Makeup Instructional Video Dataset for Fine-grained Dense Video Captioning

Xiaozhu Lin
School of Information
Renmin University of China
linxz@ruc.edu.cn

Qin Jin
School of Information
Renmin University of China
qjin@ruc.edu.cn

Shizhe Chen
School of Information
Renmin University of China
cszhe1@ruc.edu.cn

Abstract

Automatic analysis, understanding and learning from long videos remain very challenging and request more exploration. To support investigation for this challenge, we introduce a large-scale makeup instructional video dataset named iMakeup. This dataset contains 2000 videos, amounting to 256 hours, with 12,823 annotated clips in total. This dataset contains both visual and auditory modalities with a large coverage and diversity in the specific makeup domain, which is expected to support research works in various problems such as video segmentation, video dense captioning, object detection and tracking, action tracking, learning for fashion, etc.

1. iMakeup Dataset

Automatically describing images or videos with natural language sentences has received significant attention in recent years [2]. The increasing availability of large-scale image or video datasets [5][7][4] is one of the key supporting factors to the rapid progress on the challenging captioning problems. While using a single sentence cannot well recognize or articulate numerous details within long videos, like user-uploaded instructional videos of complex tasks on the internet. Hence, challenging tasks such as dense video captioning [2][8], which aims to simultaneously describe all detected contents within a long video with multiple natural language sentences, have attracted increasing attentions.

Given that few large-scale long video datasets are available for this task, we collect a large-scale instructional video dataset in the specific makeup domain, which is named iMakeup. Makeup tutorials are popular on commercial website such as Youtube which people rely on to learn how to do makeup. In such a tutorial video, the makeup artist or vlogger is always in the viewfinder and the camera is focusing on her/his face. Also, makeup sometimes requires very small, precise movements, which makes detection and tracking fine-grained actions challenging. Makeup involves explicit steps and different cosmetics used in each step, which makes it intriguing to investigate automatic techniques for procedure learning, fine-grained object detection, and dense captioning tasks. To the best of our knowledge, this dataset is the first large-scale long video dataset in makeup domain with both temporal boundaries and manual caption annotation for video segments.

1.1. Collection and Annotation

We used category "makeup" on WikiHow [6] to obtain the most popular queries that the internet users used in makeup domain. We then discarded repetitive or extraneous queries, which leads to 50 popular queries in makeup domain. With each query, we crawled YouTube and obtained the top 40 videos. Each video contains 2-20 procedure steps. We therefore target at creating annotations of temporal boundaries for each step and text descriptions of the procedure for each step. An annotated example is shown in Figure 1. For each raw video, annotators are asked to segment the whole video into clips according to the makeup procedure and annotate the start time, end time and an English sentence caption of each clip.

1.2. Dataset Statistics

The dataset contains 2000 makeup instructional videos from 50 most popular makeup topics, with 40 videos for each topic. The total video length is about 256 hours with an average duration of 7.68 min per video. There are 12,823 annotated clips in total. All video clips are temporally localized and described in complete English sentences. The average length of annotated sentences is 11.29 words. The total vocabulary size is around 2183 words.

Actions: The most frequent action word used in captions is "apply". Some specific actions like "pad", "dab", "brush", and "define" occur in less videos. Since the distribution of action vocabulary is quite biased, we then consider "Verb+Object" pairs as fine-grained actions in subsequent work. Common actions are shown in Figure 2.

Cosmetics: They are commonly occur in makeup videos as action objects (apply **mascara**) or action adverbial (define lips using **lipstick**). They pose challenges for

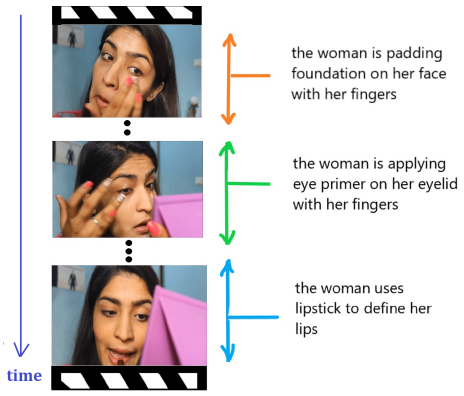

Figure 1. An annotation example of iMakeup dataset.

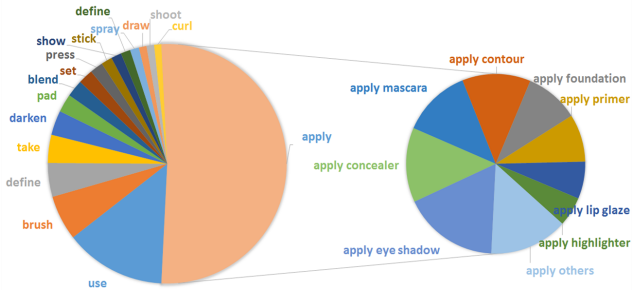

Figure 2. Common actions in iMakeup dataset.

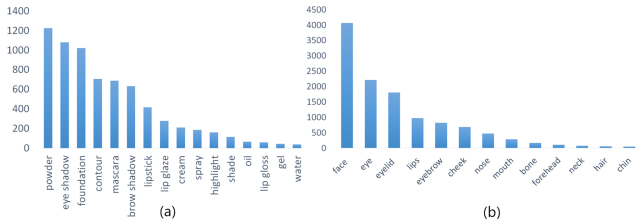

Figure 3. Common cosmetics and facial landmarks in iMakeup dataset. (a) the cosmetics, (b) the facial landmarks.

fine-grained object and action detection techniques. The commonly-used cosmetics are shown in Figure 3 (a).

Facial landmarks: To achieve fine-grained dense video captioning, the models should be able to recognize the facial landmark for detailed description. Hence the facial landmark annotation is also important. Frequent facial landmarks are shown in Figure 3 (b).

Cosmetic Applicators: Appropriate cosmetic applicators are essential for perfect application or blending of various cosmetics. Hence we emphasize this part in annotation, as well. Frequently occured applicators are "brush", "beauty blender", "sponge", "puff", etc.

Cosmetic Brands: A small portion of video annotations mentioned the cosmetic brands. For example, we can find that some cosmetics like "Estee Lauder Double Wear Foundation", "Maybelline Eraser Concealer", "Nyx Setting Spray", etc. are highly in common use. With more annotations, these can help create a knowledge base for future makeup products and facial style recommendation.

Table 1. Comparisons of large-scale video datasets. We collect their duration by hour. FAnn. is short for Fine-grained Annotation.

Name	Duration	Domain	Videos	FAnn.
YouCook [1]	2.3	Cooking	88	No
MPII-MD [4]	73.6	Movie	94	Yes
TACoS [3]	-	Cooking	123	No
YouCookII [8]	176	Cooking	2000	Yes
iMakeup	256	Makeup	2000	Yes

1.3. Comparison

We compare our dataset with several popular large-scale video datasets in Table 1.2. iMakeup is a brand-new domain-specific large-scale long video dataset with detailed annotations, which can support tasks of learning complicated information or intelligence from long videos, such as temporal action proposal, dense video captioning, etc.

References

[1] P. Das, C. Xu, R. F. Doell, and **J. J.. Corso**. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, page 6, 2017.

[3] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36, 2013.

[4] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[6] WikiHow. how to do anything. `http://www.wikihow.com`.

[7] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. *arXiv preprint arXiv: 1703.09788*, 2018.