Can Object Detectors Aid Internet Video Event Retrieval?

Davide Modolo [a] and Cees G.M. Snoek[b]

[a, b] University of Amsterdam, Science Park 904, 1098 XG, Amsterdam, The Netherlands;

ABSTRACT

The problem of event representation for automatic event detection in Internet videos is acquiring an increasing importance, due to their applicability to a large number of applications. Existing methods focus on representing events in terms of either low-level descriptors or domain-specific models suited for a limited class of video only, ignoring the high-level meaning of the events. Ultimately aiming for a more robust and meaningful representation, in this paper we question whether object detectors can aid video event retrieval. We propose an experimental study that investigates the utility of present-day local and global object detectors for video event search. By evaluating object detectors optimized for high-quality photographs on low-quality Internet video, we establish that present-day detectors can successfully be used for recognizing objects in web videos. We use an object-based representation to re-rank the results of an appearance-based event detector. Results on the challenging TRECVID multimedia event detection corpus demonstrate that objects can indeed aid event retrieval. While much remains to be studied, we believe that our experimental study is a first step towards revealing the potential of object-based event representations.

Keywords: Video-Event Recognition, Object Recognition, Video and Scene Understanding, Performance evaluation

1. INTRODUCTION

Video event retrieval in web videos has become an important research topic, due to its applicability in domains like web search, broadcast news, and sports.[1,2] The problem of event retrieval is, however, challenging for several reasons including large variance in the appearance of particular events, similarity in the appearance of different events, and ambiguity in translating semantic definitions of events into a low-level formalism for representation and recognition. Furthermore, most web videos are captured by everyday users using hand-held cameras or smart-phones, and generally contain considerable camera motion, occlusion and cluttered background. For all these reasons, there is an urgent demand to find a robust event representation for video.

We categorize state-of-the-art approaches to the problem of event representation in video as either *model-based*[3–5] or *appearance-based*.[6–10] The former attempt to estimate a set of pre-defined model parameters, such as geometric transformations, from the video data and use them to recognize the event. Most of these methods are domain specific, or impose constraints on the environment as well as the type of motion that can be detected. In addition, some assume knowledge of the scene and/or cameras. For example, Bobick *et. al.* use motion-energy images and motion-history images to recognize many types of aerobics exercises.[3] While their method is efficient, their work and many others[4] assume that the person involved is well segmented from the background and centered in the image. In contrast, appearance-based approaches perform inference directly on the observed pixel responses. The state-of-the-art event detection approach relies on bag-of-words,[8,9] where each video is represented by its vector quantized visual word frequency.[11] These methods represent videos in terms of low-level edge orientations. However, events are high-level semantic activities that humans perceive when observing a video sequence. Hence, there is a case to be made for representations which describe events semantically.

Indeed, it is well known from cognitive science that the human ability to perceive events develops through the understanding of objects.[14,15] First, humans learn about objects. Second, they learn to detect relations among individual objects. Finally, they perceive events by analyzing regularities in terms of object relationships.

Further author information: (Send correspondence to Author 1)

[a]: E-mail: D.Modolo@sms.ed.ac.uk

[b]: E-mail: cgmsnoek@uva.nl

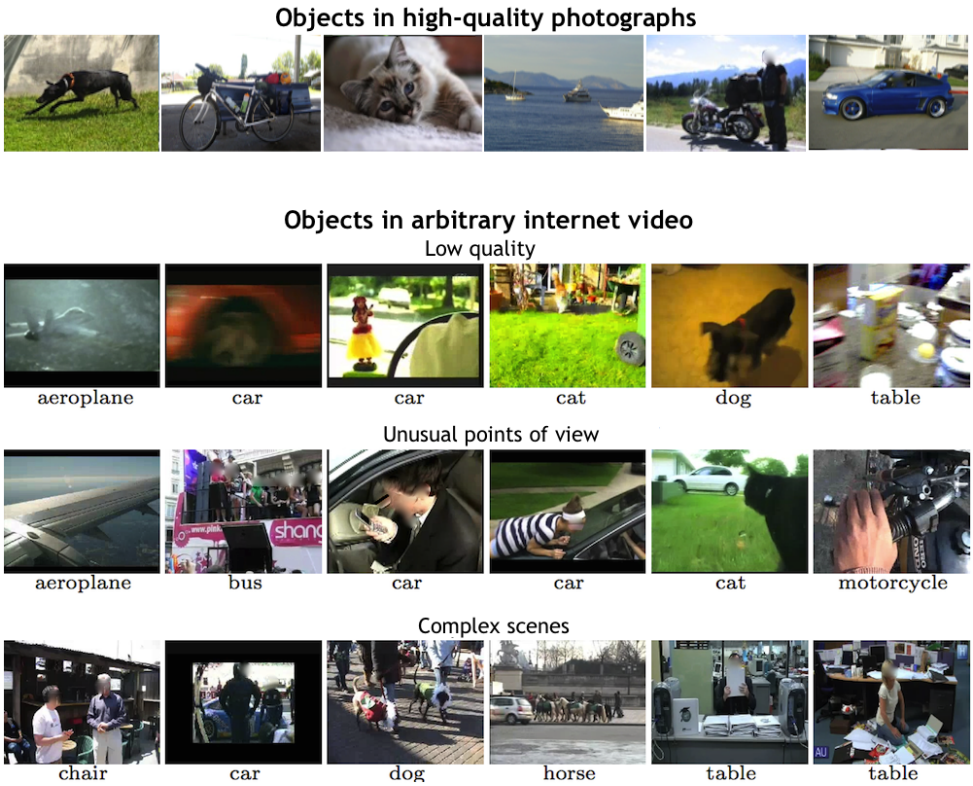


Figure 1. Objects in high-quality photographs[12] are typically items of desire and therefore the focus of attention. In contrast, objects in arbitrary Internet video[13] may appear in low quality, recorded from unusual points of view, and within complex scenes. Hence, generalizability of present-day object detectors utilized on photo collections is unclear.

We start from this theory, with the aim to make event retrieval more robust. While much progress in object detection for high-quality photographs has been reported recently,[12] their utility for low-quality web video is unclear. See Figure 1 for visual differences. This paper seeks to unravel whether present-day object detectors can aid event retrieval. To the best of our knowledge, no experimental evaluation has been proposed in the literature to quantify this utility. To shed light on the matter, we formulate two hypotheses that we address in this paper. Our first hypothesis states:

Hypothesis 1: *State-of-the art photo object detectors are suited for arbitrary Internet video.*

Once object detectors generalize, they form a potentially effective means to describe events. There is no evidence, however, that a video representation relying on object detectors only improves event retrieval and this motivates our second hypothesis:

Hypothesis 2: *Present-day object detectors can aid arbitrary Internet video event detection.*

Before testing our two hypothesis we first discuss the state-of-the-art in object detection.

2. RELATED WORK

Various object detection methods exist in the literature. We organize them by the way they classify images. We refer to *global object detection* as the ability to indicate whether an image contains a desired object. In order to do so, these detectors analyze the entire image and typically exploit the context information surrounding an object. In contrast, we refer to *local global detectors* for those methods that, in order to recognize an object, reject parts of the image that are considered non-relevant and, instead, search for local evidence. Figure 2 shows the difference between global and local object detectors.

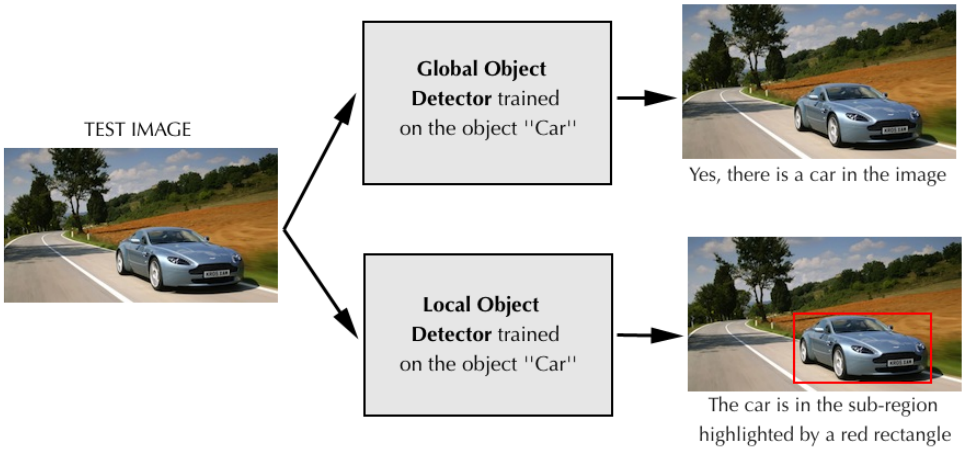


Figure 2. Global object detectors indicate whether an image contains a desired object or not, while local object detectors also define the bounds of the object in the image.

2.1 Global Object Detectors

One of the most effective approaches for global object detection is bag-of-words.[11,16] In bag-of-words models, an image is first represented by a collection of local features detected from image pixels, either sparsely (e.g., Harris Laplace) or in a regular, dense grid (e.g., Dense Sampling). Each local feature is then represented by one or more descriptors, each describing one aspect of the small region surrounding the considered feature. Typical descriptors include color, shape, and texture, but the most effective one are Lowe's SIFT[17] and its color SIFT variants. See van de Sande et al. paper[18] for a deep discussion.

Once the descriptors are extracted, they are assigned to discrete visual words predefined in a vocabulary (obtained by clustering methods like k-means or gaussian mixture models). The frequency count of the visual words is used as the image representation. See van Gemert et al. paper[19] for a detailed comparison of several visual word assignment methods. To allow for region-specific weighting of visual words, a feature pyramid is typically employed, as suggested by Lazebnik et al.[20] In the final stage of the bag-of-word approach the histogram representation of word counts are analyzed by a kernel-based classifier, like a support vector machine.[21]

State-of-the-art global object detectors[18] have evolved to an accurate level of performance for both high-quality photographs and low-quality web video. In our experimental study, we evaluate state-of-the-art global object detectors based on bag-of-words for the purpose of event representation.

2.2 Local Object Detectors

State-of-the-art approaches for local object detection are based on an exhaustive search over the image, in order to find the best object positions. However, because the number of possible positions grows exponentially with the size of an image, the search space becomes huge quickly and several heuristics have been proposed to speed up the search.

Lampert et al.[22] proposed a simple, but powerful, branch and-bound technique, called *efficient subwindow search*, which directly searches for the optimal window within an image. While they obtain impressive results for linear classifiers, for non-linear classifiers the search remains too slow.[23] To further speed-up local object detectos, Alexe et al.[23] proposed to search for any object, independently from its class. They train a classifier on the bounding boxes of those objects having a well-defined shape. They then randomly sample boxes to which they apply their classifier. The boxes with the highest likelihood of containing an object serve as a set of object hypotheses, thereby reducing the number of windows that need to be evaluated by class-specific object detectors.

Van de Sande et al.[24] reduced the huge search space by pre-defining bounding boxes using segmentation. They initially oversegment an image to obtain relatively small rectangles which describe object parts. They then hierarchically and greedily group similar regions together to define new bigger regions that aim to capture a

Table 1. 6000 Internet video data[13] used to validate hypothesis 1.

Object	Video Frames	Object	Video Frames
Airplane	264	Cow	5
Bird	168	Dog	125
Boat	309	Horse	66
Bus	46	Motorcycle	150
Car	1644	Sofa	117
Cat	166	Table	736
Chair	759	ExtraNegative	1590

more complete object representation. All the rectangles created during the hierarchical grouping form then the new, reduced, search space.

Felzenswalb et al.[25] speeded the search up using a linear SVM, HOG features and by introducing a semantic object representation. Their approach builds on the pictorial structures framework, which represents objects by a collection of parts arranged in a deformable configuration. The idea is that the parts capture local appearance features while the deformable configuration captures the spatial relationship between them.

In contrast to global detectors, local object detectors have been evaluated for photographs only, and their utility for video remains unclear. We evaluate the utility of the local object detectors by Van de Sande et al.[24] and Felzenswalb et al.[25] in our study.

3. OBJECT DETECTOR SUITABILITY

To understand what object detectors are best suited for representing events in arbitrary Internet video, we perform an experiment to test our first hypothesis.

3.1 Photo and Video Data Sets

We train detectors on the PASCAL VOC 2007 data set,[12] which is a collection of consumer photographs from the Flickr photo-sharing web-site. The data set consists of 9,963 images containing 24,640 annotated objects from 20 classes. The annotations provide what objects are present in an image and in what sub-regions. The images span the full range of consumer photographs, including indoor and outdoor scenes, close-ups and landscapes.

We test object detectors on the collection of web videos from the TRECVID 2011 Semantic Indexing task.[13] These are characterized by a high degree of diversity in creator, content, style, production qualities, recording devices, encoding, language, etc.

In our experiment we consider the object classes shared among the two data sets, which are *airplane, bird, boat, bus, car, cat, chair, cow, dog, horse, motorcycle, sofa* and *table*.

For testing, we collected and manually verified all the positively annotated keyframes for the 13 object classes. In addition, we added 1590 negative keyframes not containing any of the 13 objects to arrive at a dataset of 6000 frames. The final data set statistics are summarized in Table 1.

3.2 Experiment 1: Object Detector Suitability

To provide an appropriate evaluation, we test the performance of four state-of-the-art techniques, namely two global and two local detectors. We evaluate them on the task of video frame classification. In order to avoid confusion, we assign a discriminative name to each detector. The names are an indication of how the methods train object models (entire image vs bounding boxes) and how they score test images (entire image vs bounding boxes).

1. `TrainAll-ClassifyAll`. In order to capture information about the object of interest and its context, we use the global detector based on bag-of-words described by van de Sande et al.[18] The contextual information might prove useful in classifying some object classes. We train object models using as positive samples the images containing the object of interest.

2. `TrainBB-ClassifyAll`. In order to capture a different visual appearance, we use, again, the global detector presented by van de Sande et al.,[18] but this time we train object models considering as positive samples only the sub-regions of the positive images delimitated by the positive bounding boxes.

3. `TrainBB-SegmentationBB`. To model directly the object of interest rather than statistical regularities in the image background, we use the local detector described by van de Sande et al.[24] This approach explicitly ignores all but the object, using bounding boxes for training, and performing a selective search for testing. It has shown discriminative potential for non-rigid objects.

4. `TrainBB-DeformableBB`. We use the local detector proposed by Felzenszwalb et al.,[25] in order to model directly the objects of interest instead of capturing context information. This approach, differently from the previous one, uses an exhaustive search for testing. It has shown discriminative potential mainly for rigid objects.

For the local object detectors, we represent the likelihood of an image to contain an object, as the maximum score over all the bounding boxes evaluated by a detector. We apply all the detectors to the frames in the test set and we evaluate their ranked lists in terms of *average precision*, which is a common measure to evaluate retrieval experiments.[13]

3.3 Implementation Details

We train object models for `TrainAll-ClassifyAll`, `TrainBB-ClassifyAll` and `TrainBB-SegmentationBB` using the following settings. We use Harris-Laplace and Dense Sampling detectors to sample interest points. The latter uses a regular grid with an interval of 6 pixels and at a single scale ($\sigma = 1.2$). From these points we extract SIFT, opponentSIFT and RGB-SIFT features, as suggested in.[18] We create a codebook of 4,096 words for each of the three features using k-means clustering with hard assignment. Moreover, we use 1x1 and 1x3 spatial pyramids.[20] In this way our final feature vector has a dimension of $24 \times 4096 = 98,304$. We learn a Support Vector Machine with kernel based on histograms intersection.[21] In addition, for the segmentation procedure of `TrainBB-SegmentationBB` we use a software provided by van de Sande,[24] and for `TrainALL-DeformableBB` we use the software and the object models released by Felzenswalb.[25]

3.4 Result 1: Object Detector Suitability

Results are shown in Figure 3. They show that the detectors are able to recognize several instances of objects, considerably outperforming a random search. The detector achieving the best results is `TrainBB-DeformableBB`, with a Mean Average Precision of 0.257 and best Average Precision on 7 of the 13 objects. `TrainAll-ClassifyAll` achieves the second best MAP, 0.187, and best AP on 4 of the 13 objects. `TrainBB-SegmentationBB` achieves the third best MAP, 0.166, and best AP on 2 of the 13 objects. `TrainBB-ClassifyAll` is the least discriminative, with an MAP of 0.129.

`TrainBB-DeformableBB` is the detector least affected by the transition from photographs to video frames. As expected, the method shows potential in recognizing rigid objects, such as *car*, *motorcycle*, *bus* and *chair*, but it surprisingly achieves best results also for classes such as *horse*, *birds* and *cat*. In contrast, `TrainBB-SegmentationBB`, which in high-quality photos performs well for non-rigid objects such as *birds*, *cats*, *dogs*, and *plants*,[24] seems to be susceptible to the change of domain. We attribute this to their segmentation process,[26] which fails in segmenting low-quality frames.

`TrainAll-ClassifyAll` and `TrainBB-ClassifyAll`, which differ in the way they address the training phase (entire image vs bounding boxes), achieve considerably different results. On objects like *airplane*, *boat*, *sofa* and *table*, `TrainAll-ClassifyAll` performs particularly well, showing how important it is to capture context information in order to recognize these object classes. In contrast, `TrainBB-ClassifyAll` does not perform satisfactorily.

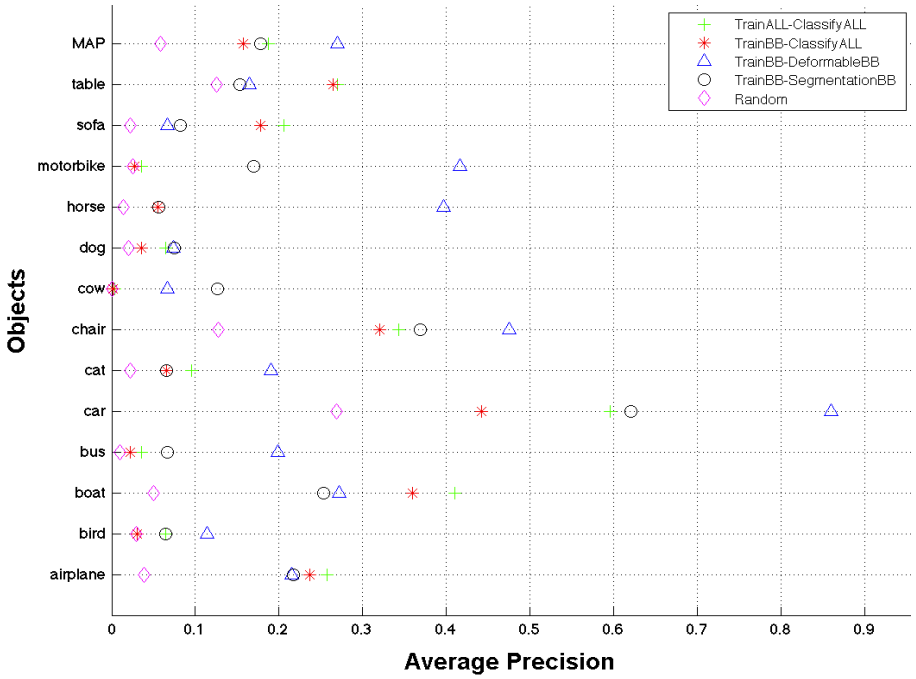


Figure 3. Results of Experiment 1: Object Detector Suitability. All the detectors clearly outperform a random search and results suggest that present-day object detectors for photos can successfully be used for recognizing objects in Internet videos.

Figure 4 present the top 6 frames for the rigid object *airplane* and the non-rigid object *cat*. The results of all the detectors on both classes are good. No relevant difference is present in the results of the class *airplane*, while it is interesting to note how well `TrainBB-DeformableBB` is able to detect the cats. By exploiting the object in "parts", it is able to match properly the face of the cat and recognize the video frame as positive instance.

The results of Experiment 1 suggest that present-day object detectors for photos can successfully be used for recognizing objects in Internet videos, which confirms our first hypothesis. In addition, the most suitable object detector is the local detector `TrainBB-DeformableBB`; to compensate the global context information not captured by this method, a global detector `TrainAll-ClassifyAll` seems suitable to be used as support.

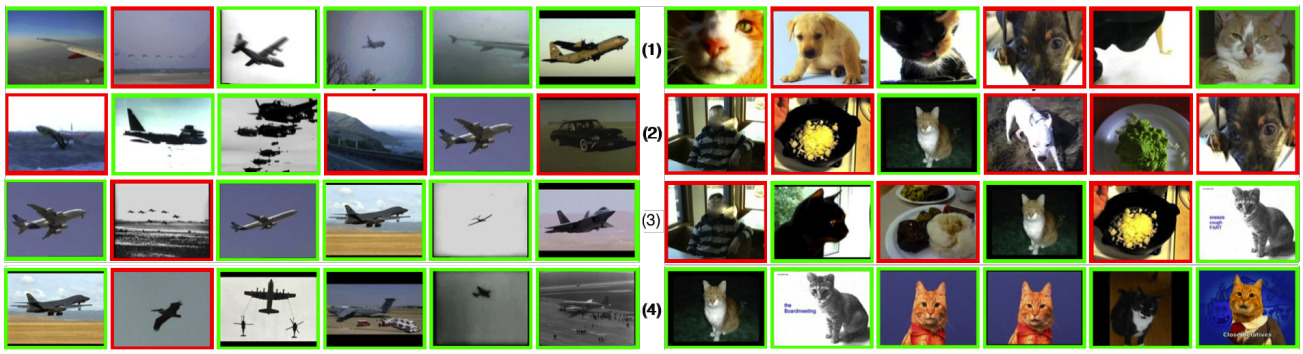


Figure 4. Experiment 1: Object Detector Suitability. Top 6 results for the four detectors trained on the rigid object *Airplane* (left) and on the non-rigid object *Cat* (right). (1) is `TrainAll-ClassifyAll`, (2) is `TrainBB-ClassifyAll`, (3) is `TrainBB-SegmentationBB` and (4) is `TrainBB-DeformableBB`. Green surrounds indicates correct detection, while a red rectangle indicates wrong prediction.

4. OBJECT EVENT-REPRESENTATION

By using object detector scores as representation, we aim to increase the robustness of event retrieval and to validate our second hypothesis. We evaluate our object-based approach in terms of how well it re-ranks a list of videos obtained by an appearance-based event detector.

4.1 Data sets

Experimentation with our object-based approach is performed on the TRECVID 2011 Multimedia Event Detection data set.[13] In this set, a huge collection of unconstrained Internet video clips together with ground truth annotations for fifteen events is provided. We restrict our analysis to the events defined in terms of objects present in experiment 1, which are "Feeding an animal", "Changing a vehicle tire", "Getting a vehicle unstuck" and "Grooming and animal". The statistics are provided in Table 2 while the event names and their designation are listed in Table 3.

Table 2. Statistics of TRECVID 2011 Event Detection data set.

Set	**Videos**	**Frame**	**Hours**
Train	2680	9.1×10^6	92
Validation	10403	32×10^6	324
Test	32061	98×10^6	991

Table 3. Training and testing events defined in the TRECVID 2011 Event Detection task. We restrict ourselves to those events that have a one-to-one correspondence with the objects evaluated in Experiment 1 (denoted with italics).

Training Events	**Testing Events**	
Attempting a board trick	Birthday Party	Making a sandwich
Feeding an animal	*Changing a vehicle tire*	Parade
Landing a fish	Flash mob gathering	Parkour
Working on woodworking project	*Getting a vehicle unstuck*	Repairing an appliance
Wedding ceremony	*Grooming an animal*	Working on a sewing project

4.2 Experiment 2: Object Event-Representation

Re-ranking. To re-rank the results of an appearance-based event detector, we simply promote in rank those videos that are considered as positive by our object-based event detector and leaves unchanged the other videos. We call this *Object Verification*. Formally, this is defined by:

$$S_{new}(\mathbf{X}_i) = \begin{cases} -P_1(\mathbf{X}_i) + M & f_2(\mathbf{X}_i) > \theta \\ -P_1(\mathbf{X}_i) & \text{otherwise} \end{cases} \tag{1}$$

where $S_{new}(\mathbf{X}_i)$ indicates the final score of the video $\mathbf{X}_i$, $P_1(\mathbf{X}_i)$ is the position of the video in the ranked list obtained by the appearance-based technique. In addition, M is the total number of videos in the test data set, $f(\dots)$ is the event detector score and θ is a threshold used to select only the videos where the detector is certain about the positive prediction.

Video Object Representation. We represent a video as follows. Let's consider an event $\mathcal{V}$. For it, we define:

$$L = \{(d_k(), l_{d_k}),\ k = 1, \dots, D\} \tag{2}$$

$$G = \{(c_j(), l_{c_j}),\ j = 1, \dots, C\} \tag{3}$$

where L is a set of D local object detectors and G is a set of C global object detectors. In addition, $d_k()$ and $c_j()$ indicate the k-th local object detector and the j-th global object detector, which take as input a frame and returns a value telling how likely the frame is to contain the object. Finally, l_{d_k} and l_{c_j} are object labels.

Then, we define the i-th video of the data set as the sequence $\{f_{i,1}, \ldots, f_{i,F}\}$, where F is the number of frames in the i-th video and $f_{i,j}$ represents the j-th frame. We apply all the detectors to all these frames and we extract the maximum score of each detector, over all the frames, in order to ensure the presence of an object:

$$L_{MAX_i} = [\max_{j=1,\ldots,F}(d_1(f_{i,j})), \ldots, \max_{j=1,\ldots,F}(d_D(f_{i,j}))] \tag{4}$$

$$G_{MAX_i} = [\max_{j=1,\ldots,F}(c_1(f_{i,j})), \ldots, \max_{j=1,\ldots,F}(c_C(f_{i,j}))] \tag{5}$$

where L_{MAX_i} contains scores for the objects trained with the local detector and G_{MAX_i} contains scores for the objects trained with the global detector.

In addition, in order to capture temporal information and to be able to distinguish between a video where an object plays an important role and a video where the object appears shortly as noise, we extract the average score of each detector, over all the frames:

$$L_{AVG_i} = [\frac{1}{F}\sum_{j=1}^{F}(d_1(f_{i,j})), \ldots, \frac{1}{F}\sum_{j=1}^{F}(d_D(f_{i,j}))] \tag{6}$$

$$G_{AVG_i} = [\frac{1}{F}\sum_{j=1}^{F}(c_1(f_{i,j})), \ldots, \frac{1}{F}\sum_{j=1}^{F}(c_C(f_{i,j}))] \tag{7}$$

Finally, we define a video in terms of an object-vector, as the concatenation of these scores:

$$\mathbf{X}_i = [L_{MAX_i}, G_{MAX_i}, L_{AVG_i}, G_{AVG_i}] \tag{8}$$

Video Classification. Finally, this video object representation is classified into an event using a simple linear support vector machine.[27] This classifier returns a score ($f_2(\mathbf{X}_i)$ from Eq. 1) indicating how likely the video is to contain the considered event.

4.3 Implementation Details

Event Detection. We implemented an appearance-based event detector using bag-of-words and we applied it to all shot-segmented keyframes from the TRECVID Multimedia Event Detection 2011 test set. The approach is based on multiple (visual) kernels using hard-assignment on SIFT, OpponentSIFT, and RGB-SIFT descriptors. Fusion is performed using an average rule combination. To relieve the computational burden of running various object detectors, we re-rank the top 400 results obtained with the appearance-based event detector.

Object Detection. As results of Experiment 1 suggest, for Eq. 2, we use the local detector `TrainBB-DeformableBB` and for Eq. 3, the global detector `TrainAll-ClassifyAll`. For the local detector we use the previously trained object models from experiment 1. For the global detector we train object models on the data set of the TRECVID 2011 Semantic Indexing task. We apply the object detectors to frames sampled every 2 seconds.

Object Selection. For each event, we pre-select from experiment 1 the set of objects described in the event definition. For example, for the event "Feeding an animal", this is represented by the objects *bird, cat, cow, dog* and *horse*. We include both the global and local object detectors. To compensate for those types of animals and vehicles for which we do not train object detectors, we also include global detectors for *animals* and *vehicle* for which annotations are provided in the TRECVID 2011 Semantic Indexing Task.

4.4 Result 2: Object Event-Representation

Results are shown in Figure 5. *Object Verification*, which simply promotes those videos that our object-based approach considers as positive, improves the results of all the events. Out of all results, it is interesting to notice the improvement for "Grooming an animal", where the average precision raises with 70%, from 0.089 to 0.151.

Furthermore, it is worth to have a look into the videos promoted to the top of the ranked lists. Figure 6(a) show some positive videos appearing in the top 10 and promoted by our *Object Verification*. In contrast, Figure

6(b) shows some negative videos appearing in the top 10 list of all the events. Even though such videos are negative, results shows that our approach is particularly good in finding videos where the objects of interest are present for a considerably long amount of time. Despite this, more detectors could be easily added in order to allow for more discrimination power, for example the inclusion of a *person* and/or a *hand* detector could potentially reject some of the videos shown in Figure 6(b).

All in all, our object-based event representation, which is a simple collection of statistics from object detectors, helps the baseline considerably. Results improve for all the events, showing good generalization. We conclude that present-day object detectors can successfully aid arbitrary Internet video event detection, which confirms our second hypothesis.

5. CONCLUSION

In this paper, we assess the potential of using object detectors to represent events. We formulate two hypotheses and experiments to evaluate this potential. Experiment 1 confirms our first hypothesis by showing that state-of-the-art local object detectors for photos can successfully be used for recognizing objects in low-quality Internet videos. Our experiments indicate that the most suitable detector is the local detector based on deformable part-based models by Felzenswalb et al.[25] (Figure 3). To compensate for the global context information ignored by this method, we advise to include in the event representation a global detector based on bag-of-words as support. Our second hypothesis states that present-day object detectors can aid arbitrary Internet video event detection. In experiment 2 we use a preselected object-based video representation to re-rank the results of an appearance-based event detector. Our results in Experiment 2 seem to confirm this hypothesis (Figure 5). Re-ranking with object verification aids event detection performance. To conclude, we believe that our experimental study reveals the potential of object-based event representations.

While we consider our results promising, a lot of research is needed to further enhance the quality of semantic representations for events. First, increasing the number of local object detectors is likely to enhance the discriminative power. Second, it is of interest to automatically learn from training data what objects are relevant for an event. Finally, it would be interesting to study semantic interactions between objects and humans. We believe that exploiting the spatio-temporal position of several objects could lead to the most semantic representation.

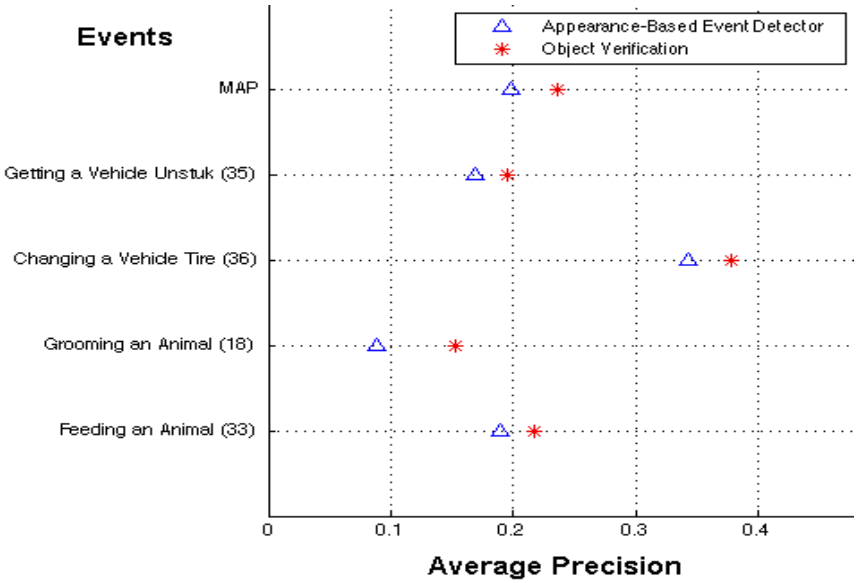
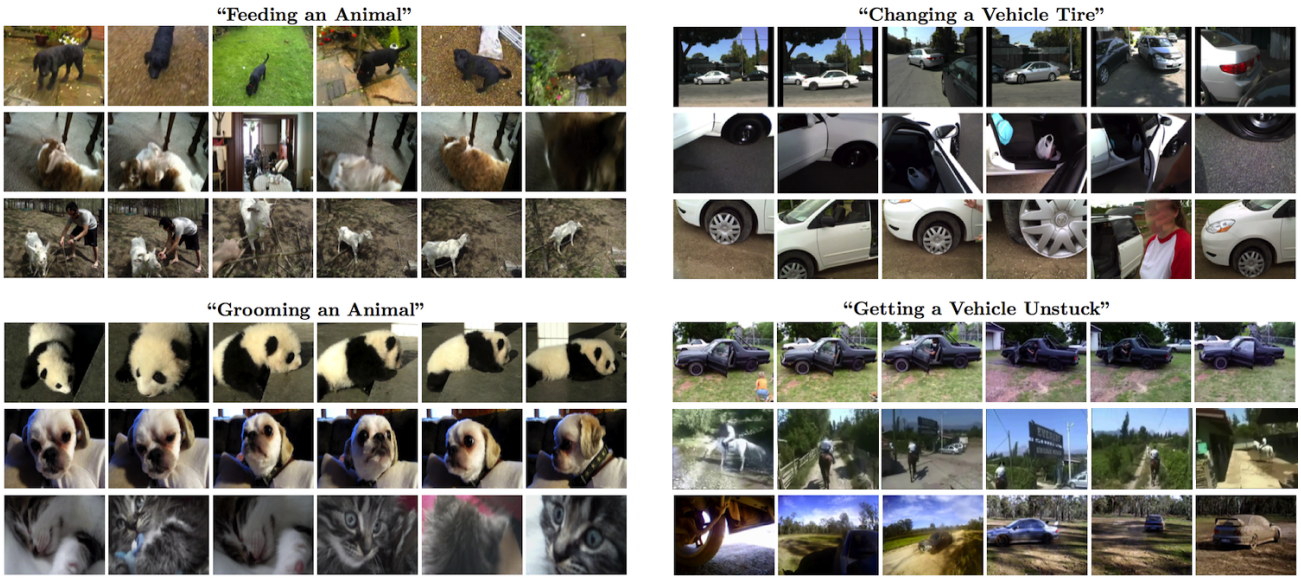


Figure 5. Results of Experiment 2: Object Event-Representation. The numbers next to the event names indicate the number of positive videos in the data set. Results show that present-day object detectors can successfully aid arbitrary Internet video event detection.

(a) Positive videos.



(b) Negative videos. It is worth noticing that all the videos contain the objects of interest and for the event "Changing a vehicle tire" two videos even contain a car with a flat tire, but not in the process of being changed.

Figure 6. Experiment 2: Object Event-Representation. Videos promoted to the top 10 by our *Object Verification*. Both positive and negative videos contain the object of interests, showing that our object-based video representation can aid Internet video event retrieval.

ACKNOWLEDGMENTS

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

REFERENCES

[1] Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., and Serra, G., "Event detection and recognition for semantic annotation of video," *Multimedia Tools and Applications* **51**, 279–302 (January 2011).

[2] Xie, L., Sundaram, H., and Campbell, M., "Event mining in multimedia streams," *Proc. IEEE* **96**(4), 623–647 (2008).

[3] Bobick, A. and Davis, J., "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(3), 257–267 (2001).

[4] Efros, A., Berg, A., Mori, G., and Malik, J., "Recognizing action at a distance," in [*Proc. 9th IEEE International Conference on Computer Vision*], 726–733 (2003).

[5] Xie, L., Chang, S.-F., Divakaran, A., and Sun, H., "Structure analysis of soccer video with hidden markov models," in [*Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*], (2002).

[6] Chang, S., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A., and Luo, J., "Large-scale multimodal semantic concept detection for consumer video," in [*Proc. ACM International Workshop on Multimedia Information Retrieval*], 255–264 (2007).

[7] Hill, M., Hua, G., Natsev, A., Smith, J., Xie, L., Huang, B., Merler, M., Ouyang, H., and Zhou, M., "Ibm research trecvid-2010 video copy detection and multimedia event detection system," in [*Proc. TRECVID Workshop*], (2010).

[8] Jiang, Y., Zeng, X., Ye, G., Bhattacharya, S., Ellis, D., Shah, M., and Chang, S., "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in [*Proc. NIST TRECVID Workshop*], (2010).

[9] Natarajan, P. and et al., "BBN VISER TRECVID 2011 multimedia event detection system," in [*Proc. 9th TRECVID Workshop*], (December 2011).

[10] Xu, D. and Chang, S., "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(11), 1985–1997 (2008).

[11] Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W., "Discovering objects and their location in images," in [*Proc. 10th IEEE International Conference on Computer Vision*], **1**, 370–377 (2005).

[12] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision* **88**, 303–338 (June 2010).

[13] Smeaton, A., Over, P., and Kraaij, W., "Evaluation campaigns and trecvid," in [*Proc. 8th ACM International Workshop on Multimedia Information Retrieval*], 321–330 (2006).

[14] Casati, R. and Varzi, A., "Event concepts," in [*Understanding events: From perception to action*], 31–53, T. F. Shipley & J. M. Zacks (Eds.) (2008).

[15] Pruden, S.M. Hirsh-Pasek, K. and Golinkoff, R., "Current events: How infants parse the world and events for language," in [*Understanding events: From perception to action*], 160–192, T. F. Shipley & J. M. Zacks (Eds.) (2008).

[16] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C., "Visual categorization with bags of keypoints," in [*Workshop on Statistical Learning in Computer Vision, ECCV*], **1**, 1–22 (2004).

[17] Lowe, D., "Object recognition from local scale-invariant features," *Proc. 7th IEEE International Conference on Computer Vision* **2**, 1150–1157 (1999).

[18] van de Sande, K., Gevers, T., and Snoek, C., "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**, 1582–1596 (September 2010).

[19] van Gemert, J., Veenman, C., Smeulders, A., and Geusebroek, J., "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(7), 1271–1283 (2010).

[20] Lazebnik, S., Schmid, C., and Ponce, J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in [*Proc. IEEE Conference on Computer Vision and Pattern Recognition*], **2**, 2169–2178 (2006).

[21] Maji, S., Berg, A., and Malik, J., "Classification using intersection kernel support vector machines is efficient," in [*Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008*], 1–8 (2008).

[22] Lampert, C., Blaschko, M., and Hofmann, T., "Beyond sliding windows: Object localization by efficient subwindow search," in [*Proc. IEEE Conference on Computer Vision and Pattern Recognition*], 1–8 (2008).

[23] Alexe, B., Deselaers, T., and Ferrari, V., "What is an object?," in [*Proc. IEEE Conference on Computer Vision and Pattern Recognition*], 73–80 (2010).

[24] van de Sande, K., Uijlings, J., Gevers, T., and Smeulders, A., "Segmentation as selective search for object recognition," in [*Proc. IEEE International Conference on Computer Vision*], (2011).

[25] Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D., "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 1627–1645 (2009).

[26] Felzenszwalb, P. and Huttenlocher, D., "Efficient graph-based image segmentation," *International Journal of Computer Vision* **59**, 167–181 (September 2004).

[27] Chang, C. and Lin, C., "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* **2**(3), 27 (2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.