**ước tính hình dáng đầu sử dụng mô hình mạng fsa được chỉnh sửa dựa trên cấu trúc mạng sinh ba**

head pose estimation using modified fsa-net based on triplet network architecture

***SVTH: Trần Văn Long1, Nguyễn Vệ Bình2, Nguyễn Khắc Ngọc Hoan3, Hồ Hoàng Hảo4, Tôn Thất Trần Vũ5***

*Lớp 19ECE135, 19ES2, 16ECE14, Khoa FAST, Trường Đại Học Bách Khoa Đà Nẵng*

*Email:* [*anhduonghnh@gmail.com*](mailto:anhduonghnh@gmail.com)*1,*

[*nguyenvebinh2001@gmail.com*](mailto:nguyenvebinh2001@gmail.com)*2,*

[*demon2442001@gmail.com*](mailto:demon2442001@gmail.com)*3,*

[*haromusio308@gmail.com*](mailto:haromusio308@gmail.com)*4,*

[*kingjune.6th2001@gmail.com*](mailto:kingjune.6th2001@gmail.com)*5*

***GVHD: TS. Nguyễn Quang Như Quỳnh6, KS. Phùng Hữu Tài7***

*Khoa FAST6, Trường Đại Học Bách Khoa – Đại Học Đà Nẵng; Email:* [*nqnquynh@dut.udn.vn*](mailto:nqnquynh@dut.udn.vn)

*Khoa Điện7, Trường Đại Học Quốc Lập Trung Chính Đài Loan; Email:* [*huutaibkdn@gmail.com*](huutaibkdn@gmail.com)

**Tóm tắt –** Bài báo này đề xuất một phương pháp cho bài toàn Ước tính dáng đầu chỉ từ một bức ảnh RGB. Những cách tiếp cận trước đây thường bắt được những đặc tính không quan trọng như hình khối khuôn mặt, … thay vì dáng đầu. Phương pháp của chúng tôi sẽ kết hợp mạng lưới gốc là FSA với cấu trúc mạng sinh ba để nâng cao mô hình mạng FSA. Để có được một mô hình hiệu quả, chúng tôi tận dụng cơ chế hồi quy mềm theo giai đoạn (soft stage-wise regression – SSR) cho mạng FSA. Hiện tại, những phương pháp tổng hợp đặc trưng xét đầu vào là một khối những đặc trưng và thường bỏ qua mối quan hệ về không gian giữa chúng. Bên cạnh đó, chúng tôi kết hợp cấu trúc mạng sinh ba để mô hình có thể học tốt hơn những đặc tính liên quan nhiều tới dáng đầu. Kết quả đạt được từ phương pháp này chưa vượt qua những phương pháp cũ vì một số vấn đề liên quan đến quá khớp, tuy nhiên cũng khá tốt và cho thấy nhiều tiềm năng. Trong tương lai, chúng tôi dự định sẽ dùng một số phương pháp điều chỉnh để có thể giải quyết bài toán này tốt hơn.

**Từ khóa –** Ước tính dáng đầu; Mạng sinh ba; Học sâu; FSA-Net; Thị giác máy tính.

**Abstract -** This paper proposes a method for Head Pose Estimation from a single RGB image. Previous approaches often catch non-essential features such as face shapes, … other than head poses. Our method will associate the original Fine-Grained Structure Aggregation Network (FSA-Net) [1] with the Triplet Network architecture to enhance the FSA-Net model. For having a compact model, we utilize the soft stage-wise regression (SSR) scheme beyond the landmark for FSA-Net. Currently, feature aggregation methods consider inputs as a block of features and often neglect their spatial relationship in a feature map. Besides that, we employ Triplet Network in order for the model to better learn features that are more relevant to head poses. The result obtained from our method is still not better than the previous method yet, as it is a bit prone to overfitting, but is good enough and shows much potential. In the future, we can apply some regularization methods to help with the overfitting problem.

**Key words -** Head pose estimation; Triplet network; Deep Learning; FSA-Net; Computer Vision.

# Introduction:

Head pose estimation is a challenging task in the field of computer vision [2, 3, 4, 24, 25]. In recent years, various facial analysis techniques have been recommended such as face recognition [4, 6] or identification, facial age estimation, landmark detection [3], and head pose estimation. Head pose estimation has many applications such as aiding in gaze estimation, modeling attention, fitting 3D models to video and performing face alignment. In addition, it is used to enhance and supply the information for identity recognition, expression recognition, or attention detection.

Head pose estimate is a 3D vector including the yaw, pitch and roll angles. Head pose estimation from a single image [2] is a complicated process which requires the knowledge of a mapping between 2D and 3D spaces. Some methods promote more modalities, namely depth image [25, 14] and temporal information in video sequences [16]. The depth images yield the “depth” of the object, an extra information of the object in real life that is lacking in 2D images. Videos are also a great preference for Head pose estimation due to the recorded continuous motion of human heads. Most single-frame pose estimation techniques apply facial landmark detection for estimating head poses [20, 3]. However, it would require more computation.

In this paper, we modify the model that requires only a single RGB frame as input, called FSA-Net using the Triplet Network architecture. Triplet Network is a method that helps obtain better embeddings for the wanted features, in this case the head pose, rather than irrelevant features like identity. Combining those techniques together, we aim to achieve better results for FSA-NET while still using only an RGB frame as input and no extra information like time or depth. The proposed method consists of 3 instances of the same feedforward network (with shared parameters). When absorbed 3 samples, the network yields 2 intermediate values - the gap between the embedded representation of two of its inputs from the representation of the third one (the anchor). Therefore, we expect that modifying FSA-Net with Triplet Network would yield more effective results.

# Related work:

FSA-Net: The approach for estimating head posture from a single image is suggested in this study [1]. Previous approaches would frequently anticipate head postures by landmark or depth estimation, which would involve extra computation. Regression and feature aggregation are the foundations of our approach. We use the soft stagewise regression method to create a compact model. The spatial link between features in a feature map is ignored by current feature aggregation algorithms since they interpret inputs as a collection of features.

Before aggregation, we suggest learning a fine-grained structural mapping for spatially grouping features. Part-based information and pooled values are provided by the fine-grained structure. Different model variants can be created by utilizing learnable and non-learnable importance across the spatial location, and these variants work together to create a complementing ensemble.Research demonstrates that our approach outperforms cutting-edge technology approaching both landmark-free and methods based on depth or landmark estimation.

Our solution even outperforms techniques using multimodality information (RGB-D, RGB-Time) on yaw angle estimation with just a single RGB frame as input.

Attention: This technique provides attention for pose estimation. Our attention can be optimized in an end-to-end manner along with the pose estimation without complex additional techniques. While other pooling methods use attention such as CBAM and Attentional Pooling, our technique has a little bit of difference with those procedures. when they focus on categorical classification problems (image classification and action recognition). On the other hand, our method is for a regression problem. While our model is capable of generating multiple spatial attention proposals which are more flexible for refining regression values, they only generate one or two spatial heat maps. And the most important, our method takes into account multi-scale information, and it’s useful to other applications.

Capsule Network: With this technique, We try to simply illustrate that a properly working capsule network should achieve higher results with a considerably lower number of parameters due to its intrinsic capability to adjust information more efficiently. In this procedure, we test the put forward methodology in an experimental context, assessing its generalization capabilities and efficiency. For this purpose, we test our proposed approach with three of the most used datasets for capsule-based networks assessment: MNIST, smallNORB and MultiMNIST. All experimentation clearly shows that a capsule network is capable of achieving higher results with a significantly lower number of parameters. Not only that, we show how a simple ensemble of a few instances of Efficient-CapsNet can easily indicate state-of-the-art results in all the three datasets. Last but not least important, using principal component analysis, we give an introspect to the inner representations of the network and its capability to encode visual information.

Metric learning: The goal of metric learning is to learn a representation function that maps objects into an embedded space. The distance in the embedded space should preserve the objects’ similarity — similar objects get close and dissimilar objects get far away. Various loss functions have been developed for metric learning. In particular, Triplet loss is implemented in different fields, which requires the distance between the anchor sample and the positive sample to be smaller than the distance between the anchor sample and the negative sample.

Triplet loss: This technique is used for face verification, recognition and clustering. The technique is to strive for an embedding f(x), from an image x into a feature space Rd, such that the squared distance between all faces, independent of imaging conditions, of the same identity is small, whereas the squared distance between a pair of face images from different identities is large. Triplet loss method is presented by the function f(x) ∈ Rd. It embeds an image x into a d-dimensional Euclidean space. In addition, we constrain this embedding to live on the d-dimensional hypersphere f(x)2 = 1. This loss is motivated in the context of nearest-neighbor classification. When applying all possible triplets would result in many triplets that are easily satisfied. These triplets would not contribute to the training and result in slower convergence, as they would still be passed through the network. It is crucial to select hard triplets that are active and can therefore contribute to improving the model.

Triplet-network. The technique, used for training on all datasets, was done by Stochastic Gradient Descent (SGD), with an initial learning-rate and a learning rate decay regime. Triplet-network used the dropout regularization technique to avoid over-fitting. After training on each dataset, the network reached a fixed error over the triplet comparisons. Triplet procedure conjectures that data augmentation techniques may provide similar benefits to those described in previous works. Another side-effect is noticed that the representation seems to be sparse - respectively 25% non-zero values. This is very helpful when used later as features for classification both computationally and with respect to accuracy, as each class is characterized by only a few non zero elements.

# Proposed method:

Our approach involves modifying FSA-Net based on the Triplet Network structure [8]. In particular, we feed 3 images at a time to a feature extractor (FSA-Net in our case) as shown in Figure 1.

Associated with the Triplet Network architecture is the Triplet loss function. xn, xa, xp are the feature matrices of the anchor, positive and negative samples, respectively. These feature matrices are considered the representative features for regression, basically the V embeddings as in the FSA-Net overview obtained from the original paper as shown below in figure 2. It contains three vectors (batch\_size\*16) for each of the three stages of FSA-Net. Meanwhile, D(x1, x2) is the Euclidean distance between two feature matrices x1 and x2:

The triplet samples are fed into the backbone, which is FSA-Net, to get 3 embeddings matrices V. Then the triplet loss is used for these samples to minimize the distance between embeddings of the anchor and positive sample as well as maximize the distance between embeddings of the negative and positive sample. The margin error is added to the triplet loss in order to separate the two types of embeddings more clearly.

The total loss used in training is the sum of average Mean Absolute Errors (MAEs) of the yaw, pitch, rollDiagram

Description automatically generated angles of each image in the triplet and the Triplet loss, as Diagram

Description automatically generatedshown below.

**Figure 1**: Our proposed architecture - a Triplet Network with FSA-Net as the backbone.

**Figure 2**. Overview of the original FSA-Net architecture obtained from its paper. V is the matrix that we use for the Triplet Loss.

where,

This loss function is used only in training, since we do not use the Triplet Network architecture in evaluation. For that phase, we only use the average Mean Absolute Errors of each image as loss.

By using this method, we expect that the network will extract more relevant features to head pose instead of identity i.e. Two frames of two different people withsimilar head poses should have more relevant features extracted than two frames of only one person but with different head poses.

# Implementation and results:

This section describes the implementation details and the result of our method.

We use a third-party Pytorch implementation repository: <https://github.com/omasaht/headpose-fsanet-pytorch> for the FSANet model as proposed by the original paper. Based on that repository, we modify it to combine FSANet with the Triplet Network architecture. For data augmentation, we use the same methods as in the original paper (random cropping, random scaling) and the cut-out method (note that by using cut-out data augmentation, the number of training samples is doubled). We modify the dataset so that each image is treated as an anchor and is grouped with a positive and a negative image just like in the Triplet Network architecture. We use a training batch size of 32. The model is trained using AdamW optimizer for 100 epochs with the initial learning rate of 75e-4. The learning rate is reduced every 10 epochs by a factor of 0.5. The margin error in triplet loss is set to be 0.2.

BIWI, a dataset popular for head pose estimation, is used for evaluation. It contains 24 videos of 20 subjects, with approximately 15000 RGB frames in total. We use 70% of the RGB frames for training and 30% for testing, which is the same as Protocol 2 used in [1].

This is the result of our method in comparison with other methods:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Yaw | Pitch | Roll | Average |
| **RGB-based** | | | | |
| DeepHeadPose | 5.67 | 5.67 | - | - |
| SSR-Net-MD | 4.24 | 4.24 | 4.24 | 4.24 |
| VGG16 | 4.24 | 4.24 | 4.24 | 4.24 |
| FSA-Caps-Fusion | 4.35 | 4.35 | 4.35 | 4.35 |
| **FSA-Caps-Fusion (Pytorch Model)** | 4.19 | 4.19 | 4.19 | 4.19 |
| **FSA-Caps-Fusion-Triplet (Pytorch Model)** | 4.26 | 4.26 | 4.26 | 4.26 |
| **RGB+Depth** | | | | |
| DeepHeadPose | 5.32 | 5.32 | 5.32 | 5.32 |
| Martin | 4.76 | 4.76 | 4.76 | 4.76 |
| **RGB+Time** | | | | |
| VGG16+RNN | 3.14 | 3.14 | 3.14 | 3.14 |

The result of the third-party Pytorch implementation that we use, denoted as FSA-Caps-Fusion (Pytorch Model), is already slightly worse than the original Tensorflow model of the authors. But for research purposes, we still use it and modify it to get our version of FSA-Caps-Fusion-Triplet (Pytorch model).

We also run a demo of our method using a webcam. The blue line represents the facing direction, the green line for the downward direction and the red one for the side direction. As we can see in figure 3, all the vectors produced by our method demonstrate the head pose accurately.

A picture containing clothing, different, spectacles

Description automatically generated

**Figure 3.** Frames captured from the demo video of our method.

# Conclusion:

In this paper, we propose a new method, which is combining the Triplet Network structure with the original Fine-Grained Structure Aggregation Network (FSA-Net). This method produces results that are comparable with the original and re-implementation model. In the future, regularization techniques should be applied to deal with the overfitting problem. With an appropriately optimized set of hyperparameters, this proposed method has the potential to achieve even better results.

**Tài liệu tham khảo**

[1] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, Yung-Yu Chuang. FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation from a Single Image. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[2] 3D Head Pose Estimation Based on Graph Convolutional Network from A Single RGB Image. IEEE International Conference on Image Processing (ICIP), 2022.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In Proceedings of International Conference on Computer Vision (ICCV), 2017.

[4] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. International Journal of Computer Vision (IJCV), 107(2):177–190, 2014.

[6] Feng Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. FacePoseNet: Making a case for landmark-free face alignment. In Proceedings of Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2017.

[7] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In Proceedings of European Conference on Computer Vision (ECCV), 2014.

[8] Hai Liu , Senior Member, IEEE, Shuai Fang , Zhaoli Zhang , Duantengchuan Li , Ke Lin , and Jia Zhang Wang. MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation. DOI: 10.1109/TMM.2021.3081873.

[9] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In Joint Pattern Recognition Symposium, pages 101–110. Springer, 2011.

[10] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From Bayesian filtering to recurrent neural network. In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[12] Martin Koestinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large scale, real-world database for facial landmark localization. In Proceedings of International Conference on Computer Vision Workshops, 2011.

[13] Amit Kumar, Azadeh Alavi, and Rama Chellappa. KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2017.

[14] Lin Liang, Rong Xiao, Fang Wen, and Jian Sun. Face alignment via component-based discriminative search. In Proceedings of European Conference on Computer Vision (ECCV), pages 72–85. Springer, 2008.

[15] Manuel Martin, Florian Van De Camp, and Rainer Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In Proceedings of The 2nd International Conference on 3D Vision (3DV), 2014.

[16] Gregory P. Meyer, Shalini Gupta, Iuri Frosio, Dikpal Reddy, and Jan Kautz. Robust model-based 3D head pose estimation. In Proceedings of International Conference on Computer Vision (ICCV), 2015.

[17] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. IEEE Transactions on Multimedia (TMM), 17(11):2094– 2107, 2015.

[18] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Finegrained head pose estimation without keypoints. In Proceedings of International Conference on Computer Vision Woskshops, 2017.

[19] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[20] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. SSR-Net: A compact soft stagewise regression network for age estimation. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018.

[21] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[22] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In Proceedings of European Conference on Computer Vision (ECCV), 2018.

[23] Elad Hoffer, Nir Ailon. Deep metric learning using Triplet networks. Workshop contribution at ICLR, 2015.

[24] Yiwen Sun, Kun Fu, Zheng Wang, Changshui Zhang and Jieping Ye. Road Network Metric Learning for Estimated Time of Arrival.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in CVPR, 2015, pp. 815–823.