Selfie Detection by Synergy-Constraint Based Convolutional Neural Network

Yashas Annadani, Vijayakrishna Naganoor, Akshay Kumar Jagadish and Krishnan Chemmangat

Department of Electrical and Electronics Engineering
National Institute Of Technology- Karnataka, India.

NITK Surathkal
Mangalore, India

Organization

Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. Results and Discussion
6. Conclusion
7. Future Work

Problem Formulation

- Binary Classification Problem

- Not a trivial task

  (a) Variations in poses.

  (b) Illumination, Proximity and Viewpoints.

  (c) Background and People involved.

Selfies	Country
2,395	United States
899	United Kingdom
590	Philippines
534	Italy
441	Malaysia
402	Indonesia
370	Mexico
366	Brazil
248	Turkey
226	Poland
215	France
208	Canada

*** Number of Selfies taken per 100,000 people.**

Source Data: http://time.com/selfies-cities-world-rankings/

Organization

Motivation

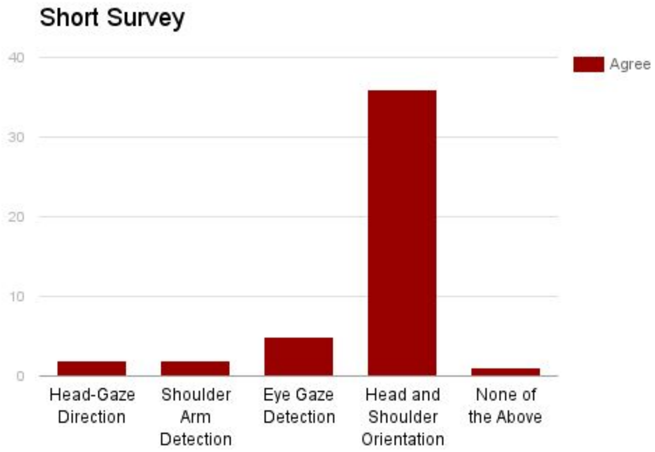
- Large scale image database segregation and retrieval [1].
- Sentiment Analysis [2] [3].
- Psychological studies [4] [5].
- Terminal Scene Understanding.
- Automatic Selfie-specific Image Processing.

Example(s)

Example(s) - Human Perception

Organization

How does the world perceive?



Organization

1. Problem Formulation
2. Motivation
3. Survey
4. **Dataset Description**
5. Results and Discussion
6. Conclusion
7. Future Work

Dataset Description

- 70,000 images - roughly equal number of selfies and non-selfies.

- Majority of selfies obtained from selfeed.com [6]

- Non-selfies (a) ImageNet [7] - People and Objects Category
  (b) INRIA persons dataset [8]

- Around 4k non-selfies were also manually collected.

- Dataset kept as realistic as possible by not including images of landscapes, animals and vehicles.

Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. **Results and Discussion**
   - **(a) Shallow Model**
   - **(b) Synergy-feature based SVM**
   - **(c) Constrained CNN for feature extraction**
6. Conclusion
7. Future Work

Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. **Results and Discussion**
   (a) **Shallow Model**
   (b) Synergy-feature based SVM
   (c) Constrained CNN for feature extraction
6. Conclusion
7. Future Work

Experimental Setup

- Fine-tuning in caffe [20], initial learning rate was set to 10e-06.
- Network was trained for 8000 iterations with Stochastic Gradient Descent with Batch Size of 16.
- The learning rate decreased by a factor of 0.5 for every 2000 iterations.
- For SVM, linear kernel was used for classification with the regularization parameter C to be 1.

Performance of Popular CNN Architectures

Architecture	**Accuracy(mAP)**
AlexNet [9]	81.9
GoogleNet [10]	82.4

- Accuracy less than 85% for both the architectures on a binary classification problem.
- How do the activations of the filters look like?

CNN Visualizations

CNN Visualizations

Can enforcing the network to learn head and shoulder orientation help?

Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. **Results and Discussion**
   (a) Shallow Model
   (b) Synergy-feature based SVM
   (c) Constrained CNN for feature extraction
6. Conclusion
7. Future Work

Handcraft feature extraction

Two features were extracted as in [11] where

(a) Hierarchical Histogram of gradients [12] - Head and shoulder alignment.

(b) Hierarchical Local Binary Descriptor [13] [17] - Face and head detection.

Synergy Feature Generation

- The synergy feature between head and shoulder orientation was learnt **to find a single descriptor that represents both features**.
- Canonical Correlation analysis [14] [15] a procedure that seeks maximal correlations between combinations of variables in both sets of data.
- As tool that finds the best projection of the feature matrices onto a common subspace such that correlation is maximized -

$$\rho_i = corr(U_i, V_i) \; \forall \; i \in min\{c, d\}$$
$$\text{where } U_i = a_i\gamma \text{ and } V_i = b_i\tau$$

Synergy feature S - Standardized Euclidean norm of the difference between U_1 and V_1.

Performance of a shallow model

Architecture	Accuracy(mAP)
AlexNet [9]	81.9%
GoogleNet [10]	82.4%
Synergy feature + SVM	52.4%

Synergy feature alone is not discriminative enough !!!

Idea

- Handcraft feature generation for head orientation and shoulder orientation - Done
- Capturing the Synergy in the two orientations - Done
- Getting the final descriptor for classification ???

Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. **Results and Discussion**
   (a) Shallow Model
   (b) Synergy based SVM
   (c) Constrained CNN for feature extraction
6. Conclusion
7. Future Work

Constrained CNN - A Feature Extractor

- CNN architecture that learns the synergy in common subspace between feature.
- We employ alexnet [9] architecture and transfer learning by fine-tuning the model pretrained on the Imagenet dataset.
- Replacing the last fully connected layer by another of dimension same as S.
- The loss function is modeled as

$$\ell = ||\sigma(\Theta) - S||_2^2$$

Θ is the vector of activations of the last modified fully-connected layer, σ is $ReLU$ non-linearity followed by a soft-max operation and S is the normalised synergy feature

Selfie Descriptor

- The trained Convolutional layers are used as feature pools.

- Features are obtained at points which are view invariant.

- SIFT [16][19] keypoint locations are determined and the corresponding locations are tracked through the network using their stride value.

$$\theta = (\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_K, y_K\})$$

- Activations in the four connected neighbourhood of these key

$$t_p^i = \frac{1}{K} \sum_{k=1}^{K} \hat{C}_p(i, \phi(\overline{r_p \times x_k}), \phi(\overline{r_p \times y_k}))$$

.

- These pooled activations are aggregated over all the layers and concatenated to get the selfie descriptor [21].

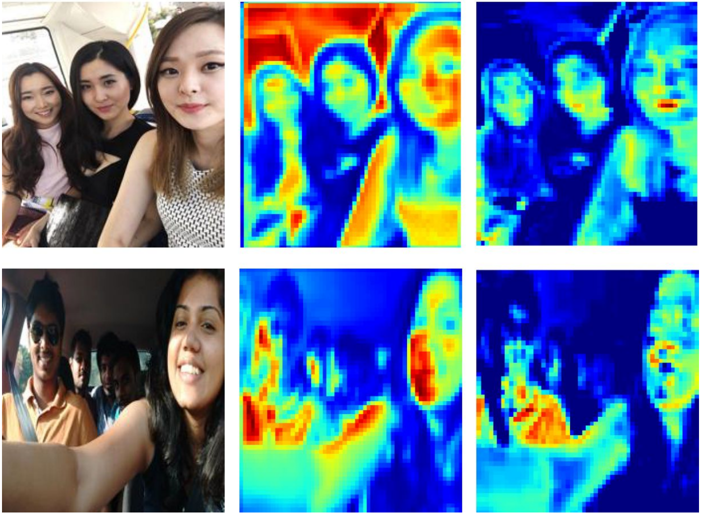
Experimental Results

Model	Accuracy(mAP)
Synergy Feature + SVM	52.4
Unconstrained AlexNet [9]	81.9
Unconstrained GoogleNet [10]	82.4
Synergy Constrained AlexNet (Proposed)	**86.3**

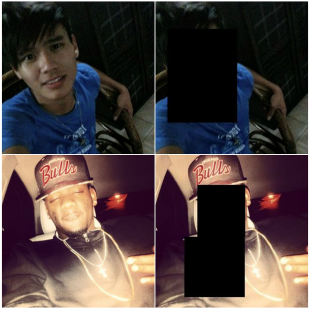
Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. **Results and Discussion**
   (a) Shallow Model
   (b) Synergy based SVM
   (c) Verification of Constrained CNN for Tractability
6. Conclusion
7. Future Work

CNN Activations



Occlusion Test [18]



Method	Accuracy (mAP)
Unconstrained AlexNet	77.1%
Unconstrained GoogleNet	75.3%
Synergy Constrained AlexNet (Proposed)	58.8%

Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. Results and Discussion
6. **Conclusion**
7. Future Work

Conclusion

- This paper presents a synergy constraint based CNN training paradigm **features discriminative for selfie detection.**

- The motivation was to mimic human perception by -

  Capturing the synergy between head orientation and shoulder arm orientation.

- Relevant features were extracted and a synergy measure was obtained using **CCA on the two sets of handcrafted features.**

- The hypothesis was tested through ablative study

  (a) **Visualization of Activations.**

  (b) **Occlusion Test.**

- Experimental evaluation prove with mAP being better by **4%.**

Organization

1. Problem Formulation
2. Motivation
3. Survey
4. Dataset Description
5. Results and Discussion
6. Conclusion
7. **Future Work**

Future Work

- Application in other subtle imaging analysis scenarios. Example: Scene understanding to separate foreground and background.
- Imposing Different Loss functions on different layer.
- Finding a method to capture synergy between more than two features.
- Making use of Tractability in medical imaging
  Example: Tumour detection

Thank you

Any Questions ?

References

[1] N. Singhai and S. K. Shandilya, “A survey on: content based image retrieval systems,” International Journal of Computer Applications, vol. 4, no. 2, pp. 22–26, 2010.

[2] L. Qiu, J. Lu, S. Yang, W. Qu, and T. Zhu, “What does your selfie say about you?” Computers in Human Behavior, vol. 52, pp. 443–449, 2015.

[3] E. B. Weiser, “# me: Narcissism and its facets as predictors of selfieposting frequency,” Personality and Individual Differences, vol. 86, pp. 477–481, 2015.

[4] P. Sorokowski, A. Sorokowska, A. Oleszkiewicz, T. Frackowiak, A. Huk, and K. Pisanski, “Selfie posting behaviors are associated with narcissism among men,” Personality and Individual Differences, vol. 85, pp. 123–127, 2015.

References

[5] K. Warfield, “Making selfies/making self: Digital subjectivities in the selfie,” 2014.

[6] M. M. Kalayeh, M. Seifu, W. LaLanne, and M. Shah, “How to take a good selfie?” in Proceedings of the 23rd Annual ACM Conference on Multimedia Conference. ACM, 2015, pp. 923–926.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 248–255.

[8] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 1. IEEE, 2005, pp. 886–893.

References

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan,
V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[11] C. Zeng and H. Ma, "Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting," in Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010, pp. 2069–2072.

References

[12] S. Wang, J. Zhang, and Z. Miao, “A new edge feature for head-shoulder detection,” in 2013 IEEE International Conference on Image Processing. IEEE, 2013, pp. 2822–2826.

[13] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 12, pp. 2037–2041, 2006.

[14] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, “Deep canonical correlation analysis.” in ICML (3), 2013, pp. 1247–1255

[15] R. L. Malacarne, “Canonical correlation analysis,” Mathematica Journal, vol. 16, p. 66, 2014.

References

[16] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2879–2886.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[18] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2879–2886.

[19] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010, pp. 1469–1472.

References

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the ACM International Conference on Multimedia. ACM, 2014, pp. 675–678.

[21] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory pooled deep-convolutional descriptors," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.