RECASPIA: RECOGNIZING CARRYING ACTIONS IN SINGLE IMAGES USING PRIVILEGED INFORMATION

Christos Smailis[1], *Michalis Vrigkas*[1,2] *and Ioannis A. Kakadiaris*[1]

[1]Computational Biomedicine Lab, Dept. of Computer Science, University of Houston, Houston, TX, USA
[2]Department of Computer Science and Engineering, University of Ioannina, Ioannina, Greece

ABSTRACT

Many approaches for action recognition focus on general actions, such as *"running"* or *"walking"*. This work presents a method for recognizing carrying actions in single images, by utilizing privileged information, such as annotations, available only during training, following the learning using privileged information paradigm. In addition, we introduce a dataset for carrying actions, formed using images extracted from YouTube videos depicting several scenarios. We accompany the dataset with a variety of different annotation types that include human pose, object and scene attributes. The experimental results demonstrate that our method, boosted sample averaged F1 score performance by 15.4% and 4.15% respectively, in the validation and testing partitions of our dataset, when compared to an end-to-end CNN model, trained only with observable information.

***Index Terms*—** Action Recognition, Static Images, Privileged Information, LUPI, Deep Learning

1. INTRODUCTION

Identifying carrying actions performed by a person in a single image is a task that has not been adequately explored in the literature. The majority of single-image action recognition methods focus on general actions such as "walking" or "riding horse". Most of these methods use several types of additional information to assist the action recognition process, by either allowing the use of more than one image region or 2D pose information. Current methods, can be categorized into three categories that use: (i) contextual regions from the image, (ii) person body parts, and (iii) pose information [1, 2, 3]. However, contextual regions do not necessarily correspond to real objects in the scene [4]. Additionally, pose cues produced by automated methods suffer from errors related to partial person visibility or misdetections of the person's body. In recent years, several datasets for single-image action recognition have been published [5, 6, 7, 8, 9]. However, most of them, contain only general action classes and a limited set of annotations, that are not suited to support the recognition of carrying actions.

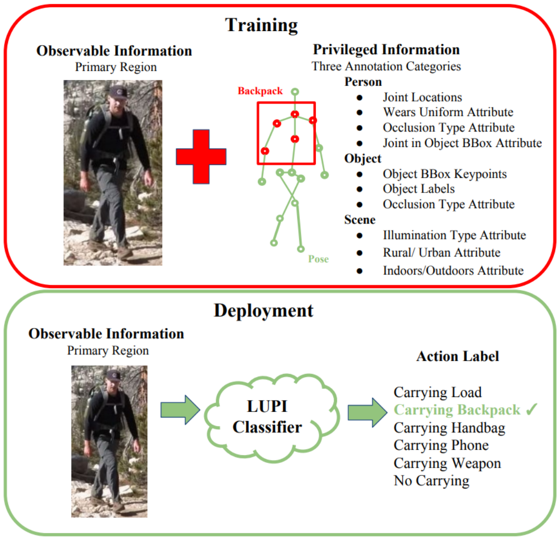

Fig. 1: Overview of the RECASPIA method. To predict the underlying action labels, our method adopts a LUPI classifier. We use as observable information a feature vector for the primary region, extracted from a CNN model. Privileged information consists of annotations about the person's pose along with object and scene attributes. We assume that privileged information is only available during training.

More specifically, since in real-life applications, data of scene attributes, person pose or objects can be hard to obtain, due to acquisition constraints, we introduce a method named RECASPIA that takes into account ground truth information from annotations, available only during training, but not during testing. For this reason, our method follows the learning using privileged information framework (LUPI) [10, 11, 12, 13]. LUPI makes the assumption that a training set is supplied with additional or *"privileged"* features that are not available during test time. An overview of our method is given in Fig. 1. We also propose a dataset for carrying actions, named UHSINICA. To create the dataset we used 2,379 single images containing six different carrying actions. These images were extracted from YouTube videos. Each person in the images is assigned multiple action labels depending on the number of objects it carries. The dataset is also accompanied

Dataset	# Images	# Classes	# Carrying Actions	Person BBoxes	Object BBoxes	Pose	Scene Attributes
Willow [5]	986	7	0	✓	✗	✗	✗
PPMI [6]	4,800	24	0	✓	✗	✗	✗
Pascal VOC 2012 [7]	4,500	10	0	✓	✗	✗	✗
Stanford 40 [8]	9,532	40	0	✓	✗	✗	✗
MPII [9]	40,522	410	3	✗	✗	✓	✗
UHSINICA	**2,379**	**6**	**6**	✓	✓	✓	✓

Table 1: A summary of datasets for action recognition in single images. Note that UHSINICA provides an extended set of annotations when compared to the other datasets.

by a rich set of annotations that describe person, object and scene attributes, such as human pose, bounding boxes and illumination conditions. We tested the performance of the RECASPIA method on the UHSINICA dataset, making the hypothesis that although privileged information is available only during training, it can still be used to assist the recognition of carrying actions. Our experimental results validated our hypothesis as the LUPI framework demonstrated higher recognition performance compared to other baselines.

The contributions of this work are: (i) Developed and implemented a multi-label action recognition method named RECASPIA, that takes into account different types of privileged information, relevant for recognizing carrying actions in single images, but unavailable during deployment, due to acquisition constraints. (ii) Developed a dataset, called UHSINICA, containing multiple annotation types, that enable the development and evaluation of methods, for recognizing carrying actions in single images, a topic not adequately explored in the literature. To the best of our knowledge, this is the first work to take into account privileged information for action recognition in single images.

2. RELATED WORK

Action recognition in single images: Most recent approaches for action recognition in single images, make use of deep learning based architectures that take into account different types of cues to assist the action recognition process, such as contextual regions in Gkioxari *et al.* [1], person body parts in Zhao *et al.* [2] and pose information in Wang *et al.* [3]. Other methods, such as the one introduced by Diba *et al.*[14], emphasize visual attention mechanisms within deep learning models to learn mid-level representations of actions. In Zhang *et al.* [15], authors attempt to perform action classification without knowledge of bounding boxes for the persons involved. However, none of the previous methods have specifically addressed the recognition of carrying actions.

Datasets for single-image action recognition: Several datasets for this problem have been published in the past. Besides the PPMI dataset [6] that focuses on persons that play musical instruments, the rest of the datasets contain images from a variety of scenarios and focus on general actions. Furthermore, aside from the MPII dataset [9], which offers pose annotations, all other datasets are limited to annotations, mainly comprised of person bounding boxes and action labels. None of the existing datasets is specializing in carrying actions. A summary of the datasets can be found in Table 1.

Privileged Information for image classification: Leveraging additional information available only while training image classification models is a concept that has been addressed in many different contexts in the literature. In one of their demonstrated uses of LUPI, Vapnik *et al.* [11], leveraged textual descriptions of hand-drawn digit images as privileged information, to further assist the recognition of handwritten characters. Wang *et al.* [16] used textual tags as privileged information to assist the recognition of objects. Vrigkas *et al.* [17] introduced a probabilistic approach that integrated privileged information such as audio and pose information into a hidden conditional random field model, in order to perform action classification in videos. Finally, in Kakadiaris *et al.* [18] the LUPI framework is used to predict a person's gender from still images using ratios of anthropometric measurements as privileged information, while in Sarafianos *et al.* [19], the authors introduced a method that used LUPI for domain adaptation, to perform animal recognition using visual attributes as privileged information. LUPI has never been applied before in the context of action recognition from single images.

3. METHODOLOGY

Problem Statement: Given an image $\boldsymbol{I}$ depicting human actions, the goal of this work is to predict the underlying subset of action labels $y \in \mathcal{Y}$, performed by a person within a primary region $\boldsymbol{R}$ where $\mathcal{Y} = \{y_1, ...y_N\}$ is a set of N actions.

Observable Information: In our method, the primary region $\boldsymbol{R}$ is used to extract observable features that are available both during training and the deployment stages of our method. Each primary region $\boldsymbol{R}$ is forwarded through a convolutional neural network (CNN) model [20]. Features for each primary region were extracted from the last fully-connected layer. The feature vector related to a primary region $\boldsymbol{R}$ is denoted as $\boldsymbol{x}$ and consists of 512 dimensions.

Privileged Information: For each primary region $\boldsymbol{R}$, we assume that a set of annotations will be available. Since annotated data can be hard to obtain during deployment, we consider it to be privileged information. We argue that this information can be used within the LUPI framework to train a classification model with enhanced performance. A feature vector of 80 dimensions, denoted as $\boldsymbol{x}^*$, is thus formed from three different information types: (i) Person attributes: 16 joint location coordinates that correspond to head, neck, thorax, pelvis, left and right shoulders, elbows, wrists, hip, knees, and ankles. Other attributes for persons include occlusion, wearing uniform and joints overlapping with object bounding boxes. (ii) Object attributes: object bounding box keypoint coordinates, object labels, and object occlusion attributes. (iii) Scene attributes: illumination, rural or urban and indoors or outdoors.

LUPI for Carrying Action Recognition: In RECASPIA, LUPI is implemented using the SVM+ algorithm, introduced in Vapnik *et al.* [10, 11], which is an extension of the original SVM introduced by Cortes *et al.*[21]. The training set is introduced in the form of N triplets $(\boldsymbol{x_i}, \boldsymbol{x_i^*}, y_i), \boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{x}^* \in \mathbb{R}^d, y \in \{-1, +1\}$. Where $\boldsymbol{x_i}$ is the feature vector corresponding to the observable information, available both at training and test time. $\boldsymbol{x_i^*}$ denotes privileged features, and $y_i \in \{-1, +1\}$ represents the class labels in a binary classification setting and d represents the dimensionality of feature vectors. The intuition behind adopting SVM+ for recognizing carrying actions, is to exploit the privileged features, $\boldsymbol{x_i^*}$ that were provided by human annotators and contain information about the persons, the objects and the scene, to further constrain the solution provided by the observable features $\boldsymbol{x_i}$ which consist of primary region features, by solving the following minimization problem during the training phase:

$$\begin{aligned} &\underset{\boldsymbol{w},b,\boldsymbol{w}^*,b^*}{\text{minimize}} \tfrac{1}{2}(||\boldsymbol{w}||^2 + \gamma||\boldsymbol{w}^*||^2) + C\textstyle\sum_{i=1}^{N} \xi_i(\boldsymbol{w}^*, b^*) \\ &\quad\text{subject to: } y_i(\langle \boldsymbol{w}, \boldsymbol{x_i}\rangle + b) \geq 1 - \xi_i(\boldsymbol{w}^*, b^*), \\ &\qquad\qquad\qquad \xi_i(\boldsymbol{w}^*, b^*) \geq 0, i = 1, ..., N \end{aligned} \tag{1}$$

In the above formulation, $\boldsymbol{w} \in \mathbb{R}^m$ represents the weight vector, $||\boldsymbol{w}||^2$ indicates the size of the margin and $b \in \mathbb{R}$ is the bias parameter. In the LUPI paradigm, the slack variables ξ are parameterized as a linear function of privileged information $\xi(\boldsymbol{w}, b) = \langle \boldsymbol{w}, \boldsymbol{x}_i^* \rangle + b$. Finally, C denotes the penalty parameter. The aforementioned framework is applicable to multi-label classification problems through the one-vs-rest approach.

4. THE UHSINICA DATASET

Previous datasets for single image action recognition focus only on generic actions but fail to present more sophisticated and complex actions. In this paper, we address this unmet critical gap by introducing a new dataset named "SINgle Image dataset for Carrying Actions" (UHSINICA). The objectives of this dataset are: (i) introduce a challenging benchmark that includes images of a person performing several types of carrying actions, (ii) to accompany the images with a rich set of annotations.

Data Assembly: All images in the dataset were extracted from 22 publicly available YouTube videos, published under the Standard YouTube Licence. The dataset contains $7,389$ person bounding boxes in $2,379$ images, performing 6 different types of carrying actions. The images depict people in a variety of scenarios such as footage from urban areas such as city squares, railway stations, and airports, people hiking in rural areas, military/police training. The videos were captured by different types of cameras, such as standard handcams, CCTV cameras, trail cameras, and near-infrared cameras, during either day or night. Thus, the dataset is composed of images with varying resolutions and quality. The images were extracted from the videos with a sampling rate of $0.5s$, as some videos had a short duration or frequent scene swaps. Links to the video frames, as well as the annotation tool we developed to create the dataset, will become publicly available upon the publication of the paper.

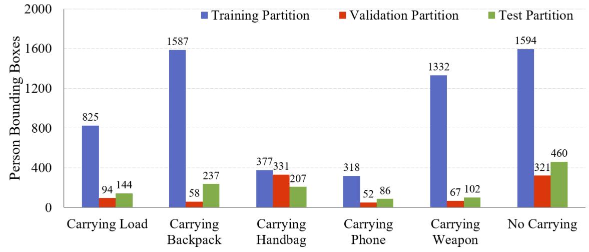

Fig. 2: UHSINICA dataset statistics: person bounding boxes per class, for the training, validation and testing partitions.

Annotation Process: For each image in the dataset, bounding box annotations are provided for the persons depicted and the objects they are carrying. Each person is associated with at least one of the following action labels: *"Carrying Load", "Carrying Backpack", "Carrying Handbag", "Carrying Phone", "Carrying Weapon", "No Carrying"*. Each object bounding box is associated with a class label from the following list: *"Person", "Load","Backpack", "Handbag", "Load", "Phone", "Weapon"*. Each person was annotated with 16 landmarks that represent the joints for the following body parts: head, neck, thorax, pelvis, shoulders, elbows, wrists, hips, knees and ankles. Additionally, we provide occlusion annotations for person bounding boxes, their joints as well as objects. Class distributions per partition can be viewed in Fig. 2.

Evaluation protocol and metrics: The dataset is split in a training partition consisting of $5,398$ primary regions from 13 videos, a validation partition with 896 primary regions from five videos, and a test partition of $1,095$ primary regions form the remaining five videos. Primary regions of each partition, come from videos that have not been used in the other partitions. We make the assumption that the bounding box of each person is known at test time. To measure action recognition performance, the sample averaged F1-score from all action classes is adopted, as it considers multiple action labels being assigned to each person, along with class imbalance.

5. EXPERIMENTS

In this section, the performance of RECASPIA is examined using the UHSINICA dataset. To better understand RECASPIA's characteristics, it is compared against five baselines. First, a comparison between an end-to-end CNN model and an SVM classifier trained with observable features is established. After justifying the selection of the SVM classifier for our task, the informative value of the privileged features

Validation Partition F1-Scores (%)						
Classes	**Primary Region - ResNet-34**	**Primary Region - SVM**	**Privileged Data - SVM**	**Low Level Fusion - SVM**	**SVM+**	**# Samples**
Carrying Load	3	6	0	4	31	94
Carrying Backpack	38	23	1	22	28	58
Carrying Handbag	39	23	15	32	41	331
Carrying Phone	0	0	67	0	8	52
Carrying Weapon	0	64	25	68	60	67
No Carrying	39	55	60	62	61	321
F1-Score (Sample Average)	30.24	34.08	31.90	39.76	45.64	-
Test Partition F1-Scores (%)						
Carrying Load	7	12	1	3	12	144
Carrying Backpack	34	37	36	41	26	237
Carrying Handbag	36	21	8	24	36	207
Carrying Phone	0	0	21	2	8	86
Carrying Weapon	0	9	28	11	1	102
No Carrying	43	47	33	52	55	460
F1-Score (Sample Average)	29.37	30.24	24.41	32.63	33.52	-

Table 2: Evaluation of the RECASPIA method and other baselines using the UHSINICA Dataset.

needs to be demonstrated. This is achieved by evaluating an SVM model trained only with privileged information. To examine the complementarity between the concatenated feature vectors of observable and privileged information, assuming the latter to be known at test time, another SVM model is also trained and evaluated. Finally, the performance of RECASPIA is assessed, by using the observable and privileged features within the context of SVM+. Results from our experiments can be found in Table 2.

Implementation Details: For all experiments that involved an SVM classifier, we used a radial basis function (RBF) kernel and performed grid search using the validation partition of the dataset to determine the optimal hyper-parameter values. The adopted CNN architecture, was based on ResNet-34 [20] and pre-trained with ImageNet-1K [22]. Training was performed through stochastic gradient descent. Since this is a multi-label classification problem, the sigmoid binary cross entropy loss was adopted. The batch size was set to 64 samples with a learning rate of 10^{-3}, that was decreased with a step decay schedule. For data augmentation, shuffling, random mirroring and random crops were applied. Primary regions were scaled to 224×224 pixels.

Primary Region - ResNet-34: To assess the performance of a CNN model with only observable features, in this experiment, the ResNet-34 CNN model is used and trained end-to-end with primary regions. This baseline achieved the lowest performance as it produced sample-averaged F1-scores of 30.24% and 29.37% for the validation and test partitions respectively.

Primary Region - SVM: As our second experiment, primary region features were used with an SVM classifier. This baseline performed better than the ResNet-34 model, as it produced sample-averaged F1-scores of 34.08% and 30.24% for the validation and test partitions respectively. Both results indicated an improvement over the end-to-end CNN model, that justified the choice of SVM as a classifier for our problem.

Privileged Data - SVM: In our third experiment, a standard SVM was used by assuming the set of privileged data to be known both at training and test time. This experiment was performed to assess the informative value of the annotations. The sample-averaged F1-scores for this case were 31.90% and 24.41% for the validation and test partitions, which is close to the models that used primary region features in the two previous experiments.

Low-Level Fusion - SVM: In this experiment, the complementarity of observable and privileged features is assessed. Again in this setting, privileged features are assumed to be known both at training and test time. The primary region feature vectors with the set of privileged feature vectors were thus concatenated. This experiment led to an increase in performance up to 5.16% for the validation partition and 2.39% for the test partition, when compared to the experiments that only primary regions with the SVM classifier were used. Therefore observable and privileged information are complementary.

SVM+: Finally, the performance of the RECASPIA method is examined. In RECASPIA, SVM+ is used with primary regions features as observable information and the annotation features as privileged information that are not available during test time. RECASPIA outperformed all previous baselines by producing sample-averaged F1-scores of 45.64% and 33.52% for the validation and test partitions respectively. Adopting the LUPI framework can thus surpass all the other baselines, despite the fact that some of them have access to privileged information at test time. The reason that the SVM+ model performed better than the low-level fusion - SVM model, is that in the latter case, the strongest modality may dominate over the other. Thus, low-level fusion may not be an optimal way to combine different types of data for this problem.

6. CONCLUSION

In this paper, a method named RECASPIA for performing multi-label carrying action recognition was presented. RECASPIA adopts the LUPI framework and leverages ground truth annotations, available only during training time, to assist the classification task. The UHSINICA dataset, that represents carrying actions from still images and provides a rich set of annotations, was also introduced. Extensive experimentation using our proposed dataset demonstrated performance gains by using the LUPI framework. Our results, indicated that using privileged information to train a model for carrying action recognition, boosts performance over models that are trained only with observable information.

Acknowledgments: This material is supported by the U.S. Department of Homeland Security under Grant Award Number 2017-STBTI-0001-0201 with resources provided by the Core facility for Advanced Computing and Data Science at the University of Houston. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. REFERENCES

[1] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, December 7-13 2015, pp. 1080–1088.

[2] Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions," in *Proc. IEEE International Conference on Computer Vision*, Venice, Italy, October 22-29 2017, pp. 3411–3419.

[3] X. Wang, K. Li, and Y. Li, "A deep model combining structural features and context cues for action recognition in static images," in *Proc. International Conference on Neural Information Processing*, Guangzhou, China, November 14-18 2017, pp. 622–632.

[4] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, Sep 2013.

[5] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: A study of bag-of-features and part-based representations," in *Proc. 21st British Machine Vision Conference*, Aberystwyth, UK, August 31-September 3 2010, pp. 97.1–11.

[6] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. 20th International Conference on Computer Vision*, San Francisco, CA, June 13-18 2010, pp. 9–16.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[8] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. 13th International Conference on Computer Vision*, Barcelona, Spain, November 2011, pp. 1331–1338.

[9] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. 27thIEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, June 24-27 2014, pp. 3686–3693.

[10] V. Vapnik, *Estimation of dependences based on empirical data*, Springer Science & Business Media, 2006.

[11] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[12] Xingyu Chen, Chen Gong, Chao Ma, Xiaolin Huang, and Jie Yang, "Privileged semi-supervised learning," in *Proc. 25th IEEE International Conference on Image Processing*, Athens, Greece, 2018, pp. 2999–3003.

[13] J. Lambert, O. Sener, and S. Savarese, "Deep learning under privileged information using heteroscedastic dropout," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 18-22 2018.

[14] A. Diba, A. M. Pazandeh, H. Pirsiavash, and L. V. Gool, "Deepcamp: Deep convolutional action attribute mid-level patterns," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 27-30 2016, pp. 3557–3565.

[15] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5479–5490, Nov 2016.

[16] S. Wang, S. Chen, T. Chen, and X. Shi, "Learning with privileged information for multi-label classification," *Pattern Recognition*, vol. 81, pp. 60 – 70, 2018.

[17] M. Vrigkas, E. Kazakos, C. Nikou, and I. A. Kakadiaris, "Inferring human activities using robust privileged probabilistic learning," in *Proc. IEEE International Conference on Computer Vision Workshops*, Venice, Italy, October 2017, pp. 2658–2665.

[18] I. A. Kakadiaris, N. Sarafianos, and C. Nikou, "Show me your body: Gender classification from still images," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, September 25 - 28 2016, pp. 3156–3160.

[19] N. Sarafianos, M. Vrigkas, and I.A. Kakadiaris, "Adaptive SVM+: Learning with privileged information for domain adaptation," in *Proc. IEEE International Conference on Computer Vision Workshops*, Venice, Italy, October 22-29 2017, pp. 2637–2644.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 29th IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 26 - July 1 2016, pp. 770–778.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] 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*, vol. 115, no. 3, pp. 211–252, 2015.