Synthetic Video Generation

Amit Kumar – B13107

Paawan Mukker – B13218

Dr. Dileep A.D.

Dr. Renu M Rameshan

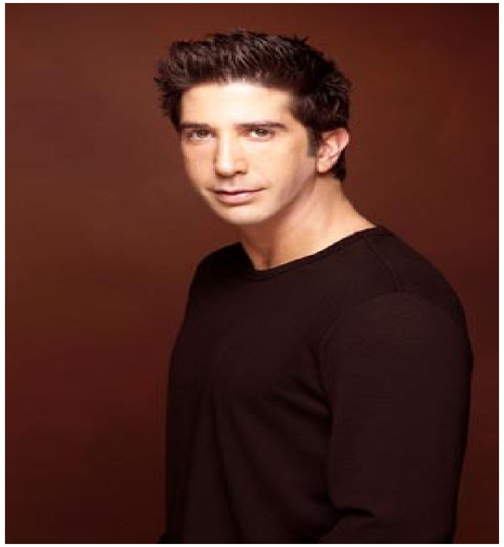
Abstract

- Given a new script and subtitles and corpus of videos.
- Generate a new video by picking matching frames from the corpus.
- Identify location of shot.
- Characters involved.
- Identify Emotion, orientation, action of characters.
- Minimize continuous frame disparity.

Friends Characters Introduction

Joey Tribbiani

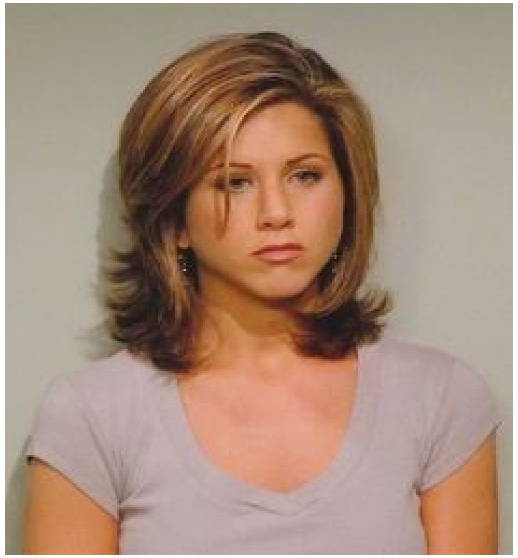
Chandler Bing



Ross Geller

Image Courtesy – www.wikipedia.org

Friends Characters Introduction



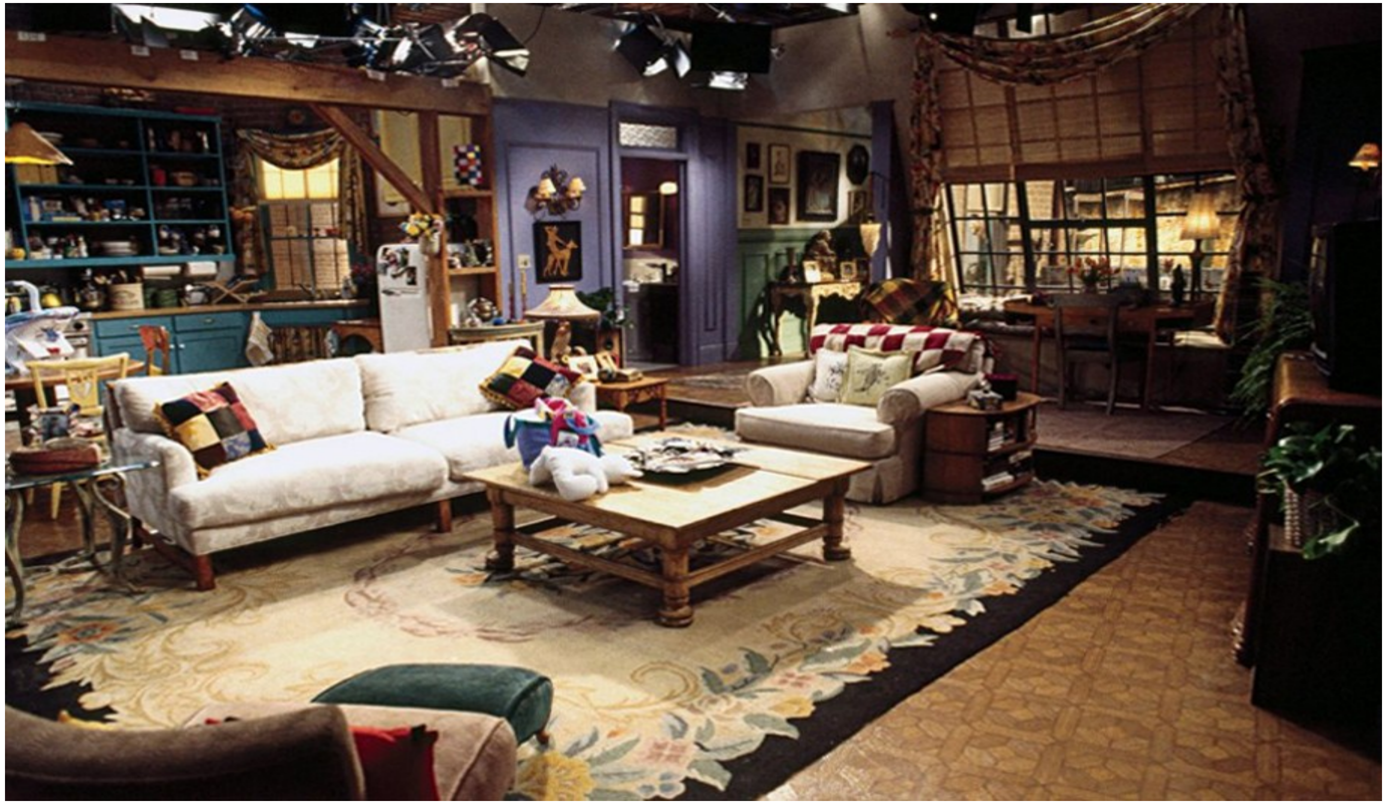
Monica Geller

Rachel Green

Phoebe Buffay

Image Courtesy – www.wikipedia.org

Friends Location Introduction



Monica and Rachel’s apartment

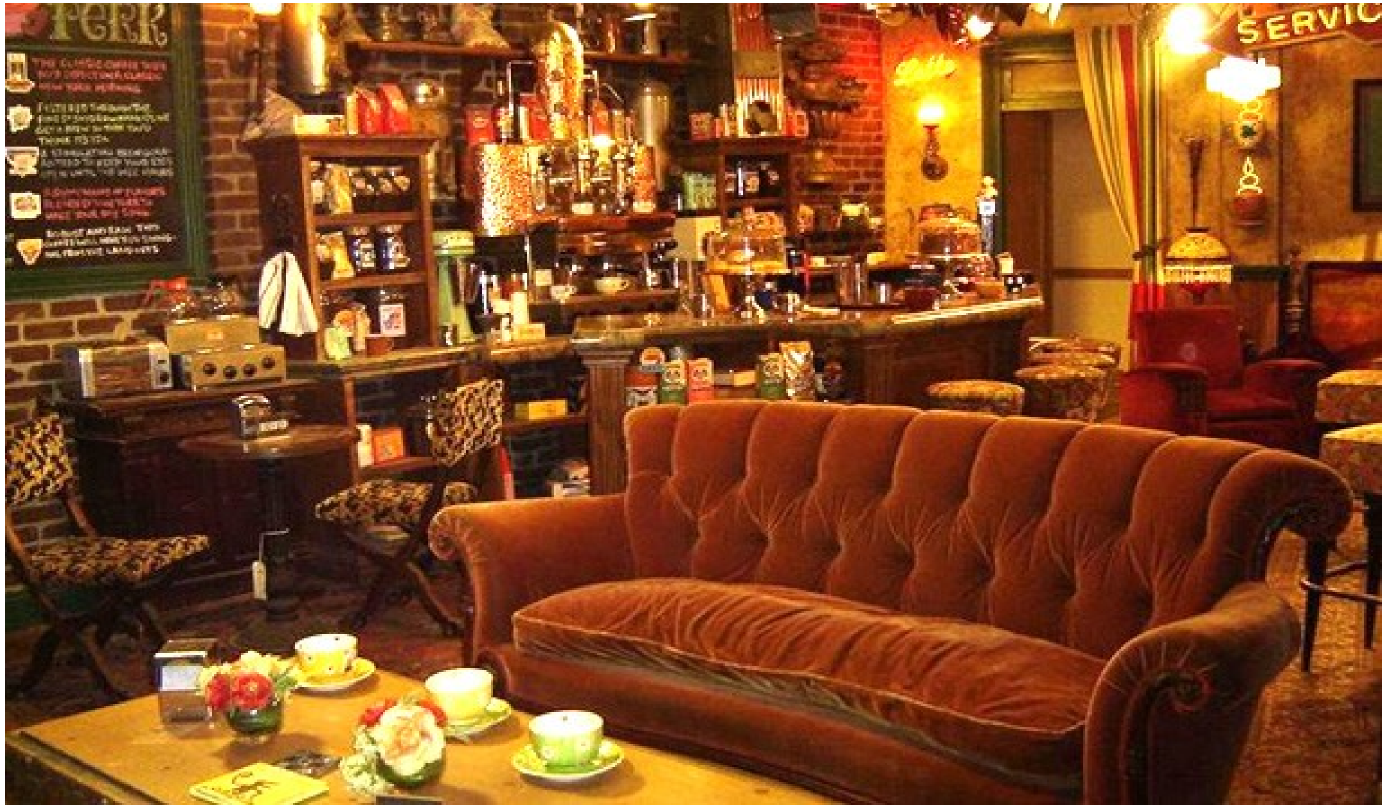
Image Courtesy –
www.hookedonhouses.net

Friends Location Introduction

Joey and Chandler’s apartment

Image Courtesy –www.friends.wikia.com

Friends Location Introduction

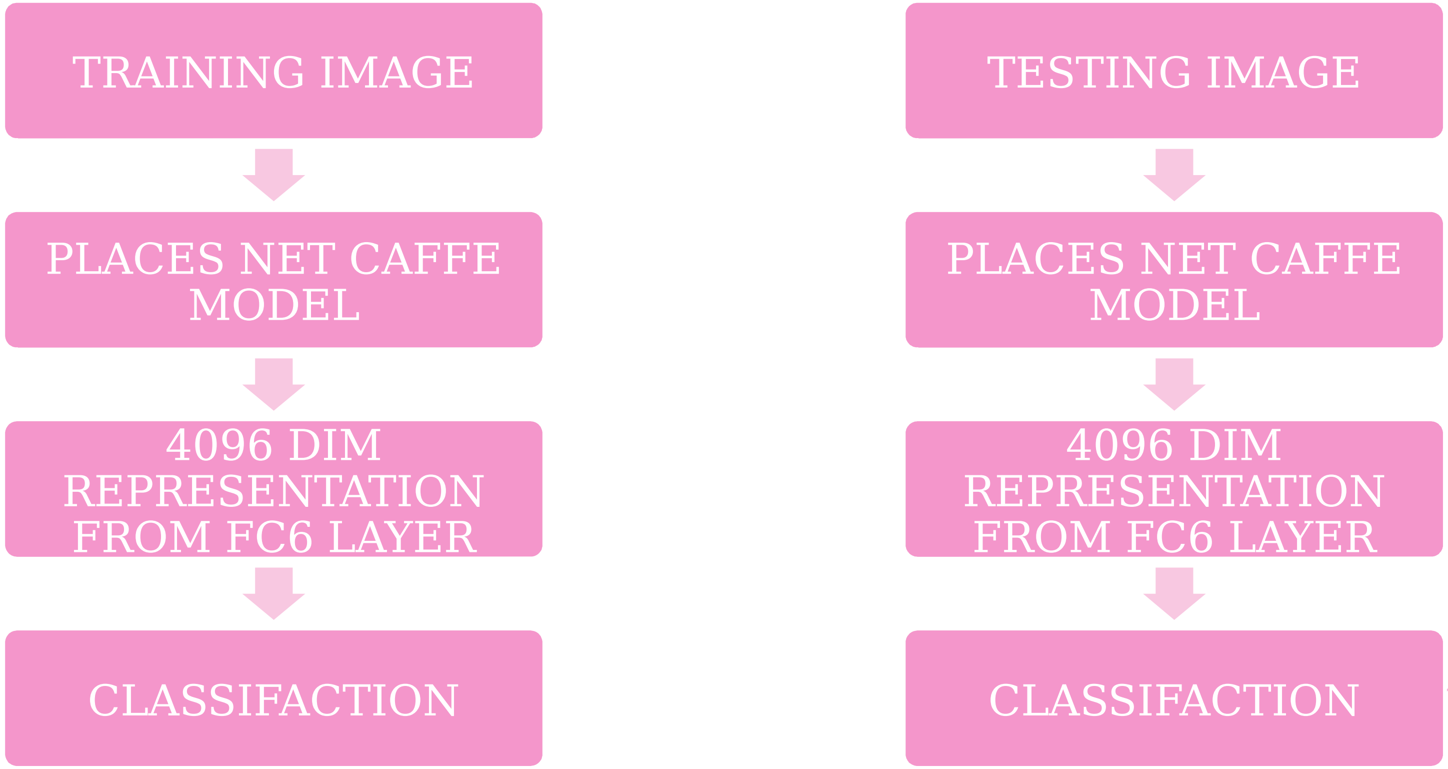


Central Perk Cafe

Image Courtesy – www.atlasobscura.com

Scene Recognition[Recap]

- PlaceNet VGG Caffe Model
- Representation from fc6 layer



TRAINING IMAGE

PLACES NET CAFFE MODEL

4096 DIM REPRESENTATION FROM FC6 LAYER

CLASSIFACTION

TESTING IMAGE

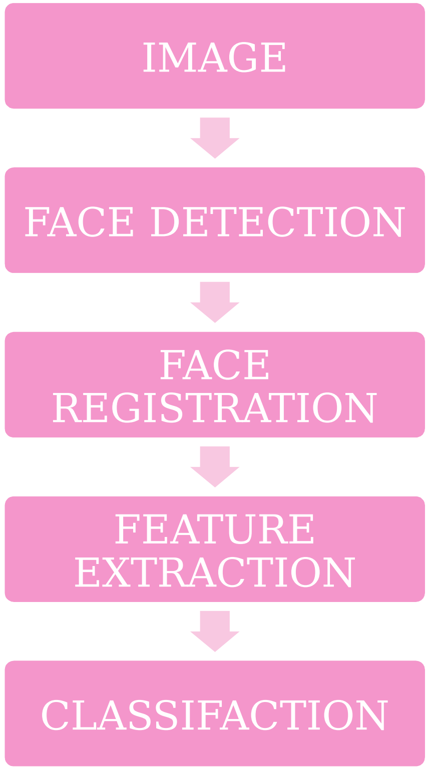
PLACES NET CAFFE MODEL

4096 DIM REPRESENTATION FROM FC6 LAYER

CLASSIFACTION

Emotion Recognition

- Predict emotion of person in an image.



IMAGE

FACE DETECTION

FACE REGISTRATION

FEATURE EXTRACTION

CLASSIFACTION

Face Detection

- Dlib for face detection
- Compare image with histogram of gradient image

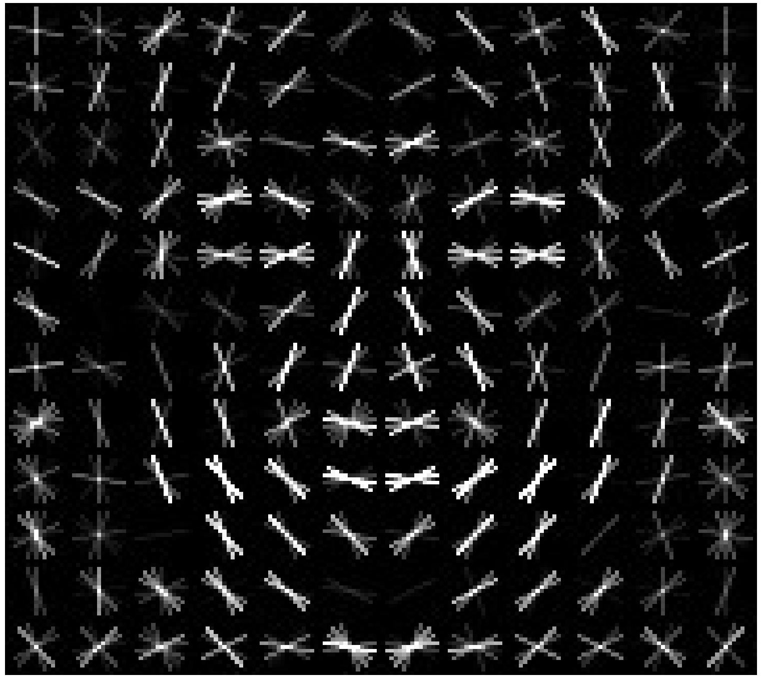


Image Courtesy – dlib.net

Finding Landmarks points

- Dlib for Landmark point detection.

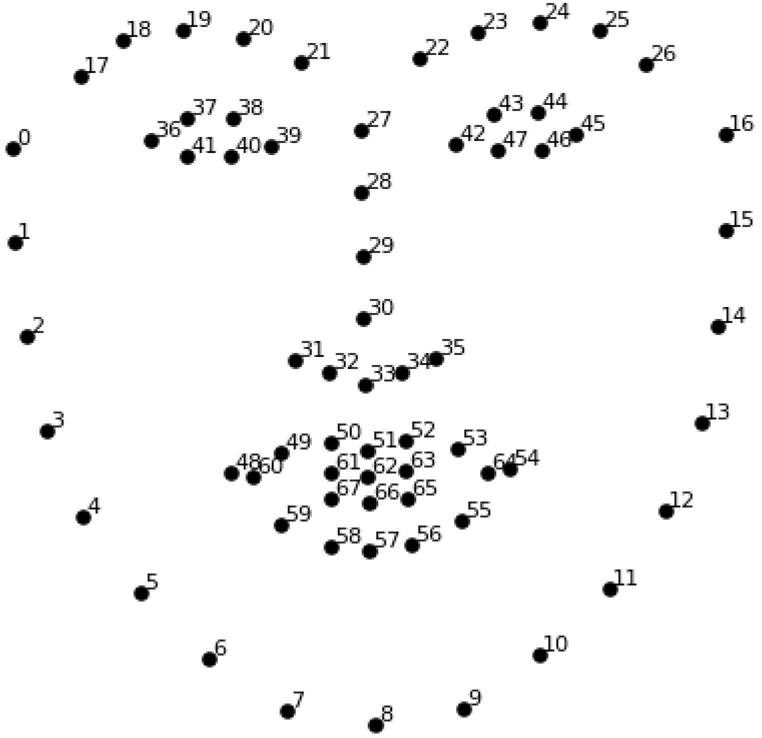


Image Courtesy – dlib.net

Face Registration based on eyes

- Eyes centre by averaging points around eyes
- Register face – Eyes on same place in each image.
  - Different width and height of face
  - Rotation of face
  - Crop extra region

Face Registration based on eyes

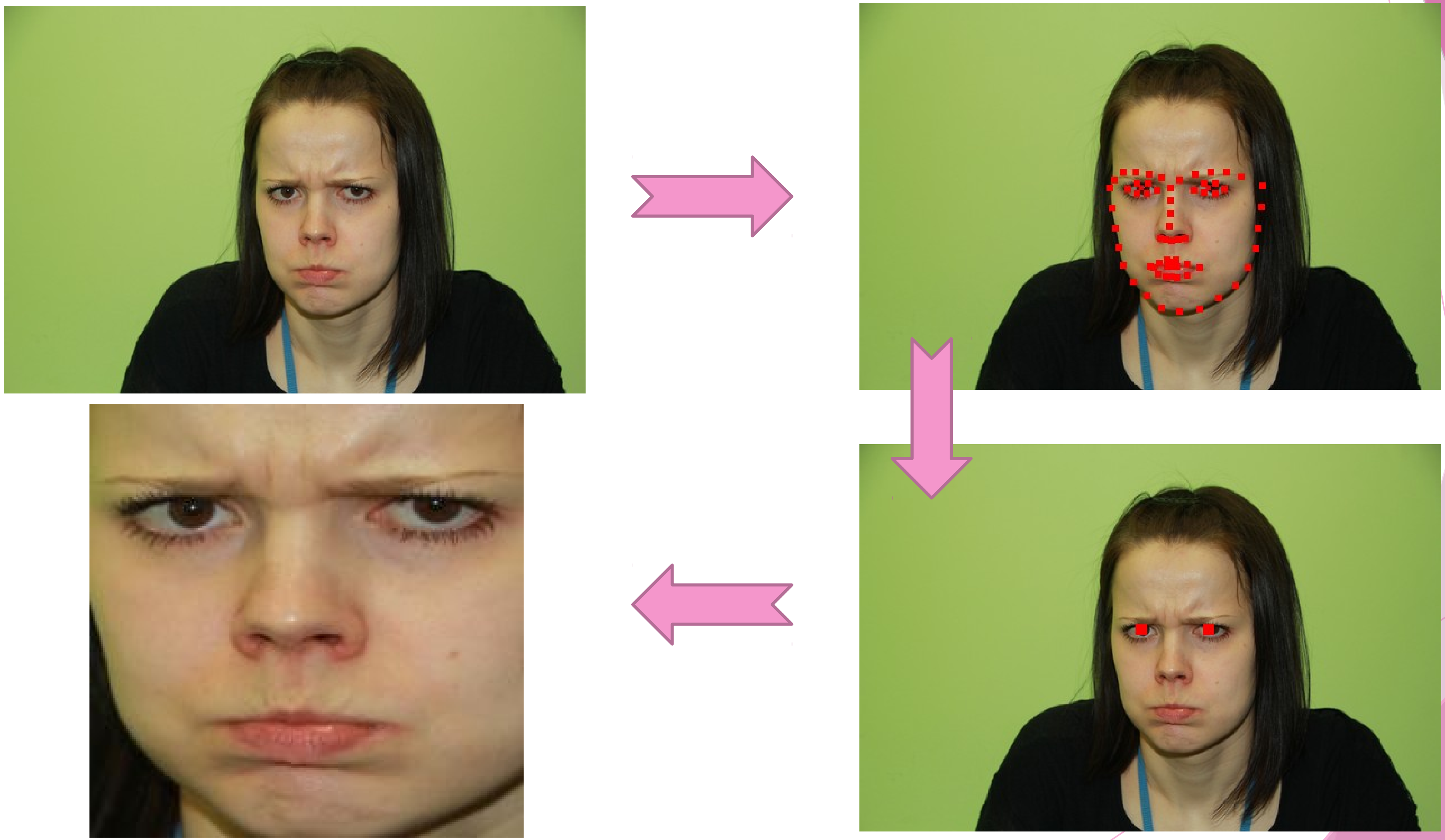
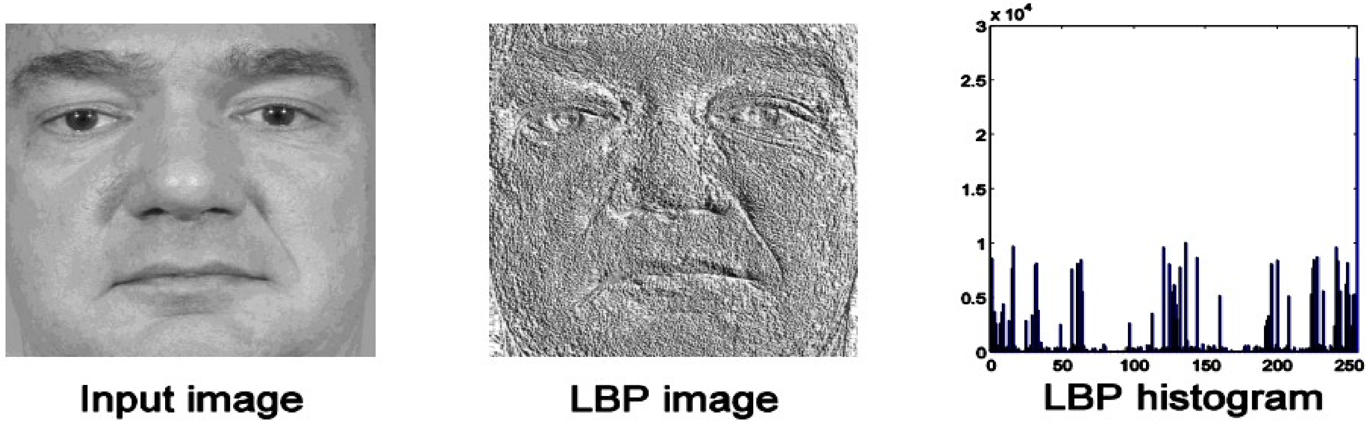


Image Courtesy – codalab.org

Feature Extraction – LBP Feature

- Took registered image
- Find out LBP Feature



example

6	5	2
7	6	1
9	8	7

thresholded

1	0	0
1	▨	0
1	1	1

weights

1	2	4
128	▨	8
64	32	16

Pattern = **11110001**

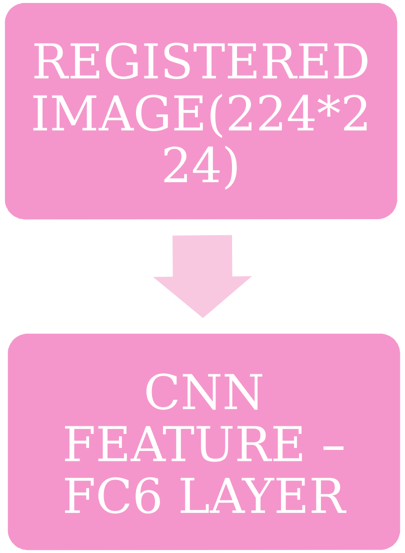
LBP = 1 + 16 +32 + 64 + 128 = **241**

Image Courtesy – ee.oulu.fi

Feature Extraction – HOG & CNN Feature

REGISTERED IMAGE(224*224)

HOG FEATURE

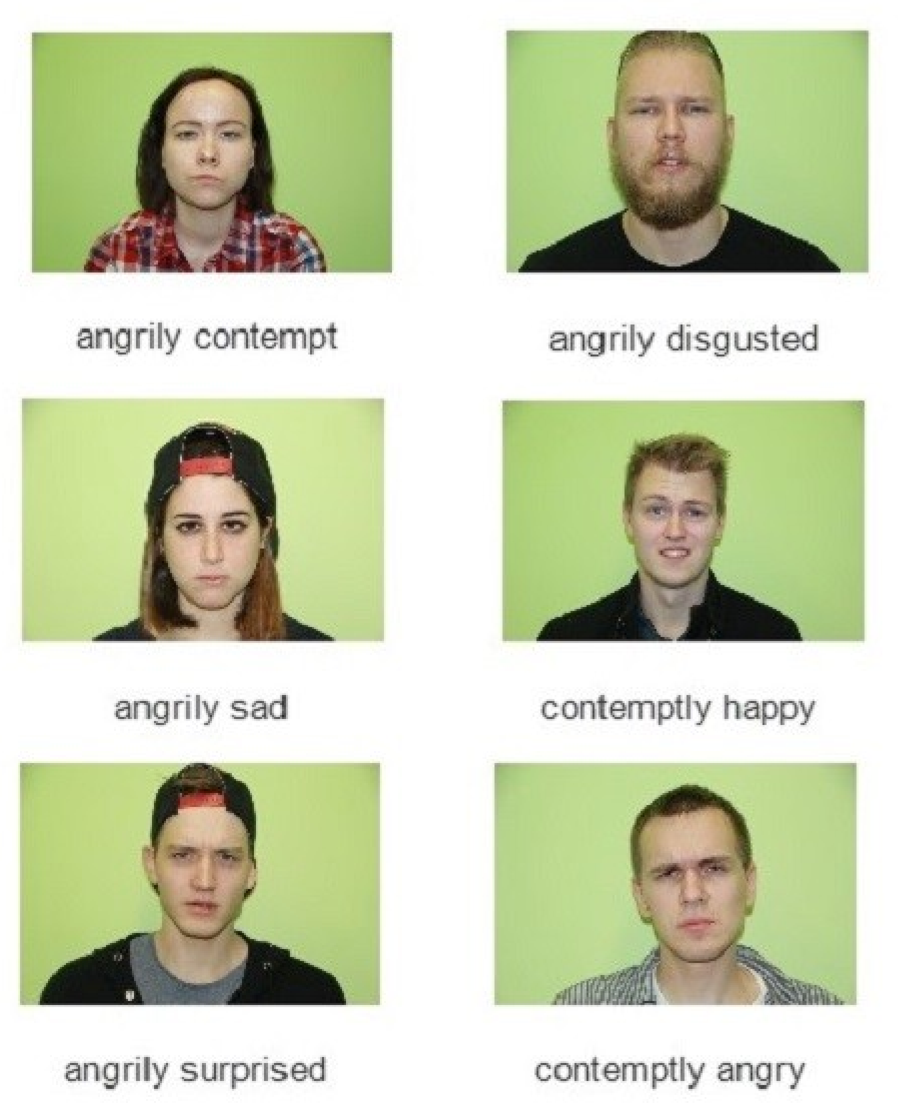


REGISTERED IMAGE(224*224)

CNN FEATURE – FC6 LAYER

Codalab Emotion Recognition Database

- 31250 facial faces with different emotions
- 125 subject
- 50 different emotions.
- Micro emotion analysis.
- Classes like Complementary emotion-Dominant emotion.



angrily contempt

angrily disgusted

angrily sad

contemptly happy

angrily surprised

contemptly angry

Image Courtesy – codalab.org

Codalab Emotion Recognition Database

- Best accuracy on this database - 83.98

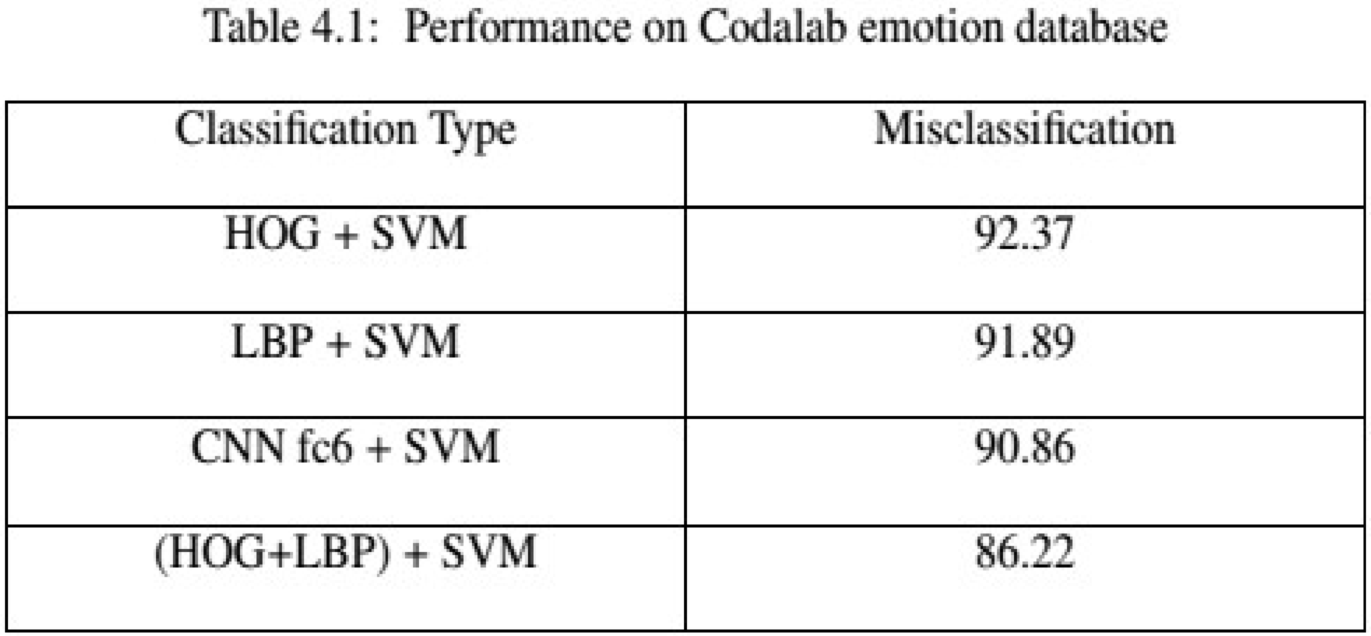
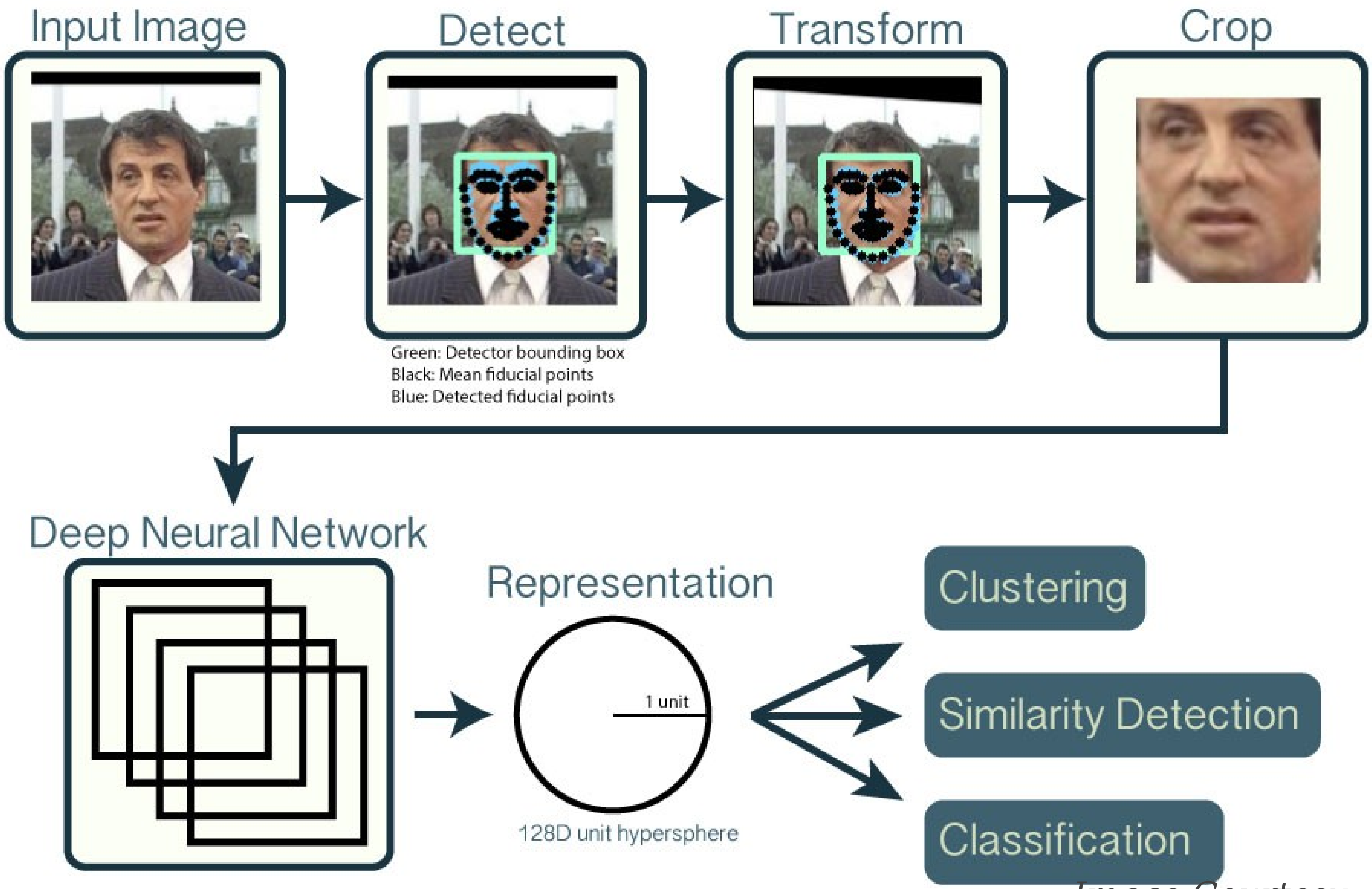


Table 4.1: Performance on Codalab emotion database

Classification Type	Misclassification
HOG + SVM	92.37
LBP + SVM	91.89
CNN fc6 + SVM	90.86
(HOG+LBP) + SVM	86.22

Character Identification - Recap



Green: Detector bounding box
Black: Mean fiducial points
Blue: Detected fiducial points

Image Courtesy – Openface, Brandon

Speaking action detection

- Use of visual features only
- Speech not used – To make it challenging

Problem specs -

- Video input – Unit used Shot
- Shot – Frames of continuous video segment (~1 sec)
- Frames Per Second ~ 24

Overview -

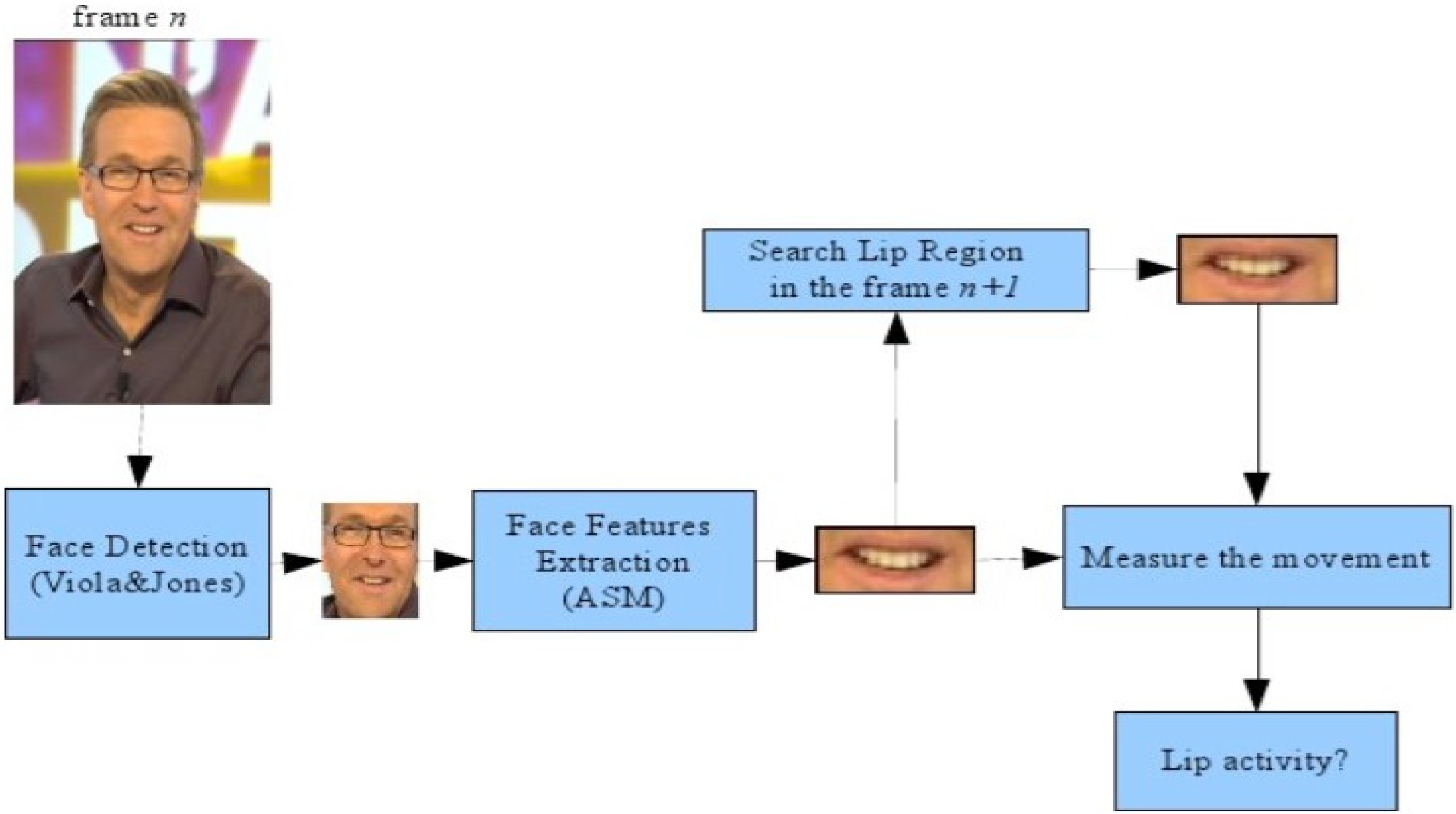
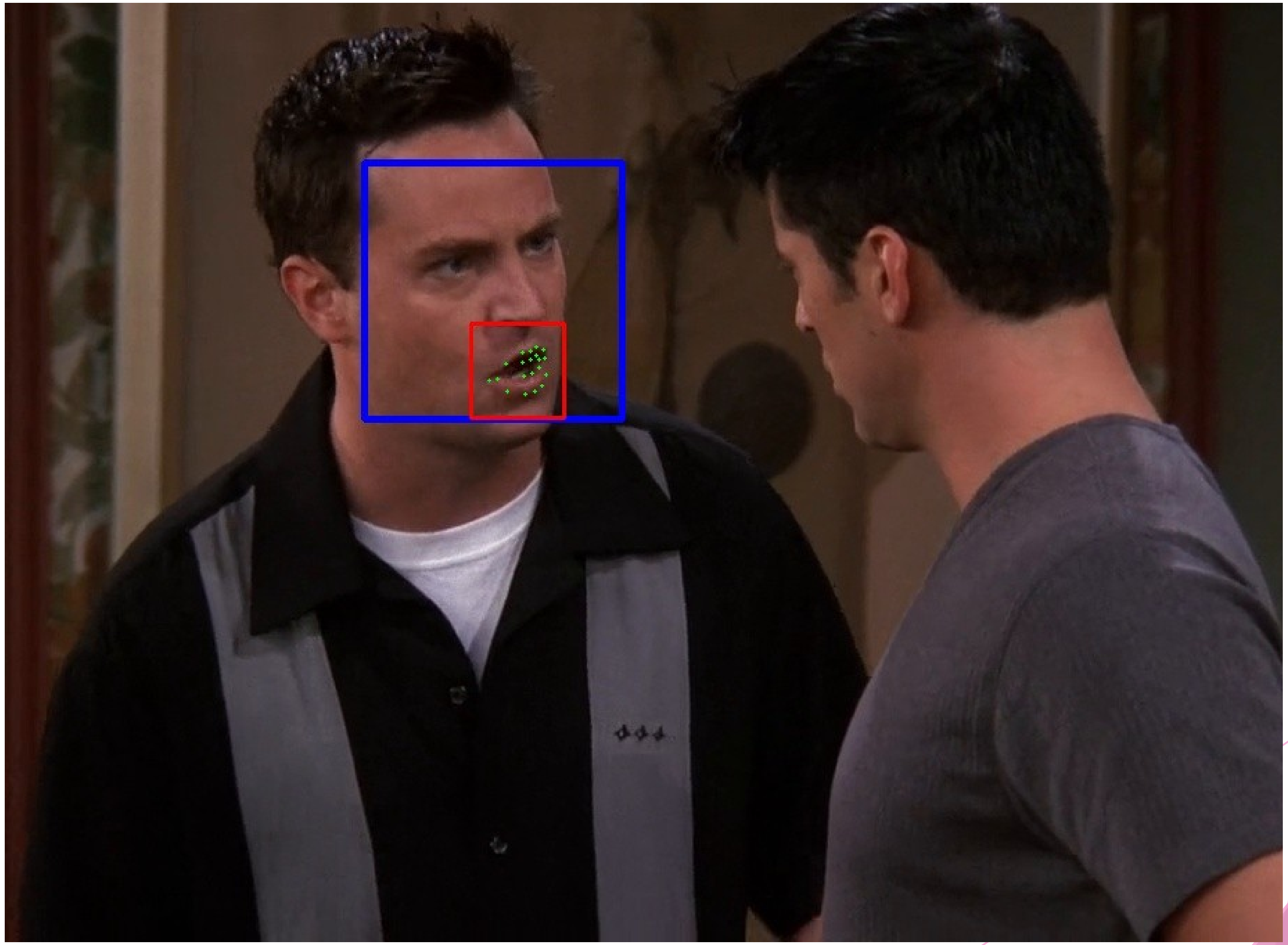

frame n
Face Detection (Viola&Jones)
Face Features Extraction (ASM)
Search Lip Region in the frame n+1
Measure the movement
Lip activity?

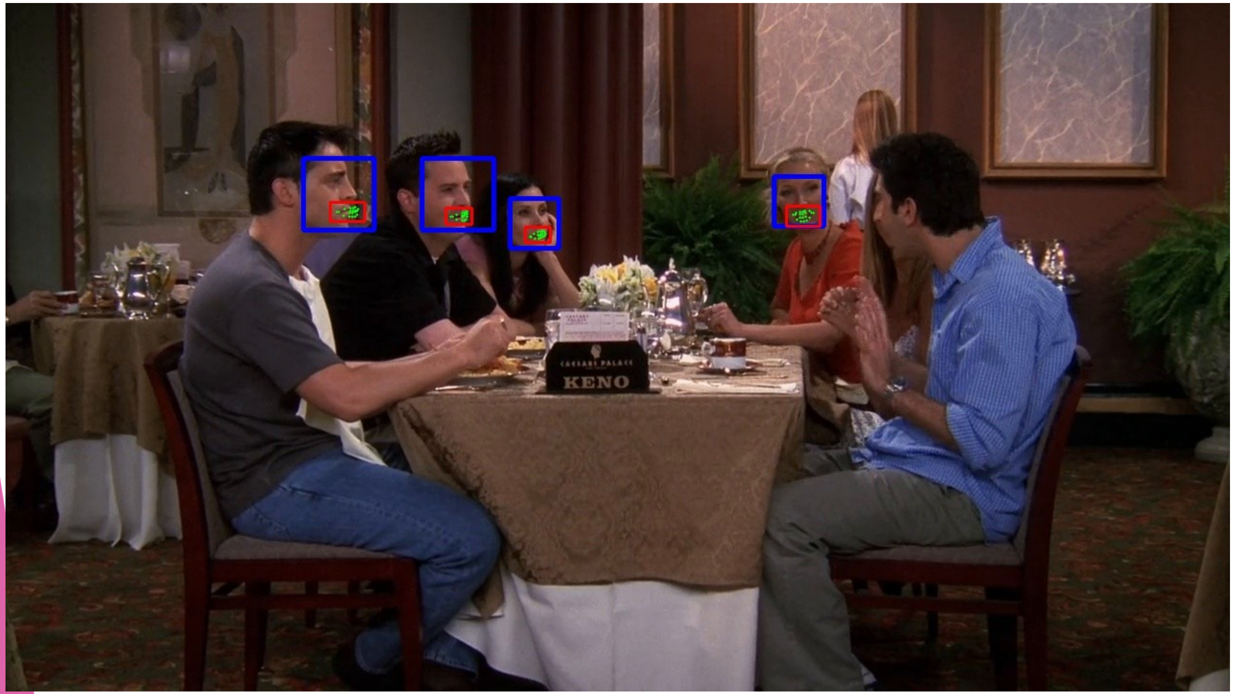
Image Courtesy – Lip activity detection. Meriem et al.

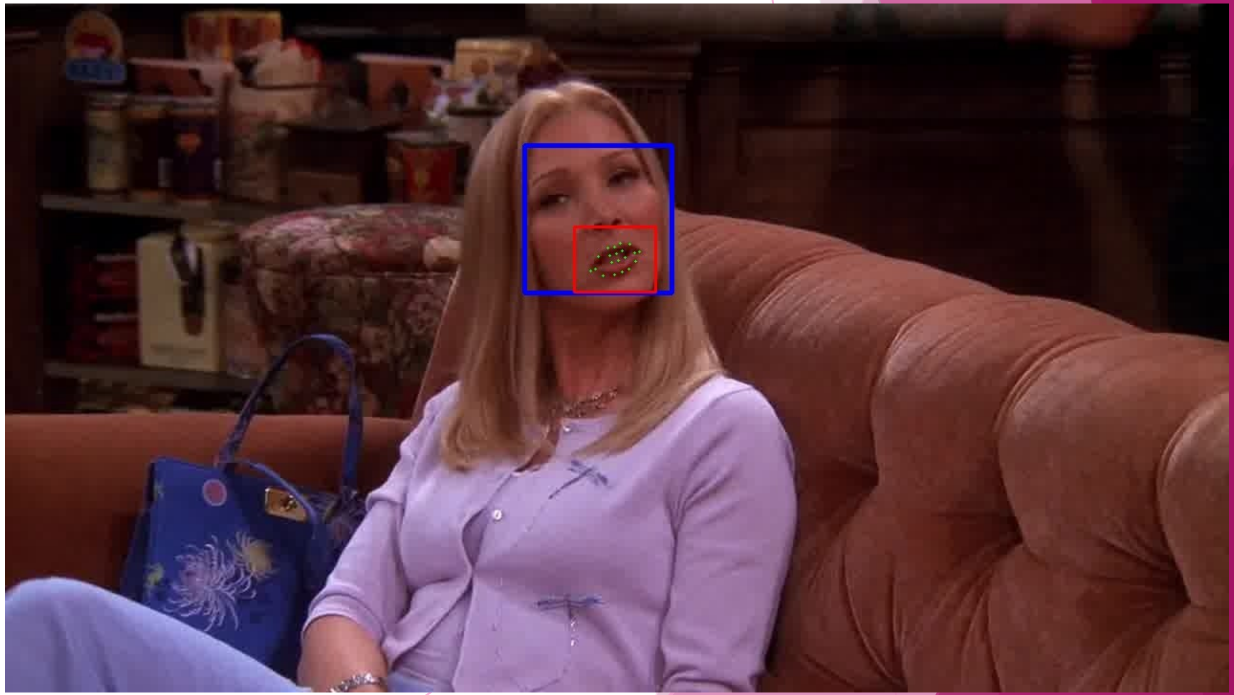
Lip localization



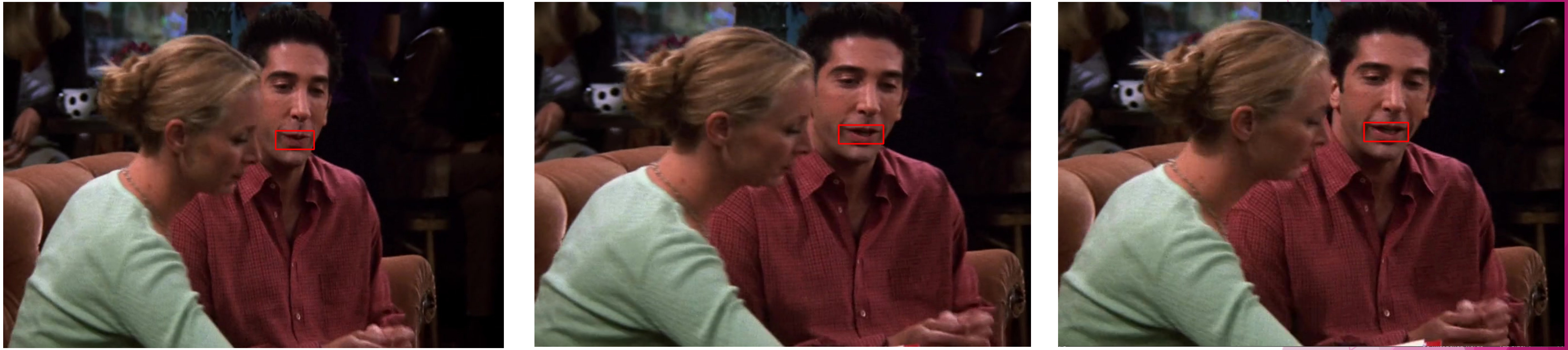
Lip bounding box

- Based on scale of lip height and width

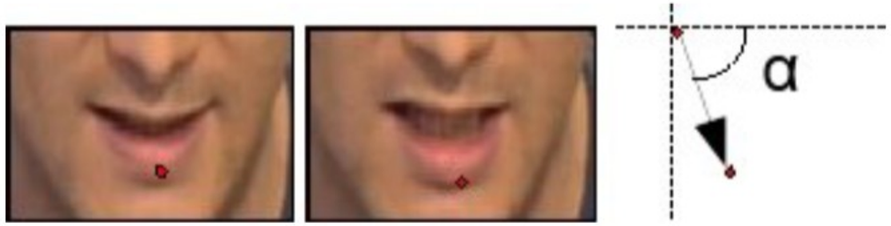




Lip tracking



Lip Activity Measure



$$\mathrm{Entropy}(\mathrm{X_t}, \mathrm{X_{t+1}}) = -\sum_i P(\alpha_i) \times log(P(\alpha_i))$$

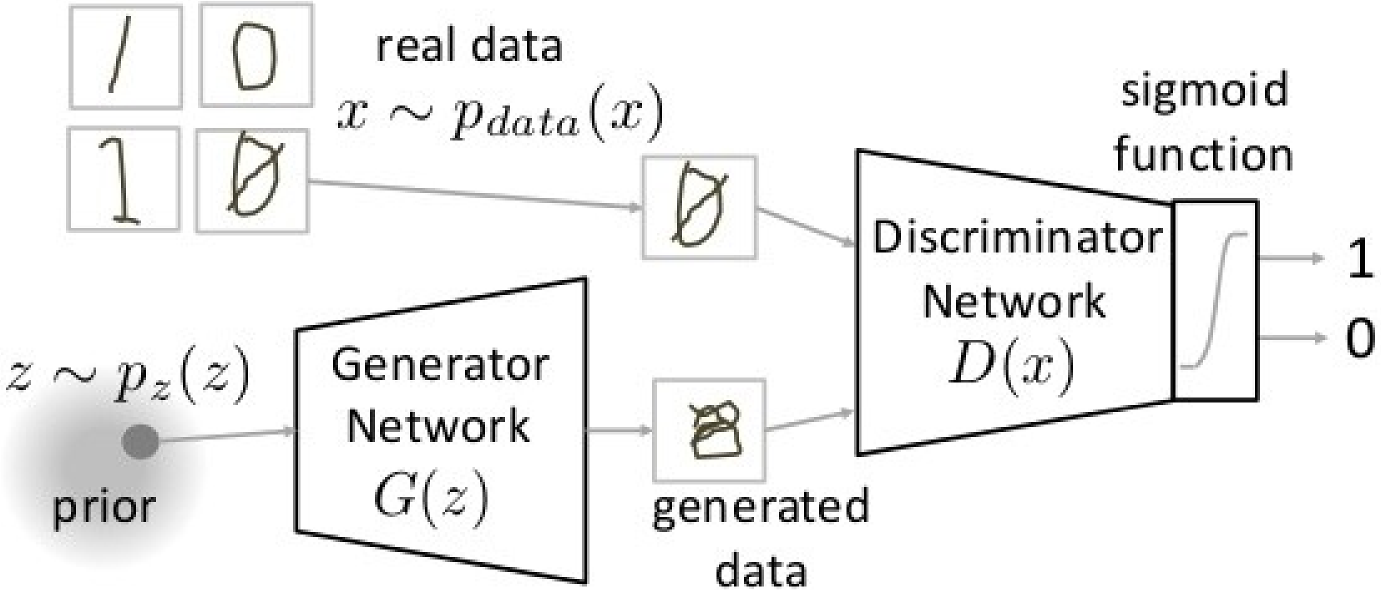
$$\mathrm{Mv}(\mathrm{X}) = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathrm{Entropy}(X_{t+i-1}, X_{t+i}) \qquad (2)$$

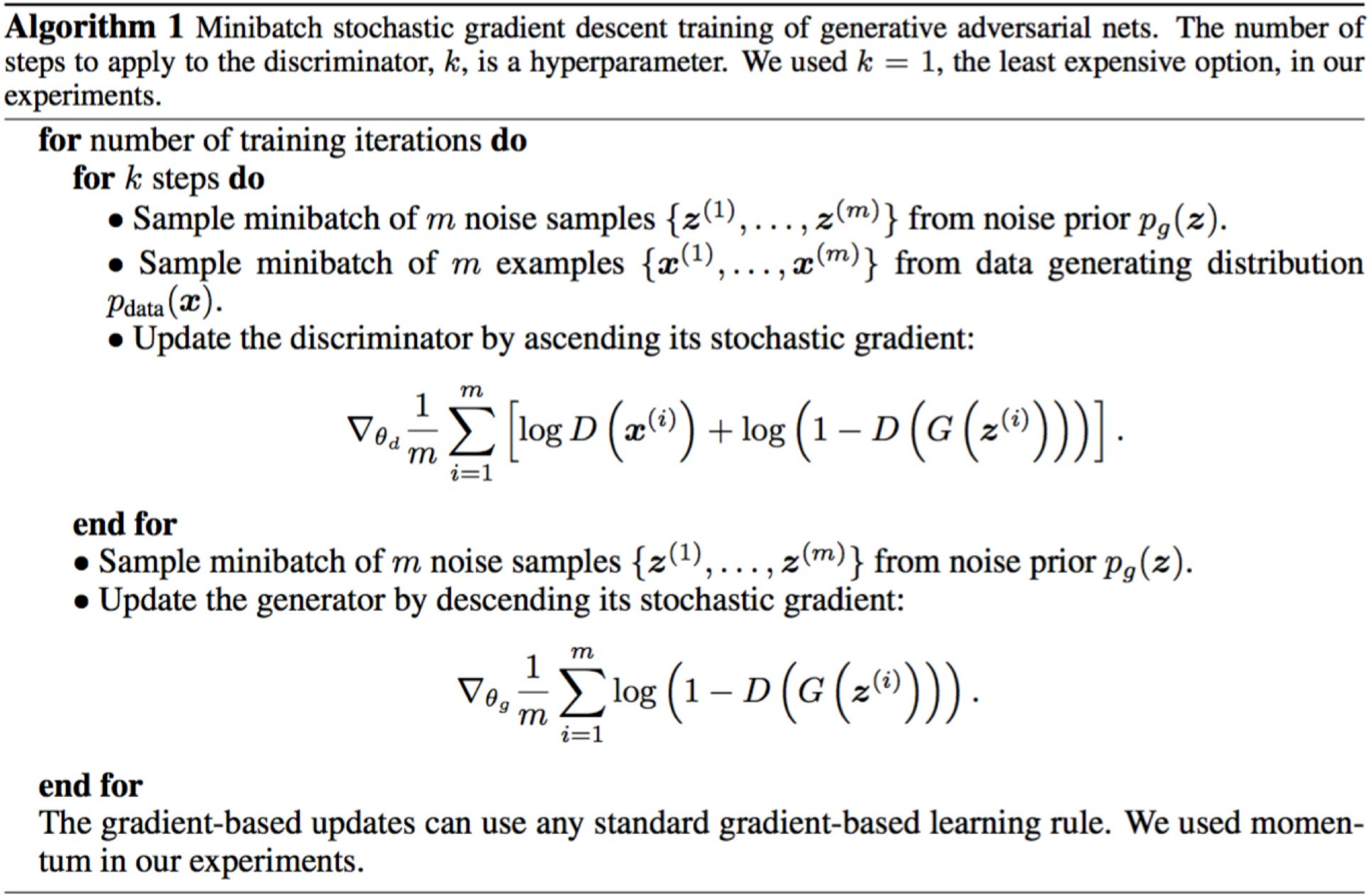
Then, the decision of talking face is taken by comparing $Mv(X)$ to a given threshold.

Generative Adversarial Networks

$$\min_G \max_D V(D, G)$$

$$V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$





Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

```
for number of training iterations do
    for $k$ steps do
        • Sample minibatch of $m$ noise samples $\{\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
        • Sample minibatch of $m$ examples $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
        • Update the discriminator by ascending its stochastic gradient:
            $$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$
    end for
    • Sample minibatch of $m$ noise samples $\{\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
    • Update the generator by descending its stochastic gradient:
        $$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$
end for
```

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

Image Courtesy – Generative adversarial Nets. Goodfellow et al.

GANs Evolution – Conditional Adversarial Networks

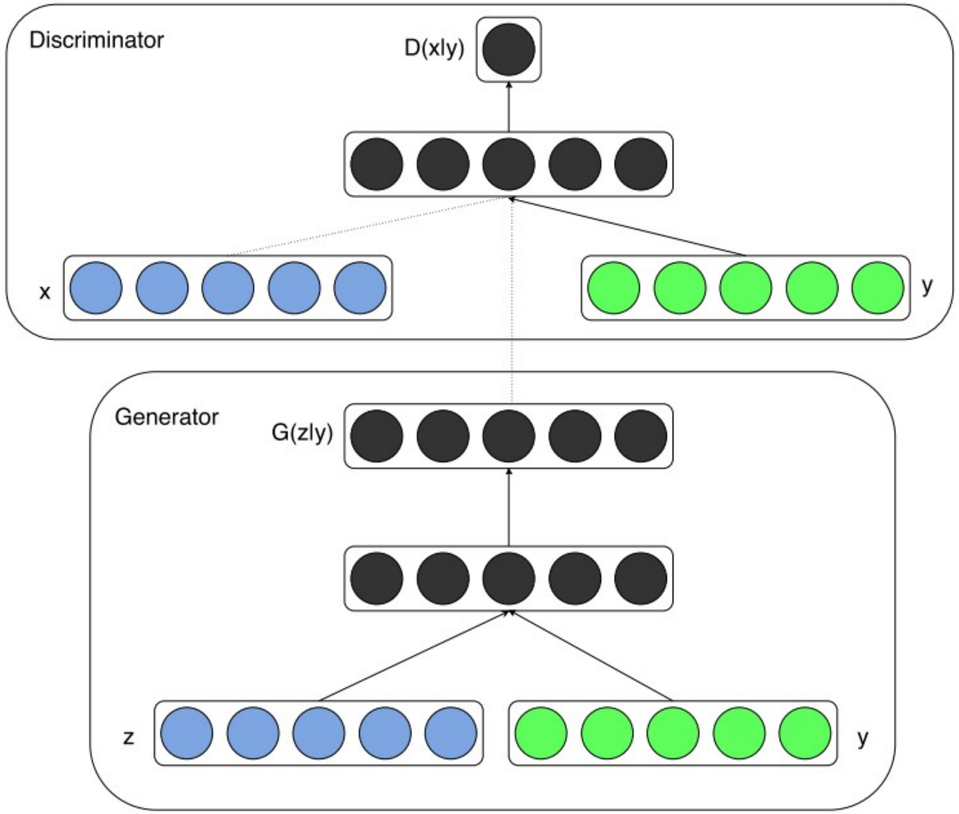


Image Courtesy – Conditional Generative adversarial Nets. Mirza et al.

GANs Evolution –
Deep Convolutional GAN

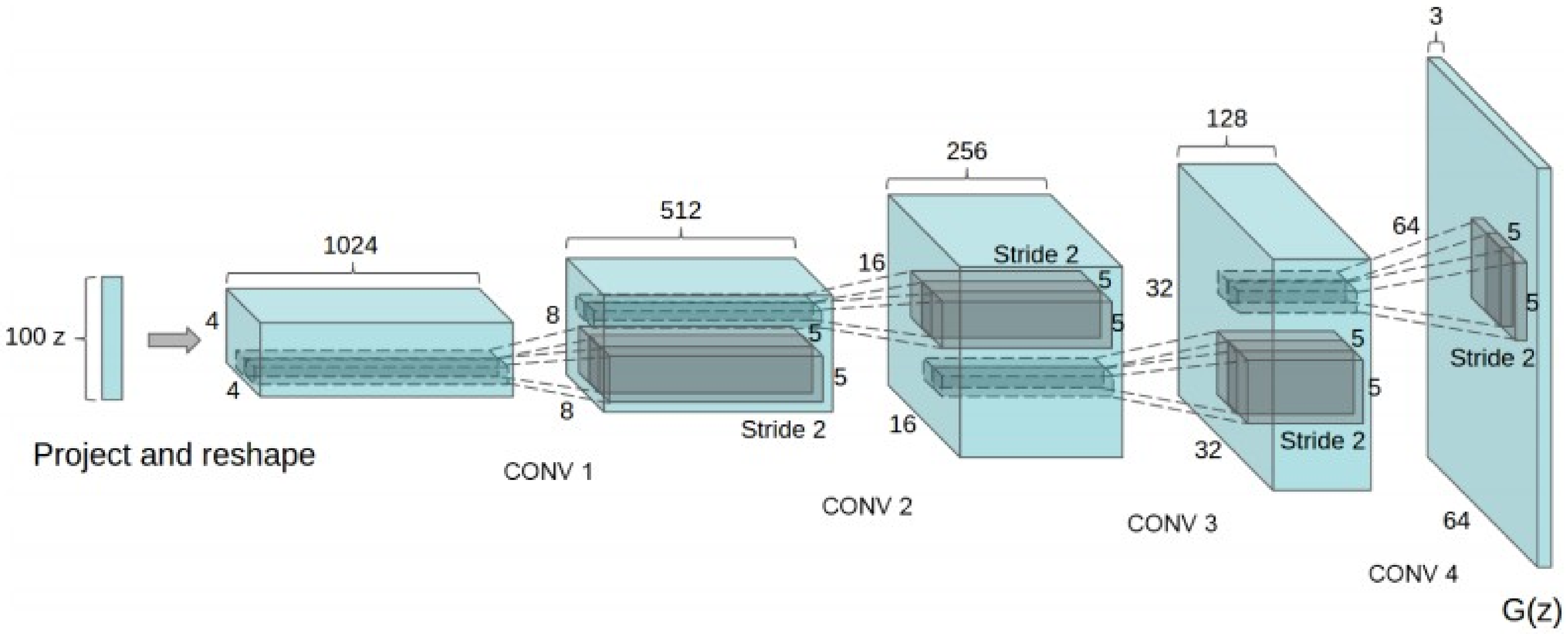


Image Courtesy – Conditional Generative adversarial Nets. Mirza et al.

Text-to-image Synthesis

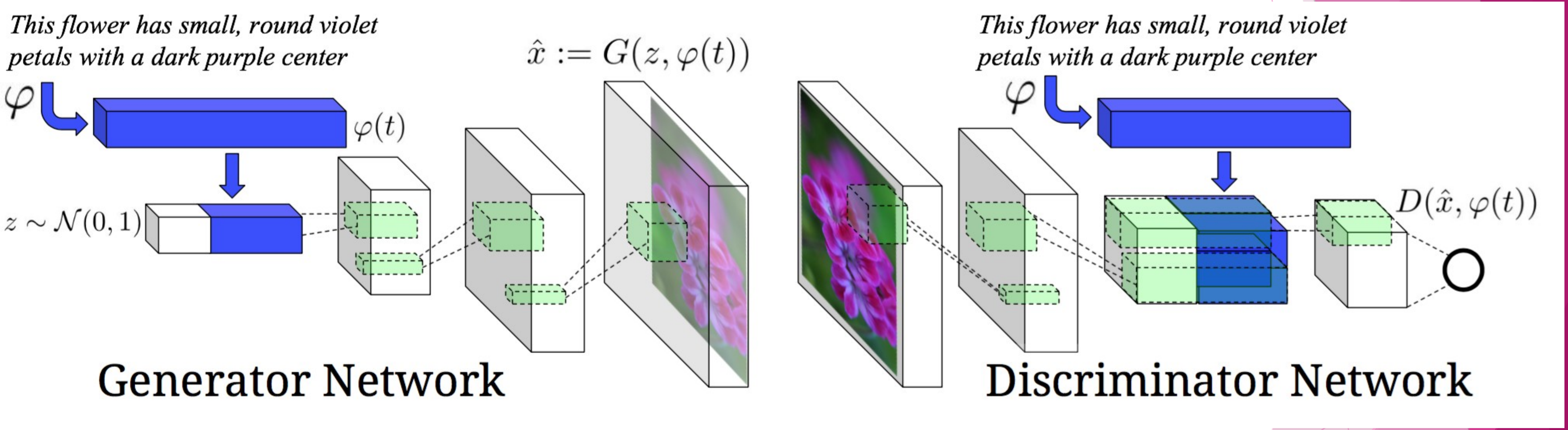
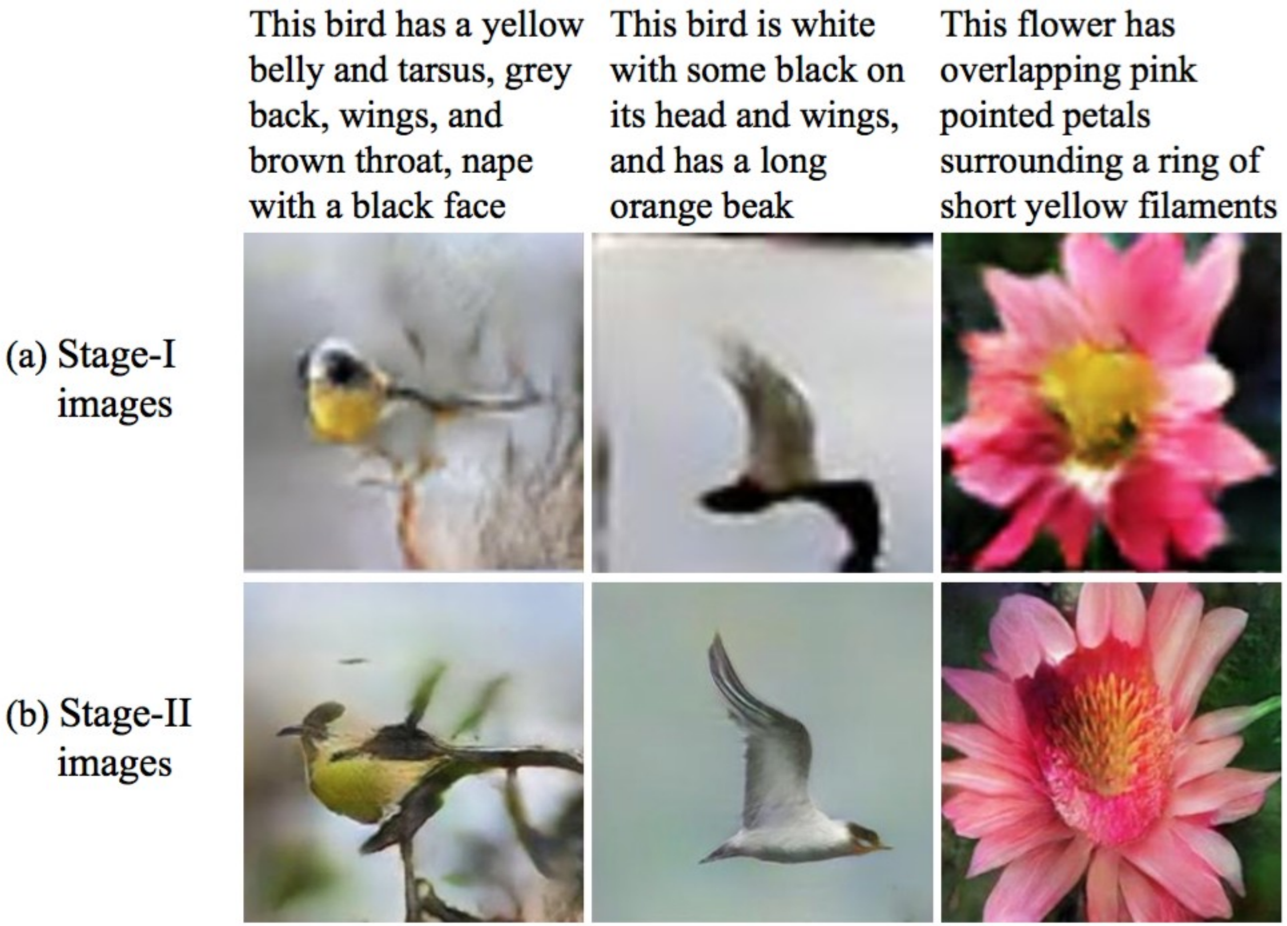

This flower has small, round violet petals with a dark purple center
φ
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
This flower has small, round violet petals with a dark purple center
φ
$D(\hat{x}, \varphi(t))$
Generator Network
Discriminator Network

Image Courtesy – Generative adversarial text to image. Reed et al.

StackGAN



This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face

This bird is white with some black on its head and wings, and has a long orange beak

This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments

Image Courtesy – StackGAN. Zhang et al.

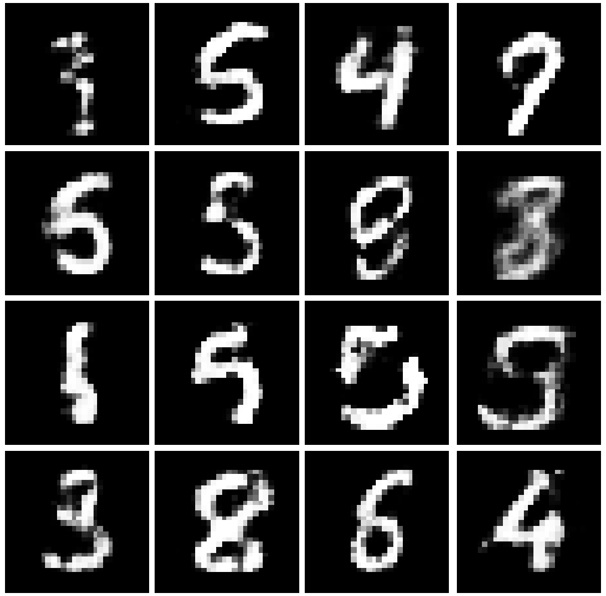
Experimentation -

- Motivation – Text-to-image

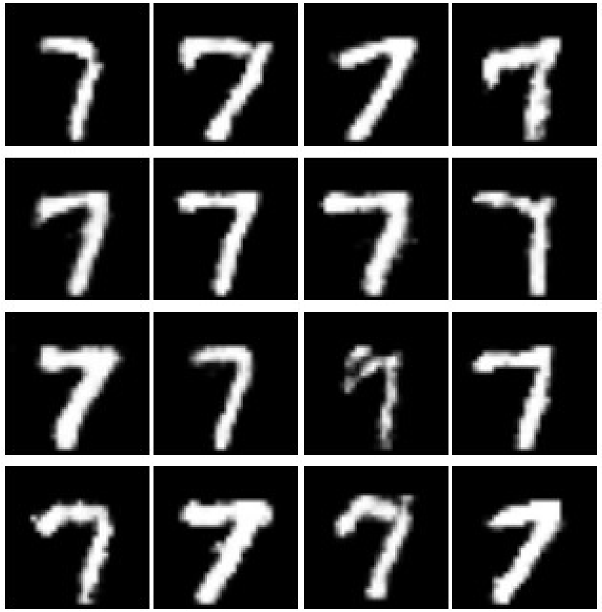
  StackGAN

- Start with Vanilla GAN.

  Generate digits using MNIST dataset.



Vanilla GAN results

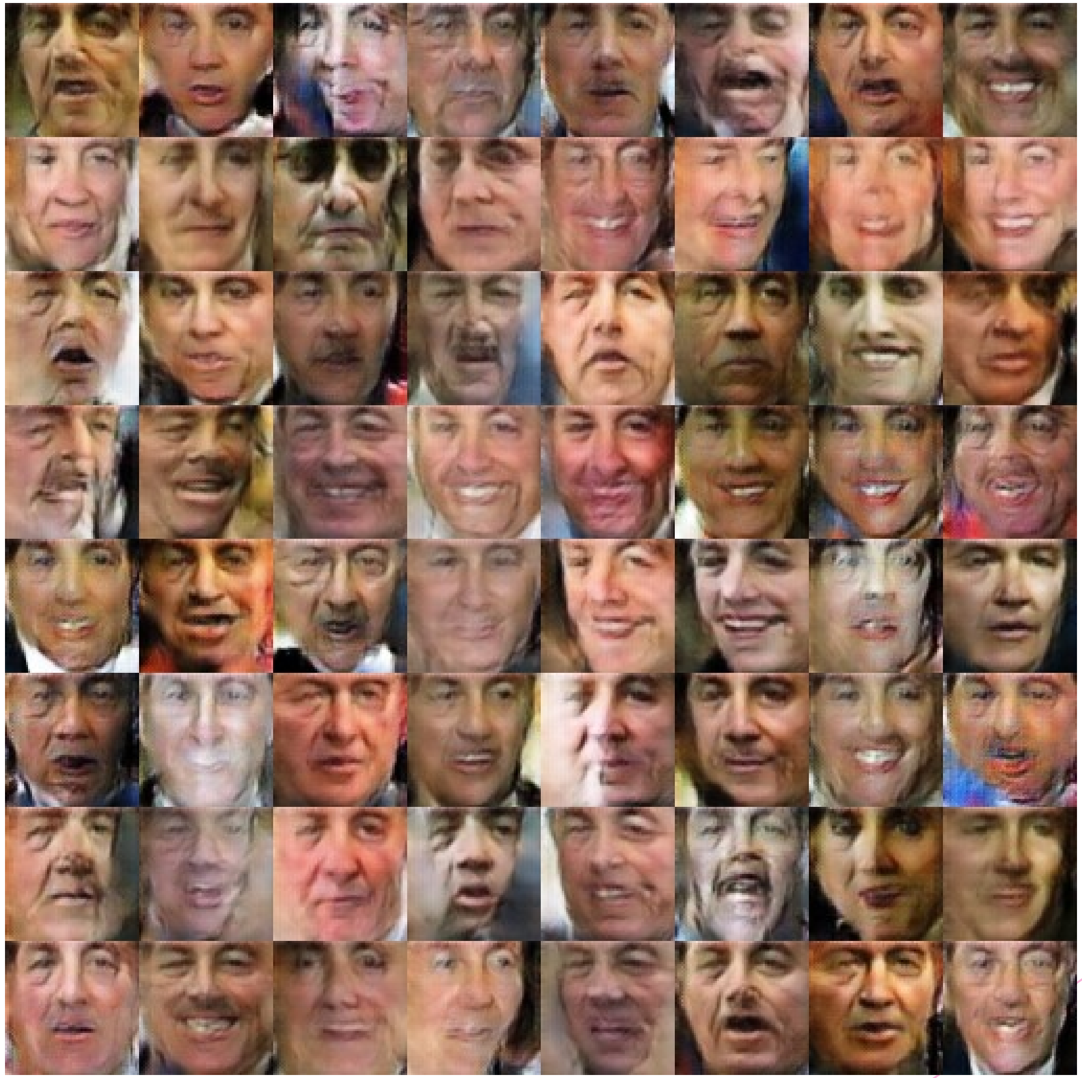


Conditional Vanilla GAN results with the 7th bit set high in conditional encoding

Experimentation -

- DCGAN

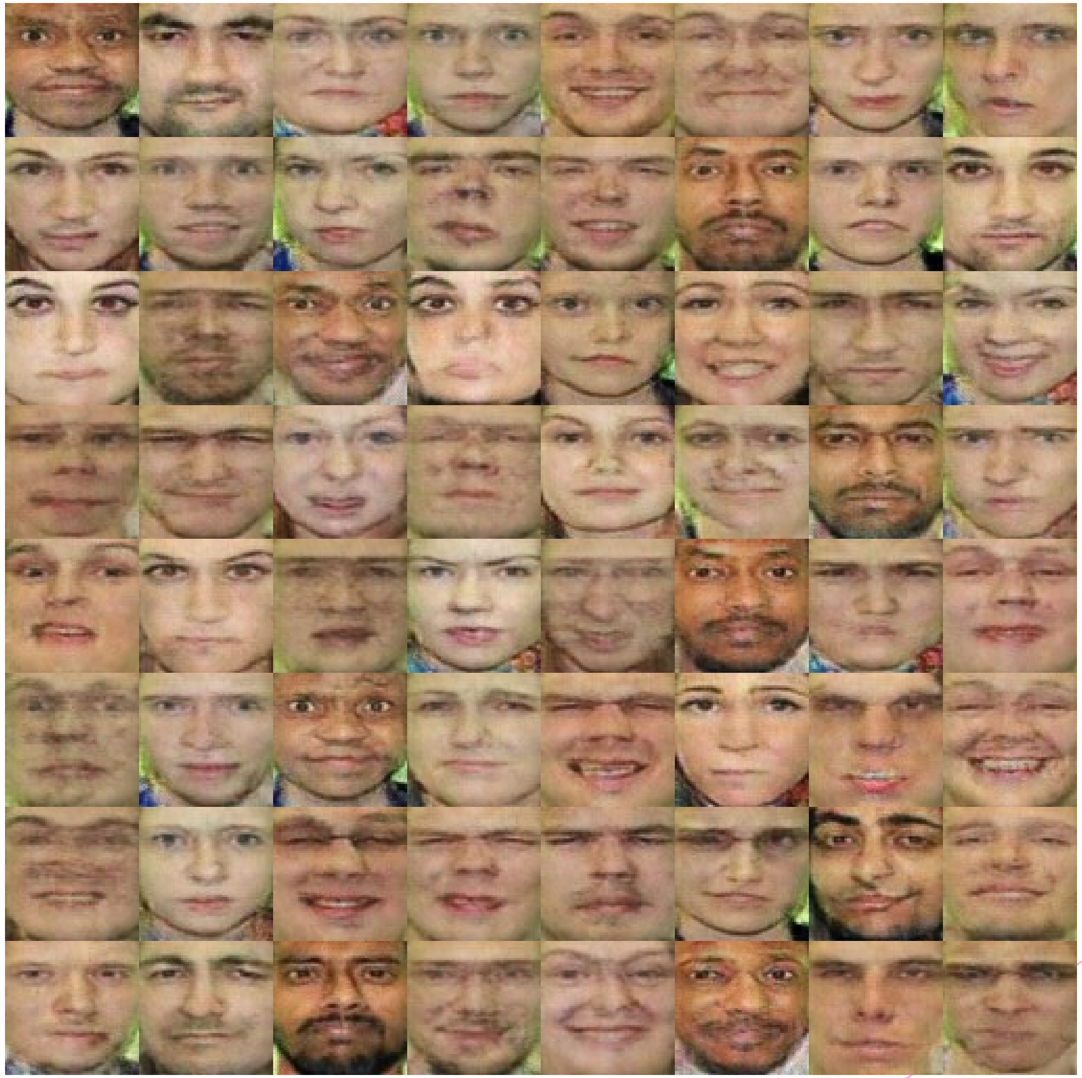
  Generate faces.



Experimentation -

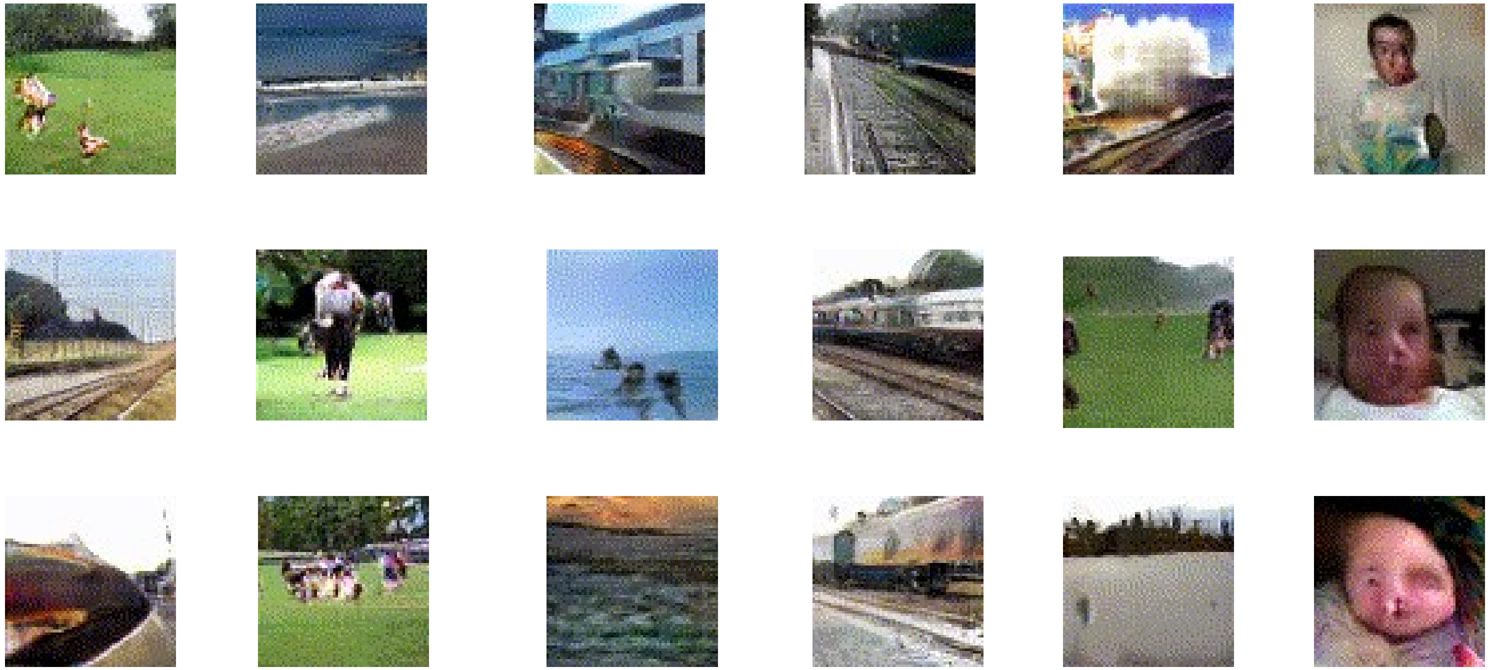
- Conditional DCGAN

Generate “Happy” faces.



Conclusion ?

- How to use in our context ?

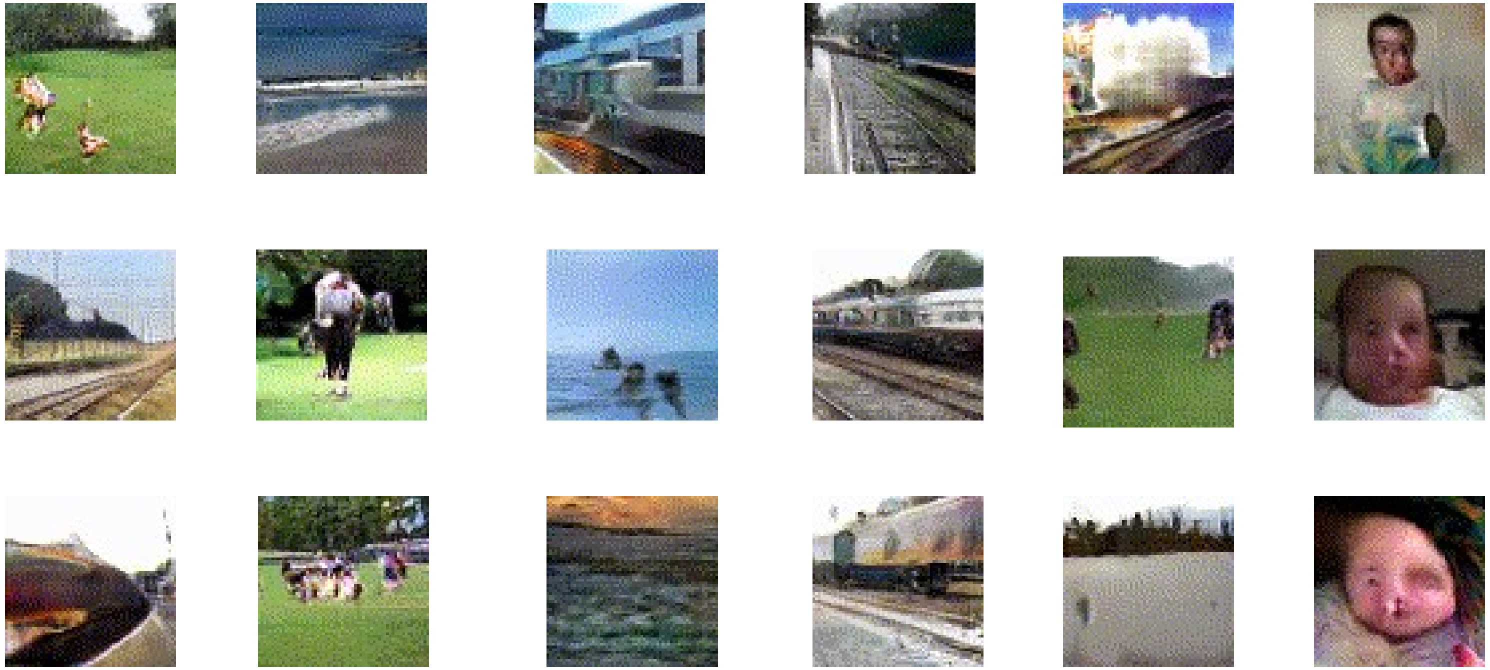


Generating Videos with Scene Dynamics
Vondrick et al - MIT

Conclusion ?

- These small 64 X 64 - 1 sec Gifs formed by training over dataset of

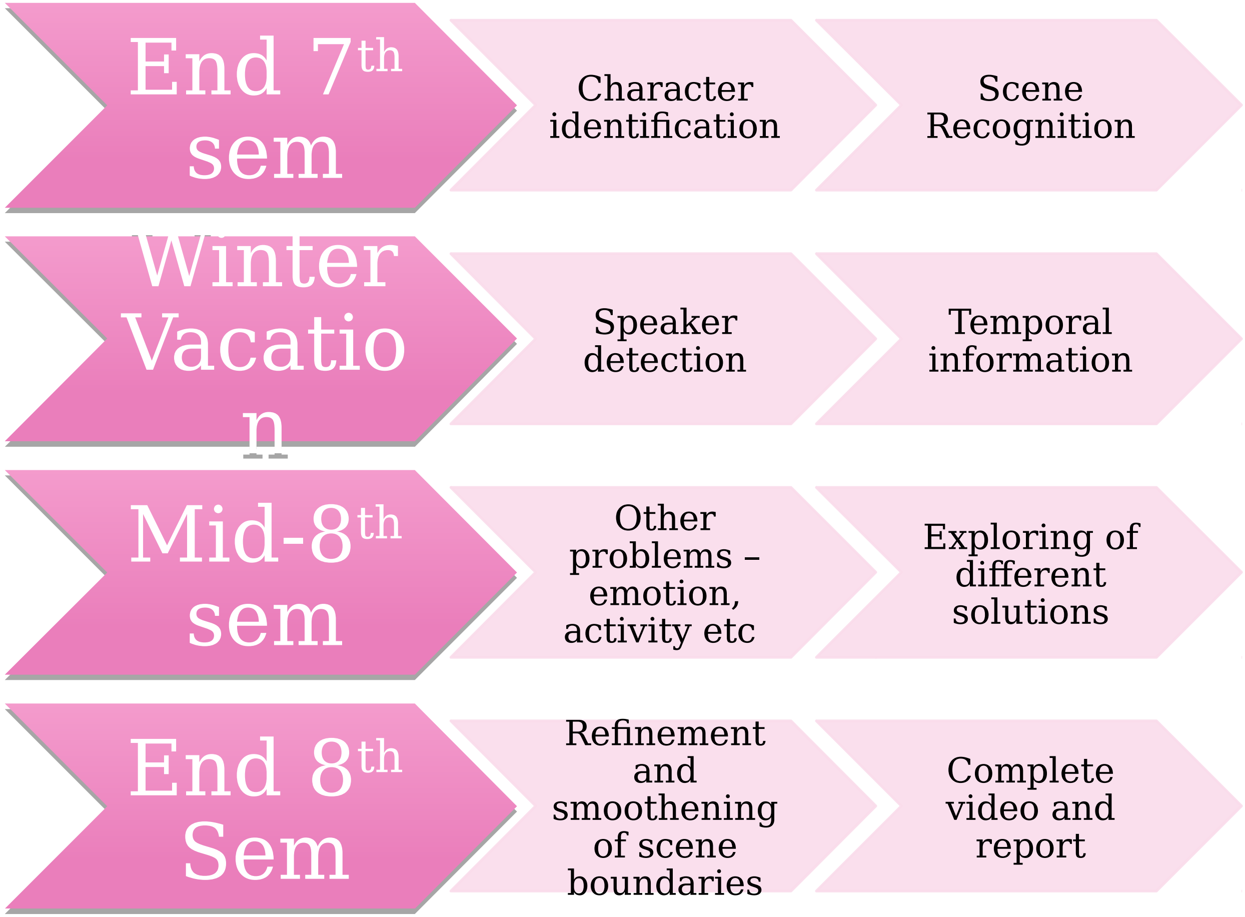
9 TB = 1024 X 9 = 9216 GB !!!



Generating Videos with Scene Dynamics
Vondrick et al - MIT

Sample Integration Videos

Timeline



End 7th sem
Character identification
Scene Recognition

Winter Vacation
Speaker detection
Temporal information

Mid-8th sem
Other problems – emotion, activity etc
Exploring of different solutions

End 8th Sem
Refinement and smoothening of scene boundaries
Complete video and report

References

- F. Chollet. (2016, Aug.) Building powerful image classification models using very little data. [Online]. Available: https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html
- K. Simonyan and A. Zisserman, “Very deep convolutional networks for largescale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available:http://arxiv.org/abs/1409.1556
- L. Wang, S. Guo, W. Huang, and Y. Qiao, “Places205-vggnet models for scene recognition,” *CoRR*, vol. abs/1508.01667, 2015. [Online]. Available: http://arxiv.org/abs/1508.01667

References

- Wikipedia, “Friends — wikipedia, the free encyclopedia,” 2016, [Online; accessed 25-November-2016]. [Online]. Available: https://en.wikipedia.org/wiki/Friends
- N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Volume 1 - Volume 01*, ser. CVPR ’05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 886–893. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.177
- Y. T. Y. R. Wolf, “Deepface: Closing the gap to human-level performance in face ver- ification,” in *Conference on Computer Vision and Pattern Recognition (CVPR), 2014*, June 2014.
- F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” *CoRR*, vol. abs/1503.03832, 2015. [Online]. Available: http://arxiv.org/abs/1503.03832
- B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “Openface: A general-purpose face recognition library with mobile applications,” CMU-CS-16-118, CMU School of Com- puter Science, Tech. Rep., 2016.
- D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Re- search*, vol. 10, pp. 1755–1758, 2009.
- V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR ’14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1867–1874. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.241