반도체디스플레이기술학회지 제20권 제3호(2021년 9월)
Journal of the Semiconductor & Display Technology, Vol. 20, No. 3. September 2021.

Improved Inference for Human Attribute Recognition using Historical Video Frames

Hoang Van Ha[*], Jong Weon Lee[*] and Chun-Su Park[†]**

[*]Department of Software, Sejong University,
[**†]Department of Computer Education, Sungkyunkwan University

ABSTRACT

Recently, human attribute recognition (HAR) attracts a lot of attention due to its wide application in video surveillance systems. Recent deep-learning-based solutions for HAR require time-consuming training processes. In this paper, we propose a post-processing technique that utilizes the historical video frames to improve prediction results without invoking re-training or modifying existing deep-learning-based classifiers. Experiment results on a large-scale benchmark dataset show the effectiveness of our proposed method.

Key Words : Computer Vision, Human Attribute Recognition, Pedestrian Attribute Recognition, Deep Learning, Soft Biometrics

1. Introduction

In surveillance scenarios, soft biometrics [1] such as age, gender, hairstyle, cloth color, etc. are usually available and can be used for recognizing individuals. Human attribute recognition (HAR) is a task that focuses on identifying people's attributes (types of soft bio-metrics) from their corresponding images (see Fig. 1 below).

Fig. 1. An example of human attribute recognition.

Recent HAR approaches [2-5] using deep convolutional neural networks (CNN) to achieved state-of-the-arts results, but they have the drawbacks of time-consuming training processes [6] for obtaining deep models. If trained deep models are available, it is beneficial to develop an approach to improve the HAR performance without retraining or modifying these trained models. Besides, most previous HAR work uses a single image, while a consecutive frame sequence can produce better HAR performance since it provides more information and is less affected by occlusions (see Fig. 2).

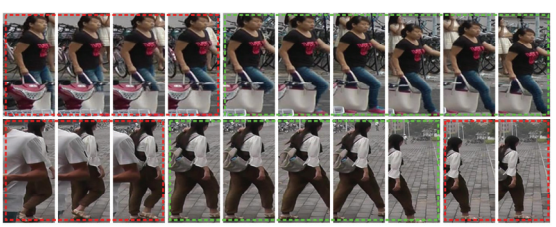

Fig. 2. Two 10-frame sequences from corresponding tracklets in MARS [7]. Due to occlusions, human attributes (top: carrying bag/bottom: backpack) in some frames (green boxes) can be easier to be recognized than in other frames (red boxes).

[†]E-mail: cspk@skku.edu

In this paper, we propose a new universal post-processing method for HAR that exploits the historical video frames for better HAR predictions. Notably, the proposed method does not require classifier modification and retraining process. We show the experimental results on a large-scale pedestrian video dataset to prove the benefits of our method.

The rest of this paper is organized as follows. Section 2 introduces related work on HAR and multi-frame fusion methods. We describe the technical details of our post-processing method in Section 3. The experimental settings and evaluation results are presented in Section 4, which is followed by conclusions in Section 5.

2. Related work

2.1 Human attribute recognition (HAR)

Recently, with the superiority of deep learning, most researchers [2-5] leverage the power of deep CNN for HAR goals. While most existing work [2, 3] on HAR is based on a single image, some recent work [4, 5] develops video-based algorithms to make use of both spatial and temporal information in the video sequence. In [4], the authors extract deep features of video sequences for training an attribute classifier with a temporal-attention strategy. The authors in [5] propose a deep network using Multiple Time steps Attention mechanism and Focal Balance Loss to address attribute imbalance issues.

2.2 Multi-Frame Fusion

Multiple frames in videos can be used for better recognition. The authors in [8] do emotion recognition by applying a multi-frame post-integration approach with choosing the maximum recognition value over sequences of frames for final recognized emotion. In [9], the authors achieve better face classification by using a multi-frame fusion method on video frames of a head rotating in a range of angles.

3. Proposed Method

Fig. 3 illustrates an overview of our method. Our key idea is to leverage the clues from frame sequences in tracklets to yield the final predictions as in multi-frame fusion. Specifically, instead of computing the attribute probabilities for a single frame F, we generate such attribute probabilities for frame F and h previous frames of F in the same tracklet. The final probability value of a single attribute α in frame F will be the result of combining the probability values of α in F and in its previous frames.

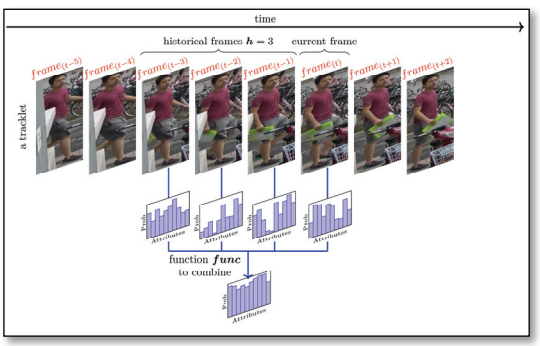

Fig. 3. Overview of the proposed method.

input :

- An test dataset $\mathcal{D}$ contains a number of K tracklet: $TL = \{tl_1, tl_2, ..., tl_K\}$; each tracklet tl_i is an image sequence of moving individuals in a short period.
- An trained HAR model $\mathcal{M}$ where $prob_{img} = \mathcal{M}(img) = [prob_{attr1}, prob_{attr2}, ..., prob_{attrL}]$ is the vector representing the probability attribute prediction for the image img, and $prob_{attr<j>}$ is the probability for attribute j in total L attributes.
- h: number of historical frames.
- $func$: a mathematical function taking a set of real numbers as input and return a single real value as output.

output:

- An result array $\mathcal{R}$ consisting of revised HAR predictions for all images in $\mathcal{D}$.

Initialization:
set values for h, $func$;
result array $\mathcal{R} = \emptyset$;
foreach *tracklet* $tl \in TL$ **do**
  tracklet prediction array $\mathcal{D} = \emptyset$;
  foreach *image* $img \in tl$ **do**
    $prob_{img} = \mathcal{M}(img)$;
    $\mathcal{D} = \mathcal{D} \cup prob_{img}$;
    $len = numberOfPredictionVector(\mathcal{D})$;
    $startPos = max(1, len - h)$;
    $matrix\ \mathcal{P} = \left[\mathcal{D}[startPos], \mathcal{D}[startPos + 1], ..., \mathcal{D}[len]\right]$;
    /* where $\mathcal{D}[x]$ is the x^{th} prediction vector in $\mathcal{D}$ as a row in the matrix $\mathcal{P}$ */
    $revised\_prob_{img} = [revised\_prob_{attr1}, ..., revised\_prob_{attrL}]$;
    /* where each value $revised\_prob_{attr<j>} = func(\mathcal{P}[j])$ with $\mathcal{P}[j]$ is the operator returning values on colum j^{th} of $\mathcal{P}$ */
    $\mathcal{R} = \mathcal{R} \cup revised\_prob_{img}$;
  end
end
return *result array* $\mathcal{R}$

Fig. 4. Post-processing method for Human Attribute Recognition.

The formal description of the proposed approach is given in Fig. 4. Notably, some video frames at the beginning of the tracklet cannot have enough h historical frames; therefore, k $(0 \leq k < h)$ previous frames are used instead. It is intuitive that a predicted probability of attribute can be either the average or the max of probability values of that attribute in concerned video frames. Thus, we propose to use the combination function $func \in \{max(), avg()\}$.

4. Experiments

4.1 Datasets, Metrics, Baselines:

Datasets: MARS [7], a large-scale pedestrian dataset, is used in the evaluation process. We test our method on the MARS test set, which contains more than 681k images belonging to 12,180 tracklets of 636 unique people. For more accurate labels, we also use the re-annotated MARS Attributes dataset [4].

Baseline: We simply use a free trained deep model [10] on Github as a baseline. We compare performance results between the trained model (baseline) and the proposed method using that model with different settings of h (number of historical frames) and $func$ (combination function). Each setting is named by following rule: $func\_h < No.ofHistoricalFrame >$. For example, setting max-h5 is the proposed method with $func = max()$ and $h = 5$ historical frames, while avg-h10 setting uses $h = 10$ historical frames, and$func = avg()$, i.e. average().

Metrics: We adopted the five most well-known HAR metrics as in [11] for evaluation: one label-level metric mean accuracy (mA) and four sample-level metrics, including accuracy (Acc.), precision (Pre.), recall (Rec.), and F1-score (F1) [12, 13].

4.2 Evaluation Results:

Table 1 below provides the evaluation results of the baseline and the proposed method with different settings.

Precisely, we test the our method with different combinations of $h \in \{1, 5, 10, 15, 20, 25, 30\}$ and $func \in \{max(), avg()\}$.

The data above shows that our proposed method using $func = max()$ and a suitable value of h can help to improve the metric values of mA and recall considerably, e.g., max-h15 outperforms baseline by 2.3% in mA and 5.8% in recall. However, using $max()$ function setting causes significant decreases in precision and only slightly improves accuracy and F1 in some cases compared to baseline. Moreover, the number of historical frames h can affect the results substantially as plotted in Fig. 5 below.

Table 1. Evaluation results of baseline and our methods with different settings of $func$ and h.

Method	mA	Acc.	Prec.	Rec.	F1
baseline	0.8392	0.8322	0.9309	0.8703	0.8996
max-h1	0.8479	0.8399	0.9215	0.8878	0.9044
max-h5	0.8582	0.8426	0.9026	0.9119	0.9072
max-h10	0.8614	0.8381	0.8889	0.9229	0.9056
max-h15	**0.8622**	0.8330	0.8795	0.9283	0.9032
max-h20	0.8621	0.8285	0.8724	0.9315	0.9010
max-h25	0.8617	0.8246	0.8667	0.9336	0.8989
max-h30	0.8612	0.8212	0.8621	**0.9351**	0.8971
avg-h1	0.8405	0.8356	0.9335	0.8723	0.9019
avg-h5	0.8424	0.8399	0.9369	0.8749	0.9048
avg-h10	0.8439	0.8429	0.9393	0.8768	0.9070
avg-h15	0.8449	0.8450	0.9408	0.8782	0.9084
avg-h20	0.8457	0.8464	0.9418	0.8791	0.9094
avg-h25	0.8463	0.8474	0.9425	0.8798	0.9101
avg-h30	0.8467	**0.8482**	**0.9430**	0.8803	**0.9106**

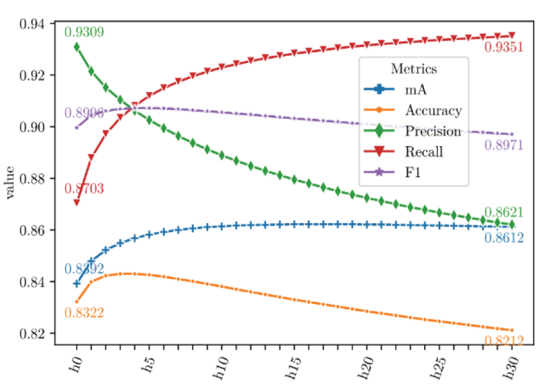

Fig. 5. Performance values when using $func = max()$ with different number of historical frames h. Note: $h0 = baseline$. Best viewed in color.

Our method with $func = avg()$ produces better results in all metrics compared to the baseline. Using $func = avg()$ consistently improves all metric values compared to inconsistent results obtained by using $func = max()$. Fig. 6 shows the correlation between the number of historical frames $h \in [0, 60]$ and the metric values of our method with $func = avg()$. The graph in Fig. 6 indicates that when using $func = avg()$, the higher number of historical frames h is, the better performance we get. However, the improvement slows down considerably as we increase the value of h.

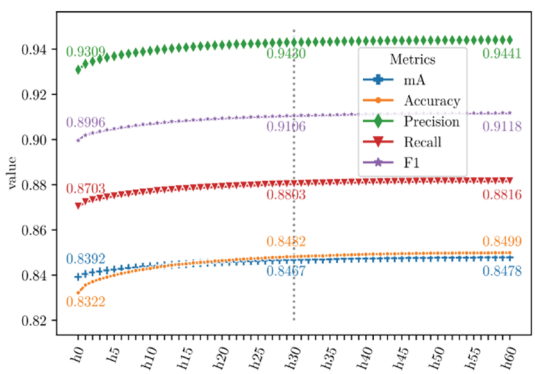

Fig. 6. Performance values when using $func = max()$ with different number of historical frames h. Note: $h0 = baseline$. Best viewed in color.

5. Conclusions

We present a new post-processing method for performance improvements on HAR trained models. Our key idea is to utilize the temporal visual information of historical frames in a tracklet to yield the prediction results. Our method not only boosts the HAR performance but also does not require any retraining or modifying the trained deep-learning-based models. Further, since some tracklet frames with no occlusion can give more useful information for HAR than others, weighing the importance of frames when computing HAR predictions can be a promising approach for future work.

Acknowledgment

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2016-0-00312) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (NRF-2019R1F1A1055593).

References

1. A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2015.
2. J. Zhu, S. Liao, Z. Lei, and S. Z. Li, "Multi-label convolutional neural network based pedestrian attribute classification," *Image and Vision Computing*, vol. 58, pp. 224– 229, 2017.
3. X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplusnet: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE international conference on computer vision*, pp. 350–359, 2017.
4. Z. Chen, A. Li, and Y. Wang, "A temporal attentive approach for video-based pedestrian attribute recognition," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 209–220, Springer, 2019.
5. Z. Ji, Z. Hu, E. He, J. Han, and Y. Pang, "Pedestrian attribute recognition based on multiple time steps attention," *Pattern Recognition Letters*, vol. 138, pp. 170–176, 2020.
6. A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020.
7. Springer, *MARS: A Video Benchmark for Large-Scale Person Re-identification*, 2016.
8. H. Gunes and M. Piccardi, "Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 102–111, Springer, 2005.
9. S. Canavan, B. Johnson, M. Reale, Y. Zhang, L. Yin, and J. Sullins, "Evaluation of multi-frame fusion based face classification under shadow," in *2010 20th International Conference on Pattern Recognition*, pp. 1265–1268, IEEE, 2010.
10. hyk1996, "A simple baseline implemented in pytorch

for pedestrian attribute recognition task", https://github.com/hyk1996/Person-Attribute-Recognition-MarketDuke.
11. D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016.
12. Y. Lee and H. Kim, "Improved Algorithm of Sectional Tone Mapping for HDR Images," *Journal of the Semiconductor & Display Technology*, vol. 20, no. 2, pp. 137-140, 2021.
13. S. Hwang, S. Hong, J. Yoon, H. Park, and H. Kim, "Deep Learning-based Pothole Detection System," *Journal of the Semiconductor & Display Technology*, vol. 20, no. 1, pp. 88-93, 2021.

접수일: 2021년 9월 2일, 심사일: 2021년 9월 13일, 게재확정일: 2021년 9월 16일