illustrated in the equations below.

$$A(t) = \mathbf{Q}(s(t)) = \mathbf{Q}([\epsilon(t), \alpha(t)]) \quad (1)$$

$$R(t) = \mathbf{H}(A(1), \ldots, A(t), F) \quad (2)$$

$$\alpha(t+1) = \mathbf{G}(A(1), \ldots, A(t)) \quad (3)$$

The function $\mathbf{Q}$ represents deep Q-network, a specific kind of DRL and will be described later. F represents the true affective label. $A(t)$ is the action at time step t. At each step, the reward is generated by function $\mathbf{H}$ based on actions taken thus far. In our case, this action represents the predicted affective labels and hence it is compared against the human annotated movie scene affective label to obtain a reward value.

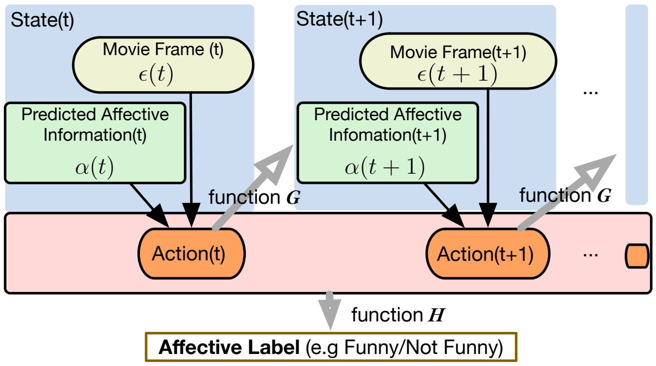

Fig. 1. Proposed RL framework on movie affective label prediction

The function $\mathbf{G}$ is similarly used to add the prediction from the previous time steps, into the current environment state, which will involve the agent's action in the environment updating process.

With the proposed RL structure, we can design different functions G and H for certain applications. In this work, the details of these functions will be described later.

2.4. Facial expression embedding pre-training

In this paper, we mainly focus on applying our proposed method on movie's visual channel. Although Deep RL algorithm can process high-dimensional inputs, directly training the agents on visual inputs is not feasible due to the large number of samples required [8]. To make the RL converge faster, we use transfer learning idea to exploit previously acquired knowledge from the pre-trained model. In our case, after extracting the cropped face from each movie frame, we will utilize the pre-trained deep neural network model on facial expression recognition task [18], and use the embedding output of neural network as the high dimension feature representative used in RL learning. More details of the configuration of our experiment will be described in the section 3.

3. EXPERIMENTS

3.1. Dataset

We use the Kaggle Facial Expression Recognition Challenge dataset [18] to pretrain our facial expression embedding model. This dataset comprises 48x48 static grayscale images of human faces, each labeled with one of the 7 emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. Face images in this dataset vary considerably in scale, pose and illumination, making the trained model more robust which is necessary considering the variety of detected face in the real movie frames. All of the images are additionally pre-processed to detect and localize the faces.

The movie dataset used in our work comes from 18 movies selected from different genres and time periods. We divide each movie into clips at the scene level and obtain annotations at the scene level indicating the mood or tone of the scene. This gave us a total of 1471 scenes. A single movie scene clip can be associated with multiple tone labels, such as funny, exciting, calm etc. In spite of having several different unique tone labels, working on a multiclass or multilabel classification problem was not feasible since the distribution of labels was extremely sparse i.e. most tone labels had very few examples. Seeking to derive better labels using data-driven methods we first analyze labels in our dataset using different unsupervised clustering methods such as spectral clustering on *k-hot* encoded tone labels using different distance metrics. This analysis revealed that most of the training data was split into half along the class "funny". As a result, we decided to simplify the task to binary classification with the labels *funny*($F = 1$) and *not funny*($F = 0$). For the purposes of this work, we consider any scenes without the tone label "funny" to be "not funny".

3.2. Movie data processing

For experiments, we created the affective label classification dataset based on the original movies and annotations. In this work, we limit our focus to the video channel in the movie, specifically the characters' faces. A sequence of these face image is pre-extracted from the video and utilized as an input to our system. To detect the faces at each frame, we employ the standard face detection library dlib [19] to extract faces from each frame of the video channel. If a frame contains multiple faces, we select the one closest to the center of the frame. The intuition behind this selection is that when multiple faces are shown on screen it is quite likely that the main character's face is located in the center of the screen to dominate the scene's affective information. We also notice that the face selected by this criteria often turns out to be the largest in area compared to other detected faces.

However, the amount of the annotated movie data is still small for training a robust model, due to the limited number of available movies and expensive human annotation process. Since our videos use a frame rate of 24fps, we notice that the difference between face posture at neighboring frames is insignificant. Thus, we augment our dataset by generating multiple movie face sequences for each movie scene in the following manner. For each movie scene clip, we use only every 18th frame. Since we have limited data we re-sample every movie clip by starting with frames 0, 2, .. 16, thus resulting in 9 sequences, each downsampled by a factor of 18 per movie clip. Based on this processing, each sequence can represent one of the original movie scene sequences, which are roughly two minute long movie clips. By this shifted temporal subsampling strategy, we can incorporate information from the full dynamic range of face changes, increase the available movie sequences, and increase the diversity of data within each sequence.

3.3. Facial Expression Embedding model

Instead of training a facial expression model from scratch on faces from movies, we use another larger facial expression dataset to first pretrain our facial expression embedding model. We train a Convolutional Neural Networks (CNNs) model with 6 layers of 2D Convolution, max-pooling and Dropout layers with ReLU nonlinearities between each layer. The output layer uses a softmax activation function to classify a face into different expression classes. Our expression classification model achieves an accuracy of 64% on the validation set. We use this pretrained model to generate embeddings using the last-but-one fully-connected layer.

3.4. RL model for Affective Label Prediction Experiment

We use deep Q-network (DQN) [15] for our proposed RL framework. The input to DQN in our model comprises two parts. The first part is the facial expression embedding (ϵ), output by the pretrained CNN of facial embedding model as described earlier. The second part is an affective information encoding vector (α) which conveys affective decision taken at the previous step. This was described in detail in section 2.3. At each time step t, the algorithm can employ human annotated ground truth label input if available, by setting $\alpha\,(t) = A_{human}(t-1)$. If no human input is available, as is often the case in real data, then $\alpha\,(t) = \mathbf{G}(A(1), \ldots, A(t{-}1))$ can be used, the belief from prior frames.

Instead of directly concatenating these two parts of input, we first transform the affective information input (α) via two fully connected layers, the output of which is then concatenated with the facial expression embedding. This fused input feature is then passed through two additional fully connected layers with ReLU activations

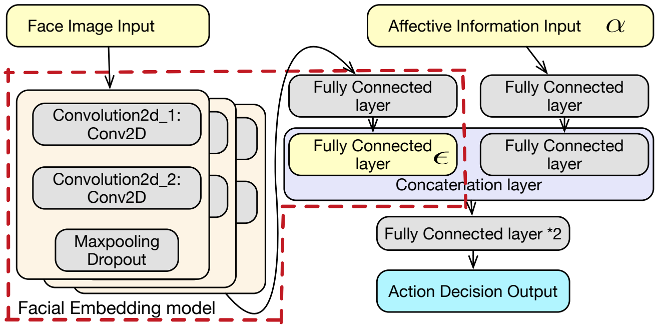

Fig. 2. Proposed structure of DQN with facial embedding model

and one Q-value output layer with the linear activation. The whole framework of our DQN plus facial expression model is shown in Figure 2. In DQN, Q value is approximated by neural networks. Thus, the output of DQN contains the Q value of every action possible, which in our model is related to the affective label decision. Recall that $\mathbf{G}$ is a function that maps the DQN's outputs to α, which will be further used as part of inputs to DQN. For our experiments, we design $\mathbf{G}$ to transform the predicted action vector at the last time step in the form of a one-hot encoding of the maximum Q-value. $\alpha(t+1) = \mathbf{G}(A(1), \ldots, A(t)) = hardmax(A(t))$. This means that we encode the action $A(t)$ with the largest Q-value at step t using a one-hot vector. This one-hot 2-dimensional vector indicating whether a scene is funny or not is indicated by $\alpha(t+1)$ and input to the DQN at step $t+1$.

Another important aspect of RL, is the choice of reward function, since the reward value is used as a feedback to update the Q value. The DQN uses the reward, which is the difference between new score and previous score, to learn the optimal policy for choosing the action. In our scenario, for each training sequence sample, we only have one human annotated label for the whole training sequence. No supervision for the affective label is available at the frame level.

The DQN needs a reward value to update the Q value at each step of action. In our framework, the function $\mathbf{H}$ is utilized to calculate the reward value R. Function $\mathbf{H}$ is designed as follows: Within each frame sequence training sample, the reward value is based on majority vote of frame-level results up to now. Based on this design, we could assign a reward at each frame step, not just at the end of the whole sequence.

In addition, we found from some initial experiments that the absolute value of the reward is critical to the convergence of the training. If the reward value is too large or too small, the system doesn't converge and the RL system fails to learn affective information. Also, a larger reward value is typically used for the terminal step than intermediate steps and the ratio of these values also needs to be carefully determined. Based on some initial experiments, we select the values $R_{inter} = \pm 0.05$, and $R_{end} = \pm 1$ for our application. A positive reward indicates that the action output is same as the true label, while a negative value is used for the reward when they differ.

4. RESULTS AND DISCUSSION

Baseline of movie scene affective classification As mentioned in Section 3.1, we generate subsampled movie face sequences and split them into three parts. The training, validation and test database has 7199, 699, 732 number of samples respectively. Within each part, the number of funny and not funny samples is roughly the same. The by-chance classification accuracy is regarded as 50%. The validation set is used to monitor the training process to avoid overfitting, and also tune the parameters of DQN.

RL framework movie scene affective classification As mentioned, we use our system to simulate annotator involved RL affective identification system. The α is designed as previous agent's action one-hot binary classification output result, in which it is either [0,1] or [1,0]. This one-hot α has a deterministic boundary value, which is similar to the case that we involve the human's decision during our system training process. We only have human annotated scene level label and two methods are employed to evaluate our model. One is using majority vote of all frames level action output results; another way is directly regarding the accumulated last frame's result as the scene level decision. The results on funny label binary classification is shown in the second row of Table 1. From these results, we note that the majority vote method performs better than the second one.

Table 1. DQN binary classification accuracy on the two datasets(%)

dataset	Majority Vote	Last acc. frame
full scene	72.95	61.43
shorter 30 fr.	84.13	74.51

RL framework movie scene affective classification with shorter sequence We perform another experiment by fixing the movie sequence length to 30 frames (represents roughly 15 seconds in the original movie) and use a sliding window approach on the downsampled sequence dataset to generate more training samples. This allows us to generate more data by training on smaller subsequences. As a result of more training data and since the samples are uniform in length, we observe a better classification accuracy as shown in the third row of Table 1. It is also possible that reducing the length of the sequences, reduces the complexity within each sequence, and the system is able to deal with this reduced complexity better. Moreover, over a shorter time duration, the variance in the movie image sequence input is significantly lesser considering the stationary property of the behavioral state of a person. Under this condition, the α at frame level would play a more important role since each frame level decision is more related to the final scene level decision. These results indicate that in real scenarios, in which occasional human annotated α is used, human's involvement in the RL training loop could potentially help the system to correct labels and the system uses that opportunistic human correction to improve its training.

5. CONCLUSION

In this paper, we address the problem of using RL method to predict funny scenes in movie clips using face images. We modified traditional RL structure in order to achieve the interaction between label prediction agent and environment state. Compared to all previous supervised sequential based affective recognition system, our DRL based work can generate the output affective label in real time and try to learn the policy through exploration and exploitation. These properties usually cannot be achieved by the regular supervised learning methods, which are usually purely exploitative. The results are promising considering the complexity of movie dataset. Further, our method shows its potential application in the real scenario, in which human involves in the RL training process to help RL correct its error and improve the performance.

For future work, we plan to employ multimodal cues including audio and video features. The rich affective content in movies is inherently multimodal. The affective information is evoked through video, human speech or even the background music. Considering all features from different modalities will apparently improve the performance of the RL affective recognition system.

Acknowledgement

The authors would like to thank Jason Lambert from *SONY Pictures* for generously providing the data and annotations.

6. REFERENCES

[1] S. Wang and Q. Ji, “Video affective content analysis: A survey of state-of-the-art methods,” *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, Oct 2015.

[2] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, “Affective audio-visual words and latent topic driving model for realizing movie affective scene classification,” *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 523–535, Oct 2010.

[3] L. H. Chen, H. W. Hsu, L. Y. Wang, and C. W. Su, “Violence detection in movies,” in *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*, Aug 2011, pp. 119–124.

[4] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis, “Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention,” *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[5] A. Goyal, N. Kumar, T. Guha, and S. S. Narayanan, “A multimodal mixture-of-experts model for dynamic emotion prediction in movies,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2822–2826.

[6] Esra Acar, “Learning representations for affective video understanding,” in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 1055–1058.

[7] Esra Acar, Frank Hopfgartner, and Sahin Albayrak, “A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material,” *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 11809–11837, 2017.

[8] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath, “A brief survey of deep reinforcement learning,” *arXiv preprint arXiv:1708.05866*, 2017.

[9] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al., “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[10] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis, “Mastering the game of go without human knowledge,” *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[11] Yoshua Bengio, “Deep learning of representations for unsupervised and transfer learning,” in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[12] Yoshua Bengio, Aaron Courville, and Pascal Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[13] Seyed Sajad Mousavi, Michael Schukat, and Enda Howley, *Deep Reinforcement Learning: An Overview*, pp. 426–440, Springer International Publishing, Cham, 2018.

[14] Yuxi Li, “Deep reinforcement learning: An overview,” *arXiv preprint arXiv:1701.07274*, 2017.

[15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.

[17] Vinay Bettadapura, “Face expression recognition and analysis: the state of the art,” *arXiv preprint arXiv:1203.6722*, 2012.

[18] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio, *Challenges in Representation Learning: A Report on Three Machine Learning Contests*, pp. 117–124, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[19] Davis E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.