International Journal of Advanced Smart Convergence Vol.11 No.2 128-134 (2022)
http://dx.doi.org/10.7236/IJASC.2022.11.2.128

IJASC 22-2-19

Cloth Product Recognition based on Siamese Network with Body Region Extraction method

Sutanto Edward Budiman†, Edwin Kurniawan††, Seung Heon Lee†, Jae Seung Lee†, and Suk-Ho Lee†††

† *Researcher, Supersell Co. Ltd, Korea*
††*Researcher, Dept. Computer Engineering, Dongseo University, Korea*
†††*Professor, Dept. Computer Engineering, Dongseo University, Korea*
E-mail petrasuk@gmail.com

Abstract

Nowadays, people consume a lot of content such as web dramas or K-pop videos through mobile devices such as smartphones, and the market for indirect advertisements through these web dramas or K-pop videos is also increasing every year. In order to lead to the immediate purchase of indirect products in web dramas, a system that allows consumers to purchase immediately at the time the products appear in the drama is needed. In this paper, we propose a system to allow viewers to purchase products worn by celebrities immediately when viewers see and click on them. When a user clicks on a video, it recognizes the product worn by the celebrity, and displays information on the screen on the most similar product corresponding to the recognized product, allowing them to go to the seller's site where they can purchase it. In order for such a system to operate stably, a pose estimation and siamese network-based system is proposed. The proposed system will primarily be released as a streaming service in the form of an app or web page that connects the products in web dramas or other K-pop video contents screened on the mobile with e-commerce. Furthermore, in the future, the technology is expected to be used globally in various industries such as smart mobility and display kiosks.

Keywords: *Mobile advertising, Video content, Pose estimation, Deep Learning, Siamese Network*

1. Introduction

With the development of smart devices and the Internet environment, the proportion of viewing content through videos rather than existing text and images is rapidly increasing. The influence of radio, and TV, which are major channels of conventional mass media, is weakening, and recently, it is changing into a two-way medium that can communicate with viewers.

In Korea, video content can be distributed and viewed through the screen anytime, anywhere due to the common smartphone, and the mobile advertising market is rapidly growing. According to a recent Internet user behavior survey, 60% of users use YouTube, as a search channel, and users perceive YouTube as a

Manuscript Received: May. 3, 2022 / Revised: May. 7, 2022 / Accepted: May. 10, 2022
Corresponding Author: petrasuk@gmail.com
Tel: +82-51-320-1744, Fax: +82-51-327-8955
Professor, Department of Computer Engineering, General graduate school, Dongseo University, Korea

platform for information search as well as watching videos. In the domestic advertising market, the mobile advertising market is higher than PC and traditional mass media advertising. Currently, Korea's mobile shopping market is expected to grow to $15 billion (about 17 trillion won) while maintaining a high growth rate in the mid-20% range. This indicates that about three-quarters of Internet users purchase on mobile at least once a year. Korea has the third highest penetration rate of digital buyers in the Asia-Pacific region after Japan and Australia, and the online shopping market has also formed the third largest market after China and Japan.

When watching TV/web dramas and web video content, viewers often wonder what products the subject is wearing, and viewers often go through the process of re-searching the Internet to obtain product information or ask for product information through comments on the web. By addressing the title of the video content or the actor/actress's name, the name of the subject like 'Shin Hyeseoun Coat' without knowing the product name, become more famous than the brand names. However, in the case of the latest dramas and videos, information acquisition is often failed, and, in particular, it is difficult to obtain information on web dramas and web entertainment even though they are popular in Korea. Therefore, it takes a long time to acquire product information which naturally reduces the need for consumption.

In this paper, we aim to develop a service that allows viewers to easily check and purchase on items in the mobile shopping market, which is showing such explosive growth. Using an artificial intelligence-based image search function, we aim to develop a technology that provides information that viewers want to obtain through simple clicks when watching web dramas or K-pop videos, expecting that this technology can drastically reduce the cumbersome process of obtaining information. By doing so, we expect that this technology contributes to the rapidly growing global video content advertising market as a development type for video-based e-commerce.

2. Overall Diagram of the Proposed System

Figure 1 shows the overall diagram of the proposed system. When a scene in a drama is clicked in conjunction with the front end, the scene is extracted as a frame. The system receives the frame of the Web drama and inputs it into the object recognizing convolutional neural network(CNN). The CNN outputs several possible object regions, which are called 'blobs'. Meaningless blobs are eliminated and overlapping blobs are merged to result in meaningful object regions. The meaningful blobs are inputted into the cloud vision API so that the objects in the frame are labelled. Meanwhile, the meta data of the video is also extracted like the title of the web drama.

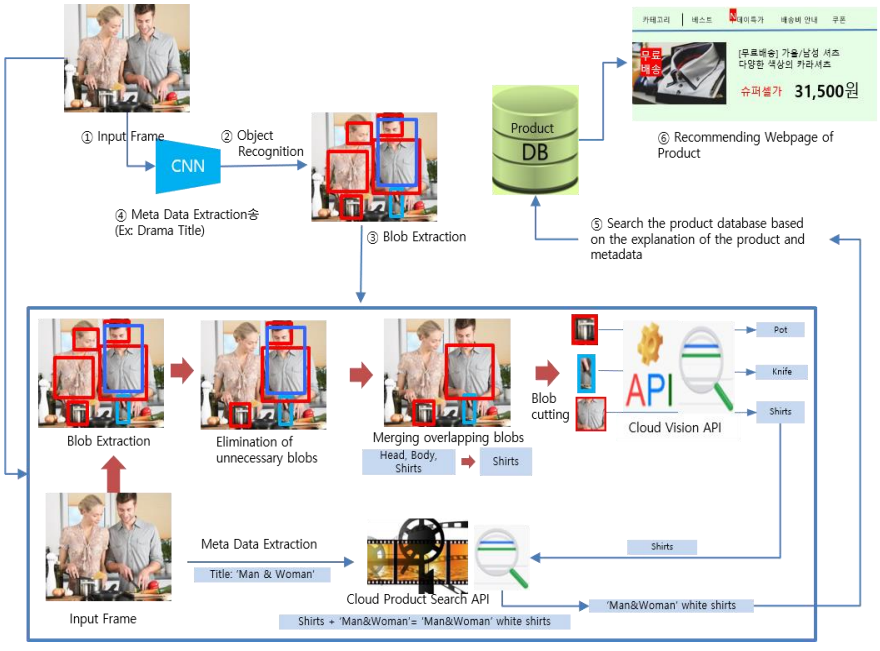


Figure 1. Overall diagram of the system

This meta data together with the label information from the cloud vision API is put into the cloud product search API which converts the input information into an object description text. For example, if the title of the drama is ‘Man & Woman’ and the object of a meaningful blob is a ‘shirts’ object, then the cloud product search API produces the product's unique nicknames, such as ‘Man & Woman shirts’ for further searching. This product’s unique nickname is then used as the keyword to search the product inside the product database which holds the seller’s webpage. If the keyword is matched the seller’s webpage is returned to the user.

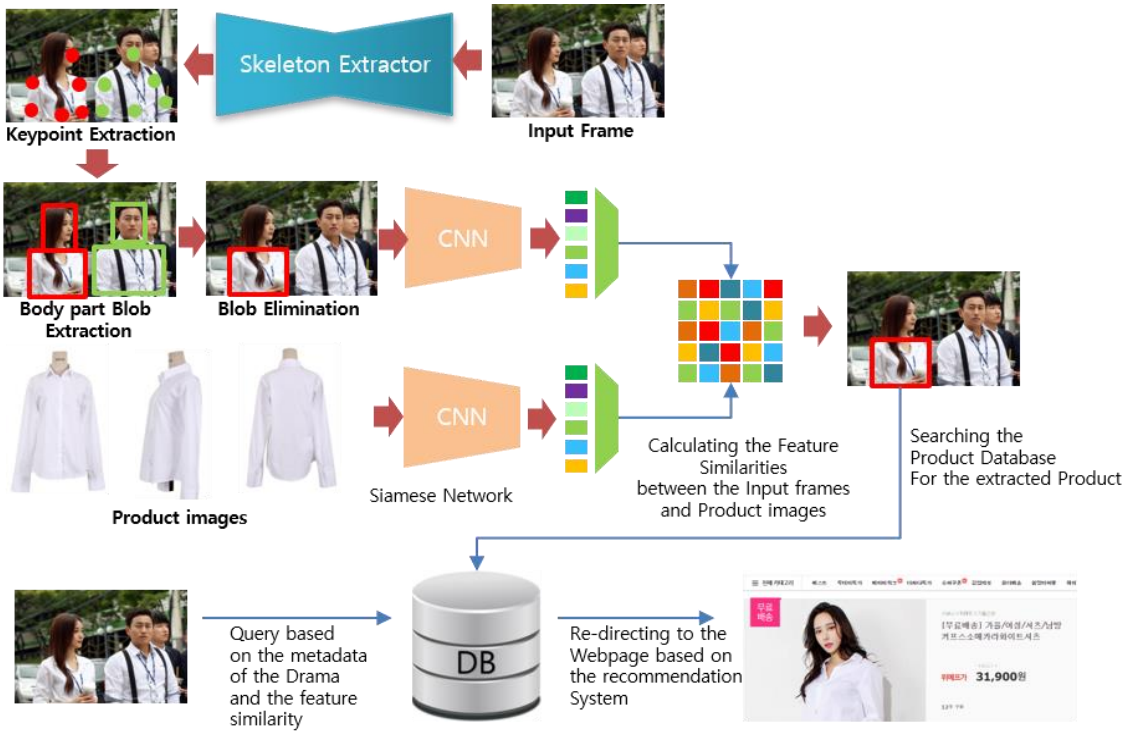


Figure 2. Product recommendation based on feature similarity

For a Web drama which is not widely known, the product cannot be searched using a unique nickname. In this case, we use a Siamese network to search for a product which has the highest feature similarity with the query image. Figure 2 shows the diagram that shows the case where we searched the product database based on the feature similarity between the product in the database and the product which the actor/actress is wearing in the web drama. In the next section, we explain in detail how the body part is extracted from the video for better extraction of the product that the actor/actress is wearing.

4. Details of the Body Extraction and Siamese Network based Comparison

In this section, we explain the details of our Siamese network based comparison. Siamese networks are a type of neural networks that are modeled by two identical network structures. Siamese networks are applied in tracking[1], face verification[2], and other applications. Siamese networks have a strong capability to measure similarities between two inputs.

We combine the Siamese network based comparison with a Human keypoints detection method, or commonly known as human pose estimation[3]. There are several method categories known in human pose

estimation techniques such as Regression based and Detection based methods. Regression based method translates the position of joints into set of variables information for human body model, while Detection based method detects from the images patch and produces heatmaps of each joint location which represents the human body keypoint. Detecting human body keypoints such as body-joints (left shoulder, right hip, etc) will help to extract the human region parts, like top part (shoulder to hip) and bottom part (hip to foot).

We use the pose estimation method as a kind of body region extraction to provide the Siamese network with more accurate region information. We assume that a person fashion is divided into two major parts, which are the upper and lower part of the body. Using the keypoints that are generated by the human pose estimation we divide the body region into the two parts, where keypoints values are x and y coordinate values that indicate the position of the keypoint in the image. We use the hips keypoint as the midpoint for dividing the upper part and lower part of the body. Furthermore, we take the y-coordinate that is the interpolated point of the y-coordinate of the nose and the y-coordinate of the highest position of the shoulder as the highest y-coordinate for the upper part and the highest y-coordinate of the hips as the lowest y-coordinate of the upper body region. The lowest y-coordinate or the upper body region becomes the highest y-coordinate of the lower body part, and the y-coordinate of the bottom of the feet becomes the lowest y-coordinate for the lower body part.

Using the extracted body region, we compare the contents in it with the Siamese network to measure the similarity between the clothes in the product database and the clothes that the K-pop celebrity wears with a similarity score α with a range of (0,1) and a defined threshold γ ($\gamma = 0.5$). If the similarity score α is more than threshold value γ, we assume that the images are similar. We then recommend every product which are similar to the query image to the user by showing the product on the display of the smartphone together with the products information and the homepage of the product. Figure 3 shows the diagram on the body region extraction and the Siamese network based product comparison workflow.

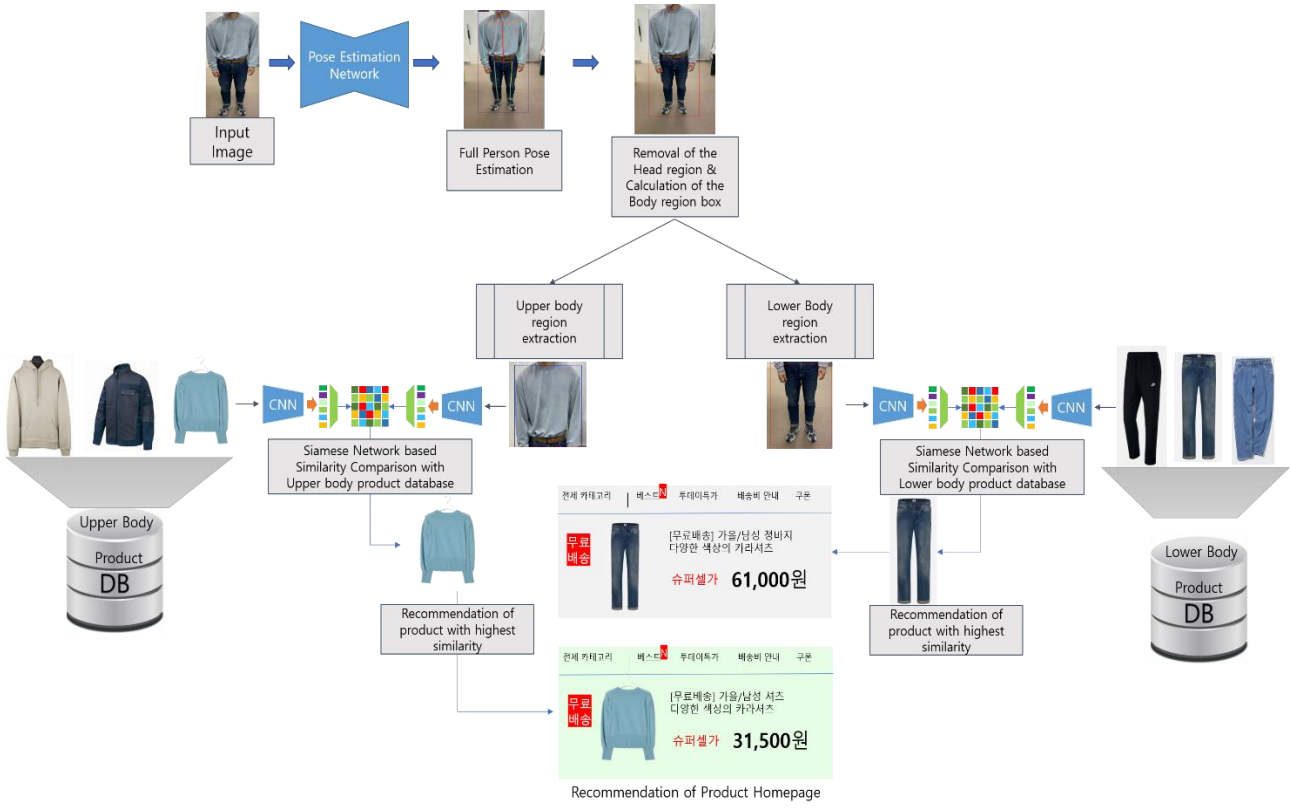
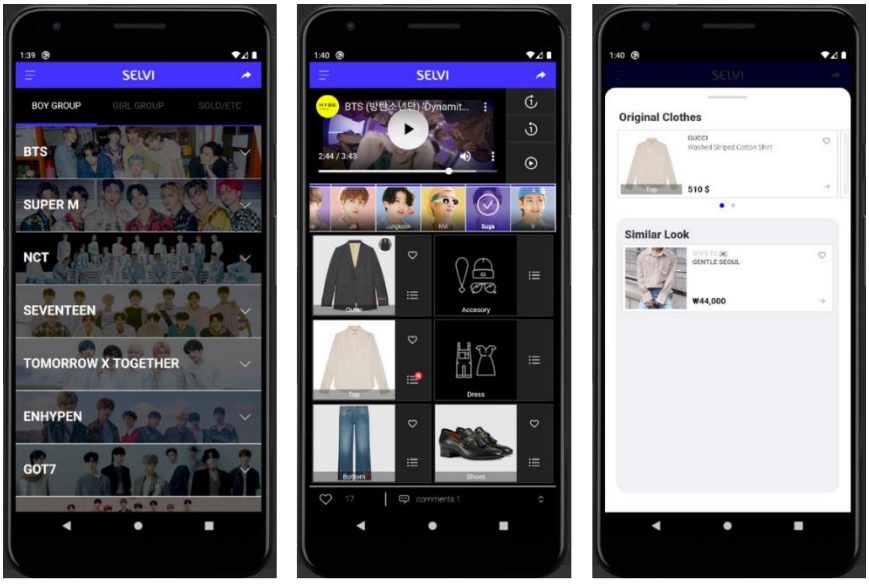


Figure 3. Diagram of the upper and lower body region extraction and the Siamese network based product comparison

5. Experimental Result

Figure 4 shows screenshots of the mobile application app we developed according the proposed product recommendation method. Figure 4(a) shows the app's starting page which shows a list of K-pop idol groups. When the user clicks a certain idol group, a list of music videos made by that group appears on the next page. Again when the user clicks on a certain video, the video is played on the screen until the user clicks on a certain scene of the video. If the user clicks on the video, the recommendation method described in this paper begins to work in the background and the upper body products or lower products which are closest to those the celebrity is wearing appear on the screen as shown in Fig. 4(b). At this time the user can click on the product he/shc is most interested in, which will lead him/her to the product's homepage as seen in Fig. 4(c).



(a) (b) (c)

Figure 4. Screenshots of the mobile application based on the proposed image based recommendation method.

We compare in Table 1, the average IoU(Intersection over Union) criterion values of 10 different video clips of Idol's music videos to show that the upper and lower body extraction method results in higher average IoU values. A higher IoU value implies a better detection of the product that the celebrity is wearing and therefore, a better performance of checking the similarity.

Table 1. Comparison of the average IoU values between the upper and lower body region extraction method and the full body method.

Average IoU	clip1	clip2	clip3	clip4	clip5	clip6	clip7	clip8	clip9	clip10
Body extraction method(53.9%)	56.2	52.7	49.0	48.9	35.7	35.3	62.7	76.2	60.3	61.9
Full body method(31.2%)	37.1	25.2	27.9	18.5	21.8	26.7	42.4	36.9	37.8	37.3

Figure 5 shows the success rate of recognition with different criterion of the IoU value. Again, it can be observed that the full body method shows a recognition rate of 100% only with IoU value lower than 0.1. In comparison, the body extraction method achieves a recognition rate is 100% with IoU values lower than 0.2, and a recognition rate higher than 80% with an IoU value as high as 0.4.

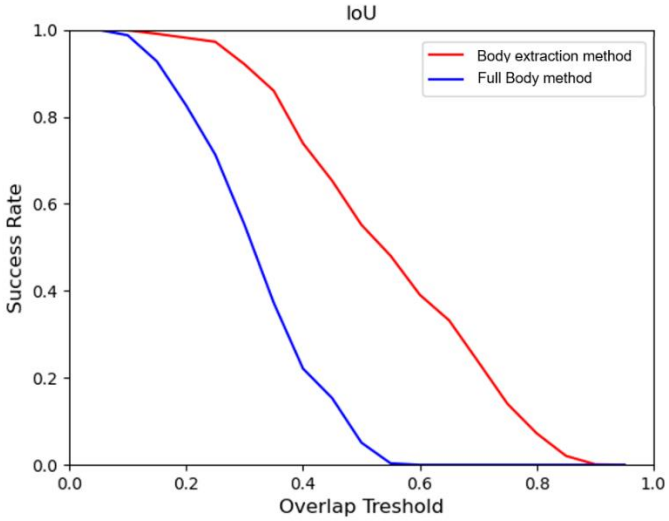


Figure 5. Graph of Success rate versus IoU value of the Body extraction method and the full body method.

6. Conclusion

In this paper, we proposed an image based recommendation system of K-pop and Web drama celebrities which is based on the combined use of the pose estimation network and the Siamese network. The proposed system suggested the products at the time the viewer clicks on the screen of a playing K-pop video or web drama and provides the information that best matches the product through image search on a pop-up or another window screen. By utilizing the Human keypoints to divide the whole body into the upper part and lower part of the body, both the IoU value and the recognition rate of the products that the K-pop artist is wearing have increased compared with mere product detection based methods. The proposed system will be released as a streaming service in the form of an app or web page that connects the products appearing in web dramas or K-pop videos, mostly screened on mobile, to e-commerce. Research to expand and apply this proposed technology to various industries such as smart mobility and display kiosks is expected to be one of the future research topics.

Acknowledgement

This work was supported by the Technology development Program(S2840023) funded by the Ministry of SMEs and Startups(MSS, Korea).

References

[1] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.H.S. Torr, “Fully-Convolutional Siamese Networks for Object Tracking,” in *European Conference on Computer Vision (ECCV) 2016*, Oct. 8-16, 2016. DOI: https://doi.org/10.1007/978-3-319-48881-3_56

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the Gap to Human-Level Performance in Face Verification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1701-1708, Jun. 23-28, 2014. DOI: https://doi.org/10.1109/CVPR.2014.220

[3] Y. Cheng, Y. Tian, and M. He, “Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods,” *Computer Vision and Image Understanding (CVIU)*, Vol. 192, 102897, March 2020. DOI: https://doi.org/10.1016/j.cviu.2019.102897

[4] Y. Wu, A. Kirillov, F. Massa, W.Y. Lo, and R.Girshick, "Detectron2," *https://github.com/facebookresearch/detectron2*

[5] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, pp.1137–1149, June, 2017. DOI: https://doi.org/10.1109/TPAMI.2016.2577031