EXPLOITATION OF SEMANTIC KEYWORDS FOR MALICIOUS EVENT CLASSIFICATION

*Hyungtae Lee[†‡], *Sungmin Eum[†‡], *Joel Levis[§], Heesung Kwon[‡], James Michaelis[‡], and Michael Kolodny[‡]

[†]Booz Allen Hamilton Inc., McLean, Virginia, U.S.A.
[‡]U.S. Army Research Laboratory, Adelphi, Maryland, U.S.A.
[§]Ohio University, Athens, Ohio, U.S.A.

ABSTRACT

Learning an event classifier is challenging when the scenes are semantically different but visually similar. However, as humans, we typically handle such tasks painlessly by adding our background semantic knowledge. Motivated by this observation, we aim to provide an empirical study about how additional information such as semantic keywords can boost up the discrimination of such events. To demonstrate the validity of this study, we first construct a novel Malicious Crowd Dataset containing crowd images with two events, benign and malicious, which look visually similar. Note that the primary focus of this paper is not to provide the state-of-the-art performance on this dataset but to show the beneficial aspects of using semantically-driven keyword information. By leveraging crowd-sourcing platforms, such as Amazon Mechanical Turk, we collect semantic keywords associated with images and then subsequently identify a subset of keywords (e.g. police, fire, etc.) unique to specific events. We first show that by using recently introduced attention models, a naïve CNN-based event classifier actually learns to primarily focus on local attributes associated with the discriminant semantic keywords identified by the Turks. We further show that incorporating the keyword-driven information into early- and late-fusion approaches can significantly enhance malicious event classification.

***Index Terms*—** malicious crowd dataset, semantic keyword, event classification, crowd-sourcing

1. INTRODUCTION

Images associated with very different events can often be represented with similar visual attributes, as shown in Figure 1.

Fig. 1. A pair of similarly looking crowd images from different events with relevant object contents.

In Figure 1, both images include crowds as main foreground attributes, yet we can immediately discern that the two images contain drastically different semantic events. The content of the right image is "malicious" as opposed to the left image being "benign" due mainly to some specific objects contained in the image, such as fire, smoke, police, etc. The above observation indicates that exploiting relevant local attributes (local objects) jointly with global attributes (the whole scene) is quite essential to better discriminating the events containing complex events. What is more important is being able to find out which objects are more correlated (semantically meaningful) than others with respect to a certain set of events.

In this paper, we introduce a new event dataset and then conduct a study, which verifies our argument that identifying highly relevant objects to associated events are crucial to improving classification performance. We first collect a set of images which contain two visually similar, yet semantically different events: benign and malicious. Since most benchmark datasets [1, 2, 3] collected for event classification do not deal with this problem, our newly constructed dataset will be useful for the community. We also collect a number of keywords that appear in each image in the dataset, as listed below each image in Figure 1. We asked the users on the Amazon Mechanical Turk to describe the semantic contents of each image in terms of keywords without providing the semantic event labels. Then we select non-overlapping distinctive and frequently occurring keywords for the malicious event, which we aim to identify and treat them as the representative "semantic keywords".

Secondly, we analyze how these human-driven sets of semantic keywords align with the visual attributes inherently

*These authors contributed equally to this work

Fig. 2. **Malicious Crowd Dataset.** Example images for the benign and malicious events are shown. The groundtruth bounding box information which corresponds to the selected malicious semantic keywords is also included in the dataset. Yellow, purple, blue, white, and red boxes correspond to fire, smoke, helmet, police, and car, respectively.

learned by the naïve event classifier. For this analysis, we leveraged the visualization approach based on the top-down neural attention maps [4] which indicate the regions on which the deep CNN-based event classifier mainly focuses. These "machine-driven" attention maps show that they clearly share high relevance with the "human-driven" semantic keywords although the classifier is not supported with any additional semantic information in the learning process.

Lastly, we carry out a study to verify the practicality of explicitly incorporating these semantic keywords using various fusion approaches, which include a novel CNN-based architecture (IOD-CNN) developed by part of the authors [5]. We show that these keyword-driven information is effective in helping out the event classification task regardless of whether the information is used in an early- or late-fusion scheme.

Our contributions are summarized as follows:

1. We introduce a new Malicious Crowd Dataset containing semantically different malicious/benign images, crowd-sourced semantic keywords, and groundtruth bounding box annotations for the objects corresponding to the semantic keywords.

2. By analyzing the attention maps, we provide valuable indications that even naïve malicious event classifier learns to focus on the regions which align with the semantic keyword set.

3. We provide the community with novel findings based on the empirical study, which verifies the practicality of explicitly using the semantic keyword information for malicious event classification.

2. MALICIOUS CROWD DATASET

We have constructed the Malicious Crowd Dataset which consists of images along with the malicious/benign event labels. It also includes the malicious semantic keywords and their corresponding bounding box information for each image.

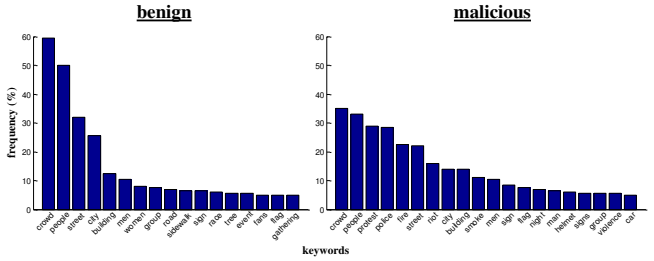

Fig. 3. Histograms of Relevant Keywords

2.1. Malicious and Benign Crowd Images

The dataset contains 1133 crowd images equally split into two classes: benign and malicious. A benign image would portray a "non-alarming" scene, while a malicious one would be alarming and potentially dangerous.

The images were collected from the web using various search terms. For benign images, search terms such as marathon, pedestrian, crowd, parade, and concert were used. Terms such as riot and protest were used to gather the malicious crowd images. Figure 2 illustrates some example images from each class.

2.2. Semantic Keywords

Malicious Semantic Keywords Selection. To describe the contents of each of the crowd images, Amazon Mechanical Turk was used. A human was responsible for assigning five keywords to each image based on what objects are observed within. To ensure the accuracy of the Mechanical Turk results, we manually removed the keywords which were incorrectly assigned.

After successfully collecting the crowd images and corresponding keywords, identifying keywords only relevant to the malicious class was necessary. We then constructed two keyword sets, each acquired by selecting the most frequently appearing keywords in the two given classes. In practice, words that are commonly annotated in 5% or more images in each class were selected. As a result of this thresholding, the numbers of selected words for the benign and malicious classes are 17 and 20, respectively. Selected words and those

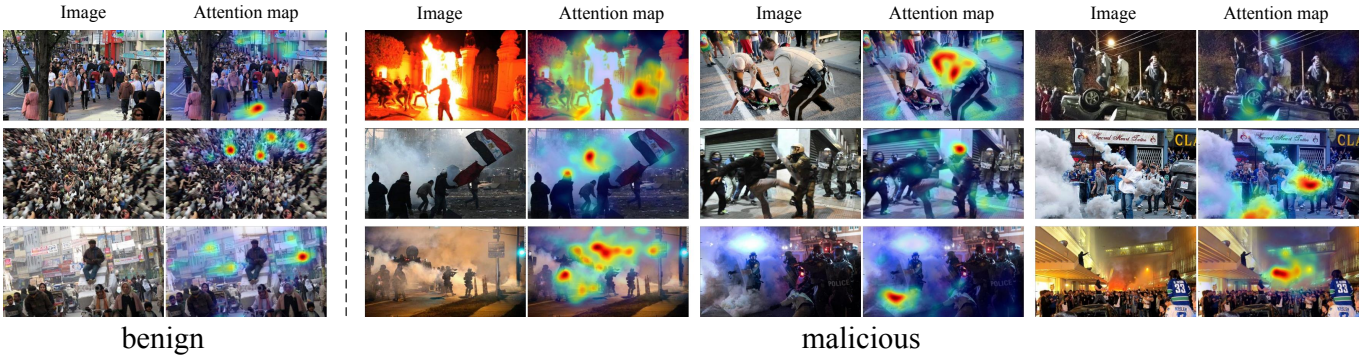

Fig. 4. **Example Images of Attention Maps.** Naïve event classifier is used. Red indicates strongly fired-up regions.

Table 1. Number of images where each keyword relevant to the malicious image appears.

class	images	police	fire	smoke	helmet	car
benign	557	8	1	2	7	57
malicious	576	205	144	150	206	65

frequency for both classes can be seen in Figure 3. We have refined the set by eliminating the common keywords that appear in both classes. This elimination resulted in nine malicious keywords. We further eliminated keywords indicating particular phenomena such as *protest*, *riot*, *night*, and *violence*. Then *police*, *fire*, *smoke*, *helmet*, and *car* are included in the final set of malicious semantic keywords.

Annotation. After finalizing the set of malicious semantic keywords, to provide the ground truths, we went through all the images and manually labeled the bounding boxes for the five objects which correspond to those keywords.

Table 1 shows the number of images where each keyword (object) actually appears. While *police*, *fire*, *smoke*, and *helmet* seem to be closely associated with the malicious event, *car* is seen in both events with a similar frequency. Note that the numbers in the table do not necessarily match the histogram of malicious semantic keywords obtained from crowd-sourcing. For example, *police* appears in 205 out of all 576 malicious images at a rate of 35.59%, but is assigned only to 28.50% of the malicious image by the Turks. This is because the visual contents associated with these keywords are not overly notable in several images. We can observe that the frequencies of the selected semantic keywords show a notable gap between the two classes, indicating that the purpose of the proposed keyword selection process is achieved.

3. VISUALIZING THE ATTENTION MAPS FOR MALICIOUS EVENTS

As mentioned in Section 2, the final set of semantic keywords is selected by carrying out a process with the human in the loop which may bring certain heuristics. In order to analyze how much these human-driven additional semantics are correlated with the actual event classifier, we have visualized the attention maps which depict the "excited" or "fired-up" regions on the images learned by the event classifier. To visualize the attention maps, we have used the top-down neural attention model [4] which uses the excitation back-propagation scheme. We have trained a deep CNN-based event classifier ('Event CNN' in Table 3) to be used for generating the attention maps. Example attention maps along with the original images are shown in Figure 4.

We can observe that this naïve event classifier ('Event CNN'), even without the additional semantic information, has learned to focus on most of the selected keyword-driven objects. This verifies that the selected semantic keywords and their corresponding objects are inherently being used in distinguishing the malicious events from the benign ones.

4. EFFECTIVENESS OF SEMANTIC KEYWORDS

4.1. Evaluation Protocol

The Malicious Crowd Dataset consists of 1133 images - 576 of 1133 are labeled as the malicious image and the rest are labeled as benign. The dataset is randomly divided into train and test sets which include 905 and 228 images, respectively. Average precision (AP) is used as an evaluation metric.

4.2. Methodology

We demonstrate the effectiveness of semantic keyword-driven information by using i) late fusion of classifiers or ii) IOD-CNN (Integrated Object Detection CNN) [5], which can be considered an early fusion.

Late Fusion of Event/object Classifiers. We have trained five different object classifiers or detectors which correspond to five selected keywords. For the rigid objects (*helmet*, *police*, and *car*), we have used the deformable part model (DPM), while two separate CNN classifiers were trained for non-rigid objects (*fire* and *smoke*). A CNN classifier ('Event

Table 2. **Malicious Event Classification Accuracy.** Late fusion approaches

	Baseline	Keyword-driven object					Late fusion						
	(EventCNN)	police	fire	smoke	helmet	car	SVM-rbf	DBF	SVM-lin	kNN	LD	LR	EC
AP	72.2	58.6	56.3	68.9	53.2	49.1	74.2	75.7	75.8	75.8	76.0	76.3	**77.1**
Gain	·	·	·	·	·	·	+2.0	+3.5	+3.6	+3.6	+3.8	+4.1	+4.9

CNN') for event classification was also trained. All the three CNN classifiers were fine-tuned from [6].

A late fusion was performed on the output of these six streams. This is to enhance the performance of the baseline event classifier. We tested several fusion methods which include Linear Discriminant Analysis (LD) [7], Logistic Regression (LR) [8], Support Vector Machines (SVM) [9], k-Nearest Neighbor Classifiers (kNN) [10], Subspace-based Ensemble Classifiers (EC) [11], and a Dynamic Belief Fusion (DBF) [12]. For SVM, we used two different kernels which are a linear kernel (SVM-lin) and RBF kernel (SVM-rbf).

IOD-CNN. Eum et al. [5] introduced a unified deep CNN architecture which integrates architecturally different, yet semantically-related object detection networks to boost event recognition task. This architecture allows the within-network sharing of the convolutional/fully connected layers across event recognition and object detection tasks. This approach can be considered as an "early fusion" which can be differentiated with the "late fusion". As the network is learned in an end-to-end fashion, the training can be performed efficiently.

4.3. Experiments

Note that the purpose of this paper is not to provide the state-of-the-art performance on this dataset but to provide an empirical study which shows the beneficial aspects of using additional information which are semantically-driven.

Late Fusion of Event/object Classifiers. Table 2 shows the performance for the baseline model ('Event CNN'), keyword-driven object detectors/classifiers, and various late fusion approaches. Note that, the numbers shown below the 'keyword-driven object' column in the table do not indicate the detection/classification performances for the corresponding objects. Whenever an object detectors/classifier finds the corresponding object in a given image, the image is classified as 'malicious'. Thus, when the *police* detector detects a police officer in a test image which is originally labeled as 'benign', this instance is counted as a false positive for that detector.

Keyword-driven object detectors/classifiers do not provide better classification accuracy than the baseline. This is because these semantically relevant objects are only seen in small portions in the dataset. However, it is interesting to notice that the accuracy can be boosted up consistently across different fusion methods when the baseline classifier is combined with these keyword-driven classifiers/detectors.

Table 3. **Overall Performance Comparison.** Early- and late fusion approaches

Method	Keyword Info.	AP	Gain
Event CNN	No	72.2	
+ Late fusion	Yes	**77.1**	+4.9
Event CNN+ [5]	No	90.2	
IOD-CNN (Early fusion) [5]	Yes	93.6	+3.4
IOD-CNN + Late fusion [5]	Yes	**94.2**	+4.0

Keyword Information Helps Consistently. Table 3 shows the event classification performance acquired by the selected fusion approaches. The second row in the table indicates the performance gained over the first baseline ('Event CNN') by using the late fusion of separately learned classifiers (see the baseline and the fusion result in Table 2). The next three rows show the classification accuracy based on a different baseline classifier ('Event CNN+'). This baseline also does not exploit any keyword information and is reported [5] to have used additional treatments such as an ROI pooling and a different training scheme. IOD-CNN [5] which embeds the keyword-driven object information by early-fusion outperforms its baseline ('Event CNN+') by 3.4 AP. When a late fusion is added on top of that, additional performance increase of 0.6 AP is acquired. See [5] for detailed description.

The results consistently show that exploiting the keyword-driven object detectors/classifiers provides beneficial information in boosting up the event classification performance.

5. CONCLUSIONS

We addressed a challenging classification problem where certain classes can be expressed by similar visual attributes but should be distinguished from each other semantically. To demonstrate, we have constructed a novel Malicious Crowd Dataset with images representing two classes (benign and malicious) that may look similar but are semantically different. To provide additional semantic information, we collected a set of semantic keywords using a crowd-sourcing platform. We have confirmed the validity of these keywords by analyzing the attention map visualizations for the naïve event classifiers ('Event CNN'). We also provide an empirical study which shows the practicality of using semantic keyword information in enhancing the malicious event classification performance.

6. REFERENCES

[1] Li-Jia Li and Li Fei-Fei, "What, where and who? classifying event by scene and object recognition," *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[2] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al., "A large-scale benchmark dataset for event recognition in surveillance video," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[3] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quéenot, Maria Eskevich, Robin Aly, and Roeland Ordelman, "TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking," in *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[4] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff, "Top-down neural attention by excitation backprop," in *European Conference on Computer Vision (ECCV)*, 2016.

[5] Sungmin Eum, Hyungtae Lee, Heesung Kwon, and David Doermann, "IOD-CNN: Integrating object detection networks for event recognition," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[7] Ronald Alymer Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[8] David A. Freedman, "Statistical models: Theory and practice," p. 128. Cambridge University Press, 2009.

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

[10] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[11] Tin Kam Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 8, pp. 832–844, 1998.

[12] Hyungtae Lee, Heesung Kwon, Ryan M. Robinson, William D. Nothwang, and Amar M. Marathe, "Dynamic belief fusion for object detection," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.