# Social Relation Trait Discovery from Visual LifeLog Data with Facial Multi-Attribute Framework

Tung Duy Dinh[1,2], Hieu Dinh Nguyen[1] and Minh-Triet Tran[1,2]
[1]*Faculty of Information Technology, University of Science – VNU-HCM, Vietnam*
[2]*Software Engineering Department, University of Science – VNU-HCM, Vietnam*

Keywords: Social Relation Trait, Visual LifeLog Data, Facial Multi-Attribute Framework.

Abstract: Social relation defines the status of interactions among individuals or groups. Although people's happiness is significantly affected by the quality of social relationships, there are few studies focusing on this aspect. This motivates us to propose a method to discover potential social relation traits, the interpersonal feelings between two people, from visual lifelog data to improve the status of our relationships. We propose Facial Multi-Attribute Framework (FMAF), a flexible network that can embed different sets of multiple pre-trained Facial Single-Attribute Networks to capture various facial features such as head pose, expression, age, and gender for social trait evaluation. We adopt the architecture of Inception-Resnet-V2 to each single attribute component to utilize the flexibility of Inception model and avoid the degradation problem with the residual module. We use a Siamese network with two FMAFs to evaluate social relation traits for two main persons in an image. Our experiment on the social relation trait dataset by Zhangpeng Zhang et.al shows that our method achieves the accuracy of 77.30%, which is 4.10% higher than the state-of-the-art result (73.20%). We also develop a prototype system integrated into Facebook to analyse and visualize the chronological changes in social traits between a user with friends in daily lives via uploaded photos and video clips.



## 1 INTRODUCTION

Within our daily activities, there is a large amount of information that can be collected for analysis to provide valuable findings to assist individuals(Gurrin, 2016), and to help us to better understand ourselves, our lives, and interpersonal relationships(Gurrin et al., 2014). For example, visual instance search (Salvador et al., 2016) can be used to find lost items or to verify certain events in daily recorded video from body worn cameras or smart glasses. Lifelog also provides materials for reminiscence of memorable moments, emotions, or locations (Nguyen et al., 2016) for social applications and interaction(Zhou et al., 2013).

Lifelog data can be recorded by different modalities, at different levels of details, and for various purposes(Gurrin et al., 2014). Visual data, such as photos and video clips, is one of the most important lifelog data sources with high availability. Photos can be taken and shared to social channels by an individual or his or her friends; video clips can be recorded from smartphones, body worn or surveillance cameras. Thus, in this paper, we analyse visual lifelog data to provide meaningful insightful person discovery: **Social relation traits** (Zhang et al., 2015).

Everyday, we interact with many people, building various social relationships, e.g. friends, business partners, lovers, etc. It is important to study the quality of those social relationships which have great influence on people's happiness and success.
This motivates our work to evaluate the social relation traits, the interpersonal feelings between two main people in a photo or a video clip captured from daily activities.

When considering two people hugging each other, our goal is not to predict their relationships, such as lovers or friends, but to determine the *traits of their social relation*, such as attached, assured, warm, trusting. Since there are many photo albums on various online image repositories such as Facebook, Flickr, Google Photos, our work can be used to help people with a better understanding of social traits from those abundant and scattered online visual data.

Although there exist many methods to analyse each independent facial attribute, such as head pose, facial expression, age, or gender, there are only few focusing on the combination of multiple facial attributes. In this paper, our key idea is that social



Dinh, T., Nguyen, D-H. and Tran, M-T.
Social Relation Trait Discovery from Visual LifeLog Data with Facial Multi-Attribute Framework.
DOI: 10.5220/0006749206650674
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 665-674
ISBN: 978-989-758-276-9

relation traits of two people can be evaluated from an appropriate set of single facial attributes of these two people. However, the suitable set of single facial attributes should be selected adaptive to different contexts, geographical regions, and cultures. For example, age prediction should be specialized for Asian people and Westerners. Therefore, we propose **Facial Multi-Attribute Framework** (FMAF), a general network model in which we can plug-in multiple pre-trained Facial Single-Attribute Networks.

Our proposed method is the generalization of the network proposed by Z. Zhang et. al. (Zhang et al., 2015) for social relation trait evaluation. We inherit the Siamese structure of the latter but replace its fixed internal structure using conventional DCNs with two FMAFs to utilize a dynamic set of pre-trained facial single attribute components in each FMAF. We use Inception-Resnet-V2 structure to train various single attribute components, such as age prediction, pose estimation, gender classification, or expression recognition. Specialized single facial attribute components can be created for specific purposes, such as age classification for Asian or Westerner people, or expression evaluation in different cultures.

By defining FMAF as a framework to integrate with different sets of facial single attribute components, we can analyse the heterogeneous contribution of each attribute in social trait evaluation in various cultures, contexts, and applications. Furthermore, due to the flexibility nature of this model, the contribution as well as the performance of each attribute could be further analyzed and utilized in future studies upon this broad subject. Moreover, additional attributes can be easily integrated into FMAF as an augmentation to boost the overall performance of the system.

We train four Facial Single-Attribute Networks to extract features for four facial attributes: head pose, facial expression, age, and gender from three datasets AFLW(Koestinger et al., 2011), Kaggle(Kaggle, 2017), and IMDB-WIKI(Rothe et al., 2016). Then we evaluate FMAF with different combinations of these four attributes on the social relation trait dataset by Zhangpeng Zhang et al. at ICCV 2015(Zhang et al., 2015). Our result on this dataset reaches 77.30%, which improves 4.10% from the published results in Zhangpeng Zhang et al. at ICCV 2015.

Using FAMF and our repository of pre-trained Facial Single Attribute Networks, we implement a prototype system integrated into Facebook to analyse the social relation traits of people in still images and video clips. We also create a diagram to visualize the progress of chronological changes in social relation traits of a user with others.

The remainder of this paper is organized as follows. Section 2 discusses backgrounds and related works. Then, our proposed method is presented in Section 3, including Facial Multi-Attribute Framework in Section 3.1 and Social Relation Trait Prediction Network in Section 3.2. Section 4 discusses results and data analysis as well as interpretation. We also briefly introduce the application for social relation trait analysis in Facebook in Section 5. Conclusion and future work are presented in Section 6.

# 2 BACKGROUND AND RELATED WORK

## 2.1 Lifelog Data and Applications

Lifelogging allows people to digitally record "*a totality of an individual's experiences*" (Dodge and Kitchin, 2007). Lifelog data can be passively collected during people's daily activities by various kinds of devices, including both wearable and environmental sensors(Gurrin et al., 2014). A lifelog may simple be a collection of photos or video clips shared by a user or his/her friends to a social network, or can be a huge data collected by inertial sensors from activity trackers or daily heart rates from smart watches.

Visual data is one of the most imporant types of lifelog data. It includes both photos and video clips recorded by individuals, stored personally or shared to online social channels. From a large collection of photos or video clips, taken by regular or body worn cameras, we can retrieve relevant shots/frames containing a certain item of interest using visual instance search (Salvador et al., 2016) or semantic query(Nguyen et al., 2017). Visual lifelog from social networks can be used for reminiscence in social applications, such as NowAndThen, a system that utilizes visual content similarity to assist reminiscence in sharing photos on social networks (Nguyen et al., 2016). In this paper, we continue the trend to exploit visual data in lifelogs for reflection on interpersonal relationship discovery.

## 2.2 Social Relation Traits

Many studies on social relation from visual data have been proposed, such as social group discovery(Ding and Yilmaz, 2010), social interacting class detection (Fathi et al., 2012), social role inference (Lan et al., 2012), and social relation traits evaluation(Zhang et al., 2015).

It is sophisticated to define social relation traits. In this paper, we consider social relation traits as the


Figure 1: The 1982 Interpersonal Circle (Kiesler, 1983) and some examples from dataset in (Zhang et al., 2015).

interpersonal feelings between two people in a particular context. Similar to (Zhang et al., 2015), we also adopt the definition of social relation traits from the interpersonal circle proposed by Kiesler (Kiesler, 1983), where interpersonal relations are divided into 16 parts as shown in Figure 1.

Since each part has its complement, we can group these 16 parts into 8 binary relation traits: (1) Dominant-Submissive, (2) Competitive-Deferent, (3) Trusting-Mistrusting, (4) Warm-Cold, (5) Friendly-Hostile, (6) Attached-Detached, (7) Demonstrative-Inhibited, and (8) Assured-Unassured. For simplicity, we use the name of the first trait in each group to represent that group. As illustrated in Figure 1 (right), these trait groups are non-exclusive so some of them can co-occur in an image.

The idea to employ multi facial attributes of a pair of persons in a photo to evaluate their social relation traits is first proposed by Z. Zhang et.al. (Zhang et al., 2015). Their work differs from previous studies as it is not based on single person but utilize attributes of both persons in a photo. Besides, while most of earlier works mainly exploit facial expressions(Valstar et al., 2011; Bettadapura, 2012; Lopes et al., 2016), Z. Zhang et.al. take advantages of multiple facial attributes, including expression, age, and gender.

Inspired by the work of Z.Zhang et.a., we also use multi-attributes from two main persons in a photo to evaluate their social relation traits. However, we make a generalization of their method by allowing a flexible set of single facial attributes to be plugged into the network. Thus, different facial attribute sets can be used to adapt to new contexts, cultures, and applications. Besides, we do not use spatial information (Zhang et al., 2015) but exploit only facial attributes for social relation trait evaluation.

### 2.3 Face Recognition and Attribute Processing

There are many learning models that can be used to recognize faces and extract facial attributes. Some architecture uses end-to-end learning for the task using a convolutional neural network (CNN) like VGG (Parkhi et al., 2015) while the other learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity (Schroff et al., 2015). CenterLose, which is a novel auxiliary supervision signal, is also considered to raise the performance of image classification and face recognition (Wen et al., 2016)

In this work, we utilize the structure of Inception-Resnet-V2 because it can overcome the problems of fixed size of filers and fixed number of layers as in DCN or VGGFace (Parkhi et al., 2015)

## 3 PROPOSE METHOD FOR SOCIAL RELATION TRAIT EVALUATION

Figure 2 illustrates the process overview of our system based on Siamese network architecture. This network inherits the idea of Z.Zhang et.al (Zhang et al., 2015) to use a Siamese network to extract facial features for two main persons, then predict their social relation traits. However, we replace the fixed network (DCN) with our proposed flexible Facial Multi-Attribute Framework to evaluate different sets of facial single attributes. Furthermore, as we aim to evaluate the effectiveness of FMAFs in social relation trait prediction, we do not use spatial information of the two faces as in (Zhang et al., 2015).

Our social relation trait prediction model, following Siamese network, consists of two parts:

- Two identical Facial Multi-Attribute Frameworks: to extract multi facial features from the two main persons in an image. The construction of a Facial Multi-Attribute Framework (FMAF) is presented in Section3.1. FMAF can be used as a container of multiple Facial Single-Attribute Networks, each of which extracts a specific facial feature from an input face. The output vector of FMAF is the concatenation of output feature vectors from its Facial Single-Attribute components.
- Social Relation Trait Prediction: to predict social relation traits from two facial multi-attribute fea-

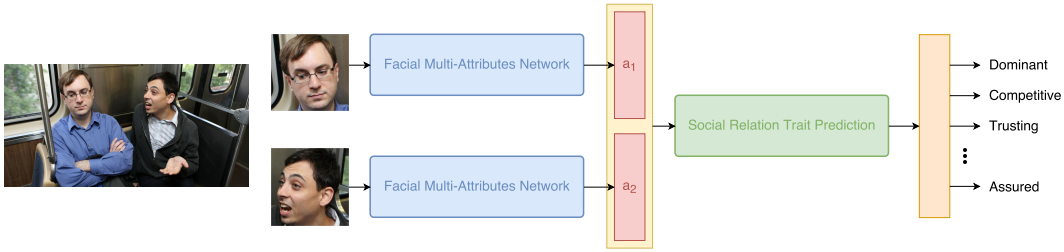


Figure 2: Propose Method.

tures of the two persons extracted from the input image (Section 3.2) . In our method, as we aim to evaluate the performance of facial attribute only, we do not use spatial information of faces as in (Zhang et al., 2015).

## 3.1 Facial Multi-Attribute Framework

Social relation traits can be determined from analyzing various facial attributes in visual contents. However, extracting and evaluating distinct properties immensely depends on specific situations. This motivates the authors to propose a generic approach for dynamically extracting numerous attributes to analyze their effectiveness (Section 3.1.2) instead of reusing conventional fixed deep networks for individual property evaluation from previous studies. Furthermore, each attribute extraction instance operates independently and can be easily embedded to a container network so that we can evaluate the accuracy of all facial attributes in the whole network to form a network bank with single forwarding operation (Section 3.1.1). For implementation, this particular approach utilizes the architecture of Inception-Resnet-V2 since its capability of overcoming problems regarding fixed size of filers as well as fixed number of layers (Szegedy et al., 2016) (section 3.1.3).

### 3.1.1 Facial Multi-Attribute Framework

The purpose of this network is to get facial attribute information. As mentioned above, we decompose the facial attributes into several types, e.g. pose, expression, etc., so that we can train the network for each of these types as a lightweight component independently. Hence, the architecture of our Facial Multi-Attribute Framework is built from multiple sub-networks called Facial Single-Attribute Network (see Section section 3.1.2).

Figure 3 illustrates how the model is deployed. The number of sub-networks is the number of attribute types we focus on. Each sub-network has the same structure based on Inception-Resnet-v2 architecture and runs independently with each other. An

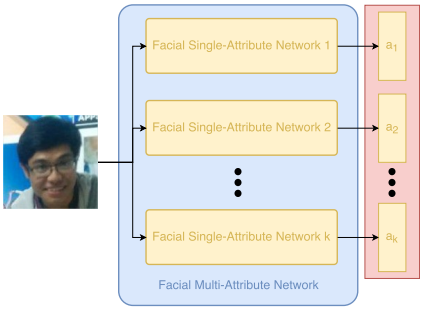


Figure 3: Facial Multi-Attribute Framework.

image is passed through all sub-networks, and each sub-network extracts the corresponding feature vector of its facial attribute information. For instance, pose attribute network extracts the pose information, while expression attribute network returns the expression information of the face. After that, we concatenate all the output vectors together. Let $k$ be the number of attribute types we use in the network, the length of the output vector is $1536 \times k$. This is all information about the face which we use as material in Section 3.2.

### 3.1.2 Facial Single-Attribute Network

The authors use Inception-Resnet-v2 architecture as a base main network to extract the facial single attribute information. The more specific about Inception-Resnet-v2 is introduced in section 3.1.3. Our study uses the "Dropout" layer whose output vector has the length of 1536. This vector is also the output vector of the Facial Single-Attribute Network. Figure 4 illustrates 2 different processes: (1) training and (2) feature extraction.

In the training process, the model is fed with a $299 \times 299$ color image which is defined in RGB format. The network is trained from scratch, which means that every layer in the model is re-calculated following the number of classes that the attribute type is classified. For example, the Pose Attribute has three classes (left, front, right) while the Expression has seven classes (angry, disgust, fear, happy, sad, surprise and neutral). We train the system in several steps

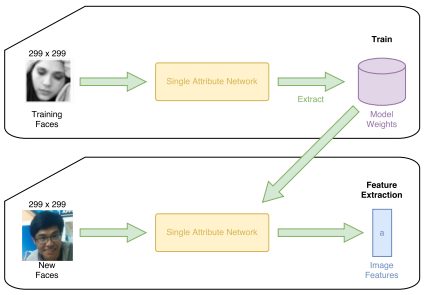


Figure 4: Facial Single-Attribute Network.

with multiple images. The number of steps we use when training is dependent on the facial attribute type (pose, expression, etc.). When the result seems acceptable, we extract the model weights so that they could be applied to the "using" process.

In the "using" process, the input size is still 299 × 299. We bind the weights in training process to the model to get the information. A face passing through sub-network will transfer into a sub-vector of values corresponding to its facial attribute type. The output vector, as we have mentioned above, is extracted with the length of 1536.

Furthermore, since there are approximately 20 attributes to be considered, teaching the machine to learning all of them simultaneously is not very efficient due to the complexity of the computation, the authors propose distributing this task into smaller ones. In other words, being able to create portable smaller version of the network by decomposing the overall work is quite beneficial for various lightweight component addition in the future. Furthermore, the ability of adapting to various specialized datasets due to the compact nature is another reason why this approach is chosen. Furthermore, only a few datasets containing enough properties or sufficient annotation, for example the AFLW dataset focuses on gender and pose while Kaggle only contains facial expressions, therefore it can be easily concluded that decomposing the tasks for specific purposes is much more efficient than generating the general result.

In addition, it is also feasible to evaluate the contribution to the overall accuracy of each group of attributes. Moreover, the contribution of each properties mainly depends on the type of the input data, therefore, the sensibility between a group of attributes and a specific category of data can be determined so that depending on the result, we could efficiently add the suitable type of input when classifying a desired properties.

### 3.1.3 Inception-ResNet-v2

There are many CNNs have been developed with a purpose that to increase the accuracy of the image recognition performance. In particularly, the authors focus on relatively low computational cost while still keep the good performance in accuracy. We have experiment many architecture to consider which one is satisfied our expectation. Thus, we choose Inception-ResNet-v2 architecture, whose main idea is to replace the filter concatenation stage of Inception architecture with residual connections, due to its outstanding performance.

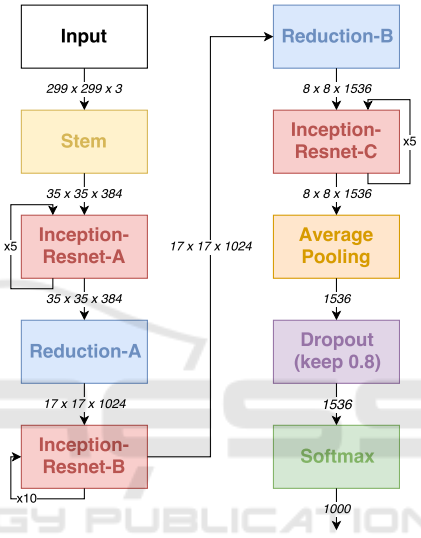


Figure 5: Schema for Inception-ResNet-v2 network.

Figure 5 just illustrates the components in Inception-ResNet-v2 network in the simple way. The main idea relating to the concept of shortcut connections which is skipping single or multiple layers. Particularly, the outcomes of the identity mapping computation are added to the outputs of the stacked layers. **Stem** is the input part of Inception-ResNet-v2 network. **Inception-resnet** is the residual inception block the replace the pure inception block in Inception network. There three types of block called A, B, and C whose number of repetition in the network is 5, 10, and 5, respectively. **Reduction** is reduction module with various number of filters.

## 3.2 Social Relation Trait Prediction Network

With the architecture proposed in Section 3.1.1, we can modify and choose the set of pre-trained attributes with the same generic approach. This method includes single forward operation which saves both time

and effort significantly. Furthermore, the results from our experiments show that the accuracy of only facial expression (75.90%) has proven to be superior than that of the full model with spatial cue (73.20%) proposed by Zhangpeng Zhang et al. (Szegedy et al., 2016).

From the image of two faces, the authors crop out two faces. Then, each face is fed through the Facial Multi-Attributes Network. For each face, the Facial Multi-Attributes Network returns the corresponding vector, $\vec{v}_1, \vec{v}_2$, consisting the information of the face. Next, the two vectors conducted from two faces are concatenated together to become new vector $\vec{v}$ with the length of $2 \times |\vec{v}_1|$. The concatenated vector is an input of a simple linear regression model to predict corresponding relation traits of the image.

The linear regression model has a formula:

$$\vec{g} = \vec{v} \times W + \vec{\varepsilon} \quad (1)$$

The weight matrix, $W$, projects the concatenated feature vectors to a space of shared representation, which is utilized to predict a set of relation traits, $\vec{g} = \{g_i\}_{i=1}^{g} \in \{0,1\}$. The weight matrix is initialized with zeros. During the training process, additive error random variable is distributed following a standard logistic distribution, $\vec{\varepsilon} \sim Logistic(0,1)$.

# 4 EXPERIMENT

## 4.1 Dataset

This section presents our method of training our own pre-trained model that can be used to recognize different facial attributes. Particularly, in order to extract head poses, facial expressions, gender, and age, the authors utilize data from three public datasets, including AFLW, Kaggle and WIKI.

### 4.1.1 Pose Attribute Extraction

AFLW is an abbreviation of Annotated Facial Landmarks in the Wild (Koestinger et al., 2011), which provides a large-scale collection of annotated face images gathered from Flickr, exhibiting a large variety in appearance such as pose, and ethnicity. AFLW also supply general imaging and environmental conditions.

The total size of dataset is 25,000 faces with different images size. However, this paper only focuses on human head poses. Because the dataset does not have any training or testing set so the authors take out 10% of images as testing set and the remaining as training set.

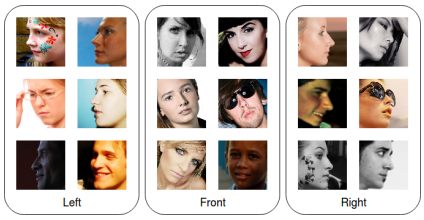


Figure 6: AFLW Dataset Sample (Koestinger et al., 2011).

### 4.1.2 Face Expression Extraction

This dataset originates from a Kaggle contest: "Challenges in Representation Learning: Facial Expression Recognition Challenge" (Kaggle, 2017). For convenience, the authors call it "Kaggle" in the whole paper. The dataset is designed for a facial expression classification problem whose purpose is strongly relevant that of our paper. In addition, the faces have been automatically registered so that the faces are evenly spread in each image.

Figure 7: Kaggle Dataset Sample.

The data consists of pixel gray-scale images of faces. Each image is labeled with a single expression that is belonged to seven categories: (1) angry, (2) disgust, (3) fear, (4) happy, (5) sad, (6) surprise, and (7) neutral. The training set consists of 28,709 examples while the size of the testing set is 7178. An example of the Kaggle dataset is shown in Figure 7

### 4.1.3 Gender & Age Extraction

This dataset originates from "Deep Expectation of Apparent Age From a single image" research (Rothe et al., 2016). The data set is built for predicting age from human face which play an important role in our research. The authors use only WIKI dataset which consists of 62328 images. An instance of this dataset is shown in Figure 8.

We classify the dataset into ten categories which is number from "under 10" (U10) to "under 100" (U100) group. Since the dataset does not have any training or testing set, we take out 10% of images as testing set and the remaining is training set.

Figure 8: WIKI Dataset Sample.

#### 4.1.4 Social Relation Trait Classification

Social Relation Dataset is a database designed for detecting social relations from a pair of faces. The dataset contains 8306 images of two faces chosen from web and movies. Each image is labeled with faces' bounding boxes and their pairwise relation. Five performing arts students label the relations independently so each label has five annotations.

In addition, due to that definition, each image can be labeled with multiple relation traits. Since some relation meanings are not clearly separated from each other, most images in dataset are collected from movies which annotators can base on the context to label more correctly. The truth value of relation is set to true if more than three annotations are consistent. Otherwise, its truth value is set to false. The dataset is divided into training and testing partitions of 7459 and 847 images, respectively.

### 4.2 Facial Multi-Attribute Framework Experiment

#### 4.2.1 Fine-tuning Pre-trained Model

The authors obtain the pre-trained Inception-Resnet-V2 model from TensorFlow homepage. The downloaded file is a model checkpoint that has been trained internally at Google. As we have mentioned above, Inception-Resnet-V2 model requires an input image size of $299 \times 299$.

When fine-tuning a model, we need to be careful about restoring checkpoint weights. In particular, when we fine-tuning a model on a new task with a different number of output labels, we are not able to restore the final Logits (Dropout) layer. When fine-tuning on a classification task using a different number of classes from the trained model, the new model will have a final "Logits" layer whose dimensions differ from the original pre-trained model from Google.

#### 4.2.2 Pre-training Full Model

As the authors introduce in section 3, Inception-Resnet-V2 model consists of multiple layers, which requires a considerable effort and rather long time to exploit all the advantages for achieving the best possible results. Furthermore, in order to train the whole network efficiently, we have to configure some important parameters that effect on our results: (1) batch size, (2) learning rate, and (3) learning rate decay type. Before going deeper, we need to understand the meanings of these parameters. Batch size is the number of samples in each batch (or called "training step"). Learning rate is a value determines how quickly the weights are adjusted. Learning rate decay type is to specify how the learning rate is decayed. For example, learning rate decay type is set to "exponential" means that the learning rate is updated based on exponential function. The default value for batch size, learning rate, and learning rate decay type is 16, 0.01, and "exponential", respectively.

In order to evaluate the best value for each parameter, the authors suggest the following scenario. The parameter we want to test is called "key" parameter. In this case, only the "key" parameter is changed, the other parameter is assigned to default value and each test is trained for 1000 steps.

First, we start with the batch size parameter. When using the default value, our testing machine is not able to build up a model because of the limited VRAM. Thus, we have to reduce batch size value from 16 to 8 for the machine to work properly. Table 1 shows the detail information of batch size test. Since batch size 16 is beyond our reach, we conduct an evaluation of two additional batch size 4 and 8. It is observed that even though the number of samples are halved from 8 to 4, the time required does not have the similar behavior, which the amount of total time needed for batch size 8 is approximately 1.76 times the amount of time for batch size 4. Because of the limitation of batch size, in two remaining test, the value of batch size is set to 8.

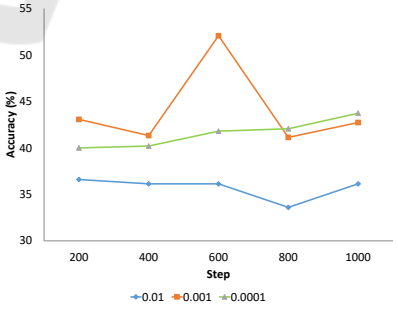


Figure 9: Test Result on Learning Rate Parameter with AFLW Dataset.

Next, the learning rate in Inception-Resnet-V2 model is set to 0.01 as default value. With this value, the accuracy does not increase smoothly. After testing with multiple values, the authors decide to take the number is 0.0001. Figure 9 and Figure 10 show the batch accuracy for each 200 steps to visualize how the

Table 1: Test Result on Batch Size.

| Batch size | Pose | | Expression | | Age | |
|---|---|---|---|---|---|---|
| | Accuracy | Time | Accuracy | Time | Accuracy | Time |
| 16 (default) | disable | disable | disable | disable | disable | disable |
| 8 | 36.60 | 828 | 17.35 | 819 | 35.51 | 823 |
| 4 | 30.27 | 469 | 15.33 | 471 | 29.67 | 472 |

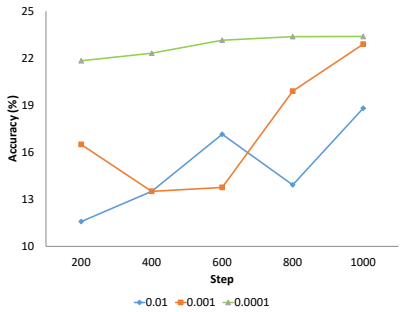


Figure 10: Test Result on Learning Rate Parameter with Kaggle Dataset.

learning rate changes. Observed from the test results on both datasets, higher learning rate tends to become unstable. In other words, the trend when learning rate is set to 0.01, represented by the blue lines, does not pose a stable increment. Same conclusion can be made on the learning rate of 0.001, represented by the red lines. The unpredictability of the two mentioned learning rates discourages the authors from training the model with such rate as the network goes deeper. On the contrary, when learning rate is set to 0.0001, the increasing trend is quite steady with the best outcome in both datasets. In addition, stable outcome is quite beneficial for concatenation step.

The learning rate decay type has 3 different types: (1) fixed – the learning rate does not change over time, (2) exponent – the learning rate change follow exponent function, and (3) polynomial – the learning rate change follow polynomial function. Exponent function is the best solution to control the change of learning rate, so that the result is converged.

Therefore, our parameter values in the paper are:

- Batch size: 8
- Learning rate: 0.0001
- Learning rate decay type: Exponent

## 4.3 Social Relation Trait Experiment

The model uses different test scenarios to evaluate the effectiveness of each attribute type. We conduct four different test cases: (1) model using only Pose Attribute , (2) model using only Expression Attribute, (3) model using only Age Attribute, and (4) model using all the aforementioned attributes.

Table 2 presents the accuracies achieved by our network as well the comparison between our results and those approximated from the chart of DCN model published by the Zhanpeng Zhang et al. in ICCV 2015(Zhang et al., 2015) on the same testing datasets. It can be observed that our pre-trained models for each attribute outperform their model in most of the relation traits as well as in average cases.

On average results, pre-training with head pose is the least accurate approach since the amount of information extracted from such attribute is quite limited and insufficient as most of the time people facing each other or looking at the same direction are usually friends or closely related. On the other hand, pre-trained model with facial expression or age or gender is more efficient as social relations are proven to be largely revealed by facial expressions along with age and gender.

While the accuracies achieved with pose, expression, and gender attributes independently only differentiate by a small percentage, the contribution of each property or a group of them towards a particular relation trait is clearly shown. In other words, age and expression attributes are most accurate when predicting “competitive”, “warm”, and “assured” while gender and age attributes possess the highest contribution when classifying “dominant”, “trusting”, and “friendly”.

Furthermore, our results also show superiority regarding predicting sophisticated attributes such as “dominant” and “assured” compared with those of the previous study as recognizing such relation traits has been proven to be quite problematic, especially the classification of “dominant” with the improvement in terms of accuracy of approximately 20% (from 60.00% to 79.82%).

The overall accuracy of 77.30% shows the efficiency of our facial multi-attributes network. In addition, our model also suggests the feasibility of improving the performance even more with the addition of supplementary attribute.

Table 2: Social Relation Trait Accuracy.

| | | Dominant | Competitive | Trusting | Warm | Friendly | Attached | Demonstrative | Assured | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| Pose | Proposed Method | 74.70 | 72.58 | 73.59 | 72.60 | 78.60 | 66.18 | 65.96 | 79.02 | 72.90 |
| | Z. Zhang et al. | 57.50 | 72.50 | 67.50 | 70.00 | 72.50 | 67.50 | 62.50 | 65.00 | 67.30 |
| Expression | Proposed Method | 76.81 | 75.51 | 76.42 | 75.68 | 82.52 | 68.68 | 68.38 | 82.83 | 75.90 |
| | Z. Zhang et al. | 60.00 | 75.00 | 72.50 | 75.00 | 75.00 | 72.50 | 67.50 | 67.50 | 70.60 |
| Age | Proposed Method | 78.92 | 76.54 | 77.74 | **76.38** | 83.70 | 68.90 | 68.65 | 84.07 | 76.90 |
| | Z. Zhang et al. | 57.50 | 72.50 | 70.00 | 70.00 | 72.50 | 67.50 | 65.00 | 60.00 | 66.80 |
| Gender | Proposed Method | 77.97 | 75.38 | 76.58 | 75.19 | 82.62 | 67.87 | 67.67 | 83.14 | 75.80 |
| | Z. Zhang et al. | 60.00 | 72.50 | 67.50 | 70.00 | 70.00 | 65.00 | 60.00 | 62.50 | 66.10 |
| All | Proposed Method | **79.82** | 76.83 | **77.96** | 76.27 | **84.83** | 68.60 | 68.52 | **85.70** | **77.30** |
| | Z. Zhang et al. | 60.00 | **77.50** | 72.50 | 75.00 | 77.50 | **72.50** | **70.00** | 72.50 | 72.50 |

## 5 SOCIAL RELATION TRAITS ANALYSIS ON ONLINE VISUAL CONTENTS

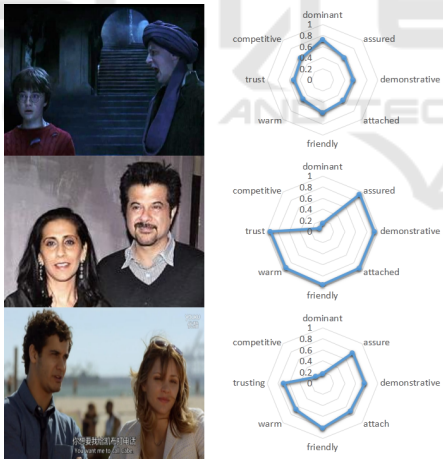


Figure 11: Social Relation Trait predicted sample images in Zhang's dataset.

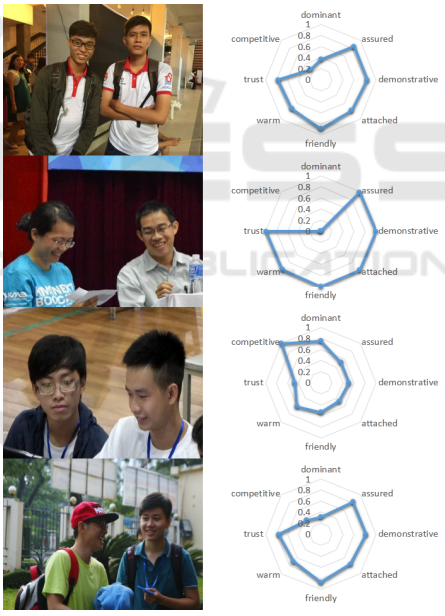


Figure 12: Social Relation Trait predicted image from Facebook.

Figure 11 represents the qualitative results of some images in the Zhang's dataset (Zhang et al., 2015). On the right side of the figure, 3 polar graphs represent the corresponding prediction for each image. In the first image, the context is that one man scream to the other so we can easily recognize the "dominant" trait is most representative. The value in the graph show the same result as we expect. There is a couple in the second image where most of positive traits can be seen clearly. The prediction also shares the same idea with through these positive traits are nearly 1. The context in last image is close to the middle one, but the man knitting his brown make the image less positive comparing to the one above.

Figure 12 illustrates the result of some random sample photos that we collected on Facebook. The first and last image are used to check-in places on Facebook where both of them seems neutral. Hence, the graphs are more rounded. In the second image, two people are in an open talk with relax smile. Thus,

the values of dominant and competitive are nealy 0. The third image, however, are taken in one competition so the expression of two students is quite serious. As a result, the value of competitive is the highest comparing to the others.

## 6 CONCLUSION

In this study, the authors show that applying our facial multi-attribute network can overcome the difficulties of predicting social relation traits from visual contents. Previous work, which mainly focuses on the conventional deep neural networks, only capable of producing generic results of all the relevant attributes without any evaluation upon the contribution of each attributes or groups of them to the overall performance. Our model has proven its availability towards different adjustments and feasible application to other lightweight systems that aim to specific purposes or specialized datasets. We will explore some feasible applications including background music recommendation for video based and photo collection clustering and visualization based on social traits.

## REFERENCES


Bettadapura, V. (2012). Face expression recognition and analysis: The state of the art. *Computer Vision and Pattern Recognition*.

Ding, L. and Yilmaz, A. (2010). *Learning Relations among Movie Characters: A Social Network Perspective*, pages 410–423. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dodge, M. and Kitchin, R. (2007). outlines of a world coming into existence: Pervasive computing and the ethics of forgetting. *Environment and Planning B: Planning and Design*, 34(3):431–445.

Fathi, A., Hodgins, J. K., and Rehg, J. M. (2012). Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233.

Gurrin, C. (2016). A guide to creating and managing lifelogs. In *ACM MM 2016 Tutorial, November 15, 2016, Amsterdam*.

Gurrin, C., Smeaton, A. F., and Doherty, A. R. (2014). Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125.

Kaggle (2017). Challenges in representation learning facial expression recognition challenge.

Kiesler, D. J. (1983). *The 1982 Interpersonal Circle: A taxonomy for complementarity in human transactions*. Psychological Review.

Koestinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*.

Lan, T., Sigal, L., and Mori, G. (2012). Social roles in hierarchical models for human activity recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1361.

Lopes, T., de Aguiar, E., Souza, A. F. D., and Oliveira-Santos, T. (2016). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*.

Nguyen, V., Le, K., Tran, M., and Fjeld, M. (2016). NowAndThen: a social network-based photo recommendation tool supporting reminiscence. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia, Rovaniemi, Finland, December 12-15, 2016*, pages 159–168.

Nguyen, V., Ngo, T. D., Le, D., Tran, M., Duong, D. A., and Satoh, S. (2017). Semantic extraction and object proposal for video search. In *MultiMedia Modeling - 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part II*, pages 475–479.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

Rothe, R., Timofte, R., and Gool, L. V. (2016). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*.

Salvador, A., i Nieto, X. G., Marqus, F., and Satoh, S. (2016). Faster R-CNN features for instance search. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 394–401.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Szegedy, C., Ioffe, S., and Vanhoucke, V. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. *Computer Vision and Pattern Recognition*.

Valstar, M. F., Jiang, B., and Mehu, M. (2011). The first facial expression recognition and analysis challenge. Automatic Face & Gesture Recognition and Workshops (FG 2011).

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). *A Discriminative Feature Learning Approach for Deep Face Recognition*, pages 499–515. Springer International Publishing, Cham.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2015). Learning social relation traits from face images. *2015 IEEE International Conference on Computer Vision*.

Zhou, L. M., Gurrin, C., Yang, C., and Qiu, Z. (2013). From lifelog to diary: a timeline view for memory reminiscence. In *Irish HCI conference 2013, Ireland, Dundalk*.