IMPLEMENTATION OF A CELEBRITY FACE RECOGNITION AI FOR VIDEO METADATA GENERATION

Yonggun Lee, Yoonjae Lee and Juhyun Oh

Korean Broadcasting System (KBS), Republic of Korea

ABSTRACT

In this paper, we introduce a celebrity face recognition AI for video metadata generation. Face recognition performance has shown significant improvement thanks to deep learning. We implemented a face recognition AI using our customized dataset composed of mostly Korean celebrity faces designed for the content analysis of KBS. Bothersome dataset labelling process was enhanced by using MTCNN face detection and face clustering. Inception-ResNet v1 model was used and test set accuracy was measured with respect to iterations. We compared our model with a commercial cloud-based celebrity recognition AI with which our celebrity database is thought to have about 26% in common. In the experiment, our model showed better performance in the precision.

INTRODUCTION

Artificial intelligence (AI) is getting more and more attention in the media industry, especially for video analysis and metadata generation. Among the possible metadata generated by AI, object labels and background information are relatively easy to acquire, since there are open datasets and pre-trained models. On the other hand, it is difficult to implement an AI engine to generate ‘face’ (or ‘person’) metadata, due to the lack of datasets and pre-trained models fit for purpose. Furthermore, face datasets are required to be constructed locally, i.e. usually for each country.

As a media corporation we decided to build a face dataset composed of about 3.6 million images of 6,690 subjects, focused especially on Korean celebrities. The celebrities were chosen from our content management system (CMS), in order of appearance counts in our contents. In order to efficiently construct the dataset and speed up the image labeling process, AI-based automation such as face detection and clustering is used.

We trained our dataset with Inception-ResNet v1 (5) as the backbone network and used softmax as the loss function. The proposed model is compared with a commercial celebrity recognition API provided as a cloud service. It is shown that the proposed model performs better in the experiments using our dataset.

KOREAN CELEBRITY DATASET

Although there are lots of open face datasets such as VGGFace2 (1), MS-Celeb-1M (2) and CASIA WebFace (3), we needed localized face dataset for our contents. As casts in our videos are mostly Koreans, we needed Korean celebrity dataset for extended video analysis. Our dataset is composed of about 3.6 million images of 6,690 identities, focused mainly on Korean celebrities. The celebrities were chosen from our content management system (CMS), in order of appearance counts in our contents. Announcers, actors, musicians, politicians, athletes and idol stars, etc., are included. We included celebrities who appeared more than once in our programs. Various poses, illumination, makeup and age diversity for each subject were encouraged for dataset generation. Our goal was to obtain dataset with label noise lower than 2%.

IMAGE LABELLING PROCESS ENHANCEMENT

Labelling process is extremely tiresome work. One has to check whether the person is labelled correctly image-by-image, and crop face image to the right size as shown in Figure 1. We used face detection and clustering method to enhance this labelling process.

First, we loosely crop face images using deep-learning-based face detection such as multi-task cascaded convolutional networks (MTCNN) (4). MTCNN is used due to its ability to detect some hard examples like partial occlusion and profile view pictures.

Second, we cluster each cropped face images using face clustering method. Chinese Whispers (6) algorithm was used for face clustering. Chinese Whispers clustering algorithm is chosen because it is fast due to linear property. Pre-trained model of an open face dataset was used for face clustering. The face clustering model turned to work quite well with the Korean faces as shown in Figure 2, although it was not trained for Korean faces. Face clustering accuracy was tested with main character clusters in dramas and resulted on average above 99%.

After face clustering, images are labelled by face clusters. Labelling face cluster is much more efficient than labelling single image one by one. The resulting face clusters are accurate enough, that one can get a bundle of dataset images by a single labelling action at most cases. Figure 2 shows face clustering result and we could get a bundle of dataset images of the target celebrity by labelling just one cluster.

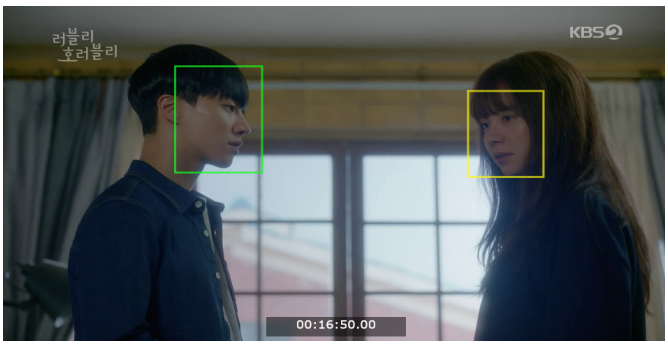

Figure 1 – Image-by-image face labelling process of the drama “Lovely Horribly” (2018)

Figure 2 – Face clustering result of actor *Haneul Kang* in the drama “When the Camellia Blooms” (2019)

CONVOLUTIONAL NEURAL NETWORK

Our backbone network based on Inception-ResNet v1(5) is trained using the dataset described above, with softmax classifier and cross-entropy loss function of [1], where L_i is *i*-th value of one-hot encoded label vector *L* and S_i is *i*-th value of softmax output vector *S*.

$$D(S,L) = -\sum_i L_i \log S_i \quad [1]$$

ADAM (7) is used for the optimization. We use the image size of 160x160. Each face is compactly represented by a 512-dimensional vector. The network is trained with a batch size of 90, on a single-GPU (Nvidia 1080 Ti) machine for 25 to 40 hours. As shown in Figure 3, there are sharp decreases in cross entropy at 100k and 200k iterations. The decrease in cross entropy loss drastically slows down after 21 hours of training. Additional training can still improve the accuracy.

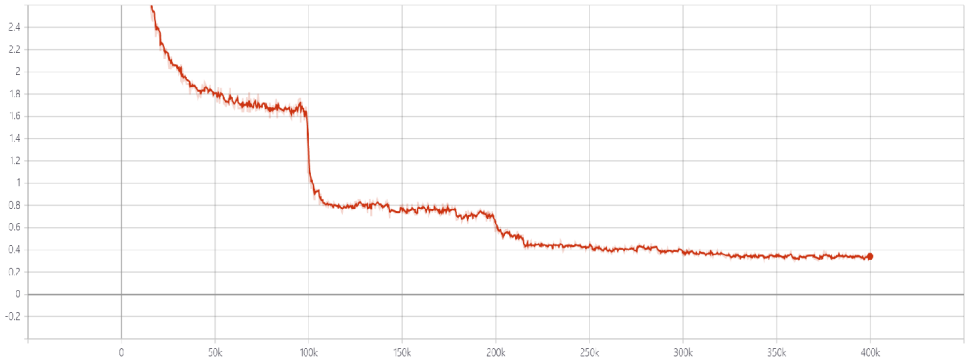

Figure 3 – Cross Entropy Loss Graph

EVALUATION RESULTS

For evaluation, the dataset was randomly shuffled and split into train and test set by 9 to 1. Test set is composed of 6,690 identities with 355,844 images. Test set accuracy was compared respect to the number of iterations as shown in Table 1. We trained the model by 250k to 400k iterations. The accuracy improved with the number of iterations and reached the maximum at 400k iterations. The highest accuracy was 96.6%.

Table 1 – Accuracy with respect to the number of iterations

Iterations	Accuracy
250,000	96.1%
275,000	96.2%
300,000	96.4%
325,000	96.5%
350,000	96.6%
400,000	96.6%

COMPARISON WITH A COMMERCIAL CLOUD SERVICE

Classification performance was compared between our model and a cloud celebrity recognition API service. We compared the cloud service's result with our test set label hand-by-hand. A celebrity was thought to be included in the cloud service, if cloud result includes at least one match to the target celebrity. From the result, about 26% of our dataset identities are thought to be included in the cloud celebrity database. Since the celebrity list of the commercial cloud service is not known to public, we estimate the intersection of the two celebrity lists by choosing only the meaningful results from the cloud celebrity recognition tests, when queried by our celebrity dataset images. In this experiment, both unknown and false returns were regarded as false.

First, our entire (6,690) celebrity test dataset was used for comparison. Our model's accuracy was 96.6% and cloud service was 19.2%. The cloud AI performed low because

the experiment was performed on our dataset, and many celebrities in our 6,690-celebrity dataset are not included in the cloud celebrity database. Moreover, commercial cloud AI services are targeted to global use, and the face recognition must have been run for a much greater number of global celebrities.

Second, 1,736 celebrity test subset which includes common identities of our celebrity database and cloud database, were used for comparison. Our model showed 97.2% and cloud service showed 48.5% accuracy. Our model showed slight improvement of 0.6% compared to 6,690 test set. Common 1,736 test set is likely to include more famous people and they have more data which results in better performance.

From the results, our model showed higher accuracy than a cloud celebrity recognition API. Although global commercial services can recognize a lot of Korean celebrities, still more celebrity data are required to analyze most Korean TV contents. Furthermore, by developing a celebrity AI engine in-house, we can quickly add or edit celebrities promptly to our needs.

Table 2 – Accuracy comparison with cloud service

Test set	Ours	Cloud Service
6,690 test set	96.6%	19.2%
1,736 test set (common)	97.2%	48.5%

APPLICATIONS

Celebrity face recognition AI can be used in many applications. Besides the general metadata generation for timeline search and analysis shown in Figure 4, we hope that it can be used for the enhancement of caption services shown in Figure 5, by analyzing the characters and showing them with the captions.

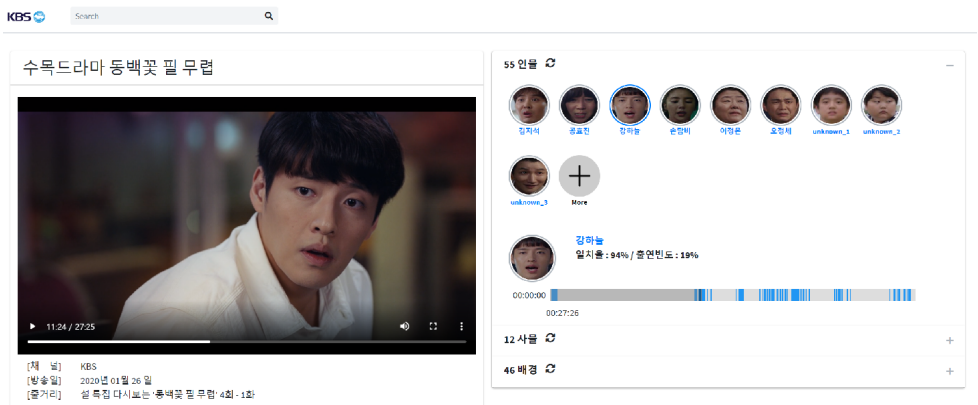

Figure 4 –Video Timeline Analysis

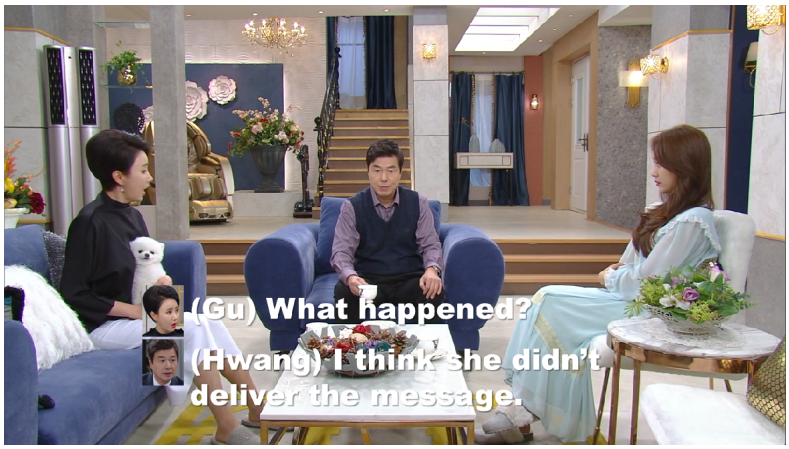

Figure 5 – Caption Enhancement

CONCLUSIONS

In this paper, an implementation of celebrity face recognition AI for video metadata generation is introduced. We constructed Korean celebrity dataset which is composed of about 3.6 million images of 6,690 identities. Image labelling process was enhanced using face detection and face clustering. Inception-ResNet v1 model was trained using ADAM. Train and test sets were split with 9 to 1 ratio. As iterations increased, accuracy went up and the highest accuracy 96.6% was obtained with 400k iterations. Our model was compared with a commercial cloud service and showed higher accuracy.

Commercial cloud-based services provide many useful AI functions with high accuracy these days. But from the results in this paper, we see the possibility that broadcasters perform better, by developing in-house AI models in some specific domains such as national celebrity recognition. We hope the proposed celebrity recognition can be used for the caption enhancement, as well as the general content search and analysis.

Our future research will focus on improving our AI engine and having rich metadata for our content archive to create a virtuous cycle, that will lead to sustainable AI-metadata improvements.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Science and ICT(MSIT, Korea) (No. 2019-0-00447, Development of emotional expression service to support hearing/visually impaired)

REFERENCES

1. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In FG, 2018.

2. Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. arXiv preprint arXiv:1607.08221, 2016.

3. D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.

4. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multi-task cascaded convolutional networks. arXiv preprint (2016). arXiv:1604.02878

5. C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. In ICLR Workshop, 2016.

6. C. Biemann. Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06, New York, USA, 2006.

7. Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980, 2014.